Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, The Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK

Third English edition 2012

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

For information on all **Elsevier** publications
visit our web site at store.elsevier.com

Printed and bound in Great Britain
12 13 14 15 16 10 9 8 7 6 5 4 3 2 1

ISBN: 978-0-444-53868-0

Preface

In May 1975, a dozen or so ecologists, mostly marine, sat during three days in a (then) dusty conference room on the first floor of a historical building of the *Station marine de Villefranche-sur-Mer* (Université Paris 6, France), metres away from the Mediterranean shore, to discuss developments concerning a new trend in the ecological literature: the statistical analysis of multivariate ecological data. We, the authors of this book, had been independently invited to participate in the seminar. On the evening of the closing day of the meeting, sitting at the terrace of a restaurant, we wrote on a paper place mat a list of subjects that was to become the table of contents of the book that we published a few years later under the title ***Écologie numérique*** (first edition, in French; Legendre & Legendre, 1979a and b).

During the 1970's, community ecology, which had traditionally been a descriptive science until then, slowly adopted the hypothesis testing approach. Testing hypotheses required the analysis of numerical data. The theoretical foundations of community ecology had been developed during the 1950's and 1960's (niche theory, succession, biodiversity concepts, food webs, etc.) and statistically inclined researchers had already suggested to analyse ecological data using multivariate methods (e.g. Odum, 1950; Goodall, 1954; Bray & Curtis, 1957; Margalef, 1958; Williams & Lambert, 1959; Dagnelie, 1960, 1965; Gower, 1966; Pielou, 1966, 1969). We felt, in 1975, that the time was ripe to inventory the available numerical methods, compare them to the array of ecological questions found in the literature, describe the correspondences between questions and methods, provide a structure to interlink the various methods, and identify methodological gaps in the edifice. This is what we did in the first editions of this book, published in French in 1979, then in English under the title ***Numerical ecology*** (Legendre & Legendre, 1983a), quickly followed by a second French edition (Legendre & Legendre, 1984a and b).

Following the inventory and educational work described above, and with the help of graduate students and research assistants, we got to work to develop new numerical methods to answer emerging ecological questions and help fill gaps in existing numerical methodologies. Similar movements towards development of numerical methods took place in several laboratories throughout the world. In the late 1990's, the

time was ripe for a new synthesis of the field, and we worked on the second English edition of the ***Numerical ecology*** textbook (Legendre & Legendre, 1998). A decade later, the field of multivariate community ecology had developed so much that a new synthesis had become necessary. We spent most of the past three years preparing the 2012 edition of ***Numerical ecology***. This edition includes numerous developments in statistical computing made available in the R statistical language, and refers to many R packages written for ecologists by researchers in several laboratories around the world.

During our teaching in universities at home and abroad, we have been repeating a key message to graduate students: *While it is important to learn about the methods developed by previous generations of scientists, do not let yourself be silenced by their aura. If you think you have a good idea, work on it, develop it, listen to criticisms, and publish it, thus contributing to the advancement of the field. Do not let people tell you that everything is known, or that you are too young or not good enough to contribute to this field — or any other field of science.*

The ***Numerical ecology*** book is written for practising scientists — graduate students and professional researchers, in classical and molecular ecology, oceanography and limnology, environmental sciences, soil science, agriculture, environmental engineering, and related fields. For that reason, it is organized both as a practical handbook and a reference textbook. Our goal is to describe and discuss the numerical methods that are successfully used for analysing ecological data, using a clear and comprehensive approach. These methods are derived from the fields of mathematical physics, parametric and nonparametric statistics, information theory, numerical taxonomy, archaeology, geography, psychometrics, sociometry, econometrics, and others. Meaningful use of most of these methods requires that their theoretical bases be mastered by users. For that reason, analyses reported in the literature have at times been carried out with methods that were not fully adapted to the question or the data under study, leading to conclusions that were sub-optimal with respect to the quality of the field observations. When we were writing the first English edition of ***Numerical ecology***, this warning mostly concerned multivariate versus elementary statistics. Nowadays, most ecologists are capable of using multivariate methods; the above remark now especially applies to the analysis of spatially or temporally correlated data (see Section 1.1; Chapters 12 to 14) and the joint analysis of several data tables (Chapter 11).

Computer packages provide easy access to the most sophisticated numerical methods. Ecologists with inadequate background often find, however, that using high-level packages leads them to dead ends. In order to efficiently use the available numerical tools, it is essential to clearly understand the principles that underlay numerical methods, and their limits. It is also important for ecologists to have guidelines for interpreting the heaps of computer-generated results. We therefore organized the present text as a comprehensive outline of methods for analysing ecological data, and also as a practical handbook pointing to the most commonly-used packages.

Our experience with graduate teaching and consulting has made us aware of the problems that ecologists may encounter when they first use advanced numerical methods. Any earnest approach to such problems requires in-depth understanding of the general principles and theoretical bases of the methods to be used. The approach followed in this book uses standardized mathematical symbols, abundant illustration, and appeal to intuition in some cases. Because the text has been used for many years for graduate teaching and greatly improved along the process, we know that, with reasonable effort, readers can get to the core of numerical ecology. In order to efficiently use numerical methods, their aims and limits must be clearly understood, as well as the conditions under which they should be employed. In addition, since most methods are well described in the scientific literature and are available in computer packages, we generally devote most of the text to the ecological interpretation of the results; computation algorithms are described only when they may help readers to understand methods. Methods described in the book are systematically illustrated with numerical examples and/or applications drawn from the ecological literature, mostly in English; references in languages other than English or French are generally of historical nature.

The expression *numerical ecology* refers to the following approach. *Mathematical ecology* covers the domain of mathematical applications to ecology. It may be divided into *theoretical ecology* and *quantitative ecology*. The latter, in turn, includes a number of disciplines, among which *modelling*, *ecological statistics*, and *numerical ecology*. *Numerical ecology* is the field of quantitative ecology devoted to the numerical analysis of ecological data sets. Community ecologists, who generally use multivariate data, are the primary users of these methods. The purpose of numerical ecology is to describe and interpret the structure of data sets by combining a variety of numerical approaches. Numerical ecology differs from descriptive or inferential *ecological statistic* in that it combines relevant multidimensional statistical methods with heuristic techniques (e.g. cluster analysis) that do not have a firm statistical foundation. In addition, it often incorporates into the analysis of multivariate data constraints that represent ecological hypotheses, e.g. spatial or temporal contiguity, or relationships between community structure and environmental variables. Numerical ecology also differs from *ecological modelling*, even though the extrapolation of ecological structures is often used to *forecast* values in space or/and time (through multiple regression or other similar approaches, which are collectively referred to as *correlative models*). When the purpose of a study is to *predict* the critical consequences of alternative solutions, ecologists must use *predictive ecological models*. The development of such models, which predict effects on some variables caused by changes in others, requires a deliberate causal structuring. This approach must be based on ecological theory and include a validation procedure. Because the ecological hypotheses that underlay causal models are often developed within the context of studies that use numerical ecology, the two fields are often in close contact.

Ecologists have used quantitative approaches since the publication by Jaccard (1900) of the first association coefficient. Floristics developed from that seed, and the method was eventually applied to all fields of ecology, often achieving high levels of

complexity. Following Spearman (1904) and Hotelling (1933), psychometricians and social scientists developed non-parametric statistical methods and factor analysis and, later, nonmetric multidimensional scaling (nMDS). During the same period, anthropologists (e.g. Czekanowski, 1909, 1913) were interested in numerical classification, and economists started to developed numerical indices (e.g. Gini, 1912). The advent of computers made it possible to analyse large data sets, using combinations of methods derived from various fields, supplemented with new mathematical developments. The first synthesis was published by Sokal & Sneath (1963), who established *numerical taxonomy* as a new discipline.

Numerical ecology combines a large number of approaches, derived from many disciplines, in a general methodology for analysing ecological data sets. Its chief characteristic is the *combined* use of treatments drawn from different areas of mathematics and statistics. Numerical ecology acknowledges the fact that many of the existing numerical methods are *complementary* of one another, each one allowing the exploration of a different aspect of the information underlying the data; it sets principles for interpreting the results in an integrated way.

The present book is organized in such a way as to encourage researchers who are interested in a method to also consider other techniques. The integrated approach to data analysis is favoured by numerous cross-references among chapters and the presence of sections devoted to syntheses of subjects. The book synthesizes a large amount of information from the literature, within a structured and prospective framework, to help ecologists take maximum advantage of the existing methods.

This third English edition of ***Numerical ecology*** is deeply revised and largely expanded compared to the second English edition (Legendre & Legendre, 1998). It contains a new chapter dealing with multiscale analysis by spatial eigenfunctions (Chapter 14). In addition, new sections have been added in several chapters and others have been rewritten. These include the sections (numbers given in parentheses) on: autocorrelation (1.1), singular value decomposition (2.11), species diversity through space (6.5.3), the double-zero problem (7.2.2), transformations for community composition data (7.7), multivariate regression trees (8.11), and matrix comparison methods (10.5). Sections 11.1 on redundancy analysis and 11.4 on canonical correlation analysis have been entirely rewritten, and a new Section 11.5 on co-inertia and Procrustes analyses has been added. New sections, found at the end of most chapters, list available computer programs, with special emphasis on R packages.

The present work reflects the input of many colleagues, to whom we express here our most sincere thanks. We first acknowledge the outstanding inputs of the late Professor Serge Frontier (Université des Sciences et Techniques de Lille) and Professor F. James Rohlf (State University of New York at Stony Brook) who critically reviewed our manuscripts for the first French and English editions, respectively. Many of their suggestions were incorporated into the texts that are at the origin of the present edition. We are also grateful to the late Professor Ramón Margalef for his support, in the form of an influential Preface to the two French and the first English editions. Over

the years, we had fruitful discussions on various aspects of numerical methods with many colleagues, whose names have sometimes been cited in the Forewords of previous editions.

During the preparation of this new edition, we benefited from the help of several colleagues. First and foremost is Daniel Borcard; after 20 years of scientific collaboration with one of us, he undertook to write a book, ***Numerical ecology with R*** (Borcard *et al*., 2011), which is the companion to the present manual. That book shows readers how to use the R language to carry out calculations for the methods described in the present book. In addition, Daniel Borcard revised several chapters and sections of this new edition, including Sections 1.1 and 10.5, Chapter 14, and all the Software sections found at the end of the chapters. He also carried out the simulations for the Dagnelie test of multivariate normality reported at the end of Section 4.6, and he developed the method of selection of rare species to be used before correspondence analysis (Box 9.2). Jari Oksanen developed an algorithm combining PCA/RDA/partial RDA and gave us permission to reproduce it in Table 11.5. We are most grateful to these two researchers for their major contributions to our book.

Other long-time collaborators and friends helped us by revising sections of the book that were either new or had been rewritten and modernized. We are most thankful to Marie-Josée Fortin who revised Section 1.1, François-Joseph Lapointe for Section 8.13, Miquel De Cáceres for Subsection 8.9.3, Stéphane Dray and Pedro Peres-Neto for Section 11.5, Patrick M. A. James for Subsection 12.5.4, and Helene H. Wagner for Subsection 13.1.4. Cajo J. F. ter Braak and Jari Oksanen commented on portions of Section 11.1. The new Chapter 14 received special attention: it was entirely revised by Daniel Borcard and Pedro Peres-Neto, whereas other colleagues revised the sections describing methods that they had contributed in developing: Stéphane Dray for Section 14.1 and 14.2, F. Guillaume Blanchet for Section 14.3, Helene H. Wagner for Section 14.4, Miquel De Cáceres for Subsection 14.5.1, and Guillaume Guénard for Subsection 14.5.2.

Graduate students in our home universities and those who participated in short courses that we gave in several countries abroad have greatly contributed to the book by raising interesting questions and pointing out weaknesses in previous versions of the text.

While writing this book, we benefited from competent and unselfish advice … which we did not always follow. We thus assume full responsibility for any gaps in the work and for all the opinions expressed therein. We shall therefore welcome with great interest all suggestions or criticisms from readers.

PIERRE LEGENDRE, Université de Montréal
LOUIS LEGENDRE, Université Pierre et Marie Curie Paris 6 April 2012

Chapter

1 *Complex ecological data sets*

1.0 Numerical analysis of ecological data

The foundation of a general methodology for analysing ecological data may be derived from the relationships that exist between the conditions surrounding ecological observations and their outcomes. In the physical sciences for example, there often are cause-to-effect relationships between the natural or experimental conditions and the outcomes of observations or experiments. This is to say that, given a certain set of conditions, the outcome may be exactly predicted. Such totally deterministic relationships are only characteristic of extremely simple ecological situations.

Probability

Generally in ecology, a number of different outcomes may follow from a given set of conditions because of the large number of influencing variables, of which many are not readily available to the observer. The inherent genetic variability of biological material is an important source of ecological variability. If the observations are repeated many times under similar conditions, the relative frequencies of the possible outcomes tend to stabilize at given values, called the *probabilities* of the outcomes. Following Cramér (1946: 148), it is possible to state that "whenever we say that the probability of an event with respect to an experiment [or an observation] is equal to P, the concrete meaning of this assertion will thus simply be the following: in a long series of repetitions of the experiment [or observation], it is practically certain that the [relative] frequency of the event will be approximately equal to P." This corresponds to the frequency theory of probability — excluding the Bayesian and likelihood approaches.

Probability distribution

In the first paragraph, the outcomes were recurring at the individual level whereas in the second, results were repeatable in terms of their probabilities. When each of several possible outcomes occurs with a given characteristic probability, the set of these probabilities is called a *probability distribution*. Assuming that the numerical value of each outcome E_i is y_i with corresponding probability p_i, a *random variable* (or

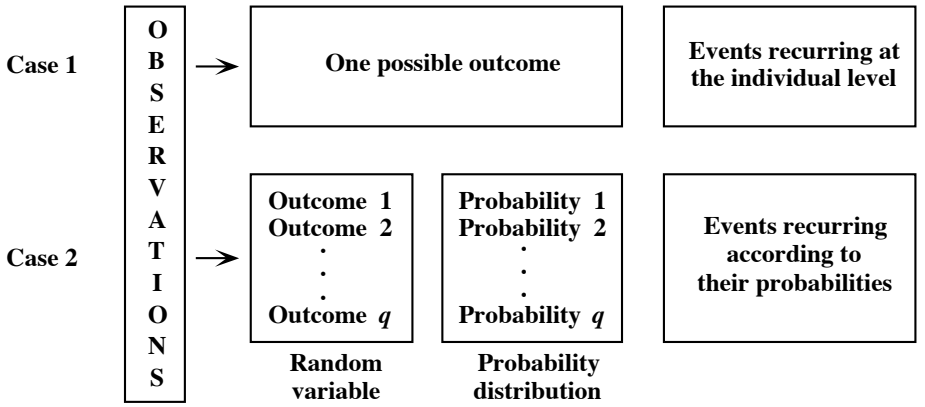


Figure 1.1 Two types of recurrence of the observations.

Random variable

variate) **y** is defined as that quantity which takes on the value y_i with probability p_i at each trial (Morrison, 1990). Figure 1.1 summarizes these basic ideas.

Of course, one can imagine other results to observations. For example, there may be *strategic* relationships between surrounding conditions and resulting events. This is the case when some action — or its expectation — triggers or modifies the reaction. Such strategic-type relationships, which are the object of *game theory*, may possibly explain ecological phenomena such as species succession or evolution (Margalef, 1968). Should this be the case, this type of relationship might become central to ecological research.

Another possible outcome is that observations bear some degree of *unpredictability*. Such data may be studied within the framework of chaos theory, which explains how deterministic processes can generate phenomena with a sensitive dependence on initial conditions that ensures dynamical behaviour with short-term predictability but long-term unpredictability (e.g. Ferriere *et al*., 1996). This is the famous "butterfly effect", whereby a butterfly flapping its wings somewhere on Earth could alter weather patterns somewhere else at a later time. The signature of chaos has been detected in a number of biological systems. For example, Beninca *et al*. (2008) used the data on a bacteria-phytoplankton-zooplankton food web that had been cultured for more than 2300 days under constant external conditions in a laboratory mesocosm to show that species interactions in that food web generated chaos. According to the authors, this result implies that the long-term prediction of species abundances could be fundamentally impossible. For an overview of chaos theory, interested readers can refer to Peitgen *et al*. (2004).

Methods of numerical analysis are determined by the four types of relationships that may be encountered between surrounding conditions and the outcome of observations (Table 1.1). The present text deals only with methods for analysing random response variables, which is the type ecologists most frequently encounter.

The numerical analysis of ecological data makes use of mathematical tools developed in many different disciplines. A formal presentation must rely on a unified approach. For ecologists, the most suitable and natural language — as will be shown in Chapter 2 — is that of *matrix algebra*. This approach is best adapted to the processing of data by computers; it is also simple, and it efficiently carries information, with the additional advantage of being familiar to many ecologists.

Other disciplines provide ecologists with powerful tools that are well adapted to the complexity of ecological data. From mathematical physics comes *dimensional analysis* (Chapter 3), which provides simple and elegant solutions to some difficult ecological problems. Measuring the association among quantitative, semiquantitative or qualitative variables is based on *parametric* and *nonparametric statistical methods* and on *information theory* (Chapters 4, 5 and 6, respectively).

These approaches all contribute to the analysis of complex ecological data sets (Fig. 1.2). Because such data usually come in the form of highly interrelated variables, the capabilities of elementary statistical methods are generally exceeded. While elementary methods are the subject of a number of excellent texts, the present manual focuses on the more advanced methods, upon which ecologists must rely in order to understand these interrelationships.

Table 1.1 Numerical analysis of ecological data.

Relationships between the natural conditions and the outcome of an observation	Methods for analysing and modelling the data
Deterministic: Only one possible result	Deterministic models
Random: Many possible results, unpredictable individually but with characteristic probabilities of occurrence	Methods described in this book (Figure 1.2)
Strategic: Results depend on the respective strategies of the organisms and of their environment	Game theory
Chaotic: Many possible results with short-term predictability and long-term unpredictability	Chaos theory
Uncertain: Many possible, unpredictable results	

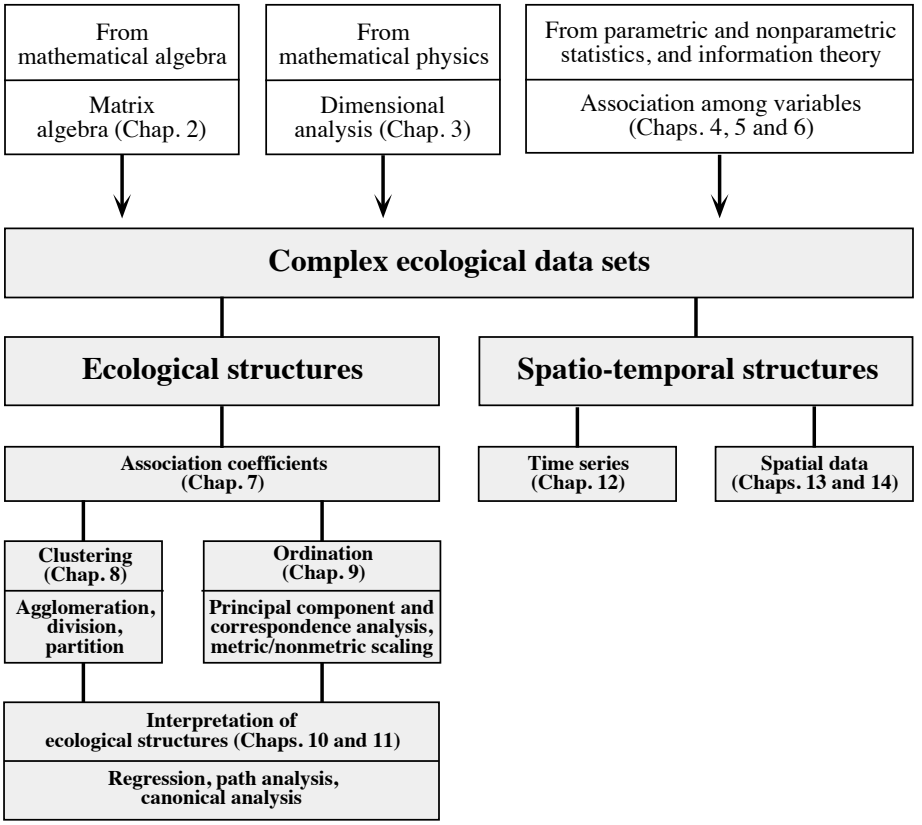


Figure 1.2 Numerical analysis of complex ecological data sets.

In ecological spreadsheets, data are typically organized in rows corresponding to sampling sites or times, and columns representing the variables; these may describe the biological communities (species presence, abundance, or biomass, for instance) or the physical environment. Because many variables are needed to describe communities and environment, ecological data matrices are, for the most part, *multidimensional* (or *multivariate*). Multidimensional data, i.e. data consisting of several variables, structure what is known in geometry as a *hyperspace*, which is a space with many dimensions. One now classical example of ecological hyperspace is the *fundamental niche* of Hutchinson (1957, 1965). According to Hutchinson, the environmental variables that are critical for a species to exist may be thought of as orthogonal axes, one for each factor, of a multidimensional space. On each axis, there are limiting conditions within which the species can exist indefinitely; this concept is

Fundamental niche

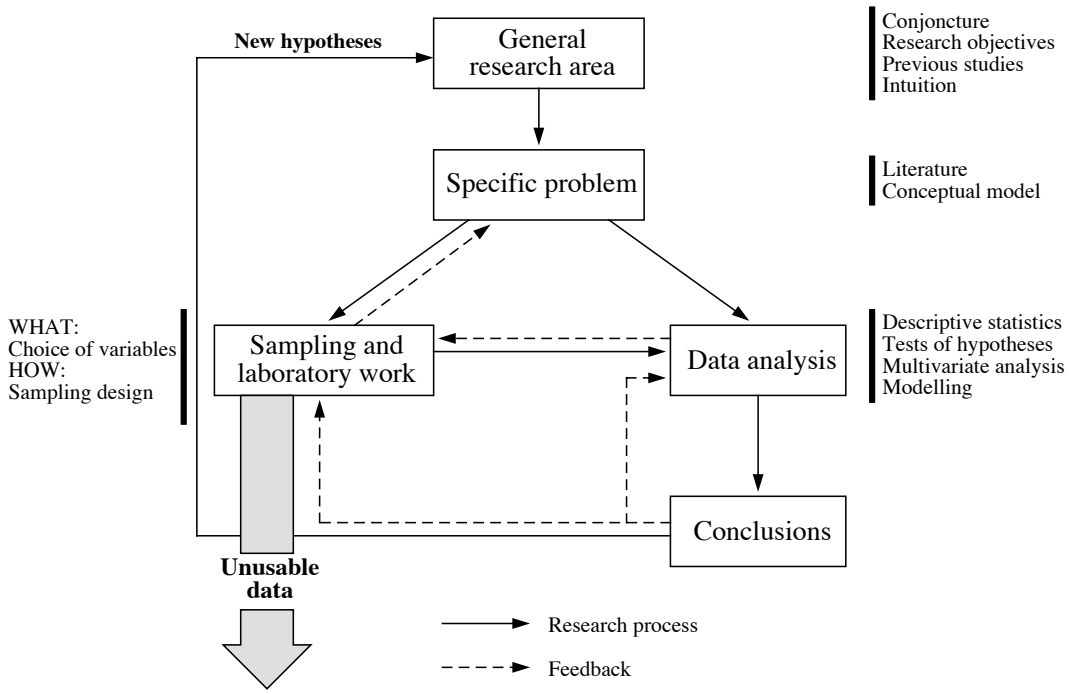


Figure 1.3 Relationships among the various phases of an ecological research.

called upon in Subsection 7.2.2, which discusses unimodal species distributions and their consequences on the choice of resemblance coefficients. In Hutchinson's theory, the set of these limiting conditions defines a hypervolume called the species' fundamental niche. The spatial axes describe the geographical distribution of the species.

The quality of the analysis and subsequent interpretation of complex ecological data sets depends, in particular, on the compatibility between data and numerical methods. It is important to take into account the requirements of the numerical techniques when planning a sampling programme, because it is obviously useless to collect quantitative data that are inappropriate to the intended numerical analyses. Experience shows that, too often, poorly planned collection of costly ecological data, for "survey" purposes, generates large amounts of unusable data (Fig. 1.3).

The search for ecological structures in multidimensional data sets is always based on *association matrices*, of which a number of variants exist, each one leading to slightly or widely different results (Chapter 7); even in so-called association-free methods, like principal component or correspondence analysis, or *K*-means partitioning, there is always an implicit resemblance measure hidden in the method.

Two main avenues are open for analysis: (1) ecological *clustering* using agglomerative, divisive or partitioning algorithms (Chapter 8), and (2) *ordination* in a space with a reduced number of dimensions, using principal component or correspondence analysis, principal coordinate analysis, or nonmetric multidimensional scaling (Chapter 9). The *interpretation of ecological structures,* derived from clustering and/or ordination, may be conducted either directly or indirectly, as will be seen in Chapters 10 and 11, depending on the nature of the problem and on the additional information available.

Ecological data may be sampled along time or space in order to study *temporal or spatial processes* driven by physics or biology (Chapters 12, 13 and 14). These data may be univariate or multivariate. Time or space sampling requires intensive field work. Time sampling can often be automated using equipment that allows the automatic recording of ecological variables. For spatial surveys, the analysis of satellite images, or of information collected by airborne or shipborne equipment, provides important support to field work, and the geographic positions of the observations can be determined using geographic positioning systems. In physical or ecological applications, a *process* is a phenomenon or a set of phenomena organized along time or through space. Mathematically speaking, such ecological data represent one of the possible realizations of a random process, also called a *stochastic process*.

Process

Two major approaches may be used for inference about the population parameters of such processes (Särndal, 1978; Koch & Gillings, 1983; de Gruijter & ter Braak, 1990; de Gruijter *et al*., 2006). In the *design-based approach*, one is interested only in the sampled population and assumes that a fixed value of the variable exists at each location in space, or point in time. A representative subset of the space or time units is selected using an appropriate (randomized) *sampling design* (for 8 different meanings of the expression "representative sampling", see Kruskal & Mosteller, 1988). Design-based (or *randomization-based*; Kempthorne, 1952) inference results from statistical analyses whose only assumption is the random selection of observations; this requires that the target population (i.e. that for which conclusions are sought) be the same as the sampled population. The probabilistic interpretation of this type of inference (e.g. confidence intervals of parameters) refers to repeated selection of observations from the same finite population using the same sampling strategy. The classical (Fisherian) methods for estimating the confidence intervals of parameters like the mean, for variables observed over a given surface or time period, are fully applicable in the design-based framework.

Design-based

Model-based

In the *model-based* (or *superpopulation*) *approach*, the assumption is that the target population is much larger than the sampled population. So, the value associated with each location, or point in time, is not fixed but random, since the geographic surface (or time period) available for sampling (i.e. the statistical population) is but one representation of the superpopulation of such surfaces or time periods — all resulting from the same generating process — about which conclusions are to be drawn. The observed population is related to the superpopulation through a *statistical model*, e.g. a variogram (Section 13.1). Under this model, even if the whole sampled population

Super-population

could be observed, uncertainty would still remain about the model parameters. So, the confidence intervals of parameters estimated over a single surface or time period are obviously too small to account for the among-surface variability, and some kind of correction must be made when estimating these intervals. The type of variability of the superpopulation of surfaces or time periods may be estimated by studying the spatial or temporal correlation of the available data (i.e. over the statistical population). This subject is discussed at some length in Section 1.1. Ecological survey data can often be analysed under either model, depending on the emphasis of the study or the type of conclusions one wishes to derive from them.

In some instances in time series analysis, the sampling design must meet the requirements of the numerical method, because some methods are restricted to data series that meet some specific conditions, such as equal spacing of observations. Inadequate planning of the sampling may render the data series useless for numerical treatment with these particular methods. There are several methods for analysing *ecological series* (Chapter 12). Regression, moving averages, and the variate difference method are designed for identifying and extracting general trends from time series. Correlogram, periodogram, and spectral analysis identify rhythms (characteristic periods) in series. Other methods can detect discontinuities in univariate or multivariate series. Variation in a series may be correlated with variation in other variables measured simultaneously. One may also develop forecasting models using the Box & Jenkins approach.

Similarly, methods are available to meet various objectives when analysing spatial data (Chapters 13 and 14). Structure functions such as variograms and correlograms, as well as point pattern analysis, may be used to confirm the presence of a statistically significant spatial structure and to describe its general features. A variety of interpolation methods are used for mapping univariate data, whereas multivariate data can be mapped using methods derived from ordination or cluster analysis. Models may also be developed that include spatial structures among their explanatory variables; in these models, spatial relationships among the study sites may be represented in a variety of ways.

For ecologists, numerical analysis of data is not a goal in itself. However, a study based on quantitative information must take data processing into account at all phases of the work, from conception to conclusion, including the planning and execution of sampling, the analysis of data proper, and the interpretation of results. Sampling, including laboratory analyses, is generally the most tedious and expensive part of ecological research, and it is therefore important that it be optimized in order to reduce to a minimum the collection of useless information. Assuming appropriate sampling and laboratory procedures, the conclusions to be drawn depend on the results of the numerical analyses. It is, therefore, important to make sure in advance that sampling and numerical techniques are compatible. It follows that numerical processing is at the heart of ecological research; the quality of the results cannot exceed the quality of the numerical analyses conducted on the data (Fig. 1.3).

Of course, the quality of ecological research is not solely a function of the expertise with which quantitative work is conducted. It depends to a large extent on creativity, which calls upon imagination and intuition to formulate hypotheses and theories (Legendre, 2004, 2008a). It is, however, advantageous for the researcher's creative abilities to be grounded into solid empirical work (i.e. work involving field data), because little progress may result from continuously building upon untested hypotheses.

Figure 1.3 shows that a correct interpretation of analyses requires that the sampling phase be planned to answer a specific question or questions. Ecological sampling programmes are designed in such a way as to capture the variation occurring along a number of axe of interest: space, time, or other ecological indicator variables. The purpose is to describe variation occurring along the given axis or axes, and to interpret or model it. Contrary to experimentation, where sampling may be designed in such a way that observations are independent of one another, ecological data are often *spatially or temporally correlated* (Section 1.1).

While experimentation is often construed as the opposite of ecological surveys, there are cases where field experiments are conducted at sampling sites, allowing one to measure rates or other processes ("manipulative experiments" *sensu* Hurlbert, 1984; Subsection 10.2.3). In aquatic ecology, for example, nutrient enrichment bioassays are a widely used approach for testing hypotheses concerning nutrient limitation of phytoplankton. In their review on the effects of enrichment, Hecky & Kilham (1988) identified four types of bioassays, according to the level of organization of the test system: cultured algae; natural algal assemblages isolated in microcosms or sometimes larger enclosures; natural water-column communities enclosed in mesocosms; whole systems. The authors discuss one major question raised by such experiments, which is whether results from lower-level systems are applicable to higher levels, and especially to natural situations. Processes estimated in experiments may be used as independent variables in empirical models accounting for survey results, while "static" survey data may be used as covariates to explain the variability observed among blocks of experimental treatments. Spatial and time-series data analysis have become an important part of the analysis of the results of ecological experiments.

1.1 Spatial structure, spatial dependence, spatial correlation

Students in elementary biostatistics courses are trained, implicitly if not explicitly, in the belief that Nature follows the assumptions of classical statistics, one of them being the independence of observations. However, field ecologists know from experience that organisms are not randomly or uniformly distributed in the natural environment, because processes such as growth, dispersal, reproduction, and mortality, which create the observed distributions of organisms, generate spatial correlation in data, as detailed below. The same applies to the physical variables that structure the environment.

Following hierarchy theory (Simon, 1962; Allen & Starr, 1982; O'Neill *et al*., 1991), we may look at the environment as primarily structured by broad-scale physical processes — orogenic and geomorphological processes on land, currents and winds in fluid environments — which, through energy inputs, create gradients in the physical environment as well as patchy structures separated by discontinuities (interfaces). These broad-scale structures lead to similar responses in biological systems, spatially and temporally. Within these relatively homogeneous zones, finer-scaled contagious biotic processes take place, causing the appearance of more spatial structuring through reproduction and death, predator-prey interactions, food availability, parasitism, and so on. This is not to say that biological processes are necessarily small-scaled and nested within physical processes; indeed, biological processes may be broad-scaled (e.g. bird and fish migrations) and physical processes may be fine-scaled (e.g. turbulence). The theory only purports that stable complex systems are often hierarchical. The concept of scale, as well as the expressions *broad scale* and *fine scale*, are discussed in Section 13.0.

In ecosystems, spatial heterogeneity is therefore functional, meaning that ecosystem functioning depends on it (Levin, 2000). It is not the result of some random, noise-generating process. So, it is important to study this type of variability for its own sake. One of the consequences is that ecosystems without spatial structuring would be unlikely to function. Let us imagine the consequences of a non-spatially-structured ecosystem: broad-scale homogeneity would cut down on diversity of habitats; feeders would not be close to their food; mates would be located at random throughout the landscape; soil conditions in the immediate surrounding of a plant would not be more suitable for its seedlings than any other location; newborn animals would be spread around instead of remaining in favourable environments; and so on. Unrealistic as this view may seem, it is a basic assumption of many of the theories and models describing the functioning of populations and communities. The view of a spatially structured ecosystem requires a new paradigm for ecologists: spatial [and temporal] structuring is a fundamental component of ecosystems (Levin, 1992; Legendre, 1993). Hence ecological theories and models, including statistical models, must be revised to include realistic assumptions about the spatial and temporal structuring of communities.

Spatial dependence, which is also called spatial correlation, is used here as the general case; temporal correlation, also called serial correlation in time series analysis, behaves essentially like its spatial counterpart but along a single sampling dimension. The difference between the spatial and temporal cases is that causality is unidirectional in time series, i.e. it proceeds from $(t - 1)$ to t and not the opposite. Temporal processes, which generate temporally correlated data, are studied in Chapter 12, whereas spatial processes are the subject of Chapters 13 and 14. The following discussion is partly inspired from the papers of Legendre & Fortin (1989), Legendre (1993), and Dray *et al*. (2012).

Spatial structures in variables may be generated by different processes. These processes produce relationships between values observed at neighbouring points in space, hence the lack of independence of values of the observed variable (Box 1.1, first

Independence Box 1.1

This word has several meanings. Five of them will be used in this book. Another important meaning in statistics concerns *independent random variables*, which refer to properties of the distributions and density functions of a group of variables; for a formal definition, see Morrison (1990, p. 7).

Independent observations. — Observations drawn from the statistical population in such a way that no observed value has any influence on any other. In the time-honoured example of tossing a coin, observing a head does not influence the probability of a head (or tail) coming out at the next toss. Spatially correlated data violate this condition because their errors are correlated across observations.

Independent descriptors. — Descriptors (variables) that are not related to one another are said to be independent. *Related* is taken here in some general sense applicable to quantitative, semiquantitative as well as qualitative data (Table 1.2).

Linear independence. — Two descriptors are said to be *linearly dependent* if one can be expressed as a linear transformation of the other, e.g. $x_1 = 3x_2$ or $x_1 = 2 - 5x_2$ (Subsection 1.5.1). Descriptors within a set are said to be *linearly dependent* if at least one of them is a linear combination of the other descriptors in the set (Section 2.7). Orthogonality (Section 2.5) is not the same as linear independence. Two vectors may be linearly independent and not orthogonal, but two orthogonal vectors are always linearly independent.

Independent variable(s) of a model. — In a regression model, the variable to be modelled is called the *dependent variable*. The variables used to model it, usually found on the right-hand side of the equation, are called the *independent variables* of the model. In empirical models, one may talk about *response* (or *target*) and *explanatory* variables for, respectively, the dependent and independent variables, whereas, in a causal framework, the terms *criterion* and *predictor* variables may be used. Some forms of canonical analysis (Chapter 11) allow the modelling of a whole matrix of dependent (target or criterion) variables in a single regression-like analysis.

Independent samples are opposed to *related* or *paired samples*. In related samples, each observation in a sample is paired with one in the other sample(s), hence the name *paired comparisons* for the tests of significance carried out on such data. Authors also talk of *independent* versus *matched* pairs of data. Before-after comparisons of the same elements also form related samples (matched pairs).

Spatial correlation

definition of independence). In many instances, observations that are closer together tend to display values that are more similar than observations that are further apart, resulting in *positive spatial dependence* also called *positive spatial correlation*. Repulsion phenomena (e.g. spatial distributions of territorial organisms that prevent other organisms from occupying neighbouring territories) may produce the opposite effect, with values of closer pairs of points being less similar than the values of pairs of observations that are further apart (*negative spatial correlation* at short distances). Closeness may be measured in a distance metric such as metres, or may be represented by counts of graph edges traversed between observations on connection networks (Subsection 13.3.1). A spatial structure may be present in data without it being caused by true autocorrelation, which is defined below. Two models for spatial structure are presented in Subsection 1.1.1; the first one (eq. 1.1 below) does not correspond to autocorrelation *sensu stricto* whereas the second does (eq. 1.2).

Because it indicates lack of independence among the observations, spatial correlation creates problems when attempting to use tests of statistical significance that assume independence of the observations. This point is developed in Subsection 1.1.2. Other types of dependencies (or, lack of independence) may be encountered in biological data. For example, related samples, discussed in more detail in Section 5.2, should not be analysed as if they were independent (Box 1.1, last definition of independence); this would result in a loss of power for the statistical test.

Spatial correlation is a very general property of ecological variables and, indeed, of most natural variables observed over geographic space (spatial correlation) or along time series (temporal correlation). Spatial [or temporal] correlation may be described by mathematical functions such as correlograms and variograms, called structure functions, which are studied in Chapters 12 and 13. The two possible approaches concerning statistical inference for spatially correlated data (i.e. the design- or randomization-based approach, and the model-based or superpopulation approach) were discussed in Section 1.0.

1 — Origin of spatial structures

A spatial structure may appear in a variable **y** because **y** depends upon one or several causal variables **X** that are spatially correlated (Model 1 below) or because the process that has produced the values of **y** is spatial and has generated correlation among the data points (Model 2 below); or some combination of these two processes. In both cases, spatial correlation will be found when analysing the data (Chapters 12 and 13). The spatially-structured causal variables **X** may be explicitly identified in the model, or not; see Table 14.1. The two models, which are also described by Fortin & Dale (2005) and Dray *et al*. (2012), are more precisely defined as follows.

Induced spatial dependence

• Model 1: induced spatial dependence — Spatial dependence may be induced by the functional dependence of the response variables (e.g. species) on explanatory variables (e.g. environmental) **X** that are themselves spatially correlated. We talk about *induced spatial dependence* in that situation where **y** has acquired the spatial structure of **X**.

This phenomenon is a restatement, in the spatial context, of the classical environmental control model (Whittaker, 1956; Bray and Curtis, 1957), which ecologists call upon when they use regression to analyse the variation of a response variable **y** by a table of environmental variables **X**. That model is the foundation of niche theory (Hutchinson, 1957). On the one hand, if all important spatially-structured explanatory variables are included in the analysis, the following model correctly accounts for the spatial structure induced in **y**:

$$y_j = f(\mathbf{X}_j) + \varepsilon_j \qquad \textbf{(1.1)}$$

where y_j is the value of the dependent variable **y** at site j and ε_j is an error term whose value is independent from site to site. On the other hand, if the function is misspecified, for example through the omission of key explanatory variables with spatial patterning such as a broad-scale linear or polynomial trend, or through inadequate functional representation, one may end up incorrectly interpreting the spatial patterning of the residuals as autocorrelation, which is described in the next paragraph.

Autocorrelation

• Model 2: spatial autocorrelation — Spatial dependence may appear in species distributions as the result of "neutral processes" of population and community dynamics (see for instance Hubbell, 2001, and Alonso *et al*., 2006). Neutral processes include ecological drift (variation in species demography due to random reproduction and survival of individuals due to competition, predator-prey interactions, etc.) and random dispersal (migration in animals, propagule dispersion in plants). These processes create *spatial autocorrelation* (*sensu stricto*) in response variables. The value y_j observed at site j on the geographic surface is assumed to be the overall mean of the process (μ_y) in the study area plus a weighted sum of the centred values $(y_i - \mu_y)$ at surrounding sites i, plus an independent error term ε_j:

$$y_j = \mu_y + \Sigma\, w_i\,(y_i - \mu_y) + \varepsilon_j \qquad \textbf{(1.2)}$$

The y_i's are the values of **y** at other sites i located within the zone of spatial influence of the process generating the autocorrelation (Fig. 1.4). The influence of neighbouring sites may be given, for instance, by weights w_i which are function of the distances between sites i and j (eq. 13.20); other functions may be used. The total error term is $[\Sigma\, w_i\,(y_i - \mu_y) + \varepsilon_j]$; it contains the autocorrelated component of variation $[\Sigma\, w_i\,(y_i - \mu_y)]$, which is noted SA_j below. The model assumes spatial stationarity (Subsection 13.1.1). Its equivalent in time series analysis is the autoregressive (AR) response model (eq. 12.29) where each observation in the time series is modelled as a function of preceding observations.

The term autocorrelation is sometimes loosely used to designate any type of spatial dependence; in that case, one would refer to spatial dependence resulting from neutral processes of population and community dynamics as "true autocorrelation", "inherent autocorrelation", or "autogenic autocorrelation" (Fortin & Dale, 2005), or as the "interaction model" (meaning: interaction among the sites) by Cliff & Ord (1981,

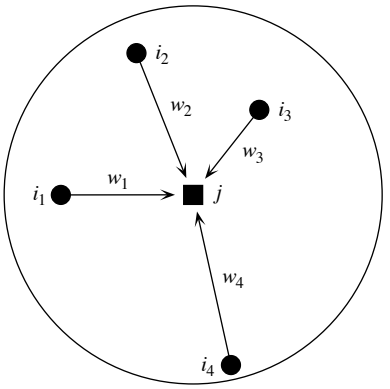


Figure 1.4 The value at site j may be modelled as a weighted sum (with weights w_i) of the influences of other sites i located within the zone of influence of the process generating the autocorrelation (large circle).

p. 141). In statistics, spatial autocorrelation is the spatial dependence found in the error component of a response variable **y** observed trough space after the effect of all important spatially-structured explanatory variables **X** has been accounted for.

The full model describing the value y_j of a response variable **y** at site j is written as follows:

$$y_j = f(\mathbf{X}_j) + u_j \qquad \text{with } u_j = \mathrm{SA}_j + \varepsilon_j$$

where **y** is modelled as a function of the explanatory (e.g. environmental) variables **X,** and u is the spatially autocorrelated residual, which has two components: the spatial autocorrelation (SA_j) in the residual and a random error component (ε_j).

For illustration, Fig. 1.5 describes the two processes that can be at the origin of a spatial structure (i.e. Model 1, induced spatial dependence, and Model 2, spatial autocorrelation) in a simplified system consisting of 4 ponds (large circles) connected by a stream; a light current is flowing from left to right. Five cases of increasing complexity are shown. In each case, circles in the upper row describe the values of an environmental variable **x** whereas the lower row concerns a response variable **y**, for example the abundances of a zooplankton species.

- Case 1 represents the null situation: there are no relationships among the values of **x** nor among those of **y** and no relationship between **x** and **y**. In a simulation program, the values of **y** corresponding to this case could be simulated as $y_j = \varepsilon_j$ where ε_j is a random normal deviate generated independently for each pond j.

- Case 2 is more interesting: it depicts functional dependence of the response variable **y** on the explanatory variable **x**. This is the classical environmental control model mentioned in the description of eq. 1.1 (Model 1). It can be implemented in simulations by equation $y_j = \beta_0 + \beta_x x_j + \varepsilon_j$ where β_0 is a constant and the functional dependence of **y** on **x** is represented by a regression parameter β_x. There is no spatial dependence (spatial correlation) among the values of **x** nor among those of **y** here.

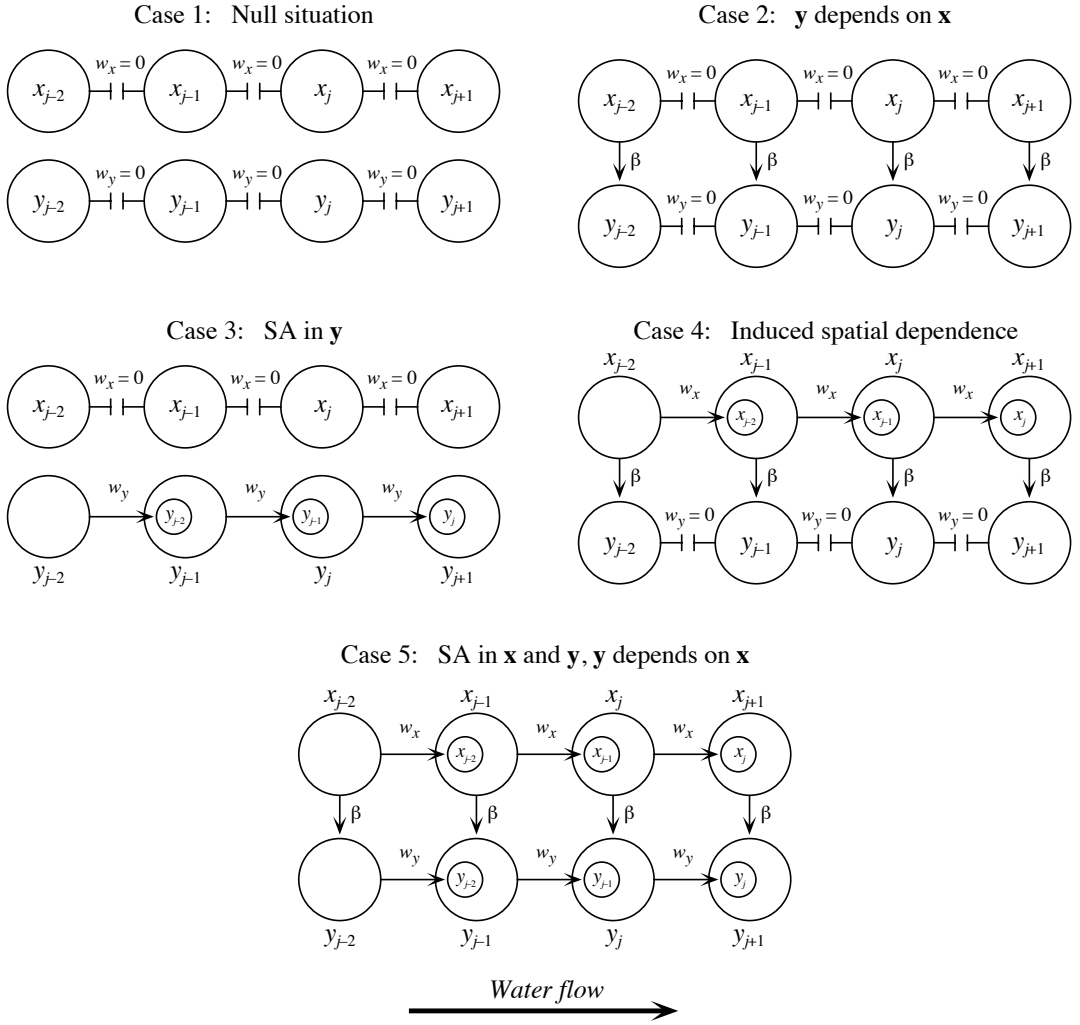


Figure 1.5 Five cases illustrating the origin of spatial structures through different types of relationships between an explanatory variable **x** and a response variable **y** observed across space. Of special interest are case 3 (spatial autocorrelation (SA) in **y**, Model 2) and case 4 (induced spatial dependence, Model 1). Modified from Fortin & Dale (2005, Chapter 5).

• Case 3 describes the process producing *spatial autocorrelation* (SA) in the response variable **y**. The arrows indicate that a random fraction of the zooplankton from pond $(j - 2)$ moves near the outflow stream and is transferred to pond $(j - 1)$ (the small circle inside the second large circle), and so on down the chain of ponds. There is no river-

like strong current moving water across the chain of ponds. As a result, zooplankton abundances in neighbouring ponds are more similar than expected in case 1. This similarity in the values of **y** due to proximity in space is called spatial autocorrelation. In numerical simulations, this process can be simulated by generating a random deviate in the first pond, $y_1 = \varepsilon_1$, and propagating it down the chain of ponds with the equation $y_j = w_y y_{j-1} + \varepsilon_j$. Equation 1.2 (Model 2) describes a similar process for sites on a 2-dimensional map with bidirectional exchanges between sites. In case 3, there is no autocorrelation in explanatory variable **x** and no functional dependence of **y** on **x**.

• Case 4 describes *induced spatial dependence*. A spatial structure is observed in **y** because that variable reflects the autocorrelated spatial structure of **x** through functional dependence of **y** on **x**. Two equations are necessary to represent this process in numerical simulations: the first describes the autocorrelation in **x** along the chain of ponds: $x_j = w_x x_{j-1} + \zeta_j$, and the second describes the spatial dependence of **y** on **x**: $y_j = \beta_0 + \beta_x x_j + \varepsilon_j$. A more general form for surfaces is eq. 1.1 (Model 1).

• Case 5 is the most complex as it combines the processes of cases 3 and 4. This is a situation often encountered in nature. There is spatial autocorrelation (SA) in **x** and in **y**, plus functional dependence of **y** on **x**. The equations describing this case in a simulation program would be: $x_j = w_x x_{j-1} + \zeta_j$ for the spatial autocorrelation (SA) in **x** and $y_j = \beta_0 + \beta_x x_j + w_y y_{j-1} + \varepsilon_j$ for the spatial dependence and autocorrelation in **y** (combination of Models 1 and 2). Methods described in Chapter 14 will show how to disentangle the two processes, using the fact that they often correspond to different spatial scales. More complex cases could be explored, e.g. the simultaneous autoregressive (AR) model and the conditional AR model (Cliff & Ord, 1981, Sections 6.2 and 6.3; Griffith, 1988, Chapter 4).

Figure 1.6 shows an example of simulated data corresponding to case 5. In the upper half of the figure, an environmental variable **x** is constructed on a map (400-point grid) as the sum of: a deterministic structure (here a unimodal distribution, upper-left map), plus spatial autocorrelation (SA) in **x**, plus random error at each point (ζ_j term in the first equation of case 5). The response variable **y** is constructed in the lower half of the figure. The effect of **x** on **y** is obtained by transporting the **x** surface (upper-right map), weighted by a regression coefficient $\beta_x = 0.3$ causing a change in the range of values in this example, to the lower-left corner where it becomes the first element in the construction of **y**. To that map, we add spatial autocorrelation (SA) in **y** and random error at each point (ε_j term in the second equation of case 5). The sum of these three surfaces produces the response variable **y** in the lower-right map. In this example, the **x** and **y** variables are sampled using a cross-shaped sampling design, represented in grey on the surface, containing 39 sampling units; any other sampling design appropriate to the study could have been used.

When there is a significant spatial structure in the data (Chapters 13 and 14), a hypothesis of induced spatial dependence (Model 1) can be examined by multiple regression (Subsection 10.3.3) or canonical analysis (Sections 11.1 and 11.2). Variation partitioning (Sections 10.3.5 and 11.1.11) and multiscale ordination (MSO,

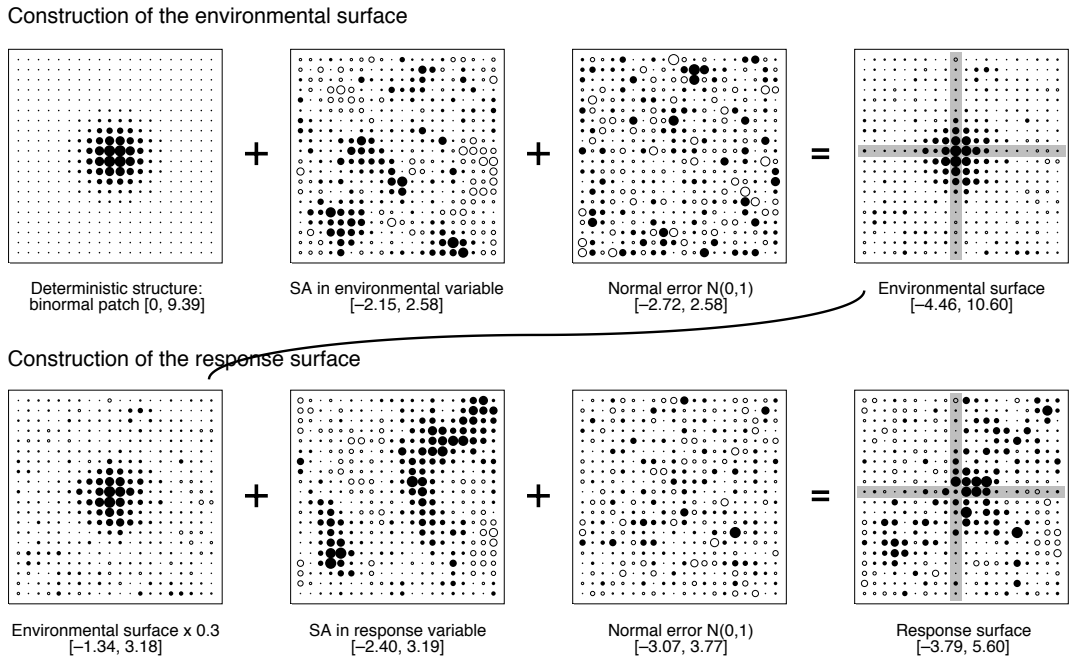


Figure 1.6 Construction of an explanatory (environmental) surface **x** and a response surface **y** in a simulation study. Each square is a bubble map of the study area. Large empty bubbles represent large negative values, and large filled bubbles, large positive values. The range of values in each map is shown in brackets underneath. The sampling design, shown in grey, is a cross with 39 sampled points in this example. Modified from Legendre *et al*. (2002, Fig. 1).

Section 14.4) can be used to determine whether or not the entire spatial structure detectable in the response data can be explained by the environmental variables (case 4) or if there remains an unexplained portion of spatial variation that would support a hypothesis of spatial autocorrelation in **y** (case 5).

Trend

A broad-scale spatial structure larger than the extent of the study area is called a *trend*. When there is a trend in the data, methods of spatial analysis detect spatial correlation due to the trend irrespective of the presence, or not, of finer-scaled sources of spatial correlation. In order to study the finer-scaled spatial structures, the trend must be removed from the data by an operation called *detrending*. One can then proceed with the analysis of the multi-scale spatial structure, for instance by spatial eigenfunction analysis (Sections 14.1 to 14.3). Linear detrending is done by regressing the response data on the geographic coordinates of the study sites (Section 13.2.1). Likewise, detrending must be done on time series before periodic or spectral analysis (Section 12.2).

Detrending

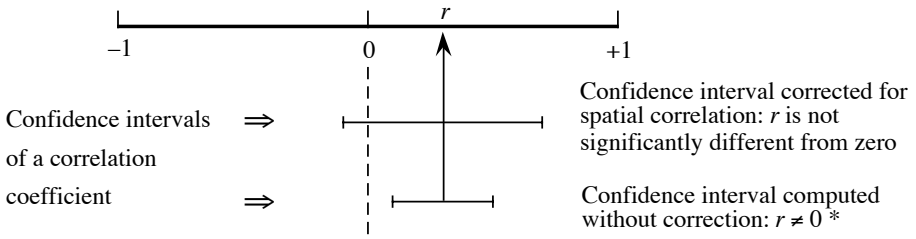


Figure 1.7 Effect of positive spatial correlation on tests of correlation coefficients; * means that the coefficient is (incorrectly) declared significantly different from zero in this example.

It is difficult to determine whether a given observed variable has been generated under Model 1 (eq. 1.1) or Model 2 (eq. 1.2). That question is further discussed in Subsection 13.1.2 in the case of gradients ("false gradients" and "true gradients") and in Chapter 14.

2 — Tests of significance in the presence of spatial correlation

Spatial correlation in a variable brings with it a statistical problem in the model-based approach (Section 1.0): it impairs the ability to perform standard statistical tests of hypotheses (Section 1.2). Let us consider an example of spatially autocorrelated data. The observed values of an ecological variable of interest — the abundances of a species for example — are most often influenced, at any given site, by the spatial distribution of the variable at surrounding sites, because of contagious biotic processes such as growth, dispersion, reproduction, and mortality. Make a first observation at site A and a second one at site B located near A. Since the ecological process is understood to some extent, one can assume that the data are spatially correlated. Using this assumption, one can anticipate to some degree the value of the variable at site B before the observation is made. Because the value at any one site is influenced by, and may be at least partly forecasted from the values observed at neighbouring sites, these values are not stochastically independent of one another.

The influence of spatial correlation on statistical tests may be illustrated using the correlation coefficient (Pearson r, Section 4.2). The problem lies in the fact that, when the two variables under study are positively spatially correlated, the confidence interval, estimated by the classical procedure around a Pearson correlation coefficient (whose calculation assumes independent and identically distributed error terms for all observations), is narrower than it is when calculated correctly, i.e. taking spatial correlation into account. The consequence is that one would declare too often that Pearson r coefficients are significantly different from zero (Fig. 1.7).

An important point is that in correlation or regression analysis, spatial correlation has a deleterious effect on tests of significance only when it is present in both variables. Simulation studies have shown that when spatial correlation was present in only one of the two variables, the test had a correct rate of type I error (Bivand, 1980; Legendre *et al.,* 2002). These simulations have also shown that deterministic spatial structures present in both variables have the same effect as spatial autocorrelation. For example, with a deterministic structure in one of the variables and spatial autocorrelation in the other, tests of significance had inflated rates of type I error.

All the usual statistical tests, nonparametric and parametric, have the same behaviour: in the presence of positive spatial correlation, the computed test statistics are too often declared significant. Negative spatial correlation may produce the opposite effect, for instance in analysis of variance (ANOVA).

The effects of spatial correlation on statistical tests may also be examined from the point of view of the *degrees of freedom*. As explained in Box 1.2, in classical statistical testing, one degree of freedom is counted for each independent observation, from which the number of estimated parameters is subtracted. The problem with spatially correlated data is their lack of independence or, in other words, the fact that new observations do not each bring with them one full degree of freedom, because the values of the variable at some sites give the observer some prior knowledge of the values the variable will take at other sites. The consequence is that new observations cannot be counted for one full degree of freedom. Since the size of the fraction they bring with them is difficult to determine, it is not easy to know what the proper reference distribution for the test should be. All that is known for certain is that positive spatial correlation at short distance distorts statistical tests (references in the next paragraph), and that this distortion is on the "liberal" side. This means that, when positive spatial correlation is present in the small distance classes, the usual statistical tests lead too often to the decision that Pearson or Spearman correlations, regression coefficients, or differences among groups are significant, when in fact they may not be.

This problem has been well documented in correlation analysis (Bivand, 1980; Cliff & Ord, 1981, §7.3.1; Clifford *et al.*, 1989; Haining, 2003, Section 8.2.1; Dutilleul, 1993a; Legendre *et al.,* 2002), linear regression (Cliff & Ord, 1981, §7.3.2; Chalmond, 1986; Griffith, 1988, Chapter 4; Haining, 1990, pp. 330-347), analysis of variance (Crowder & Hand, 1990; Legendre *et al.*, 1990, Legendre *et al.,* 2004), and tests of normality (Dutilleul & Legendre, 1992). The problem of estimating the confidence interval of the mean when the sample data are spatially correlated has been studied by Cliff & Ord (1975, 1981, §7.2) and Legendre & Dutilleul (1991).

When the presence of spatial correlation has been demonstrated, one may wish to remove the spatial dependency among observations; it would then be valid to compute the usual statistical tests. This might be done, in theory, by removing observations until spatial independence is attained; this solution is not recommended because it entails a net loss of information that was often costly to obtain. Another solution is detrending (Subsection 1.1.1) if the spatial structure is a broad-scale trend in the data; if spatial

Degrees of freedom — Box 1.2

Statistical tests of significance often call upon the concept of degrees of freedom. A formal definition is the following: "The degrees of freedom of a model for expected values of random variables is the excess of the number of variables [observations] over the number of parameters in the model" (Kotz & Johnson, 1982).

In practical terms, the number of degrees of freedom associated with a statistic is equal to the number of its independent components, i.e. the total number of components used in the calculation minus the number of parameters one had to estimate from the data before computing the statistic. For example, the number of degrees of freedom associated with a variance is the number of observations minus one (noted $\nu = n - 1$): n components $(x_i - \bar{x})$ are used in the calculation, but one degree of freedom is lost because the mean of the statistical population $(\bar{x})$ is estimated from the sample data; this is a prerequisite before estimating the variance.

There is a different t-distribution for each number of degrees of freedom. The same is true for the F and χ^2 families of distributions, for example. So, the number of degrees of freedom determines which statistical distribution, in these families (t, F, or χ^2), should be used as the reference for a given test of significance. Degrees of freedom are discussed again in Chapter 6 with respect to the analysis of contingency tables.

correlation is part of the process under study, however, this would amount to throwing out the baby with the water of the bath. It is better to analyse the spatially correlated data as such (Chapters 13 and 14), acknowledging the fact that spatial correlation in a variable may result from various causal mechanisms (physical or biological, see Subsection 1.1.1), acting simultaneously and additively.

The alternative for testing statistical significance is to modify the statistical method in order to take spatial correlation into account, as described in the following paragraphs. When such a correction is available, this approach is to be preferred if one assumes that spatial correlation is an intrinsic part of the ecological process to be analysed or modelled.

Corrected tests rely on modified estimates of the variance of the statistic, and on corrected estimates of the effective sample size and of the number of degrees of freedom. Simulation studies have been used to demonstrate the validity of the modified tests. In these studies, a large number of spatially correlated data sets are generated under the null hypothesis (e.g. for testing the difference between two means, pairs of observations are drawn at random from *the same* simulated, spatially

correlated statistical distribution, which corresponds to the null hypothesis of no difference between population means) and tested using the modified procedure; this experiment is repeated a large number of times to demonstrate that the modified testing procedure leads to the nominal rate of rejection of H_0, e.g. 0.05.

Cliff & Ord (1973) proposed a method for correcting the standard error of parameter estimates for the simple linear regression in the presence of spatial correlation. This method was extended to linear correlation, multiple regression, and *t*-test by Cliff & Ord (1981, Chapter 7: approximate solution) and to the one-way analysis of variance by Griffith (1978, 1987). Bartlett (1978) perfected a previously proposed method of correction for the effect of spatial correlation due to an autoregressive process in randomized field experiments, adjusting plot values by covariance on neighbouring plots before the analysis of variance; see also the discussion by Wilkinson *et al*. (1983) and the papers of Cullis & Gleeson (1991) and Grondona & Cressie (1991). Cook & Pocock (1983) suggested another method for correcting multiple regression parameter estimates by maximum likelihood, in the presence of spatial correlation. Using a different approach, Legendre *et al*. (1990) proposed a permutational method for the analysis of variance of spatially correlated data, in the case where the classification criterion is a division of a territory into nonoverlapping regions and one wants to test for differences among the means of these regions. Numerical simulations showed that, using this permutation method, ANOVA was insensitive to spatial correlation and effectively provided a test with a correct rate of type I error. They illustrated the method with an ecological application.

Clifford *et al*. (1989) tested the significance of the correlation coefficient between two spatial processes by estimating a modified number of degrees of freedom, using an approximation of the variance of the correlation coefficient computed from the data. Empirical results showed that their method worked fine for positive spatial correlation in large samples. Dutilleul (1993a) generalized the procedure and proposed an exact method to compute the variance of the sample covariance; the new method is valid for any sample size. In a simulation study, Legendre *et al*. (2002) showed that Dutilleul's modified *t*-test for the correlation coefficient effectively corrects for any kind of spatial correlation in the data: deterministic structures or spatial autocorrelation.

A general method to control for spatial correlation in tests of significance involving univariate or multivariate data was proposed by Peres-Neto & Legendre (2010). It involves partialling out the effect of spatial structures in partial regression (for univariate response data **y**) or partial canonical analysis (for multivariate response data **Y**). Spatial structures are represented in these analyses by spatial eigenfunctions. This method is described in Subsection 14.5.3.

Other major contributions to this topic are found in the literature on time series analysis, especially in the context of regression modelling. Important references are Cochrane & Orcutt (1949), Box & Jenkins (1976), Beach & MacKinnon (1978), Harvey & Phillips (1979), Chipman (1979), and Harvey (1981).

When methods specifically designed to handle spatial correlation are not available, it is sometimes possible to rely on permutation tests, where the significance is determined by random reassignment of the observations (Section 1.2). For some analytical situations, special permutational schemes have been developed that leave spatial correlation invariant; examples are found in Besag & Clifford (1989), Legendre *et al*. (1990) and ter Braak (1990, Section 8). The difficulty encountered in these complex problems is to design a permutation procedure that preserves the spatial or temporal correlation of the data.

The methods of clustering and ordination described in Chapters 8 and 9 to study ecological structures do not rely on tests of statistical significance. So, they are not affected by the presence of spatial correlation. The impact of spatial correlation on numerical methods will be stressed wherever appropriate.

3 — Classical sampling and spatial structure

Random or systematic sampling designs have been advocated as a way of controlling the dependence among observations (Cochran, 1977; Green, 1979; Scherrer, 1982). This was then believed to be a necessary and sufficient safeguard against violations of the independence of errors, which is a basic assumption of classical statistical tests. It is adequate, of course, when one is trying to estimate the parameters of a well-localized statistical population, for example the total number of trees in a forest plot. In such a case, a random or systematic sample is suitable to obtain unbiased estimates of the parameters since, *a priori*, each point has the same probability of being included in the sample. Of course, the variance and, consequently, also the standard error of the mean increase if the distribution is patchy, but their estimates remain unbiased.

Even with random or systematic allocation of observations through space, observations may retain some degree of spatial dependence if the average distance between first neighbours is shorter than the zone of spatial influence of the underlying ecological phenomenon. In the case of broad-scale spatial gradients, no point is far enough to lie outside this zone of spatial influence. Correlograms and variograms (Chapter 13), combined with maps, are used to assess the magnitude and shape of spatial correlation present in data sets.

Classical books such as Cochran (1977) adequately describe the rules that should govern sampling designs. Such books, however, only emphasize design-based inference (Section 1.0) and do not discuss the influence of spatial correlation on sampling designs. At the present time, most of the literature on this subject is from the field of geostatistics, where important references are: David (1977, Ch. 13), McBratney & Webster (1981), McBratney *et al*. (1981), Webster & Burgess (1984), Borgman & Quimby (1988), and François-Bongarçon (1991). In ecology, see Legendre *et al*. (2002).

Ecologists interested in designing field experiments should read the paper of Dutilleul (1993b), who discusses how to accommodate an experiment to spatially

heterogeneous conditions. Legendre *et al*. (2004) have also shown how one can effectively control for the effect of spatial correlation by the design of the experiment, and which experimental designs lead to tests of significance that have greater power.

Heterogeneity

The concept of spatial heterogeneity is discussed at some length in the multi-author book edited by Kolasa & Pickett (1991), in the review paper of Dutilleul & Legendre (1993), in the book of Dutilleul (2011), and in Section 13.0.

1.2 Statistical testing by permutation

The role of a statistical test is to decide whether some *parameter* of the reference population may take a value assumed by hypothesis, given the fact that the corresponding statistic, whose value is estimated from a sample of objects, may have a somewhat different value. A *statistic* is any quantity that may be calculated from the data and is of interest for the analysis (examples below); in tests of significance, a statistic is called *test statistic* or *test criterion*. The assumed value of the statistic, in the reference population, is given by the statistical null hypothesis (written H_0), which translates the biological null hypothesis into numerical terms; it often negates the existence of the phenomenon that the scientist is hoping to evidence. The reasoning behind statistical testing directly derives from the scientific method; it allows the confrontation of experimental or observational findings to intellectual constructs that are called hypotheses, with the explicit purpose of determining whether or not the data support the null hypothesis (see below) at some predetermined confidence level.

Statistic

Testing is the central step of inferential statistics. It allows one to generalize the conclusions of statistical estimation to the reference population from which the observations have been drawn and that they are supposed to represent. Within that context, the problem of multiple testing is too often ignored (Box 1.3). Another legitimate section of statistical analysis, called descriptive statistics, does not rely on testing. The methods of clustering and ordination described in Chapters 8 and 9, for example, are descriptive multidimensional statistical methods. The interpretation methods described in Chapters 10 and 11 may be used in either descriptive or inferential mode.

1 — Classical tests of significance

Consider, for example, a correlation coefficient (which is the statistic of interest in correlation analysis) computed between two variables (Section 4.2). When inference to the statistical population is sought, the null hypothesis is often that the value of the correlation parameter (ρ, rho) is zero in the statistical population; the null hypothesis may also be that ρ has some value other than zero, value provided by the ecological hypothesis. To judge of the validity of the null hypothesis, the only information available is an *estimate* of the correlation coefficient, *r*, obtained from a sample of objects drawn from the statistical population. (Whether the observations adequately

Null hypothesis

Multiple testing Box 1.3

When several tests of significance are carried out simultaneously, the probability of a type I error becomes larger than the nominal value α. Consider for example a correlation matrix among 5 variables: 10 tests are carried out simultaneously. For randomly generated data, there is a probability $p = 0.401$ (computed from the binomial distribution) of rejecting the null hypothesis at least once over 10 tests at the nominal $\alpha = 0.05$ level; this is called the *familywise* or *experimentwise* error rate. So, when conducting multiple tests, one should perform a global test of significance to determine whether there is any significant value at all in the set.

A general approach is the Bonferroni (1935) correction for k independent tests: replace the significance level, say $\alpha = 0.05$, by an adjusted level $\alpha' = \alpha/k$, and compare the probabilities p_i to α'. This is equivalent to adjusting individual p-values p_i to $p'_i = kp_i$ and comparing p'_i to the unadjusted significance level α. In the Sidák (1967) correction, α is replaced by an adjusted level $\alpha' = 1 - (1 - \alpha)^{1/k}$; or one can compare individual corrected values $p'_i = 1 - (1 - p_i)^k$ to the original α significance level. Although the Bonferroni and Sidák methods are appropriate to test the null hypothesis for the whole set of simultaneous hypotheses (i.e. reject H_0 for the family of k hypotheses if the smallest unadjusted p-value in the set is less than or equal to α'), these two methods are overly conservative and often lead to rejecting too few individual hypotheses in the set k, resulting in tests with low power.

Several alternatives have been proposed in the literature; see Wright (1992) for a review. For non-independent tests, Holm's procedure (1979) is nearly as simple to carry out as the Bonferroni adjustment and it is much more powerful, leading to rejecting the null hypothesis more often. It is computed as follows. (1) Order the p-values from left to right so that $p_1 \leq p_2 \leq \ldots \leq p_i \ldots \leq p_k$. (2) Compute adjusted probability values $p'_i = (k - i + 1)p_i$; adjusted probabilities may be larger than 1. (3) Proceeding from left to right, if an adjusted p-value in the ordered series is smaller than the one occurring at its left, make the smallest equal to the largest one. (4) Compare each adjusted p'_i to the unadjusted α significance level and make the statistical decision. The procedure could be formulated in terms of successive corrections to the α significance level, instead of adjustments to individual probabilities.

An even more powerful solution is that of Hochberg (1988), which has the desired overall ("experimentwise") error rate α only for independent tests, i.e. tests that do not share part of their data (Wright, 1992). This procedure is identical to Holm's except for step 3: proceeding this time from right to left, if an adjusted p-value in the series is smaller than the one at its left, make the largest equal to the smallest value. Because the adjusted p-values form a nondecreasing series, both procedures present the properties (1) that a hypothesis in the ordered series cannot be rejected unless all previous hypotheses in the series have also been rejected and (2) that equal p-values receive equal adjusted p-values. Hochberg's method has the further characteristic that no adjusted p-value can be larger than the largest unadjusted p-value or exceed 1. More complex and powerful procedures are described by Wright (1992).

Fisher's *combined probability test* allows one to combine probabilities p_i from k tests computed on independent data sets (meta-analysis). The value $-2\sum \log_e(p_i)$ is distributed as χ^2 with $2k$ degrees of freedom if H_0 is true in all k tests (Fisher, 1954; Sokal & Rohlf, 1995).

represent the statistical population is another question, for which the readers are referred to the literature on sampling design.) We know, of course, that a sample is quite unlikely to produce a parameter estimate that is exactly equal to the value of the parameter in the statistical population. A statistical test tries to answer the following question: given a hypothesis stating, for example, that $\rho = 0$ in the statistical population and the fact that the estimated correlation is, say, $r = 0.2$, is it justified to conclude that the difference between 0.2 and 0.0 is due to sampling variation?

Pivotal statistic

The choice of the statistic to be tested depends on the problem at hand. For instance, in order to find whether two samples may have been drawn from the same statistical population or from populations with equal means, one would choose a statistic measuring the difference between the two sample means ($\bar{x}_1 - \bar{x}_2$) or, preferably, a *pivotal* form like the usual *t*-statistic used in such tests; a pivotal statistic has a distribution under the null hypothesis that remains the same for any value of the measured effect (here, $\bar{x}_1 - \bar{x}_2$) because the difference of means statistic is divided by its standard error. In the same way, the slope of a regression line is described by the slope parameter of the linear regression equation, which is assumed, under the null hypothesis, to be either zero or some other value suggested by ecological theory. The test statistic describes the difference between the observed and hypothesized values of the slope; the pivotal form of this difference is a *t* or *F*-statistic.

Alternative hypothesis

Another aspect of a statistical test is the alternative hypothesis (H_1), which is also imposed by the ecological problem at hand. H_1 is the opposite of H_0, but there may be several statements that represent some opposite of H_0. In correlation analysis for instance, if one is satisfied to determine that the correlation coefficient in the reference population (ρ) is significantly different from zero in either the positive or the negative direction, meaning that *some* linear relationship exists between two variables, then a *two-tailed* alternative hypothesis is stated about the value of the parameter in the statistical population: $\rho \neq 0$. On the contrary, if the ecological phenomenon underlying the hypothesis imposes that a relationship, if present, should have a given sign, one formulates a *one-tailed* hypothesis. For instance, studies on the effects of acid rain are motivated by the general paradigm that acid rain, which lowers the pH, has a negative effect on terrestrial and aquatic ecosystems. In a study of the correlation between pH and diversity, one would formulate the following hypothesis H_1: pH and diversity are positively correlated (i.e. low pH is associated with low diversity; H_1: $\rho > 0$). Other situations would call for a different alternative hypothesis, symbolized by H_1: $\rho < 0$.

The expressions *one-tailed* and *two-tailed* refer to the fact that, in a two-tailed test, one would look in both tails of the reference statistical distribution for values as extreme as, or more extreme than the observed value of the statistic (i.e. the one computed from the actual data). In a correlation study for instance, where the reference distribution (*t*) for the test statistic is symmetric about zero, the probability of the data under the null hypothesis in a two-tailed test is given by the proportion of values in the *t*-distribution that are, *in absolute value*, as large as, or larger than the *absolute value* of the observed *t*-statistic. In a one-tailed test, one would look only in the tail corresponding to the sign given by the alternative hypothesis. For instance, for a test in

the right-hand tail (H_1: $\rho > 0$), look for the proportion of values in the *t*-distribution that are as large as or larger than the *signed value* of the observed *t*-statistic.

In standard statistical tests, the *test statistic* computed from the data is referred to one of the usual statistical distributions printed in books or computed by some appropriate computer software; the best-known are the *z*, *t*, *F* and χ^2 distributions. This, however, can only be done if certain assumptions are met by the data, depending on the test. The most commonly encountered are the assumptions of normality of the variable(s) in the reference population, normality of the regression residuals, homoscedasticity (Box 1.4), and independence of the observations (Box 1.1). Refer to Siegel (1956, Chapter 2), Siegel & Castellan (1988, Chapter 2), or Snedecor & Cochran (1967, Chapter 1), for concise yet clear classical exposés of the concepts related to statistical testing.

2 — Permutation tests

The method of *permutation*, also called *randomization*, is a very general approach to testing statistical hypotheses. Following Manly (1997), permutation and randomization are considered synonymous in the present book, although *permutation* may also be considered to be the technique by which the principle of *randomization* is applied to data during permutation tests. Other points of view are found in the literature. For instance, Edgington (1995) considers that a randomization test is a permutation test based on randomization, by opposition to restricted permutations in a loop for time series or by toroidal shift for grid data on a map. A different although related meaning of *randomization* refers to the random assignment of replicates to treatments in experimental designs.

Randomization

Permutation testing can be traced back to at least Fisher (1935, Chapter 3). Instead of comparing the actual value of a test statistic to a standard statistical distribution, the reference distribution is generated from the data themselves, as described below; other randomization methods are mentioned at the end of the present section. Permutation provides an efficient approach to testing when the data do not conform to the distributional assumptions of the statistical method one wants to use (e.g. normality). Permutation testing is applicable to very small samples, like nonparametric tests. It *does not,* however, solve problems of independence of the observations, including those caused by spatial correlation. Nor does the method solve distributional problems that are linked to the hypothesis subjected to a test*. Permutation remains the method of choice to test novel or other statistics whose distributions are poorly known.

* For instance, when studying the differences among sample means (two groups: *t*-test; several groups: *F*-test of ANOVA), the classical Behrens-Fisher problem (Robinson, 1982) reminds us that two null hypotheses are tested simultaneously by these methods, i.e. equality of the means and equality of the variances. Testing the *t* or *F*-statistics by permutations does not change the dual aspect of the null hypothesis; in particular, it does not allow one to unambiguously test the equality of the means without checking first the equality of the variances using another, more specific test (two groups: *F* ratio; several groups: Bartlett's test of equality of variances).

Furthermore, results of permutation tests are valid even with observations that are not a random sample of some statistical population; this point is further discussed in Subsection 1.2.4. Edgington (1995) and Manly (1997) have written excellent introductory books about the method. A short account is given by Sokal & Rohlf (1995) who use the expression "randomization test". Permutation tests are used in several chapters of the present book.

The speed of modern computers would allow users to perform any statistical test using the permutation method. The chief advantage is that one does not have to worry about the distributional assumptions of classical testing procedures; the disadvantage is the extra computer time required to actually perform a large number of permutations, each one being followed by recomputation of the test statistic. Permutation tests are fairly easy to program and are increasingly available in computer packages. As an example, let us consider the situation where the significance of a correlation coefficient between two variables, $\mathbf{x}_1$ and $\mathbf{x}_2$, is to be tested.

Hypotheses

- H_0: The correlation between the variables in the reference population is zero ($\rho = 0$).
- For a two-tailed test, H_1: $\rho \neq 0$.
- Or for a one-tailed test, either H_1: $\rho > 0$, or H_1: $\rho < 0$, depending on the ecological hypothesis.

Test statistic

- Compute the Pearson correlation coefficient r. Calculate the pivotal statistic $t = \sqrt{n-2}\,[r/\sqrt{1-r^2}]$ (eq. 4.13; n is the number of observations) and use it as the observed value of the test statistic in the remainder of the test.

In this specific case, the permutation test results would be the same using either r or t as the test statistic, because t is a monotonic function of r for any constant value of n; r and t are "equivalent statistics for permutation tests", *sensu* Edgington (1995). This is not always the case. For example, when testing a partial regression coefficient in multiple regression, the test should not be based on the distribution of permuted partial regression coefficients because they are not monotonic to the corresponding partial t-statistics. The partial t should be preferred because it is pivotal and, hence, it is expected to produce correct type I error.

Considering a pair of equivalent test statistics, one could choose the statistic which is the simplest to compute if calculation time would otherwise be longer in an appreciable way. This is not the case in the present example: calculating t involves a single extra line in the computer program compared to r. So the test is conducted using the usual t-statistic.

Distribution of the test statistic

The argument invoked to construct a null distribution for the statistic is that, if the null hypothesis is true, all possible pairings of the two variables are equally likely to occur. The pairing found in the observed data is just one of the possible, equally likely pairings, so that the value of the test statistic for the unpermuted data should be typical, i.e. located in the central part of the permutation distribution.

- It is always the null hypothesis that is subjected to testing. Under H_0, the rows of $\mathbf{x}_1$ are exchangeable with one another if the rows of $\mathbf{x}_2$ are fixed, or conversely, and the observed pairing of $\mathbf{x}_1$ and $\mathbf{x}_2$ values is due to chance alone; accordingly, any value of $\mathbf{x}_1$ could have been paired with any value of $\mathbf{x}_2$.

- A realization of H_0 is obtained by permuting at random the values of $\mathbf{x}_1$ while holding the values of $\mathbf{x}_2$ fixed, or the opposite (which would produce, likewise, a random pairing of values). Recompute the value of the correlation coefficient and the associated *t*-statistic for the randomly paired vectors $\mathbf{x}_1$ and $\mathbf{x}_2$, obtaining a value t^*.

- Repeat this operation a large number of times (say, 999 or 9999 times). The different permutations produce a set of values t^* obtained under H_0.

- Add to these the observed value of the *t*-statistic, computed for the unpermuted vectors. Since H_0 is being tested, this value is considered to be one of those that could be obtained under H_0 and, consequently, it should be added to the distribution of *t* values (Hope, 1968; Edgington, 1995; Manly, 1997). Together, the unpermuted and permuted values form an estimate of the sampling distribution of *t* under H_0, which will be used as the reference distribution in the next step.

Statistical decision

- As in any other statistical test, the decision is made by comparing the observed value of the test statistic (*t*) to the reference distribution obtained under H_0. If the observed value of *t* is typical of the values obtained under the null hypothesis (which states that there is no relationship between $\mathbf{x}_1$ and $\mathbf{x}_2$), H_0 cannot be rejected; if it is unusual, being too extreme to be considered a likely result under H_0, H_0 is rejected and the alternative hypothesis is considered to be a more likely explanation of the data.

- Compute the associated p-value, which is the proportion of values in the reference distribution that are as extreme as, or more extreme than the observed value of the test statistic. The p-value is either computed from the reference distribution obtained by permutations, or found in a table of the appropriate statistical distribution. The p-value is a statement about the probability of obtaining a result as extreme as, or more extreme than the one actually obtained from the sample data, assuming that H_0 is true for the reference population. Researchers often write in short that it is *the probability of the data under the null hypothesis*. Fisher (1954) saw the p-value as a measure of the strength of evidence against the null hypothesis; the smaller the p-value, the stronger the evidence against H_0.

Significance level

• Compare the p-value to a predetermined significance level α. Following the Neyman-Pearson (or *frequentist*) approach (Neyman & Pearson, 1966), one rejects H_0 if $p \leq \alpha$, and does not reject it if $p > \alpha$. Or one can use the Fisher approach: Fisher left the interpretation of the p-value and the ensuing statistical decision to the researcher.

3 — Numerical example

Let us consider the following case of two variables observed over 10 objects:

$\mathbf{x}_1$	–2.31	1.06	0.76	1.38	–0.26	1.29	–1.31	0.41	–0.67	–0.58
$\mathbf{x}_2$	–1.08	1.03	0.90	0.24	–0.24	0.76	–0.57	–0.05	–1.28	1.04

These values were drawn at random from a positively correlated bivariate normal distribution, as shown in Fig. 1.8a. Consequently, they would be suitable for parametric testing. So, it is interesting to compare the results of a permutation test to the usual parametric *t*-test of the correlation coefficient. The statistics and associated probabilities for this pair of variables, for $\nu = (n - 2) = 8$ degrees of freedom, are:

$r = 0.70156$, $t = 2.78456$, $n = 10$:
prob (one-tailed) = 0.0119, prob (two-tailed) = 0.0238.

There are $10! = 3.6288 \times 10^6$ possible permutations of the 10 values of variable $\mathbf{x}_1$ (or $\mathbf{x}_2$). Here, 999 of these permutations were generated using a random permutation algorithm; they represent a random sample of the 3.6288×10^6 possible permutations. The computed values for the test statistic (t) between permuted $\mathbf{x}_1$ and fixed $\mathbf{x}_2$ have the distribution shown in Fig. 1.8b; the observed value, $t = 2.78456$, has been added to this distribution. The permutation results are summarized in the following table, where $|t|$ is the (absolute) observed value of the *t*-statistic ($|t| = 2.78456$) and t^* is a value obtained after permutation. The absolute value of the observed t is used in the following table to make it a general example since there are cases where t is negative.

	$t^* < -\lvert t \rvert$	$t^* = -\lvert t \rvert$	$-\lvert t \rvert < t^* < \lvert t \rvert$	$t^* = \lvert t \rvert$	$t^* > \lvert t \rvert$
Statistic t	8	0	974	1[†]	17

[†] This count corresponds to the observed t value that was added to the reference distribution.

For a one-tailed test (in the right-hand tail in this case, since H_1: $\rho > 0$), one counts how many values in the permutational distribution of the statistic are equal to, or larger than, the observed value ($t^* \geq t$; there are $1 + 17 = 18$ such values in this case). This is the only one-tailed hypothesis worth considering, because the objects are known in this case to have been drawn from a positively correlated distribution. A one-tailed test in the left-hand tail (H_1: $\rho < 0$) would be based on how many values in the permutational distribution are equal to, or smaller than, the observed value ($t^* \leq t$, which are $8 + 0 + 974 + 1 = 983$ in the example). For a two-tailed test, one counts all values that are as extreme as, or more extreme than the observed value *in both tails of the distribution* ($|t^*| \geq |t|$, which are $8 + 0 + 1 + 17 = 26$ in the example).

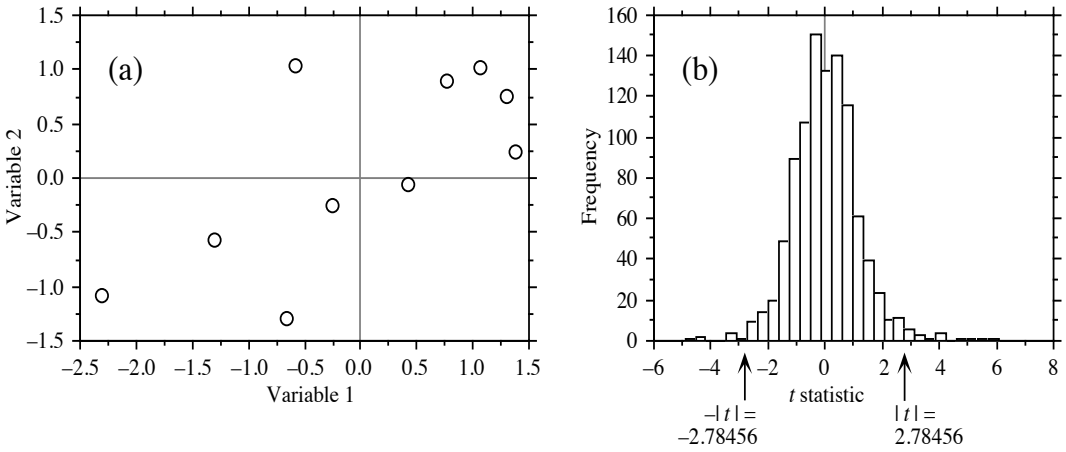


Figure 1.8 (a) Scatter diagram of the 10 points of the numerical example with respect to variables $\mathbf{x}_1$ and $\mathbf{x}_2$. (b) Frequency histogram of the (1 + 999) permutation results (t-statistics for correlation coefficients). The observed value of t, $|t| = 2.78456$, is shown, as well as $-|t| = -2.78456$.

Probabilities associated with these distributions are computed as follows, for a one-tailed and a two-tailed test (results using the r statistic would be the same):

One-tailed test [H_0: $\rho = 0$; H_1: $\rho > 0$]:
prob ($t^* \geq 2.78456$) = (1 + 17)/1000 = 0.018

Two-tailed test [H_0: $\rho = 0$; H_1: $\rho \neq 0$]:
prob($|t^*| \geq 2.78456$) = (8 + 0 + 1 + 17)/1000 = 0.026

Note how similar the permutation results are to the results obtained from the classical test, which referred to a table of the Student t-distribution. The observed difference is partly due to the small number of pairs of points ($n = 10$) sampled at random from the bivariate normal distribution, with the consequence that the data set does not quite conform to the hypothesis of normality. It is also due, to a certain extent, to the use of only 999 permutations, sampled at random among the 10! possible permutations.

4 — Remarks on permutation tests

In permutation tests, the reference distribution against which the statistic is tested is obtained by randomly permuting the data under study, without reference to any statistical population. The test is valid as long as the reference distribution has been generated by a procedure related to a null hypothesis that makes sense for the problem at hand, irrespective of whether or not the data set is representative of a larger statistical population. This is the reason why the data do not have to be a random

sample from some larger statistical population. The only information the permutation test provides is whether the pattern observed in the data is likely, or not, to have arisen by chance. For this reason, one may think that permutation tests are not as "good" or "interesting" as classical tests of significance because they might not allow one to infer conclusions that apply to a statistical population.

A more pragmatic view is that the conclusions of permutation tests may be generalized to a reference population if the data set is a random sample of that population. Otherwise, they allow one to draw conclusions only about the particular data set, measuring to what extent the value of the statistic is "usual" or "unusual" with respect to the null hypothesis implemented in the permutation procedure. Edgington (1995) and Manly (1997) further argue that data sets are very often not drawn at random from statistical populations, but simply consist of observations that happen to be available for the study. The generalization of results, in classical as well as permutation tests, depends on the degree to which the data were actually drawn at random, or are equivalent to a sample drawn at random, from a reference population.

Complete permutation test

For small data sets, one can compute all possible permutations in a systematic way and obtain the complete permutation distribution of the statistic; an *exact* or *complete permutation test* is obtained. For large data sets, only a sample of all possible permutations may be computed because there are too many. When designing a *sampled permutation test*, it is important to make sure that one is using a *uniform random generation algorithm*, capable of producing all possible permutations with equal probabilities (Furnas, 1984). Computer programs use procedures that produce random permutations of the data; these in turn call the 'Random' function of computer languages. Such a procedure is described in Section 5.8 of Manly's book (1997). Random permutation functions are available in subroutine libraries and in R.

Sampled permutation test

The case of the correlation coefficient has shown how the null hypothesis guided the choice of an appropriate permutation procedure, capable of generating realizations of this null hypothesis. A permutation test for the difference between the means of two groups would involve random permutations of the objects between the two groups instead of random permutations of one variable with respect to the other. The way of permuting the data depends on the null hypothesis to be tested.

Some tests may be reformulated in terms of some other tests. For example, the *t*-test of equality of means is equivalent to a test of the correlation between the vector of observed values and a vector assigning the observations to group 1 or 2. The same value of *t* and probability (classical or permutational) are obtained using both methods.

Simple statistical tests such as those of correlation coefficients or differences between group means may be carried out by permuting the original data, as in the example above. Problems involving complex relationships among variables may require permuting the residuals of some *model* instead of the raw data; *model-based permutation* is discussed in Subsection 11.1.8. The effect of a nominal covariable may be controlled for by *restricted permutations*, limited to the objects within the groups

Restricted permutations

defined by the covariable. This method is discussed in detail by Manly (1997). Applications are found in Brown & Maritz (1982; restrictions within replicated values in a multiple regression) and in Sokal *et al*. (1987; Mantel test), for instance.

In sampled permutation tests, adding the observed value of the statistic to the distribution has the effect that it becomes impossible for the test to produce no value "as extreme as, or more extreme than the observed value", as the standard expression goes. This way of computing the probability is biased, but it has the merit of being statistically valid (Edgington, 1995, Section 3.5). The precision of the probability estimate is the inverse of the number of permutations performed; for instance, after (999 + 1) permutations, the precision of the probability statement is 0.001.

How many permutations?

The number of permutations one should perform is always a trade-off between precision and computer time. The more permutations the better, since probability estimates are subject to error due to sampling the population of possible permutations (except in the rare cases of complete permutation tests), but it may be tiresome to wait for the permutation results when studying large data sets. Jackson & Somers (1989) recommend to compute 10000 to 100000 permutations in order to ensure the stability of the probability estimates in Mantel tests (Subsection 10.5.1). The following recommendation can be made. In exploratory analyses, 500 to 1000 permutations may be sufficient as a first contact with the problem. If the computed probability is close to the preselected significance level, run more permutations. In any case, use more permutations (e.g. 10000) for final results submitted for publication.

Interestingly, tables of critical values in nonparametric statistical tests for small samples are based on permutations. The authors of these tables computed how many cases can be found, in the complete permutation distribution, that are as extreme as, or more extreme than the computed value of the statistic. Hence, probability statements obtained from small-sample nonparametric tests are exact probabilities (Siegel, 1956).

Monte Carlo

Named after the city that hosts the famous casino in the principality of Monaco, Monte Carlo methods use random numbers to study either real data sets or the behaviour of statistical methods through simulations. Permutation tests are Monte Carlo methods because they use random numbers to randomly permute data. Other such methods are based on computer-intensive resampling. Among these are the jackknife (Tukey, 1958; Sokal & Rohlf, 1995) and the bootstrap (Efron, 1979; Efron & Tibshirani, 1993; Manly, 1997). In the latter methods, the values used in each iteration to compute a statistic are a subsample of the original data. In the jackknife, each subsample leaves out one of the original observations and sampling is done *without replacement*. In the bootstrap, each subsample is obtained by resampling the original sample *with replacement*; the justification is that resampling the original sample approximates a resampling of the original population.

Jackknife
Bootstrap

As an exercise, readers are invited to figure out how to perform a permutation test for the difference between the means of two groups of objects on which a single variable has been measured, using the *t*-statistic, and create a permutational *t*-test R

function*. Other types of permutation tests are discussed in Sections 5.4, 7.3, 8.9, 10.2, 10.3, 10.5, 10.6, 11.1, 11.4, 11.5, 12.6, 13.1 and 13.3.

1.3 Computer programs and packages

Processing complex ecological data sets almost always requires the use of computers, as much for the amount of data to be processed as for the fact that the operations to be performed are complex and often repetitious.

Powerful statistical packages such as SAS®, SPSS®, Statistica® and others are commercially available for general statistical analysis. Many other programs are either commercially or freely available on the Web pages of researchers or research institutions; some of these programs will be mentioned in *Software* sections in the following chapters.

This book will pay special attention to statistical functions available in the R language, which was developed in 1990 by Ross Ihaca and Robert Gentleman at the University of Auckland. R is a dialect of the S language. The S freeware was created in 1976 by John Chambers and colleagues at *AT&T Bell Laboratories*. R became freeware in 1995 and an international project in 1997. Its source code is freely available under the GNU General Public License. For most users, R is a powerful environment to carry out statistical analyses. R is also a programming language that allows scientists to easily write new functions. For computationally-intensive tasks, R functions can call compiled code written in C, C++ and Fortran.

The main features of the R language are described on the Web page http://en.wikipedia.org/wiki/R_(programming_language). Other computer languages such as S-PLUS® (a commercial implementation of S) and MATLAB® offer features comparable to R; however, they are not free.

The use of R has grown tremendously among researchers during the past 15 years and it has become a *de facto* standard for software development and computing in most fields of science. The fact that it is free and multi-platform explains in part its success: functions can be used in the same way on all major personal computer operating systems (presently Microsoft Windows, Mac OS X, and Linux). R is also available for a wide variety of Unix platforms. The other part of the explanation holds in the fact that the *R Development Core Team* has encouraged contributions from the community of users and methods developers, who have joined in the movement wholeheartedly. As a result, thousands of R packages are now available on the *Comprehensive R Archive Network* (CRAN) main site (http://cran.r-project.org/) and on mirror sites.

* Readers can compare their solution to the R function ***t.perm()*** available on the Web page http://numericalecology.com/rcode.

Thousands more packages and individual functions are distributed by researchers on their Web pages or are attached to scientific papers describing new numerical methods. All functions found in R packages come with documentation files, called by the ***help()*** function or by a question mark, and they are all presented in the same format.

There are many reference books published about the R language and its application to various fields. A good starting point to learn about R is *The R book* of Crawley (2007). The Venables & Ripley (2002) textbook is the acknowledge reference for many functions found in the R and S languages. In several of the following chapters, we will refer to the book *Numerical ecology with R* by Borcard *et al*. (2011), which was written as a companion to the 1998 and the present editions of *Numerical ecology*. The Borcard *et al*. (2011) book is of particular interest to readers who wish to implement the methods described in this book using available R software.

Here is an example of how R packages and functions will be referred to in this book: package VEGAN, function ***rda()***. The parentheses after function names contain data file names and other parameters necessary to run functions.

Ecologists should bear in mind that easy computation has two pitfalls: the fact that computations are done and results are produced does not ensure (1) that the data satisfy the conditions required by the method, or (2) that the results produced by the computer are interpreted correctly in ecological terms. This book provides colleagues with the theoretical and practical information they need to avoid these pitfalls.

1.4 Ecological descriptors

Descriptor
Variable

Any ecological study, classical or numerical, is based on *descriptors*. In the present text, the terms *descriptor* and *variable* will be used interchangeably. These refer to the attributes, or characters (also called items in the social sciences, and profiles or features in the field of pattern recognition) used to describe or compare the *objects of the study*.

Object

The *objects* that ecologists compare are the sites, quadrats, observations, sampling units, individual organisms, or subjects; these are defined *a priori* by the sampling design, before making the observations (Section 2.1). Observation units are often called "samples" by ecologists. The term *sample* is only used in its statistical sense in this book; it refers to *a set of observations* resulting from a sampling action or campaign. Objects may be called individuals or OTUs (*Operational taxonomic units*) in numerical taxonomy, OGUs (*Operational geographic units*) in biogeography, cases, patterns or items in the field of pattern recognition, etc.

The descriptors, used to describe or qualify the objects, are the physical, chemical, ecological, or biological characteristics of these objects that are of interest for the study. In particular, biological species are *descriptors* of sites for ecologists; in (numerical) taxonomy on the contrary, the species are the *objects* of the study, and the sites where the species are observed or collected may be used by the taxonomist as

descriptors of the species. It all depends on the variable, defined *a priori,* that specifies the objects of a study. In ecology, sites are compared using the species they contain, there being no possibility of choosing the species, whereas taxonomists compare populations or other taxonomic entities obtained from a number of different sites.

Descriptor

A *descriptor is a law of correspondence established by the researcher* to describe and compare, on the same basis, all the objects of the study. This definition applies to all types of descriptors discussed below (Table 1.2). The fundamental property of a descriptor, as understood in the present book, is that it distributes the objects among non-overlapping states. Each descriptor must, therefore, operate like *a law that associates with each object in the group under study one and only one element of a set of distinguishable states that belong to the descriptor.*

Descriptor state

The *states* that constitute a descriptor *must necessarily be mutually exclusive*. In other words, two different states of the same descriptor must not be applicable to the same object. Descriptors, on the contrary, do not have to be independent of one another (see Box 1.1: independent descriptors). In Chapter 6, it will be seen that the information contained in one descriptor may partially or totally overlap with the information in another descriptor.

1 — Mathematical types of descriptors

The states that form a descriptor — i.e. the qualities observed or determined on the objects — may be of a qualitative or quantitative nature, so that descriptors may be classified into several types. In ecology, a descriptor may be biological (presence, abundance, or biomass of different species), physical, chemical, geological, geographical, temporal, climatic, etc. Table 1.2 presents a classification of descriptors according to their mathematical types. That classification is independent of the particular discipline to which the descriptors belong. The mathematical type of a descriptor determines the type of numerical processing that can be applied to it. For example, parametric correlations (Pearson's r) may be calculated between quantitative descriptors, while nonparametric correlations (such as Kendall's τ) may be used on ordered but not necessarily quantitative descriptors, as long as their relationship is monotonic. To measure the dependence among descriptors that are not in monotonic relationship, or among qualitative descriptors, requires the use of other methods based on contingency tables (Chapter 6). Subsection 1.5.7 will show how descriptors of different mathematical types can be made compatible, in order to use them together in ecological studies.

Quantitative descriptors, which are the most usual type in ecology, are found at the bottom of Table 1.2. They include all descriptors of abundance and other quantities that can be plotted on a continuous axis of real numbers. They are called quantitative, or *metric* (Falconer, 1960), because they measure changes in a phenomenon in such a way that the difference between 1 and 2, for example, is quantitatively the same as the difference between, say, 6 and 7. Such descriptors may be further subdivided into *relative-scale* quantitative variables, where value 'zero' means the absence of the characteristic of interest, and *interval-scale* variables where the 'zero' is chosen arbitrarily. For the latter type, the fact that the 'zero' reference is chosen

Relative scale
Interval scale

Table 1.2 The different mathematical types of descriptors.

Descriptor types	Ecological examples
Binary (two states, presence-absence)	Species present or absent
Multi-state (many states)	
Nonordered (qualitative, nominal, attributes)	Geological group
Ordered	
Semiquantitative (rank-ordered, ordinal)	Importance or abundance scores
Quantitative (metric, measurement)	
Discontinuous (meristic, discrete)	Equidistant abundance classes
Continuous (metric)	Temperature, length

arbitrarily prevents comparisons of the type "this temperature (°C) is twice as high as that one". Species abundance data, or temperatures measured in Kelvin, are examples of the first type, while temperature measured in degrees Celsius, dates, or geographic directions (of wind, currents, etc.) in degrees, are examples of the second.

Continuous quantitative descriptors are usually processed as they are. If they are divided into a small number of *equidistant* classes of abundance (further discussed below), the discontinuous descriptors that are obtained may usually be processed as if they were continuous, because the distortion due to grouping is negligible for the majority of distribution types (Sneath & Sokal, 1973). Before the advent of computers, it was usual practice, in order to facilitate calculations, to divide continuous descriptors into a small number of classes. This transformation is still necessary when, due to low precision of the measurements, only a small number of classes can be distinguished in practice, or when comparisons are sought between quantitative and semiquantitative descriptors.

Meristic variables (the result of enumeration, or counting) theoretically should be considered as discontinuous quantitative. In ecology, however, these descriptors are most often counts of the number of specimens belonging to the various species, whose range of variation is so large that they behave, for all practical purposes, as continuous variables. When they are transformed (Sections 1.5 and 7.7), as is often the case, they become real numbers instead of integers.

In order to speed up field observations or counts in the laboratory, it is often interesting for ecologists to record observations in the form of *semiquantitative* descriptors. Usually, it is possible to estimate environmental characteristics very rapidly by ascribing them a score using a small number of ordered classes: score 1 < score 2 < score 3, etc. Ecologists may often proceed

in this way without losing pertinent information, whereas precise counts would have necessitated more considerable efforts than required by the ecological phenomenon under study. For example, in studying the influence of the unevenness of the landscape on the fauna of a given area, it may be enough to describe the relief using ordered classes such as flat, undulated, rough, hilly and mountainous. In the same way, counting large numbers of organisms may be done using abundance scores instead of precise numbers of individuals. Frontier (1973), for example, established such a scoring scale to describe the variability of zooplankton. Another score scale, also developed by Frontier (1969) for counting zooplankton, was used to estimate biomass (Dévaux & Millerioux, 1976b) and diversity of phytoplankton (Dévaux & Millerioux, 1977) as well as to evaluate schools of cetaceans at sea (Frontier & Viale, 1977). Frontier & Ibanez (1974) as well as Dévaux & Millerioux (1976a) have shown that this rapid technique is as informative as classical enumeration for principal component analysis (Section 9.1). It must be noted that nonparametric statistical tests of significance, which are used on such semiquantitative descriptors, have a discriminatory power almost equal to that of their parametric equivalent. Naturally occurring semiquantitative descriptors, which give *ranks* to the objects under study, as well as quantitative descriptors divided into non-equidistant classes (which is done either to facilitate data collection or to evidence holes in frequency distributions), are included among the semiquantitative descriptors. Method 6.4 in Subsection 1.5.6 shows how to normalize semiquantitative descriptors if they have to be used in methods of data analysis that perform better in the presence of normality. Normalized semiquantitative descriptors should only be interpreted in terms of the ordinal value that they really represent. In addition, methods designed for quantitative data analysis may often be adapted to ranked data. This is the case, for example, with principal component analysis (Lebart *et al*., 1979; Subsection 9.1.7) and linear regression (Iman & Conover, 1979).

Qualitative descriptors often present a problem to ecologists, who are tempted to discard them, or reduce them to a series of binary variables (Subsection 1.5.7). Let us forget the cases where descriptors of this kind have been camouflaged as ordered variables by scientists who did not quite know what to do with them …Various methods based on contingency tables (Chapter 6) may be used to compare such descriptors with one another, or to ordered descriptors divided into classes. Special resemblance coefficients (Chapter 7) allow these descriptors to be used as a basis for clustering (Chapter 8) or ordination (Chapter 9). The first paragraph of Chapter 6 gives examples of qualitative descriptors. An important class is formed by classifications of objects, which may in turn become descriptors of these objects for subsequent analyses, since the definition of a classification (Section 8.1) corresponds to the definition of a descriptor given above.

Binary or *presence-absence* descriptors may be noted + or –, or 1 or 0. In ecology, the most frequently used type of binary descriptors is the presence or absence of a species, when reliable quantitative information is not available. It is only for historical reasons that they are considered as a special class: programming the first computers was greatly facilitated by such descriptors and, as a result, several methods have been developed for processing them. Sneath & Sokal (1973) present various methods to recode variables into binary form; see also Subsection 1.5.7. Binary descriptors encountered in ecology may be processed either as qualitative, semiquantitative or quantitative variables. Even though the mean and variance parameters of binary descriptors are difficult to interpret, such descriptors may be used with methods originally designed for quantitative variables — in a principal component or correspondence analysis, for instance, or as independent variables in regression or canonical analysis models.

When collecting ecological data, the level of precision with which descriptors are recorded should be selected with consideration of the problem at hand. Quantitative descriptors may often be recorded either in their original form or in semiquantitative or qualitative form. The degree of precision should be chosen with respect to the following factors: (1) What is the optimal degree of precision of the descriptor for analysing this particular ecological phenomenon? (2) What type of mathematical treatment will be used? This choice may determine the mathematical types of the descriptors. (3) What additional cost in effort, time or money is required to raise the level of precision? Would it not be more informative to obtain a larger number of less precise data?

2 — Intensive, extensive, additive, and non-additive descriptors

There are other useful ways of looking at variables. Margalef (1974) classified ecological variables as either *intensive* or *extensive*. These notions are derived from thermodynamics (Glansdorff & Prigogine, 1971). A variable is said to be *intensive* if its value is defined independently of the size of the sampling unit in which it is measured. For example, water temperature is defined independently of the size of the bucket of water in which a thermometer is placed: we do not say "12°C per litre" but simply "12°C". This does not mean that the *measured value* of temperature may not vary from place to place in the bucket; it may indeed, unless water is well-mixed and therefore homogeneous. Concentration of organisms (number per unit surface or volume), productivity, and other rate variables (e.g. birth, death) are also intensive because, in a homogeneous system, the same value is obtained whether the original measurements are made over 1 m^2 or over 100 m^2. In contrast, an *extensive* variable is one whose value, in a homogeneous system, changes proportionally (in linear relationship) to the size of the sampling unit (which may consist in a line, a surface, or a volume). It is formally defined as an integral over the sampling unit. Number of individuals and biomass in a quadrat or volume, at a given point in time, are examples of extensive variables.

Intensive

Extensive

Extensive variables have the property that the values they take in two sampling units can be added to provide a meaningful estimate of the value in the combined unit: they are additive (next paragraph). Other variables do not have this property; either they do not vary at all (e.g. temperature in a homogeneous bucket of water, which is an intensive variable), or they vary in a nonlinear way with the size of the sampling unit. For example, species richness in a sampling unit (surface or volume) cannot be computed as the sum of the numbers of species found in two sub-units; that sum would usually be larger than the number of species actually found in the combined unit because some species are common to the two sub-units. Species diversity (Chapter 5) also has this property. The relationship of such variables to scale is complex and depends on the distribution patterns of the species and the size of the sampling units (grain size of the measurements; Section 13.0).

Additive

Another, more statistical point of view concerns additivity. This notion is well-known in geostatistics (Olea, 1991, p. 2; Journel & Huijbregths, 1978). A variable is

said to be *additive* if its values can be added while retaining the same meaning as the original variable. A good example is the number of individuals in a quadrat. Concentrations, which are intensive variables, are additive if they are referred to the same linear, surface or volume unit measure (e.g. individuals m^{-2}; kg m^{-3}) (Journel & Huijbregths, 1978, p. 199); values may be added to compute a mean for example.

Extensive variables (e.g. number of individuals) are, by definition, additive; a sum or a mean has the same meaning as the original data although, if the sampling units differ in size, the values must be weighted by the sizes of the respective sampling units for their mean to be meaningful. For intensive additive variables (e.g. temperature or concentration), only the (weighted) mean has the same meaning as the original values. Variables may be additive over either time or space (Walliser, 1977); numbers of individuals in quadrats, for example, are additive over space, but not over time if the time lag between observations is shorter than the generation time of the organisms (the same individuals would be counted several times).

Non-additive

Examples of *non-additive variables* are pH values, logarithms and ratios of random variables, indices of various kinds, and directions of vectors (wind direction, aspect of a slope, etc.). Values of non-additive variables must be transformed in some way before (and if) they can be meaningfully combined. Logarithms of counts of organisms, for instance, have to be back-transformed using antilogarithms before values can be added. For ratios, the numerator and denominator must be added separately, and the ratio recomputed from these sums. Other non-additive variables, such as species richness and diversity, cannot be numerically combined; these indices for combined sampling units must be recomputed from the combined raw data.

These notions are of prime importance when analysing spatial data (Chapters 13 and 14). To appreciate their practical usefulness, let us consider a study in which the following variables have been measured at a site in a lake or in the ocean, at different times: incident solar energy at water surface (W m^{-2}), temperature (°C), pH, O_2 concentration (g m^{-3}), phytoplankton production (g C m^{-3} s^{-1}), and concentration of zooplankton (individuals m^{-3}). All these variables are intensive; they all have complex physical units, except temperature (simple unit) and pH (no unit). Assuming that some form of random sampling has been conducted with constant-sized observation units, how could estimates be obtained for the whole study area? This question may be viewed from two different angles, i.e. one may be looking for a mean or for an integral value over the study area. For additive variables (i.e. all except pH), values can be computed that represent the mean over the study area. However, integrating over the study area to obtain values for total incident solar energy, zooplankton, etc. is not that simple, because it requires the variables to be extensive. No extensive variable can be derived from temperature or pH. In the case of variables with complex physical units, new variables may be derived with units that are appropriate for integration:

- Consider O_2 concentration. Its physical dimensions (Section 3.1) are $[ML^{-3}]$, with units g m^{-3}. This indicates that the "mass" part (dimension [M], with unit g), which is extensive, may be integrated over a volume, for example that of the surface mixed

layer over the whole study area. Also, values from different depths in the mixed layer may be vertically integrated, to provide areal concentrations (dimensions $[ML^{-2}]$, with units g m^{-2}). The same applies to the concentration of zooplankton.

- Flux variables can be turned into variables that are additive over both space and time. Phytoplankton production (dimensions $[ML^{-3}T^{-1}]$, with units g C m^{-3} s^{-1}) is a flux variable since it is expressed per unit space and time. The extensive "mass" part may be integrated over a volume or/and over time, e.g. the euphotic zone over the whole study area or/and for the duration of the study. Values from different depths in the euphotic zone may be vertically integrated, thus providing areal concentrations (dimensions $[ML^{-2}T^{-1}]$, with units g C m^{-2} s^{-1}), which can then be integrated over time.

- Incident solar energy (W m^{-2}) represents a more complex case. The "power" part (W) can be integrated over space (m^2) only. However, because W = J s^{-1} (Table 3.2), it is possible to integrate the "energy" part (J) over both space and time. Since incident solar energy is either in W m^{-2} or J m^{-2} s^{-1}, the "power" part may be integrated over space or, alternatively, the "energy" part may be integrated over both surface (m^2) and time (s). For example, one can compute solar energy over a given area during 24 h.

1.5 Coding

Coding is a technique by which original data are transformed into other values, to be used in the numerical analysis. All types of descriptors may be coded, but nonordered descriptors must necessarily be coded before they may be analysed numerically. The functions or laws of correspondence used for coding qualitative descriptors are generally discontinuous; positive integers are usually associated with the various states.

Consider the case where one needs to compute the dependence between a variable with a high degree of precision and a less precisely recorded descriptor. Two approaches are available. In the first approach, the precision of the more precise descriptor is lowered, for example by dividing continuous descriptors into classes. Computers can easily perform such transformations. Dependence is then computed using a mathematical method adapted to the descriptor with the *lowest* level of precision. In the second approach, the descriptor with the lower precision level will be given a numerical scale adjusted to the more precise one. This operation is called *quantification* (Cailliez & Pagès, 1976; Gifi, 1990); one method of quantification through canonical correspondence analysis is described in Subsection 11.2.2. Other transformations of variables, that adjust a descriptor to another, have been developed in the regression framework discussed in Section 10.3.

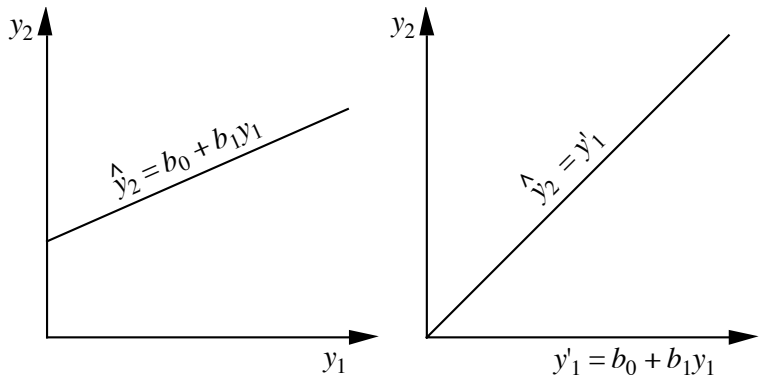


Figure 1.9 The regression parameters (b_0 and b_1) found by regressing y_2 on y_1 (left panel) may be used (right panel) to transform y_1 into y'_1 such that y'_1 is now on the same scale as y_2.

1 — Linear transformation

In a study where there are quantitative descriptors of different types (metres, litres, mg L^{-1}, …), it may be useful to put them all on the same scale in order to simplify the mathematical forms of relationships. It may be difficult to find an ecological interpretation for a relationship that includes a high level of artificial mathematical complexity, where scale effects are intermingled with functional relationships. Such changes of scale may be linear (of the first order), or of some higher order.

A linear change of scale of variable y is described by the transformation $y' = b_0 + b_1 y$ where y' is the value after transformation. Two different transformations are actually included in this equation. The first one, *translation*, consists in adding or subtracting a constant (b_0 in the equation) to all data. Graphically, this consists in sliding the scale beneath the data distribution. Translation is often used to bring to zero the mean, the modal class, the weak point of a bimodal distribution, or another point of interest in the distribution. The second transformation, *expansion*, is a change of scale obtained by multiplying or dividing all observed values by a constant (b_1 in the equation). Graphically, this operation is equivalent to contracting or expanding the scale beneath the distribution of a descriptor.

Two variables that are linearly related can always be put on the same scale by a combination of an expansion followed by a translation, or the opposite, the values of parameters b_0 and b_1 being found by linear regression (model I or model II: Chapter 10). For example (Fig. 1.9), if a linear regression analysis shows the equation relating y_2 to y_1 to be $\hat{y}_2 = b_0 + b_1 y_1$ (where $\hat{y}_2$ represents the values estimated by the regression equation for variable y_2), then transforming y_1 into $y'_1 = b_0 + b_1 y_1$ successfully puts variable y_1 on the same scale as variable y_2, since $\hat{y}_2 = y'_1$. If one

wishes to transform y_2 instead of y_1, the regression equation should be computed the other way around.

2 — Nonlinear transformations

The methods of multidimensional analysis described in this book are often based on covariances or linear correlations. Using them requires that the relationships among variables be made linear by an appropriate transformation. When two variables are not linearly related, their relationship may be described by a second- or higher-degree equation, or by other functional forms, depending on the situation. If the nonlinear form of the equation is derived from ecological theory, as it is often the case in population dynamics models, interpretation of the relationship poses no problem. If, however, a nonlinear transformation is chosen empirically, for reasons of mathematical elegance and without grounding in ecological theory, it may be difficult to find an ecological meaning to it.

The relationship between two variables may be determined with the help of a scatter diagram of the objects in the plane formed by the variables. The principles of analytical geometry may then be used to recognize the type of relationship (Fig. 1.10), which in turn determines the most appropriate type of transformation. A relationship frequently found in ecology is the exponential function, in which a variable y_2 increases in geometric progression with respect to y_1, according to one of the following equations:

$$y_2 = b^{(y_1)} \text{ or } y_2 = b_0 b_1^{(y_1)} \text{ or } y_2 = b_0 b_1^{(y_1 + b_2)} \text{ or else } y_2 = b_0 b_1^{(b_2 y_1)} \qquad \textbf{(1.3)}$$

Logarithmic transformation

depending on the number of constants b that shift or amplify the function. Such relationships can easily be linearized by using the logarithm of variable y_2 (called y'_2 below) instead of y_2 itself. The above relationships then become:

$$y'_2 = \log(y_2) = b' y_1, \text{ or } y'_2 = b'_0 + b'_1 y_1,$$

$$\text{or } y'_2 = b'_0 + b'_1 (y_1 + b_2), \text{ or } y'_2 = b'_0 + b'_1 b_2 y_1 \qquad \textbf{(1.4)}$$

where the b's are the logarithms of constants b in eq. 1.3.

If two variables display a logarithmic relationship of the form

$$y_2 = \log_b (y_1) \qquad \textbf{(1.5)}$$

where b is the base of the logarithm, their relationship can be made linear by applying a $\log^{-1}$ transformation to y_2:

$$y'_2 = b^{(y_2)} = y_1 \qquad \textbf{(1.6)}$$

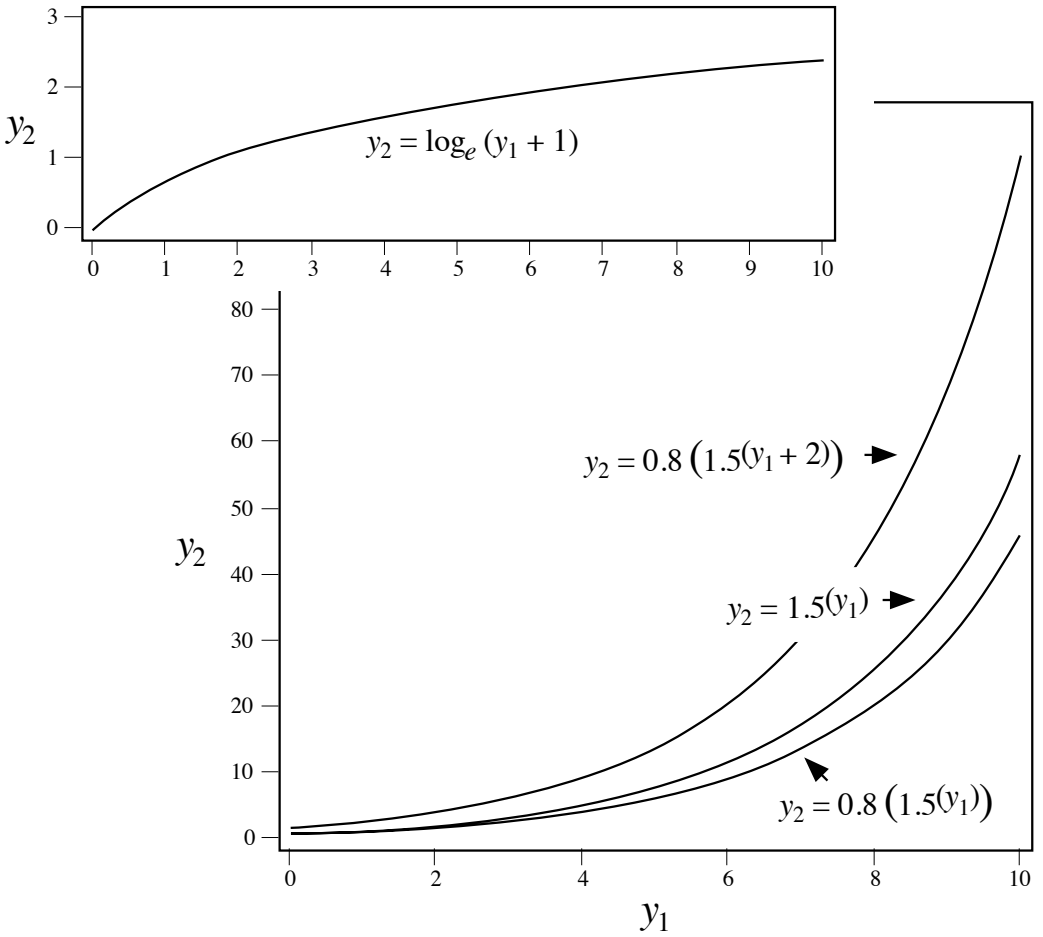


Figure 1.10 The relationship between variables may often be recognized by plotting them one against the other. In the upper panel, y_2 varies as the natural logarithm of y_1. In the lower panel, y_2 is an exponential function of y_1. These curves (and corresponding equations) may take different forms, depending on the modifying constants b, b_0, b_1 and b_2 (eq. 1.3).

When a nonlinear form can be assumed from knowledge of the ecological process involved, the corresponding equation can be used as the basis for a linearizing transformation. For instance, the nonlinear equation

$$N_t = N_0 e^{rt} \tag{1.7}$$

describes the exponential growth of a population, as observed in population explosions. In this equation, the independent variable is time (t); N_0 and N_t are the

population sizes at times 0 and t, respectively; r is the Malthus parameter describing the intrinsic rate of increase of the population. This nonlinear equation indicates that N_t should be transformed into its natural logarithm to make the relationship linear. After this transformation, $\log_e(N_t)$ is linearly related to t: $\log_e(N_t) = \log_e(N_0) + rt$.

3 — Combining descriptors

Another transformation that is often used consists in combining different descriptors by addition, subtraction, multiplication or division. In limnology, for example, the ratio (surface O_2/ bottom O_2) is often used as a descriptor. So is the Pearsall ionic ratio, all ions being in the same physical units:

$$y = \frac{\mathrm{Na} + \mathrm{K}}{\mathrm{Mg} + \mathrm{Ca}} \tag{1.8}$$

Beware, however, of the spurious correlations that may appear when comparing a ratio variable $\mathbf{y}/\mathbf{z}$ to $\mathbf{z}$, or two ratio variables $\mathbf{y}_1/\mathbf{z}$ and $\mathbf{y}_2/\mathbf{z}$ (Pearson 1897). Jackson & Somers (1991a) illustrate the problem using simulated data and recommend that such correlations be tested using permutation tests (Section 1.2) involving permutation of the parent variables, followed by construction of the ratios from the permuted variables and computation of the correlation coefficient under permutation.

Permutation test

One may want to take into account a factor of magnitude or size. For example, when observation units are of different sizes, the number of specimens of each species may be divided by the area or the volume of the unit (depending on whether the units come from an area or a volume), or by some other measure of the sampling effort. One must exert care when interpreting the results, however, since large observation units are more representative of populations and have smaller variances than small ones.

4 — Ranging and standardization

Quantitative variables, used in ecology as environmental descriptors, are often expressed in incompatible units such as metres, mg L^{-1}, pH units, etc. In order to compare such descriptors, or before using them together in a classification or ordination procedure, they must be brought to some common scale. Among the methods available, some only eliminate size differences while others reduce both the size and variability to a common scale.

Translation, a method previously discussed, allows one to *centre* the data, eliminating size differences due to the position of the zero on the various scales. Centring is done by subtracting the mean of the observations $(\bar{y})$ from each value y_i:

$$y'_i = y_i - \bar{y} \tag{1.9}$$

For relative-scale variables (Subsection 1.4.1), dividing each y_i by the largest observed value is a way, based on expansion, to bring all values in the range [0, 1] (Cain & Harrison, 1958):

$$y'_i = y_i / y_{max} \quad \textbf{(1.10)}$$

For interval-scale variables, whose range may include negative values, the absolute value of the largest positive or negative value is used as divisor. The transformed values are in the interval [–1, +1].

Ranging

Other methods allow the simultaneous adjustment of the magnitude and the variability of the descriptors. The method of *ranging*, proposed by Sneath & Sokal (1973), reduces the values of a variable to the interval [0, 1] by first subtracting the minimum observed for each variable and then dividing by the range:

$$y'_i = \frac{y_i - y_{min}}{y_{max} - y_{min}} \quad \textbf{(1.11)}$$

For relative-scale variables (Subsection 1.4.1) for which y_{min} is always zero, ranging can be achieved as well with eq. 1.10.

Standardization

The most widely used method for making descriptors compatible is to *standardize* the data (transformation into so-called "z-scores"). This method will be fully discussed in Section 4.2, which deals with correlation. Principal components (Section 9.2) are frequently computed using standardized data. Standardization is achieved by subtracting the mean (translation) and dividing by the standard deviation (s_y) of the variable (expansion):

$$z_i = \frac{y_i - \bar{y}}{s_y} \quad \textbf{(1.12)}$$

The position of each object on the transformed variable z_i is expressed in standard deviation units; as a consequence, it refers to the group of objects from which s_y has been estimated. The new variable z_i is called a *standardized variable*. Such a variable has three interesting properties: its mean is zero ($\bar{z} = 0$); its variance and hence its standard deviation are 1 ($s_z^2 = s_z = 1$); it is also a *dimensionless variable* (Chapter 3) since the physical dimensions (metres, mg L^{-1}, etc.) in the numerator and denominator cancel out. Transformations 1.8, 1.10 and 1.11 also produce dimensionless variables.

Beware of the "default options" of computer programs that may implicitly or explicitly suggest to standardize all variables before data analysis. Milligan & Cooper (1988) report simulation results showing that, for clustering purposes, if a transformation is needed, the ranging transformation (eqs. 1.10 and 1.11) gives results that are in general better to those obtained using standardization (eq. 1.12).

5 — Implicit transformation in association coefficients

When descriptors with different scales are used together to compare objects, the choice of the association coefficient (Section 7.6) may partly determine the type of transformation that must be applied to the descriptors. Some coefficients give equal weights to all variables independently of their scales while others take into account the magnitude of variation of each one. Since the amount of information (in the sense of information theory; Chapter 6) in a quantitative descriptor increases as a function of its variance, equalizing the variances before the association coefficient is computed is a way to ensure that all descriptors have the same weight. It is for ecologists to decide the kind of contribution they expect from each descriptor; again, beware of the "default options" of computer programs.

Some association coefficients require that the data be expressed as integers. Depending on the capabilities of the computer program and the degree of discrimination required, ecologists may decide to use the closest integer value, or to multiply first all values by 10 or 100, or else to apply some other simple transformation to make the data compatible with the coefficient to be computed.

6 — Normalization

Another type of transformation, called *normalizing transformation*, is performed on descriptors to make the frequency distributions of their data values look like the normal curve of errors — or, at least, as unskewed as possible. Indeed, several of the methods used in multivariate data analysis have been developed under the assumption that the variables are normally distributed. Although most of these methods do not actually require full normality (i.e. no skewness nor kurtosis), they may perform better if the distributions of values are, at least, not skewed. Skewed distributions, as in Fig. 1.11, are such that the variance of the distribution is controlled mostly by the few points in the extreme right tail; so, variance-partitioning methods such as principal component analysis (Chapter 9) or spectral analysis (Chapter 12) would bring out components expressing the variation of these few data points first instead of the variation of the bulk of data values. Normalizing transformations also have the property of reducing the *heteroscedasticity* of descriptors (Box 1.4). The data analysis phase of research should always start by looking at the distributions of values for the different variables, i.e. computing basic distribution statistics (including skewness and kurtosis, eqs. 4.41 and 4.42), drawing histograms of frequency distributions, and testing for normality (described in Section 4.6). A normalizing transformation may have to be found for each variable separately; in other cases, one is looking for the best transformation that would normalize several variables.

• 6.1 — Ecologists often encounter distributions where a species is abundant in a few observation units (quadrats, etc.), fairly abundant in more, present in even more, and absent in many; this is in agreement with the concept of ecological niche briefly explained in Section 1.0, if the sampling programme covers a large enough area or environmental gradient. Distributions of this type are clearly not normal, being

Homoscedasticity Box 1.4

Homoscedasticity, also called *homogeneity* or *equality of the variances*, technically means that the variances of the error terms are equal for all observations. Its antonym is **heteroscedasticity** or *heterogeneity of the variances*. Homoscedasticity may actually refer to different properties of the data.

- *For a single variable*, homoscedasticity of the distribution means that, when the statistical population is sampled repeatedly, the expected value of the variance remains the same, whatever the value of the mean of the data sample. Data drawn from a normal distribution possess this property whereas data drawn from a Poisson distribution, for instance, do not, since the variance is equal to the mean in this type of distribution.

- *In regression analysis*, homoscedasticity means that, for all values of the independent variable, the variances of the corresponding values of the response variable (called error variances or variances of the residuals) are the same.

- *In t-test, analysis of variance and discriminant analysis*, homoscedasticity means that variances are equal in all groups, for each variable.

strongly skewed to the right (long tail in the higher values). Needless to say, environmental variables may also have non-normal distributions. For instance, the scales on which chemical variables are measured are conventions of chemistry which have no relation whatsoever with the processes generating these values in nature. So, any normalizing transformation is as good as the scale on which these data were originally measured.

Skewed data are often transformed by taking logarithms (below) or square roots. *Square root* is the least drastic transformation and is used to normalize data that have a Poisson distribution, where the variance is equal to the mean, whereas the *logarithmic transformation* is applicable to data that depart more widely from a normal distribution (Fig. 1.11). Several intermediate transformations have been proposed between these two extremes (Fig. 1.12): cubic root, $\log^2$, $\log^p$, etc. The *hyperbolic transformation* is useful for one particular type of data, which share the two extreme types at the same time (when the standard deviation is proportional to the mean, with many observations of a very small size which follow a Poisson distribution: Quenouille, 1950; Barnes,

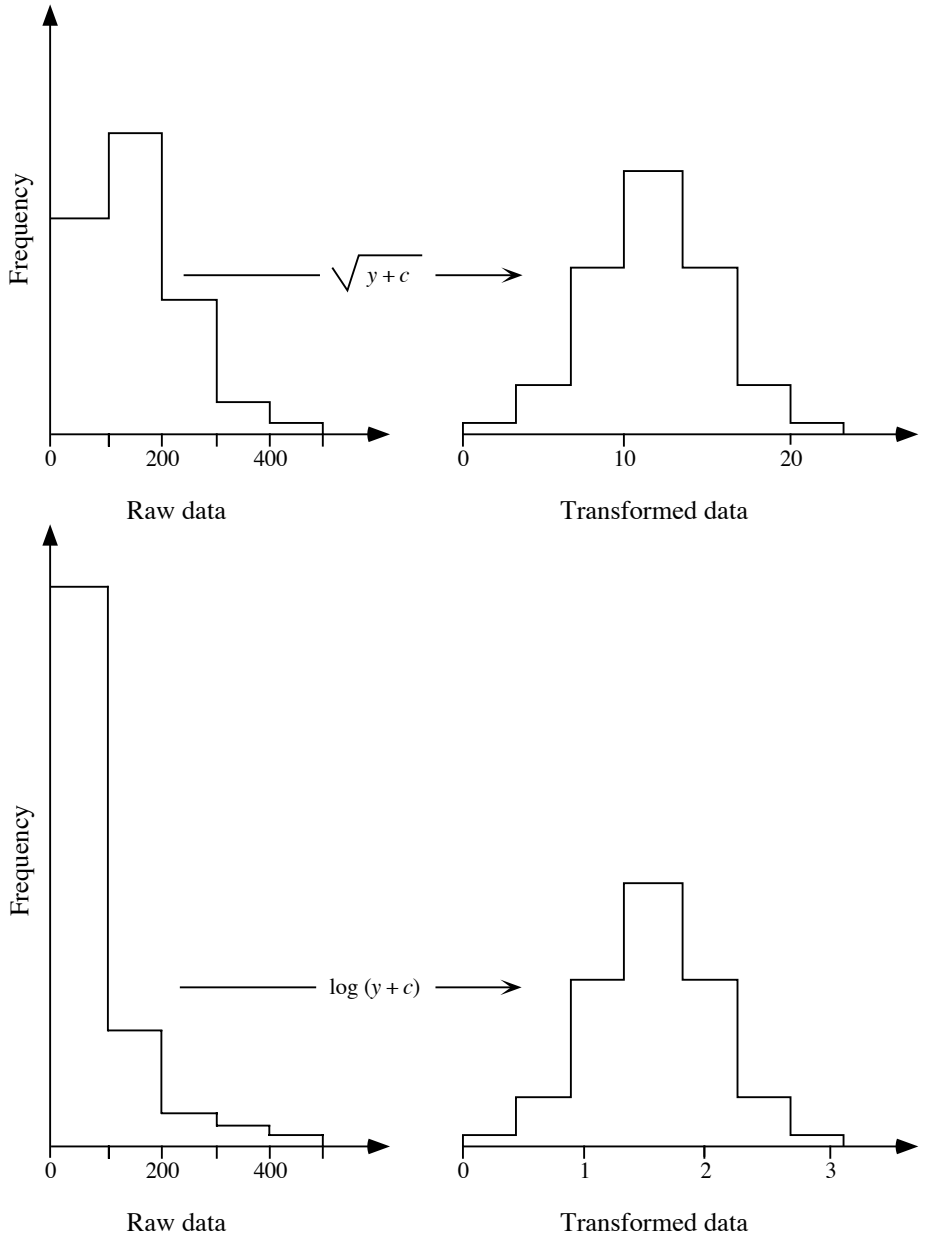


Figure 1.11 Numerical examples. Upper panel: Data that follow a Poisson distribution (left) can be normalized by the square root transformation (right). For a given species, these frequencies may represent the number of quadrats (ordinate) occupied by the number of specimens shown along the abscissa. Lower panel: Data distribution (left) that can be normalized by a logarithmic transformation (right).

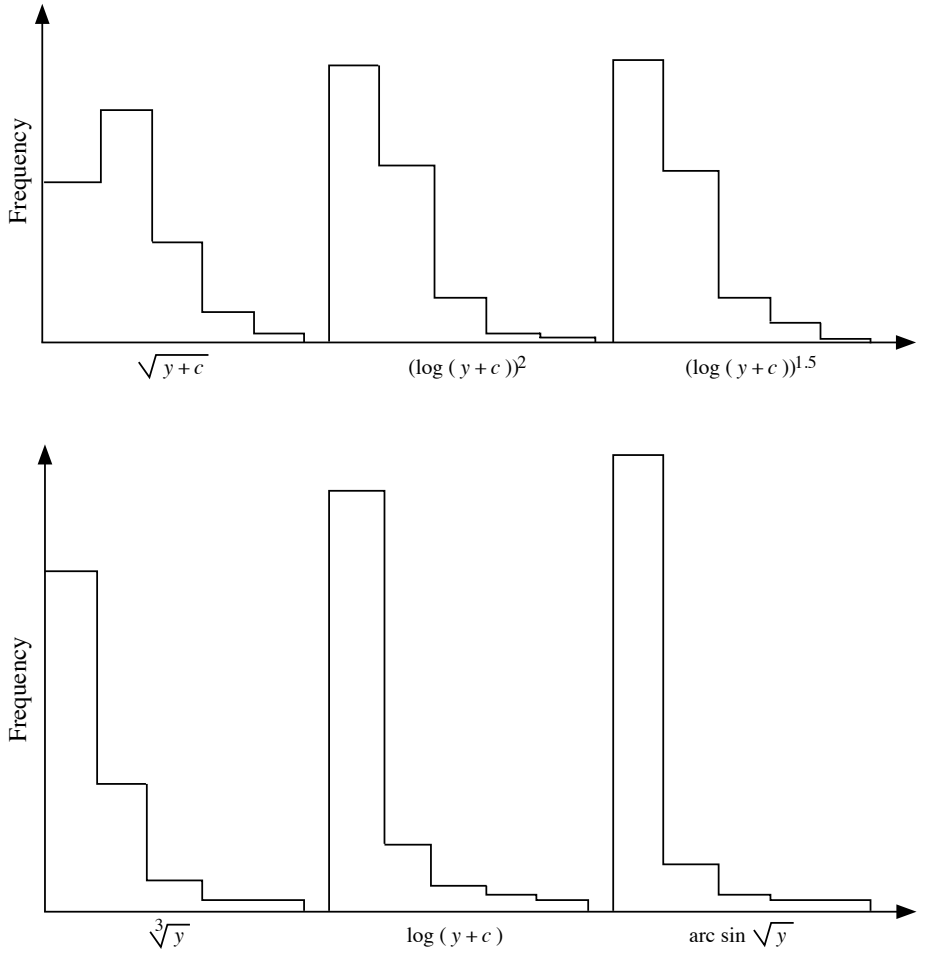


Figure 1.12 Numerical examples. Each histogram is labelled by the normalizing transformation to be used in that case. The bottom rightmost histogram refers to a simplified version of the hyperbolic transformation.

1952). The *angular* or *arcsine transformation* is appropriate for percentages and proportions (Sokal & Rohlf, 1981, 1995):

$$y_i' = \arcsin\sqrt{y_i} \qquad \textbf{(1.13)}$$

In case of doubt, one may try several of these transformations and perform a test of normality (Section 4.6), or compute the skewness of the transformed data, retaining the transformation that produces the most desirable results. Alternatively, the Box-Cox method (point 6.2, below) may be used to find the best normalizing transformation.

A logarithmic transformation is computed as follows:

Logarithmic transformation

$$y'_i = \log(b_0 + b_1 y_i) \qquad \textbf{(1.14)}$$

The base of logarithm chosen has no influence on the normalising power, since transformation from one base (c) to another (d) is a linear change of scale (expansion, see Subsection 1.5.1: $\log_d y_i = \log_c y_i / \log_c d$). When the data to be transformed are all strictly positive (all $y_i > 0$), it is not necessary to carry out a translation ($b_0 = 0$ in eq. 1.14). When the data contain fractional values between 0 and 1, one may multiply all values by some appropriate constant in order to avoid negative transformed values: $y'_i = \log(b_1 y_i)$. When the data to be transformed contain negative or null values, a translation must be applied first, $y'_i = \log(b_0 + y_i)$, since the logarithmic function is defined over the set of positive real numbers only. One should choose for translation a constant b_0 that is of the same order of magnitude as the significant digits of the variable to be transformed; for example, $b_0 = 0.01$ for data between 0.00 and 0.09 (the same purpose would have been achieved by selecting $b_0 = 1$ and $b_1 = 100$ for these data). For species abundance data, this rule produces the classical transformation $y'_i = \log(y_i + 1)$.

Box-Cox method

• 6.2 — When there is no *a priori* reason for selecting one or the other of the above transformations, the Box-Cox method allows one to empirically estimate the most appropriate exponent of the following general transformation function:

$$y'_i = (y_i^{\gamma} - 1)/\gamma \qquad (\text{for } \gamma \neq 0) \qquad \textbf{(1.15)}$$

and

$$y'_i = \log_e(y_i) \qquad (\text{for } \gamma = 0)$$

As before, y'_i is the transformed value of observation y_i. In this transformation, the value γ is used that maximizes the following log likelihood function:

$$L = -(\nu/2)\log_e(s^2_{y'}) + (\gamma - 1)(\nu/n)\sum_i \log_e(y_i) \qquad \textbf{(1.16)}$$

since it is this value that yields the best transformation to normality (Box & Cox, 1964; Sokal & Rohlf, 1995). The value L that maximizes the likelihood function is found by iterative search. In this equation, $s^2_{y'}$ is the variance of the *transformed* values y'_i. When analysing several groups of observations at the same time (below), $s^2_{y'}$ is estimated instead by the within-group, or residual variance computed in a one-way ANOVA. The group size is n and ν is the number of degrees of freedom ($\nu = n - 1$ if the computation is made for a single group). All y_i values must be strictly positive numbers since logarithms are taken in the likelihood function L (eq. 1.16); all values may easily be made strictly positive by translation, as discussed in Subsection 1.5.1. It is interesting to note that, if $\gamma = 1$, the function is a simple linear transformation; if $\gamma = 1/2$, the function becomes the square root transformation; when $\gamma = 0$, the transformation is logarithmic; $\gamma = -1$ yields the reciprocal transformation.

Readers are invited to take a value (say 150) and transform it, using eq. 1.15, with a variety of values of γ gradually tending toward 0 (say 1, 0.1, 0.01, 0.001, etc.). Comparing the results to the logarithmic transformation will make it clear that the natural logarithm is indeed the limit of eq. 1.15 when γ tends towards 0.

Another log likelihood function L' is proposed by Sokal & Rohlf (1995) to achieve homogeneity of the variances for several groups of observations of a given variable, together with the normality of their distributions. This generalized Box-Cox transformation may also be applied to the identification of the best normalizing transformation for several species, for a given set of sampling sites.

• 6.3 — When the data distribution includes several groups, or when the same transformation is to be applied to several quantitative and dimensionally homogeneous descriptors (Chapter 3; for instance, a species abundance data table), Taylor's (1961) power law provides the basis for another general transformation that stabilizes the variances and thus makes the data *more likely* to conform to the assumptions of parametric analysis, including normality (Southwood, 1966; see also Downing, 1979 on this subject). This law relates the means and variances of the k groups through the equation

Taylor's power law

$$s^2_{y_k} = a\,(\bar{y}_k)^b \qquad \textbf{(1.17)}$$

from which constants a and b can be computed by nonlinear regression (Subsection 10.3.6). When the latter is not available, an approximation of b may be calculated by linear regression using the logarithmic form

$$\log s^2_{y_k} = \log a + b \log \bar{y}_k \qquad \textbf{(1.18)}$$

Having found the value of b, the variance stabilizing transformations

$$y'_i = y_i^{\left(1-\frac{b}{2}\right)} \qquad (\text{for } b \neq 2) \qquad \textbf{(1.19)}$$

or

$$y'_i = \log_e(y_i) \qquad (\text{for } b = 2)$$

are applied to the data.

• 6.4 — The following method represents an *omnibus normalizing procedure*, which is able to normalize most kinds of data. The procedure is easy to carry out in R or using a standard statistical packages. The package must have a pseudo-random number generator for random normal deviates, i.e. values drawn at random from a normal distribution.

Omnibus procedure

(1) Write the quantitative or semiquantitative descriptor to be normalized into a vector or a column of a spreadsheet. Sort the vector in order of increasing values.

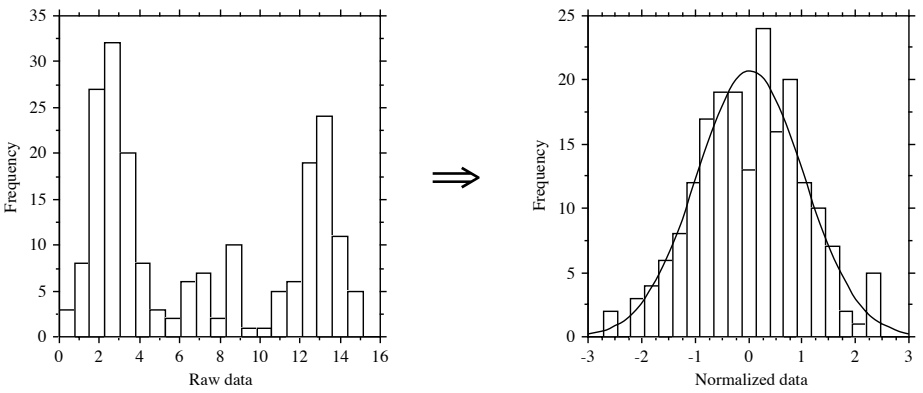


Figure 1.13 The *omnibus* procedure is used here to normalize a set of 200 data values with tri-modal distribution (left). A normal curve is fitted to the normalized data (right). The normalized data could be rescaled to the approximate range of the original data through the linear transformation $y_{rescaled} = 8 + (y_{normalized} \times 16/5.5)$ where 16 is the approximate range of the raw data and 5.5 that of the normalized data; the constant 8 makes all rescaled values positive.

(2) Create a new descriptor with the same number of values, using a pseudo-random normal deviate generator (***rnorm()*** is the function to use in R). Sort this new vector in order of increasing values. (3) Bind the two vectors, or copy the sorted normal deviate values besides the first sorted vector in the spreadsheet. Sort the bound vectors or the spreadsheet back into the original order if necessary. (4) Use the normal deviates as a monotonic proxy for the original descriptor. Figure 1.13 shows an example of this transformation. (5) It may be useful in some cases to rescale the normalized data to the approximate range of the original data through a linear transformation.

This procedure may be modified to handle *ex aequo* (tied) values (Section 5.3). Tied values may either receive the same normal deviate value, or they may be sorted in some random order and given neighbouring normal deviate values; one should select a solution that makes sense considering the data at hand.

Data transformed in this way may be used in methods of data analysis that perform better in the presence of normally distributed data. Several such methods will be studied in Chapters 9 and 11. The main disadvantage is that a back-transformation is difficult. If the study requires that values of the transformed descriptor be forecasted by a model, the database itself will have to be used to find the original descriptor values that are the closest to the forecasted normal deviate. An interpolation may have to be made between observed data values.

7 — Dummy variable coding

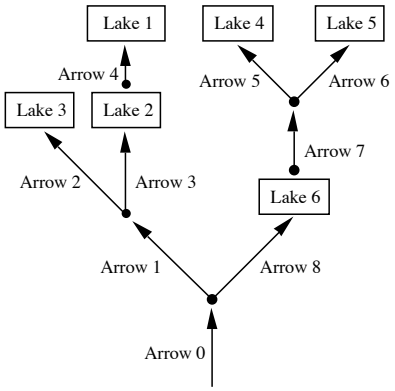
Multistate qualitative descriptors may be binary-coded as *dummy variables*. This coding is interesting because it allows the use of qualitative descriptors in procedures such as multiple regression, discriminant analysis or canonical analysis, which have been developed for quantitative variables and in which binary variables may also be used. A multistate qualitative descriptor with s states can be decomposed into $(s - 1)$ dummy variables V_j, as shown by the following example of a four-state descriptor:

States	Dummy variables			
	V_1	V_2	V_3	V_4
1	1	0	0	0
2	0	1	0	0
3	0	0	1	0
4	0	0	0	1

In this example, three dummy variables, e.g. V_1 to V_3, are sufficient to code for the four states of the nominal descriptor, excluding V_4. Had dummy variable V_4 been included (shaded column above), its information would have been totally *linearly dependent* (Box 1.1 and Section 2.7) on the first three variables, meaning that it would have been entirely predictable from the sum of the other three variables and the intercept represented by a column vector of 1: $V_4 = \mathbf{1}_{\text{intercept}} - (V_1 + V_2 + V_3)$. This shows that the first three dummy variables are sufficient to determine the four states of the multistate qualitative descriptor. Actually, any one of the four dummy variables may be eliminated to return to the condition of linear independence among the remaining ones. Using the coding table above, the objects are coded by three dummy variables instead of a single 4-state descriptor. An object with state 1, for instance, would be recoded as [1 0 0], an object with state 2 as [0 1 0], and so on.

There are other methods to code for a qualitative variable or a factor of an experiment. Helmert contrasts are now briefly described. Consider an experimental factor with s levels. The first Helmert variable contrasts the first and second levels; the second variable contrasts the third level to the first two; the third variable contrasts level 4 to the first three; and so on. The coding rule for Helmert contrasts is illustrated by the following examples:

2 groups: 1 variable	**3 groups: 2 variables**	**4 groups: 3 variables**	**5 groups: 4 variables**	**etc.**
$\begin{bmatrix} -1 \\ +1 \end{bmatrix}$	$\begin{bmatrix} -1 & -1 \\ +1 & -1 \\ 0 & +2 \end{bmatrix}$	$\begin{bmatrix} -1 & -1 & -1 \\ +1 & -1 & -1 \\ 0 & +2 & -1 \\ 0 & 0 & +3 \end{bmatrix}$	$\begin{bmatrix} -1 & -1 & -1 & -1 \\ +1 & -1 & -1 & -1 \\ 0 & +2 & -1 & -1 \\ 0 & 0 & +3 & -1 \\ 0 & 0 & 0 & +4 \end{bmatrix}$	



	Arrows							
	1	2	3	4	5	6	7	8
Lake 1	1	0	1	1	0	0	0	0
Lake 2	1	0	1	0	0	0	0	0
Lake 3	1	1	0	0	0	0	0	0
Lake 4	0	0	0	0	1	0	1	1
Lake 5	0	0	0	0	0	1	1	1
Lake 6	0	0	0	0	0	0	0	1

Figure 1.14 Lakes interconnected by a river network (left) can be binary-coded as shown in the table to the right. Numbers are assigned in an arbitrary order to the directional edges (arrows) of the network. It is not useful to code the root of the network (arrow 0) in the matrix since all lakes would be coded '1' for that arrow. This example is revisited in Subsection 14.3.1.

Contrasts can be constructed based on some quantitative variable of interest associated with the objects, instead of the levels of a qualitative variable. Polynomial contrasts are based on an orthogonal polynomial of the quantitative variable of interest. The reference variable may be the position of the observations along a transect or a time series, or along an ecological gradient of altitude, pH, humidity, and so on. The contrasts are the successive monomials of the polynomial of the variable of interest, centred and made orthogonal to the lower-degree monomials; the monomials are then usually standardized to have a sum-of-squares of 1. Polynomial contrasts are used as explanatory variables in analyses in the same way as Helmert contrasts.

Other forms of coding have been developed for special types of variables. In phylogenetic analysis, the states of multistate characters are sometimes related by a hypothesized transformation series, going from the single hypothesized ancestral state to all the advanced states; such a series can be represented by a directed network where the states are connected by arrows representing evolutionary progression. A transformation series may be coded into binary variables using a method proposed by Kluge & Farris (1969).

River network

This same method may be applied to code the spatial relationships among localities in a geographic network. An example in freshwater ecology is a group of lakes connected by a river network (Fig. 1.14). In this example, a pseudo-map containing rivers and lakes is drawn to represent the network. A number is assigned to each river segment (the river segments are the edges of the connected graph) while nodes represent the furcation points. In Fig. 1.14, the coding is based on the river segments; it could just as well be based on the nodes if one felt that the nodes were the important

carriers of geographic information (as in Magnan *et al*., 1994). If the phenomenon to be modelled is, for example, fish dispersal from downstream, the arrows can be drawn going upstream, as in Fig. 1.14. In the lake-by-arrow matrix, a value '1' is assigned to each arrow found downstream from a lake, representing the fact that the corresponding river segment may allow fish to travel from the root to that lake. All other arrows are coded '0' for that lake. The resulting matrix is a complete numerical coding of the hydrographic network information: knowing the coding procedure, one can entirely reconstruct the network topology from the matrix entries.

The coding method may be tailored to the ecological problem at hand. For a dispersion phenomenon going downstream, arrows could point the other way around; in that case, a lake would be coded '1' in the table for arrows arriving in that lake from upstream. The pattern of interconnections does not even need to be a tree-like structure; it may form a more general type of directed network, but no cycle is allowed. Coding the information allows the use of this type of geographical information in different types of numerical models, like multiple regression (Chapter 10) or canonical analysis (Chapter 11). In many of these methods, zeros and ones are interchangeable. This coding method for directional spatial processes will be further developed in Section 14.3 where it will serve as the basis for the *Asymmetric Eigenvector Maps* (AEM) method of spatial analysis.

1.6 Missing data

Ecological data matrices are often plagued by missing data. The latter do not necessarily result from negligence on the part of the field team; most often, they are caused by the breakdown of measuring equipment during field surveys, weather events that prevented sampling sites from being visited on a given date, lost or incorrectly preserved specimens, improper sampling procedures, and so on.

Three families of solutions are available to cope with this problem for the analysis of field survey data, if one can make the assumption that the missing values occur at random in the data set. Most of the approaches mentioned below are discussed by Little & Rubin (1987), who also proposed methods for estimating missing values in controlled experiments (when the missing values are only found in the outcome variable; their Chapter 2) as well as valid model-based likelihood estimation of missing values for situations where the distribution of missing values does not meet the randomness assumption stated above.

Missing values may be represented in data matrices by numbers that do not correspond to possible data values. Codes such as –1 or –9 are often used when the real data in the table are all positive numbers, as it is the case with species abundance data; otherwise, –99 or –999, or other such unambiguous codes, may be used. In spreadsheets, missing values are often represented by bullets or 'NA' symbols.

1 — Delete rows or columns

Delete any row or column of the data matrix (Section 2.1) containing missing values. If a few rows contain most of the missing values, proceed by *rowwise* (also called *listwise*) *deletion*; conversely, if most missing values are found in a few variables only, proceed by *columnwise deletion*. This is the simplest, yet the most costly method, as it throws away the valuable information present in the remainder of these rows or columns.

2 — Accommodate algorithms to missing data

Accommodate the numerical method in such a way that the missing values are skipped during calculations. For instance, when computing resemblance coefficients among rows (Q-mode) or columns (R-mode) of the data matrix (Chapter 7), a simple method is *pairwise deletion* of missing values. This means, for example, that when computing a correlation coefficient between variables $\mathbf{y}_1$ and $\mathbf{y}_2$, if the value of the tenth object is missing for $\mathbf{y}_2$, object $\mathbf{x}_{10}$ is skipped in the computation of this correlation value. When it comes to comparing $\mathbf{y}_1$ and $\mathbf{y}_3$, if $\mathbf{x}_{10}$ has no missing data for these variables, it is then kept in the calculation for this pair of variables. However, one must be aware that covariance and correlation matrices computed in this way may be indefinite (i.e. they may have negative eigenvalues; Table 2.2).

3 — Estimate missing values

Estimate the missing values, a method called *imputation* by Little & Rubin (1987). This is the best strategy when missing values are located all over the data matrix — contrary to the situation where the missing values are found in a few rows or columns only, in which case deletion of these rows or columns may be the strategy of choice. The assumption one has to make when estimating missing values is that the missing data are not grossly atypical compared to those present in the data set. Methods for estimating missing data are interesting in cases where the numerical algorithm required for analysing the data cannot accommodate missing values. Ecologists should never imagine, however, that the estimated values are ecologically meaningful; as a consequence, they should refrain from attempting to interpret these numbers in ecological terms. Ecologists should also keep in mind that the estimation procedure has not created the missing degrees of freedom that would have accompanied observations carried out in nature or in the laboratory.

Three groups of methods are available for replacing quantitative missing values.

• 3.1 — The easiest way, which is often used in computer programs, is to replace missing values by the mean of the variable, estimated from the values present in the data table. When doing so, one assumes that nothing is known about the data, outside of the weak assumption mentioned above that the missing value comes from the same population as the non-missing data. Although this solution produces covariance and correlation matrices that are positive semidefinite (Section 2.10), the variances and

covariances are systematically underestimated. One way around this problem is to select missing value estimates at random from some distribution with appropriate mean and variance. This is not recommended, however, when the relative positions of the objects are of interest (principal component analysis; Section 9.1). A variant of the same method is to use the median instead of the mean; it is more robust in the sense that it does not assume the distribution of values to be unskewed. It is also applicable to semiquantitative descriptors. For qualitative descriptors, use the most frequent state instead of the mean or median.

• 3.2 — Estimate the missing values by regression. Multiple linear regression (Section 10.3), with rowwise deletion of missing values, may be used when there are only a few missing values to estimate. The dependent (response) variable of the regression is the descriptor with missing value(s) while the independent (explanatory) variables are the other descriptors in the data table. After the regression equation has been computed from the objects without missing data, it can be used to estimate the missing value(s). Using this procedure, one assumes the descriptor with missing values to be linearly related to the other descriptors in the data table (unless some form of nonparametric or nonlinear multiple regression is being used) and the data to be approximately multivariate normal. This method also leads to underestimating the variances and covariances, but less so than in 3.1. An alternative approach is to use a regression program allowing for pairwise deletion of missing values in the estimation of the regression coefficients, although, in that case, a maximum likelihood estimation of the covariance matrix would be preferable (Little & Rubin, 1987, p. 152 *et seq*.).

If such a method cannot be used for estimating the covariance matrix and if the missing values are scattered throughout the data table, an approximate solution may be obtained as follows. Compute a series of simple linear regressions with pairwise deletion of missing values, and estimate the missing value from each of these simple regression equations in turn. The mean of these estimates is taken as the working estimated value. The assumptions are basically the same as in the multiple regression case (above). Other methods of imputation are available in specialized R packages; see Section 1.7.

To estimate missing values in qualitative (nominal) descriptors, use logistic regression (Section 10.3) instead of linear regression.

• 3.3 — Interpolate missing values in spatially correlated data. Positive spatial correlation (Section 1.1) means that near points in time or space are similar. This property allows the interpolation of missing or otherwise unknown values from the values of near points in the series. With spatial data, interpolation is the first step of any mapping procedure, and it may be done in a variety of ways (Subsection 13.2.2), including the kriging method developed by geostatisticians. The simplest such method is to assign to a missing data the value of its nearest neighbour. In time series, interpolation of missing values may be performed using the same methods; see also Shumway & Stoffer, 1982, and Mendelssohn & Cury, 1987, for a maximum likelihood method for estimating missing data in a time series using a state-space model.

Myers (1982, 1983, 1984) proposed a method, called co-kriging, that combines the power of principal component analysis (Section 9.1) with that of kriging. It allows the estimation of unknown values of a data series using both the values of the same variable at neighbouring sites and the known values of other variables, correlated with the first one, observed at the same or neighbouring points in space; the spatial inter-relationships of these variables are measured by a cross-variogram. This method is interesting for the estimation of missing data in broad-scale ecological surveys and to compute values at unobserved sites on a geographic surface.

1.7 Software

The methods presented in this introductory chapter are implemented in the R language.

1. Corrections for multiple testing (Box 1.1) can be done using the ***p.adjust()*** function of the STATS package.

2. Several R functions use permutation tests. They will be identified in later chapters where permutation-based statistical methods are presented. For R functions that do not rely on compiled code for intensive calculations, permutations are produced by the ***sample()*** function of the STATS package. That function can also carry out resampling with replacement (bootstrapping).

3. All standard statistical distributions, and many others, are available in the STATS package. To find out about them, type in the R console: ***help.search(****"distribution", package="stats"****)***. Additional statistical distributions are available in other R packages.

4. Ranging and standardization (Subsection 1.5.4), as well as other transformations, are available in the ***decostand()*** function of VEGAN. Variable standardization is also available through the ***scale()*** function of STATS. The Box-Cox transformation (Subsection 1.5.6) can be done using the ***boxcox.fit()*** function of the GEOR package.

5. Helmert contrasts are available in the ***contr.helmert()*** function of the STATS package; polynomial contrasts can be computed using the ***contr.poly()*** function of the same package. Contrast matrices corresponding to actual data files are generated using the ***model.matrix()*** function of the STATS package; this function calls ***contr.helmert()*** or ***contr.poly()*** for calculation of the contrasts.

6. Imputation of missing values using a principal component analysis model is available in function ***imputePCA()*** of MISSMDA. Function ***mice()*** of package MICE carries out multivariate imputation by chained equations.

Chapter

2 *Matrix algebra: a summary*

2.0 Matrix algebra

Matrix language is the algebraic form best suited to the present book. The following chapters will systematically use the flexible and synthetic formulation of *matrix algebra*, with which many ecologists are already acquainted.

There are many reasons why matrix algebra is especially well suited for ecology. The format of computer files, including spreadsheets, in which ecological *data sets* are now most often recorded, is a *matrix* format. The use of *matrix notation* thus provides an elegant and compact representation of ecological information and *matrix algebra* allows operations on whole data sets to be performed. Last but not least, *multidimensional methods*, discussed in following chapters, are nearly impossible to conceptualise and explain without resorting to matrix algebra.

Matrix algebra goes back more than one century: "After Sylvester had introduced matrices [...], it is Cayley who created their algebra [in 1851]" (translated from Bourbaki, 1960). Matrices are of great conceptual interest for theoretical formulations, but it is only with the increased use of *computers* that matrix algebra became truly popular with ecologists. The use of computers naturally enhances the use of matrix notation. Most scientific programming languages are adapted to matrix logic. All matrix operations described in this chapter can be carried out using advanced statistical languages such as R, S-PLUS® and MATLAB®.

Ecologists who are familiar with matrix algebra could read Sections 2.1 and 2.2 only, where the vocabulary and symbols used in the remainder of this book are defined. Other sections of Chapter 2 may then be consulted whenever necessary.

The present chapter is only a *summary* of matrix algebra. Readers looking for more complete presentations should consult Bronson (2011), where numerous exercises are found. Graybill (2001) and Gentle (2007) provide numerous applications in general

Table 2.1 Ecological data matrix.

	Descriptors						
Objects	$\mathbf{y}_1$	$\mathbf{y}_2$	$\mathbf{y}_3$	⋯	$\mathbf{y}_j$	⋯	$\mathbf{y}_p$
$\mathbf{x}_1$	y_{11}	y_{12}	y_{13}	⋯	y_{1j}	⋯	y_{1p}
$\mathbf{x}_2$	y_{21}	y_{22}	y_{23}	⋯	y_{2j}	⋯	y_{2p}
$\mathbf{x}_3$	y_{31}	y_{32}	y_{33}	⋯	y_{3j}	⋯	$_{3p}$
.	.	.	.		.		.
.	.	.	.		.		.
.	.	.	.		.		.
$\mathbf{x}_i$	y_{i1}	y_{i2}	y_{i3}	⋯	y_{ij}	⋯	y_{ip}
.	.	.	.		.		.
.	.	.	.		.		.
.	.	.	.		.		.
$\mathbf{x}_n$	y_{n1}	y_{n2}	y_{n3}	⋯	y_{nj}	⋯	y_{np}

statistics. There are also a number of recent books, such as Vinod (2011), explaining how to use matrix algebra in R. The older book of Green & Carroll (1976) stresses the geometric interpretation of matrix operations commonly used in statistics.

2.1 The ecological data matrix

Descriptor
Object

As explained in Section 1.4, ecological data are obtained as object-observations or sampling units, which are described by a set of state values corresponding to as many descriptors, or variables. Ecological data are generally recorded in a table where each column j corresponds to a descriptor y_j (species present in the sampling unit, physical or chemical variable, etc.) and each object i (sampling site, sampling unit, locality, observation) occupies one row. In each cell (i,j) of the table is found the state taken by object i for descriptor j (Table 2.1). Objects will be denoted by a boldface, lower-case letter $\mathbf{x}$, with a subscript i varying form 1 to n, referring to object $\mathbf{x}_i$. Similarly, descriptors will be denoted by a boldface, lower case letter $\mathbf{y}$ subscripted j, with j taking values from 1 to p, referring to descriptor $\mathbf{y}_j$. When considering two sets of descriptors, members of the second set will generally have subscripts k from 1 to m.

Following the same logic, the different values in a data matrix will be denoted by a doubly-subscripted y, the first subscript designating the object being described and the second subscript the descriptor. For example, y_{83} is the value taken by object 8 for descriptor 3.

As mentioned in Section 1.4, it is not always obvious which are the objects and which are the descriptors. In ecology, for example, the different sampling sites (objects) may be studied with respect to the species found therein. In contrast, when studying the behaviour or taxonomy of organisms belonging to a given taxonomic group, the objects are the organisms themselves, whereas one of the descriptors could be the types of habitat found at different sampling sites. To unambiguously identify objects and descriptors, one must decide which is the variable defined *a priori* (i.e. the objects). When conducting field or laboratory observations, the variable defined *a priori* is totally left to the researcher, who decides how many observations will be included in the study. Thus, in the first example above, the researcher could choose the number of sampling sites needed to study their species composition. What is observed, then, are the descriptors, namely the different species present and possibly their abundances. Another approach to the same problem would be to ask which of the two sets of variables the researcher could theoretically increase to infinity; this identifies the variable defined *a priori*, or the objects. In the first example, it is the number of samples that could be increased at will — the samples are therefore the objects — whereas the number of species is limited and depends strictly on the ecological characteristics of the sampling sites. In the second example, the variable defined *a priori* corresponds to the organisms themselves, and one of their descriptors could be their different habitats (states).

The distinction between objects and descriptors is not only theoretical. One may analyse either the relationships among descriptors for the set of objects in the study (R mode analysis), or the relationships among objects given the set of descriptors (Q mode study). It will be shown that the mathematical techniques that are appropriate for studying relationships among objects are not the same as those for descriptors. For example, correlation coefficients can only be used for studying relationships among descriptors, which are vectors of data observed on samples extracted from populations with a theoretically infinite number of elements; vector lengths are actually limited by the sampling effort. It would be incorrect to use a correlation coefficient to study the relationship between two objects across the set of descriptors; other measures of association are available for that purpose (Section 7.3). Similarly, when using the methods of multidimensional analysis that will be described in this book, it is important to know which are the descriptors and which are the objects, in order to avoid methodological errors. The results of incorrectly conducted analyses — and there are unfortunately many in the literature — are not necessarily wrong because, in ecology, phenomena that are easily identified are usually sturdy enough to withstand considerable distortion. What is a pity, however, is that the more subtle phenomena, i.e. the very ones for which advanced numerical techniques are used, could very well not emerge at all from a study based on inappropriate methodology.

Linear algebra

The table of ecological data described above is an array of numbers known as a *matrix*. The branch of mathematics dealing with matrices is *linear algebra*.

Matrix **Y** is a rectangular, ordered array of numbers y_{ij}, set out in rows and columns as in Table 2.1:

$$\mathbf{Y} = [y_{ij}] = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{bmatrix} \qquad \textbf{(2.1)}$$

Order

There are n rows and p columns. When the *order* of the matrix (also known as its *dimensions* or *format*) must be stated, a matrix of order $(n \times p)$, which contains $n \times p$ elements, is written $\mathbf{Y}_{np}$. As above, any given element of **Y** is denoted y_{ij}, where subscripts i and j identify the row and column, respectively (always in that conventional order).

In linear algebra, ordinary numbers are called *scalars*, to distinguish them from *matrices*.

The following notation will be used hereinafter: a matrix will be symbolised by a capital letter in boldface, such as **Y**. The same matrix could also be represented by its general element in italics and in brackets, such as $[y_{ij}]$, or alternatively by an enumeration of all its elements, also in italics and in brackets, as in eq. 2.1. Italics will only be used for algebraic symbols, not for actual numbers. Occasionally, other notations than brackets may be found in the literature, i.e. (y_{ij}), (y_i^j), $\{y_{ij}\}$, $\left\|y_i^j\right\|$, or $\langle iyj \rangle$.

Any subset of a matrix can be explicitly recognized. In the above matrix (eq. 2.1), for example, the following submatrices could be considered:

Square matrix

a *square matrix* $\begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{bmatrix}$

a *row matrix* $\begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1p} \end{bmatrix}$, or a *column matrix* $\begin{bmatrix} y_{12} \\ y_{22} \\ \cdot \\ \cdot \\ \cdot \\ y_{n2} \end{bmatrix}$

Matrix notation simplifies the writing of data sets. It also corresponds to the way computers work. Indeed, most programming languages are designed to input data as matrices (arrays) and manipulate them either directly or through a simple system of subscripts. This greatly simplifies programming the calculations. Accordingly, computer packages generally input data as matrices. In addition, many of the statistical models used in multidimensional analysis are based on linear algebra, as will be seen later. So, it is convenient to approach them with data already set in matrix format.

2.2 Association matrices

Two important matrices may be derived from the ecological data matrix: the association matrix among objects and the association matrix among descriptors. An association matrix is denoted **A**, and its general element a_{ij}. Although Chapter 7 is entirely devoted to association matrices, it is important to mention them here in order to better understand the purpose of methods presented in the remainder of the present chapter.

Using data from matrix **Y** (eq. 2.1), one may examine the relationship between the first two objects $\mathbf{x}_1$ and $\mathbf{x}_2$. In order to do so, the first and second rows of matrix **Y**

$$\begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1p} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} y_{21} & y_{22} & \cdots & y_{2p} \end{bmatrix}$$

are used to calculate a measure of association (similarity or distance: Chapter 7), to assess the degree of resemblance between the two objects. This measure, which quantifies the strength of the association between the two rows, is denoted a_{12}. In the same way, the association of $\mathbf{x}_1$ with $\mathbf{x}_3$, $\mathbf{x}_4$, …, $\mathbf{x}_p$, can be calculated, as can also be calculated the association of $\mathbf{x}_2$ with all other objects, and so on for all pairs of objects. The coefficients of association for all pairs of objects are then recorded in a table, ordered in such a way that they could be retrieved for further calculations. This table is the association matrix **A** among objects:

$$\mathbf{A}_{nn} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \tag{2.2}$$

A most important characteristic of any association matrix is that it has a number of rows equal to its number of columns, this number being equal here to the number of objects n. The number of elements in the above square matrix is therefore n^2.

Similarly, one may wish to examine the relationships among descriptors. For the first two descriptors, $\mathbf{y}_1$ and $\mathbf{y}_2$, the first and second columns of matrix $\mathbf{Y}$

$$\begin{bmatrix} y_{11} \\ y_{21} \\ \cdot \\ \cdot \\ \cdot \\ y_{n1} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} y_{12} \\ y_{22} \\ \cdot \\ \cdot \\ \cdot \\ y_{n2} \end{bmatrix}$$

are used to calculate a measure of dependence (Chapter 7) which assesses the degree of association between the two descriptors. In the same way as for the objects, $p \times p$ measures of association can be calculated among all pairs of descriptors and recorded in the following association matrix:

$$\mathbf{A}_{pp} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{bmatrix} \qquad \textbf{(2.3)}$$

Association matrices are most often (but not always, see Section 2.3) symmetric, with elements in the upper right triangle being equal to those in the lower left triangle ($a_{ij} = a_{ji}$). Elements a_{ii} on the diagonal measure the association of a row or a column of matrix $\mathbf{Y}$ with itself. In the case of objects, the measure of association a_{ii} of an *object* with itself usually takes a value of either 1 (similarity coefficients) or 0 (distance coefficients). Concerning the association between *descriptors* (columns), the correlation a_{ii} of a descriptor with itself is 1, whereas the (co)variance provides an estimate a_{ii} of the variability among the values of descriptor i.

At this point of the discussion, it should thus be noted that the data, to which the models of multidimensional analysis are applied, are not only matrix $\mathbf{Y}_{np}$ = [objects × descriptors] (eq. 2.1), but also the two association matrices $\mathbf{A}_{nn}$ = [objects × objects] (eq. 2.2) and $\mathbf{A}_{pp}$ = [descriptors × descriptors] (eq. 2.3), as shown in Fig. 2.1.

2.3 Special matrices

Matrices with an equal number of rows and columns are called *square* matrices (Section 2.1). These, as will be seen in Sections 2.6 *et seq*., are the only matrices for

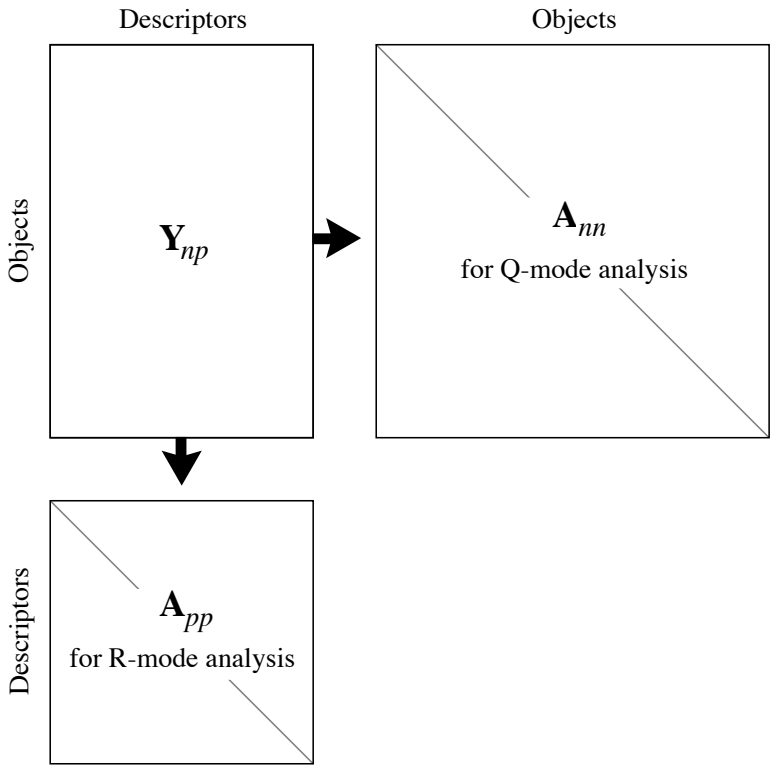


Figure 2.1 Data analysed in numerical ecology include matrix $\mathbf{Y}_{np}$ = [objects × descriptors] (eq. 2.1) as well as the two association matrices $\mathbf{A}_{nn}$ = [objects × objects] (eq. 2.2) and $\mathbf{A}_{pp}$ = [descriptors × descriptors] (eq. 2.3). The Q and R modes of analysis are defined in Section 7.1.

which it is possible to compute a determinant, an inverse, and eigenvalues and eigenvectors. As a corollary, these operations can be carried out on association matrices, which are square matrices.

Some definitions pertaining to square matrices now follow. In matrix $\mathbf{B}_{nn}$, of order $(n \times n)$ (often called "square matrix of order n" or "matrix of order n"),

$$\mathbf{B}_{nn} = [b_{ij}] = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ b_{n1} & b_{n2} & \cdots & b_{nn} \end{bmatrix} \quad \textbf{(2.4)}$$

the *diagonal elements* are those with identical subscripts for the rows and columns (b_{ii}). They are located on the *main diagonal* (simply called *the diagonal*) which, by convention, goes from the upper left to the lower right corners. The sum of the diagonal elements is called the *trace* of the matrix.

Trace

Diagonal matrix

A *diagonal matrix* is a square matrix where all non-diagonal elements are *zero*. Thus,

$$\begin{bmatrix} 3 & 0 & 0 \\ 0 & 7 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

is a diagonal matrix. Diagonal matrices that contain on their diagonal values coming from a vector $[x_i]$ are noted $\mathbf{D}(x)$. Special examples used later in the book are the diagonal matrix of standard deviations $\mathbf{D}(\sigma)$, the diagonal matrix of eigenvalues $\mathbf{D}(\lambda_i)$, also noted $\mathbf{\Lambda}$, and the diagonal matrix of singular values $\mathbf{D}(w_i)$ also noted $\mathbf{W}$.

Identity matrix

A diagonal matrix where all diagonal elements are equal to unity is called a *unit matrix* or *identity matrix*. It is denoted $\mathbf{D}(1)$ or $\mathbf{I}$:

$$\mathbf{D}(1) = \mathbf{I} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ 0 & 0 & \dots & 1 \end{bmatrix} \qquad \textbf{(2.5)}$$

This matrix plays the same role, in matrix algebra, as the number 1 in ordinary algebra, i.e. it is the neutral element in multiplication (e.g. $\mathbf{I}\,\mathbf{B} = \mathbf{B}$, or $\mathbf{B}\,\mathbf{I} = \mathbf{B}$).

Scalar matrix

Similarly, a *scalar matrix* is a diagonal matrix of the form

$$\begin{bmatrix} 7 & 0 & \dots & 0 \\ 0 & 7 & \dots & 0 \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ 0 & 0 & \dots & 7 \end{bmatrix} = 7\mathbf{I}$$

All the diagonal elements are identical since a scalar matrix is the unit matrix multiplied by a scalar (here, of value 7).

Null matrix

A matrix, square or rectangular, whose elements are all zero is called a *null matrix* or *zero matrix*. It is denoted **0** or [0].*

Triangular matrix

A square matrix with all elements above (or below) the diagonal being zero is called a *lower* (or *upper*) *triangular* matrix. For example,

$$\begin{bmatrix} 1 & 2 & 3 \\ 0 & 4 & 5 \\ 0 & 0 & 6 \end{bmatrix}$$

is an upper triangular matrix. These matrices are very important in matrix algebra because their determinant (Section 2.6) is equal to the product of all terms on the main diagonal (i.e. 24 in this example). Diagonal matrices are also triangular matrices.

Transpose

The *transpose* of a matrix **B** with format $(n \times p)$ is denoted **B**' and is a new matrix of format $(p \times n)$ in which $b'_{ij} = b_{ji}$. In other words, the rows of one matrix are the columns of the other. Thus, the transpose of matrix

$$\mathbf{B} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \end{bmatrix}$$

is matrix

$$\mathbf{B}' = \begin{bmatrix} 1 & 4 & 7 & 10 \\ 2 & 5 & 8 & 11 \\ 3 & 6 & 9 & 12 \end{bmatrix}$$

Transposition is an important operation in linear algebra, and also in ecology where a data matrix **Y** (eq. 2.1) may be transposed to study the relationships among descriptors after the relationships among objects have been analysed (or conversely).

---

* Although the concept of zero was known to Babylonian and Mayan astronomers, inclusion of the zero in a decimal system of numeration finds its origin in India, in the eighth century A.D. at least (Ifrah, 1981). The ten Western-world numerals are also derived from the symbols used by ancient Indian mathematicians. The word *zero* comes from the Arabs, however. They used the word *sifr*, meaning "empty", to refer to a symbol designating nothingness. The term turned into *cipher*, and came to denote not only zero, but all 10 numerals. *Sifr* is at the root of the medieval latin *zephirum*, which became *zefiro* in Italian and was then abbreviated to *zero*. It is also the root of the medieval latin *cifra*, which became *chiffre* in French where it designates any of the 10 numerals.

Symmetric matrix

A square matrix that is identical to its transpose is *symmetric*. This is the case when corresponding terms b_{ij} and b_{ji}, on either side of the diagonal, are equal. For example,

$$\begin{bmatrix} 1 & 4 & 6 \\ 4 & 2 & 5 \\ 6 & 5 & 3 \end{bmatrix}$$

is symmetric since $\mathbf{B}' = \mathbf{B}$. All symmetric matrices are square.

Non-symmetric matrix

It was mentioned in Section 2.2 that association matrices are generally symmetric. *Non-symmetric* (or *asymmetric*) matrices may be encountered, however. This happens, for example, when each coefficient in the matrix measures the ecological influence of an organism or a species on another, these influences being asymmetrical (e.g. A is a predator of B, B is a prey of A). Asymmetric matrices are also found in behaviour studies, serology, DNA pairing analysis, etc.

Skew-symmetric matrix

Matrix algebra tells us that any *non-symmetric* matrix may be expressed as the sum of two other matrices, one *symmetric* and one *skew-symmetric*, without loss of information. Consider for instance the two numbers 1 and 3, found in opposite positions (1,2) and (2,1) of the first matrix in the following numerical example:

$$\begin{bmatrix} 1 & 1 & 2 & 2 \\ 3 & 1 & 0 & -1 \\ 1 & 2 & 1 & 0 \\ 0 & -4 & 3 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 2.0 & 1.5 & 1.0 \\ 2.0 & 1 & 1.0 & -2.5 \\ 1.5 & 1.0 & 1 & 1.5 \\ 1.0 & -2.5 & 1.5 & 1 \end{bmatrix} + \begin{bmatrix} 0 & -1.0 & 0.5 & 1.0 \\ 1.0 & 0 & -1.0 & 1.5 \\ -0.5 & 1.0 & 0 & -1.5 \\ -1.0 & -1.5 & 1.5 & 0 \end{bmatrix}$$

Non-symmetric Symmetric (average) Skew-symmetric

The *symmetric* part is obtained by averaging these two numbers: $(1 + 3)/2 = 2.0$. The *skew-symmetric* part is obtained by subtracting one from the other and dividing by 2: $(1 - 3)/2 = -1.0$ and $(3 - 1)/2 = +1.0$ so that, in the skew-symmetric matrix, corresponding elements on either side of the diagonal have the same absolute values but opposite signs. When the symmetric and skew-symmetric components are added, the result is the original matrix: $2 - 1 = 1$ for the upper original number, and $2 + 1 = 3$ for the lower one. Using letters instead of numbers, one can derive a simple algebraic proof of the additivity of the symmetric and skew-symmetric components. The symmetric component can be analysed using the methods applicable to symmetric matrices (for instance, metric or non-metric scaling, Sections 9.3 and 9.4), while analysis of the skew-symmetric component requires methods especially developed to assess asymmetric relationships. Basic references are Coleman (1964) in the field of sociometry and Digby & Kempton (1987, Ch. 6) in numerical ecology. An application to biological evolution is found in Casgrain *et al*. (1996). Relevant biological or ecological information may be found in the symmetric portion only and, in other instances, in the skew-symmetric component only.

2.4 Vectors and scaling

Vector

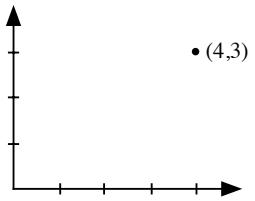
Another matrix of special interest is the *column matrix*, with format ($n \times 1$), which is also known as a *vector*. Some textbooks restrict the term "vector" to *column matrices*, but the expression *row vector* (or simply *vector*, as used in some instances in Chapter 4) may also designate *row matrices*, with format ($1 \times p$).

A (column) vector is noted as follows:

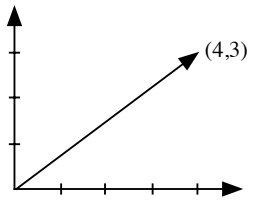
$$\mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \cdot \\ \cdot \\ \cdot \\ b_n \end{bmatrix} \tag{2.6}$$

A vector graphically refers to a directed line segment. It also forms a mathematical entity on which operations can be performed. More formally, a vector is defined as an ordered n-tuple of real numbers, i.e. a set of n numbers with a specified order. The n numbers are the coordinates of a point in a n-dimensional Euclidean space, which may be seen as the end-point of a line segment starting at the origin.

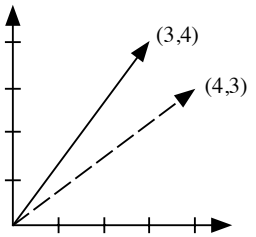
For example, (column) vector [4 3]' is an ordered doublet (or 2-tuple) of two real numbers (4, 3), which may be represented in a two-dimensional Euclidean space:



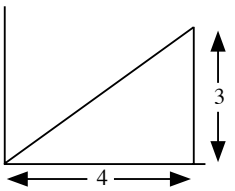
This same point (4, 3) may also be seen as the end-point of a line segment starting at the origin:



These figures illustrate the two possible representations of a vector; they also stress the *ordered* nature of vectors, since vector [3 4]' is different from vector [4 3]'.



Using the Pythagorean theorem, it is easy to calculate the length of any vector. For example, the length of vector [4 3]' is that of the hypotenuse of a right triangle with base 4 and height 3:



Length
Norm

The length (or *norm*) of vector [4 3]' is therefore $\sqrt{4^2 + 3^2} = 5$; it is also the length (norm) of vector [3 4]'. The norm of vector b is noted $\|\mathbf{b}\|$.

Scaling
Normalization

The comparison of different vectors, as to their directions, often requires an operation called *scaling*. In the scaled vector, all elements are divided by the same characteristic value. A special type of scaling is called *normalization*. In the normalized vector, each element is divided by the length of the vector:

$$\begin{bmatrix} 4 \\ 3 \end{bmatrix} \xrightarrow{\text{normalization}} \begin{bmatrix} 4/5 \\ 3/5 \end{bmatrix}$$

Normalized vector

The importance of normalization lies in the fact that the length of a *normalized vector* is equal to unity. Indeed, the length of vector [4/5 3/5]', calculated by means of the Pythagorean formula, is $\sqrt{(4/5)^2 + (3/5)^2} = 1$.

The example of doublet (4, 3) may be generalized to any n-tuple $(b_1, b_2, \ldots, b_n)$, which specifies a vector in n-dimensional space. The length of the vector is $\sqrt{b_1^2 + b_2^2 + \ldots + b_n^2}$, so that the corresponding normalized vector is:

$$\begin{bmatrix} b_1 / \sqrt{b_1^2 + b_2^2 + \ldots + b_n^2} \\ b_2 / \sqrt{b_1^2 + b_2^2 + \ldots + b_n^2} \\ \cdot \\ \cdot \\ \cdot \\ b_n / \sqrt{b_1^2 + b_2^2 + \ldots + b_n^2} \end{bmatrix} = \frac{1}{\sqrt{b_1^2 + b_2^2 + \ldots + b_n^2}} \begin{bmatrix} b_1 \\ b_2 \\ \cdot \\ \cdot \\ \cdot \\ b_n \end{bmatrix} \qquad \textbf{(2.7)}$$

The length of any normalized vector, in n-dimensional space, is 1.

2.5 Matrix addition and multiplication

Recording the data in table form, as is usually the case in ecology, opens the possibility of performing operations on these tables. The basic operations of matrix algebra (*algebra*, from the Arabic "al-jabr" which means *reduction*, is the theory of addition and multiplication) are very natural and familiar to ecologists.

Numerical example. Fish (3 species) were sampled at five sites in a lake, once a month during the summer (northern hemisphere). In order to get a general idea of the differences among sites, total numbers of fish caught at each site are calculated over the whole summer:

	July				August				September				Whole summer		
Site 1	1	5	35	+	15	23	10	+	48	78	170	=	64	106	215
Site 2	14	2	0		54	96	240		2	0	0		70	98	240
Site 3	0	31	67		0	3	9		0	11	14		0	45	90
Site 4	96	110	78		12	31	27		25	13	12		133	154	117
Site 5	0	0	0		8	14	6		131	96	43		139	110	49
	*sp*1	*sp*2	*sp*3		*sp*1	*sp*2	*sp*3		*sp*1	*sp*2	*sp*3		*sp*1	*sp*2	*sp*3

This operation is known as *matrix addition*. Note that only matrices of the same order can be added together. This is why, in the first matrix, site 5 was included with abundances of 0 to indicate that no fish had been caught there in July although site 5 had been sampled. Adding two matrices consists in a term-by-term addition. Matrix addition is associative and commutative; its neutral element is the null matrix **0**.

To study seasonal changes in fish productivity at each site, one possible approach would be to add together the terms in each row of each monthly matrix. However, this makes sense only if the selectivity of the fishing gear (say, a net) is comparable for the three species. Let us imagine that the efficiency of the net was 50% for species 2 and 25% for species 3 of what it was for species 1. In such a case, values in each row must be corrected before being added. Correction factors would be as follows: 1 for species 1, 2 for species 2, and 4 for species 3. To obtain

estimates of total fish abundances, correction vector [1 2 4]' is first multiplied by each row of each matrix, after which the resulting values are added. Thus, for the first site in July:

Site 1 July	Correction factors	Total fish abundance Site 1, July
$\begin{bmatrix}1 & 5 & 35\end{bmatrix}$	$\begin{bmatrix}1\\2\\4\end{bmatrix}$	$(1 \times 1) + (5 \times 2) + (35 \times 4) = 1 + 10 + 140 = 151$

Scalar product

This operation is known in linear algebra as a *scalar product* because this product of two vectors produces a scalar.

In physics, there is another product of two vectors, called the *external* or *vector product*, where the multiplication of two vectors results in a third one, which is perpendicular to the plane formed by the first two. This product is not used in multidimensional analysis. It is however important to know that, in the literature, the expression "vector product" may be used for either that product or the scalar product of linear algebra, and that the scalar product is also called "inner product" or "dot product". The vector product (of physics) is sometimes called "cross product". This last expression is also used in linear algebra, for example in "matrix of sum of squares and cross products" (SSCP matrix), which refers to the product of a matrix with its transpose.

In matrix algebra, and unless otherwise specified, multiplication follows a convention that is illustrated by the scalar product above: in this *product* of a column vector *by* a row vector, the row vector *multiplies* the column vector or, which is equivalent, the column vector *is multiplied by* the row vector. This convention, which should be kept in mind, will be followed in the remainder of the book.

The result of a scalar product is a number, which is equal to the sum of the products of those elements with corresponding order numbers. The scalar product is designated by a dot, or is written **<a**,**b>**, or else there is no sign between the two terms. For example:

$$\mathbf{b'c} = \mathbf{b'} \bullet \mathbf{c} = \begin{bmatrix} b_1 & b_2 & \cdots & b_p \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \cdot \\ \cdot \\ \cdot \\ c_p \end{bmatrix} = b_1c_1 + b_2c_2 + \ldots + b_pc_p = \text{a scalar.} \quad \textbf{(2.8)}$$

The rules for computing scalar products are such that only vectors with the same numbers of elements can be multiplied.

In analytic geometry, it can be shown that the scalar product of two vectors obeys the relationship:

$$\mathbf{b}' \bullet \mathbf{c} = (\text{length of } \mathbf{b}) \times (\text{length of } \mathbf{c}) \times \cos\theta \qquad \textbf{(2.9)}$$

Orthogonal vectors

When the angle between two vectors is $\theta = 90°$, then $\cos\theta = 0$ and the scalar product $\mathbf{b}' \bullet \mathbf{c} = 0$. As a consequence, two vectors whose scalar product is zero are *orthogonal* (i.e. at right angle). This property will be used in Section 2.9 to compute eigenvectors. A matrix whose (column) vectors are all orthogonal to one another is called *orthogonal*. For any pair of vectors **b** and **c** with values centred on their respective mean, $\cos\theta = r(\mathbf{b}, \mathbf{c})$ where r is the correlation coefficient (eq. 4.7).

Gram-Schmidt orthogonalization is a procedure to make a vector **c** orthogonal to a vector **b** that has first been normalized (eq. 2.7); **c** may have been normalized or not. The procedure consists of two steps: (1) compute the scalar product $sp = \mathbf{b}'\mathbf{c}$. (2) Make **c** orthogonal to **b** by computing $\mathbf{c}_{ortho} = \mathbf{c} - sp\mathbf{b}$. Proof that $\mathbf{c}_{ortho}$ is orthogonal to **b** is obtained by showing that $\mathbf{b}'\mathbf{c}_{ortho} = 0$: $\mathbf{b}'\mathbf{c}_{ortho} = \mathbf{b}'(\mathbf{c} - sp\mathbf{b}) = \mathbf{b}'\mathbf{c} - sp\mathbf{b}'\mathbf{b}$. Since $\mathbf{b}'\mathbf{c} = sp$ and $\mathbf{b}'\mathbf{b} = 1$ because **b** has been normalized, one obtains $sp - (sp \times 1) = 0$. In this book, in the iterative procedures for ordination algorithms (Tables 9.5 and 9.8), Gram-Schmidt orthogonalization will be used in the step where the vectors of new ordination object scores are made orthogonal to previously found vectors.

Numerical example. Returning to the above example, it is possible to multiply each row of each monthly matrix with the correction vector (scalar product) in order to compare total monthly fish abundances. This operation, which is the *product of a vector by a matrix*, is a simple extension of the scalar product (eq. 2.8). The product of the July matrix **B** with the correction vector **c** is written as follows:

$$\begin{bmatrix} 1 & 5 & 35 \\ 14 & 2 & 0 \\ 0 & 31 & 67 \\ 96 & 110 & 78 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix} = \begin{bmatrix} 1(1) + 5(2) + 35(4) \\ 14(1) + 2(2) + 0(4) \\ 0(1) + 31(2) + 67(4) \\ 96(1) + 110(2) + 78(4) \\ 0(1) + 0(2) + 0(4) \end{bmatrix} = \begin{bmatrix} 151 \\ 18 \\ 330 \\ 628 \\ 0 \end{bmatrix}$$

The product of a vector by a matrix involves calculating, for each row of matrix **B**, a scalar product with vector **c**. Such a product of a vector by a matrix is only possible if the number of *elements in the vector* is the same as the number of *columns in the matrix*. The result is no longer a scalar, but a column vector with dimension equal to the number of rows in the matrix on the left. The general formula for this product is:

$$\mathbf{B}_{pq} \bullet \mathbf{c}_q = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1q} \\ b_{21} & b_{22} & \cdots & b_{2q} \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ b_{p1} & b_{p2} & \cdots & b_{pq} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \cdot \\ \cdot \\ \cdot \\ c_q \end{bmatrix} = \begin{bmatrix} b_{11}c_1 + b_{12}c_2 + \cdots + b_{1q}c_q \\ b_{21}c_1 + b_{22}c_2 + \cdots + b_{2q}c_q \\ \cdot \\ \cdot \\ \cdot \\ b_{p1}c_1 + b_{p2}c_2 + \cdots + b_{pq}c_q \end{bmatrix}$$

Using summation notation, this equation may be rewritten as:

$$\mathbf{B}_{pq} \bullet \mathbf{c}_q = \begin{bmatrix} \sum_{k=1}^{q} b_{1k} c_k \\ \vdots \\ \sum_{k=1}^{q} b_{pk} c_k \end{bmatrix} \qquad \textbf{(2.10)}$$

The product of two matrices is the logical extension of the product of a vector by a matrix. Matrix **C**, to be multiplied by **B**, is simply considered as a set of column vectors $\mathbf{c}_1, \mathbf{c}_2, \ldots$; eq. 2.10 is repeated for each column. Following the same logic, the resulting column vectors are juxtaposed to form the result matrix. Matrices to be multiplied must be *conformable*, which means that the number of columns in the matrix on the left must be the same as the number of rows in the matrix on the right. For example, given

$$\mathbf{B} = \begin{bmatrix} 1 & 0 & 2 \\ 3 & 1 & 1 \\ 1 & 2 & 1 \\ -1 & 3 & 2 \end{bmatrix} \qquad \text{and} \qquad \mathbf{C} = \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 3 & -1 \end{bmatrix}$$

$$\mathbf{C} = [\,\mathbf{d} \quad \mathbf{e}\,]$$

the product of **B** with each of the two columns of **C** is:

$$\mathbf{Bd} = \begin{bmatrix} 1(1) + 0(2) + 2(3) \\ 3(1) + 1(2) + 1(3) \\ 1(1) + 2(2) + 1(3) \\ -1(1) + 3(2) + 2(3) \end{bmatrix} = \begin{bmatrix} 7 \\ 8 \\ 8 \\ 11 \end{bmatrix} \quad \text{and} \quad \mathbf{Be} = \begin{bmatrix} 1(2) + 0(1) + 2(-1) \\ 3(2) + 1(1) + 1(-1) \\ 1(2) + 2(1) + 1(-1) \\ -1(2) + 3(1) + 2(-1) \end{bmatrix} = \begin{bmatrix} 0 \\ 6 \\ 3 \\ -1 \end{bmatrix}$$

so that the product matrix is:

$$\mathbf{BC} = \begin{bmatrix} 7 & 0 \\ 8 & 6 \\ 8 & 3 \\ 11 & -1 \end{bmatrix}$$

Thus, the product of two conformable matrices **B** and **C** is a new matrix with the same number of rows as **B** and the same number of columns as **C**. Element d_{ij}, in row i and column j of the resulting matrix, is the scalar product of row i of **B** with column j of **C**.

The only way to master the mechanism of matrix products is to go through some numerical examples. As an exercise, readers could apply the above method to two cases which have not been discussed so far, i.e. the *product* (**bc**) *of a row vector* **c** *by a column vector* **b**, which gives a *matrix* and not a scalar, and the *product* (**bC**) *of a matrix* **C** *by a row vector* **b**, which results in a *row vector*. This exercise would help to better understand the rule of conformability.

As supplementary exercises, readers could calculate numerical examples of the eight following properties of matrix products, which will be used later in the book:

(1) $\mathbf{B}_{pq}\,\mathbf{C}_{qr}\,\mathbf{D}_{rs} = \mathbf{E}_{ps}$, of order $(p \times s)$.

(2) The existence of product **BC** does not imply that product **CB** exists, because matrices are not necessarily conformable in the reverse order; however, **C'C** and **CC'** always exist.

(3) **BC** is generally not equal to **CB**, i.e. matrix products are not commutative.

(4) $\mathbf{B}^2 = \mathbf{B} \times \mathbf{B}$ exists only if **B** is a square matrix.

(5) [**AB**]' = **B'A'** and, more generally, [**ABCD…**]' = **…D'C'B'A'**.

(6) The products **XX'** and **X'X** always give rise to symmetric matrices.

(7) In general, the product of two symmetric but different matrices **A** and **B** is not a symmetric matrix.

(8) If **B** is an orthogonal matrix (i.e. a rectangular matrix whose column vectors are orthogonal to one another), then **B'B** = **D**, where **D** is a diagonal matrix. All non-diagonal terms are zero because of the property of orthogonality, while the diagonal terms are the squares of the lengths of the column vectors. That **B'B** is diagonal does not imply that **BB'** is also diagonal. **BB'** = **B'B** only when **B** is square and symmetric.

Hadamard product

The *Hadamard* or *elementwise* product of two matrices of the same order $(n \times p)$ is the cell-by-cell product of these two matrices. For example,

$$\text{for } \mathbf{A} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix} \text{ and } \mathbf{B} = \begin{bmatrix} 7 & 8 \\ 9 & 10 \\ 11 & 12 \end{bmatrix}, \qquad \mathbf{A}*\mathbf{B} = \begin{bmatrix} 7 & 16 \\ 27 & 40 \\ 55 & 72 \end{bmatrix}$$

The Hadamard product may be noted by different operator signs, depending on the author. The sign used in this book is $*$, as in the R language.

The last type of product to be considered is that of a matrix or vector *by a scalar*. It is carried out according to the usual algebraic rules of multiplication and factoring, i.e. for matrix $\mathbf{B} = [b_{jk}]$ or vector $\mathbf{c} = [c_j]$, $d\mathbf{B} = [db_{jk}]$ and $d\mathbf{c} = [dc_j]$. For example:

$$3\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 3 & 6 \\ 9 & 12 \end{bmatrix} \qquad \text{and} \qquad \begin{bmatrix} 5 \\ 6 \end{bmatrix} 2 = \begin{bmatrix} 10 \\ 12 \end{bmatrix}$$

The terms *premultiplication* and *postmultiplication* may be encountered in the literature. Product **BC** corresponds to *premultiplication* of **C** by **B**, or to *postmultiplication* of **B** by **C**. Unless otherwise specified, it is always premultiplication which is implied and **BC** simply reads: **B** multiplies **C**, or **C** is multiplied by **B**.

2.6 Determinant

It is often necessary to transform a matrix into a new one, in such a way that the information of the original matrix is preserved, while new properties that are essential for subsequent calculations are acquired. Such new matrices, which are linearly derived from the original matrix, will be studied in following sections under the names *inverse matrix*, *canonical form*, etc.

The new matrix must have a minimum number of characteristics in common with the matrix from which it is linearly derived. The connection between the two matrices is a matrix function $f(\mathbf{B})$, whose properties are the following:

(1) The determinant function must be *multilinear*, which means that it should respond linearly to any change taking place in the rows or columns of matrix **B**.

(2) Since the order of the rows and columns of a matrix is specified, the function should be able to detect, through *alternation of signs*, any change in the positions of rows or columns. As a corollary, if two columns (or rows) are identical, $f(\mathbf{B}) = 0$; indeed, if two identical columns (or rows) are interchanged, $f(\mathbf{B})$ must change sign but it must also remain identical, which is possible only if $f(\mathbf{B}) = 0$.

(3) Finally, there is a scalar associated with this function; it is called its *norm* or *value* of the determinant function. For convenience, the norm is calibrated in such a way that the value associated with the unit matrix **I** is 1, i.e. $f(\mathbf{I}) = 1$.

It can be shown that the determinant, as defined below, is the only function that has the above three properties, and that it only exists for square matrices. Therefore, it is not possible to calculate a determinant for a rectangular matrix. The determinant of matrix **B** is denoted *det* **B**, det(**B**), or, more often, $|\mathbf{B}|$:

$$|\mathbf{B}| \equiv \begin{vmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ b_{n1} & b_{n2} & \cdots & b_{nn} \end{vmatrix}$$

The value of function $|\mathbf{B}|$ is a scalar, i.e. a number.

What follows is the formal definition of the value of a determinant. The way to compute it in practice is explained later. The *value of a determinant* is calculated as the sum of all possible products containing one, and only one, element from each row and each column; these products receive a sign according to a well-defined rule:

$$|\mathbf{B}| = \sum \pm (b_{1j_1} b_{2j_2} \ldots b_{nj_n})$$

where indices $j_1, j_2, \ldots, j_n$, go through the $n!$ permutations of the numbers $1, 2, \ldots, n$. The sign depends on the number of inversions, in the permutation considered, relative to the sequence $1, 2, \ldots, n$: if the number of inversions is even, the sign is (+) and, if the number is odd, the sign is (–).

The determinant of a matrix of order 2 is calculated as follows:

$$|\mathbf{B}| = \begin{vmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{vmatrix} = b_{11}b_{22} - b_{12}b_{21} \qquad \textbf{(2.11)}$$

In accordance with the formal definition above, the scalar so obtained is composed of $2! = 2$ products, each product containing one, and only one, element from each row and each column.

Expansion by minors

The determinant of a matrix of order higher than 2 may be calculated using different methods, among which is the *expansion by minors*. When looking for a determinant of order 3, a determinant of order $3 - 1 = 2$ may be obtained by crossing out one row (i) and one column (j). This lower-order determinant is the *minor* associated with b_{ij}:

$$\begin{vmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{vmatrix} \rightarrow \begin{vmatrix} b_{21} & b_{23} \\ b_{31} & b_{33} \end{vmatrix} \qquad \textbf{(2.12)}$$

crossing out row 1 *and column* 2

minor of b_{12}

Cofactor

The minor being here a determinant of order 2, its value is calculated using eq. 2.11. When multiplied by $(-1)^{i+j}$, the minor becomes a *cofactor*. Thus, the cofactor of b_{12} is:

$$\text{cof } b_{12} = (-1)^{1+2} \begin{vmatrix} b_{21} & b_{23} \\ b_{31} & b_{33} \end{vmatrix} = - \begin{vmatrix} b_{21} & b_{23} \\ b_{31} & b_{33} \end{vmatrix} \qquad \textbf{(2.13)}$$

The expansion by minors of a determinant of order n is:

$$|\mathbf{B}| = \sum_{i=1}^{n} b_{ij} \text{cof } b_{ij} \qquad \text{for any column } j \qquad \textbf{(2.14)}$$

$$|\mathbf{B}| = \sum_{j=1}^{n} b_{ij} \text{cof } b_{ij} \qquad \text{for any row } i$$

The expansion may involve the elements of any row or any column, the result being always the same. Thus, going back to the determinant of the matrix on the left in eq. 2.12, expansion by the elements of the first row gives:

$$|\mathbf{B}| = b_{11} \text{cof } b_{11} + b_{12} \text{cof } b_{12} + b_{13} \text{cof } b_{13} \qquad \textbf{(2.15)}$$

$$|\mathbf{B}| = b_{11}(-1)^{1+1} \begin{vmatrix} b_{22} & b_{23} \\ b_{32} & b_{33} \end{vmatrix} + b_{12}(-1)^{1+2} \begin{vmatrix} b_{21} & b_{23} \\ b_{31} & b_{33} \end{vmatrix} + b_{13}(-1)^{1+3} \begin{vmatrix} b_{21} & b_{22} \\ b_{31} & b_{32} \end{vmatrix}$$

Numerical example. Equation 2.15 is applied to a simple numerical example:

$$\begin{vmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 10 \end{vmatrix} = 1(-1)^{1+1} \begin{vmatrix} 5 & 6 \\ 8 & 10 \end{vmatrix} + 2(-1)^{1+2} \begin{vmatrix} 4 & 6 \\ 7 & 10 \end{vmatrix} + 3(-1)^{1+3} \begin{vmatrix} 4 & 5 \\ 7 & 8 \end{vmatrix}$$

$$\begin{vmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 10 \end{vmatrix} = 1(5 \times 10 - 6 \times 8) - 2(4 \times 10 - 6 \times 7) + 3(4 \times 8 - 5 \times 7) = -3$$

The amount of calculations required to expand a determinant increases very quickly with increasing order n. This is because the minor of each cofactor must be expanded, the latter producing new cofactors whose minors are in turn expanded, and so forth until cofactors of order 2 are reached. Another, faster method is normally used to calculate determinants by computer. Before describing this method, however, some properties of determinants must be examined; in all cases, *column* may be substituted for *row*.

(1) The determinant of a matrix is equal to that of its transpose since a determinant may be computed from either the rows or columns of the matrix: $|\mathbf{A}'| = |\mathbf{A}|$.

(2) If two rows are interchanged, the sign of the determinant is reversed.

(3) If two rows are identical, the determinant is null (corollary of the second property; see beginning of the present section).

(4) If a scalar is a factor of *one* row, it becomes a factor of the determinant (since it appears *once* in each product).

(5) If a row is a multiple of another row, the determinant is null (corollary of properties 4 and 3, i.e. factoring out the multiplier produces two identical rows).

(6) If all elements of a row are 0, the determinant is null (corollary of property 4).

(7) If a scalar c is a factor of *all* rows, it becomes a factor c^n of the determinant (corollary of property 4), i.e. $|c\mathbf{B}| = c^n|\mathbf{B}|$.

(8) If a multiple of a row is added to another row, the value of the determinant remains unchanged.

(9) The determinant of a triangular matrix (and therefore also of a diagonal matrix) is the product of its diagonal elements.

(10) The sum of the products of the elements of a row with the corresponding cofactors of a *different* row is equal to zero.

(11) For two square matrices of order n, $|\mathbf{A}| \bullet |\mathbf{B}| = |\mathbf{AB}|$.

Pivotal condensation

Properties 8 and 9 can be used for rapid computer calculation of the value of a determinant; the method is called *pivotal condensation*. The matrix is first reduced to triangular form using property 8. This property allows the stepwise elimination of all terms on one side of the diagonal through combinations of multiplications by a scalar, and addition and subtraction of rows or columns. Pivotal condensation may be performed in either the upper or the lower triangular parts of a square matrix. If the lower triangular part is chosen, the upper left-hand diagonal element is used as the first pivot to modify the other rows in such a way that their left-hand terms become zero. The technique consists in calculating by how much the pivot must be multiplied to cancel out the terms in the rows below it; when this value is found, property 8 is used with this value as multiplier. When all terms under the diagonal element in the first column are zero, the procedure is repeated with the other diagonal terms as pivots, to cancel out the elements located under them in the same column. Working on the pivots from left to right insures that when values have been changed to 0, they remain so. When the whole lower triangular portion of the matrix is zero, property 9 is used to compute the determinant which is then the product of the modified diagonal elements.

Numerical example. The same numerical example as above illustrates the method:

$$\begin{vmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 10 \end{vmatrix} = \begin{vmatrix} 1 & 2 & 3 \\ 0 & -3 & -6 \\ 7 & 8 & 10 \end{vmatrix} = \begin{vmatrix} 1 & 2 & 3 \\ 0 & -3 & -6 \\ 0 & -6 & -11 \end{vmatrix} = \begin{vmatrix} 1 & 2 & 3 \\ 0 & -3 & -6 \\ 0 & 0 & 1 \end{vmatrix}$$

a b c

a: (row 2 – 4 × row 1) b: (row 3 – 7 × row 1) c: (row 3 – 2 × row 2)

The determinant is the product of the diagonal elements: $1 \times (-3) \times 1 = (-3)$.

2.7 Rank of a matrix

A square matrix contains n vectors (rows or columns), which may be *linearly independent* or not (for the various meanings of "independence", see Box 1.1). Two vectors are *linearly dependent* when the elements of one are proportional to the elements of the other. For example:

$$\begin{bmatrix} -4 \\ -6 \\ -8 \end{bmatrix} \text{ and } \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix} \text{ are linearly dependent, since } \begin{bmatrix} -4 \\ -6 \\ -8 \end{bmatrix} = -2 \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix}$$

Similarly, a vector is linearly dependent on two others, which are themselves linearly independent, when its elements are a linear combination of the elements of the other two. For example:

$$\begin{bmatrix} -1 \\ 3 \\ 4 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} \text{ and } \begin{bmatrix} 1 \\ -2 \\ -3 \end{bmatrix}$$

illustrate a case where a vector is linearly dependent on two others, which are themselves linearly independent, since

$$(-2) \begin{bmatrix} -1 \\ 3 \\ 4 \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} + 3 \begin{bmatrix} 1 \\ -2 \\ -3 \end{bmatrix}$$

Rank of a square matrix

The *rank* of a *square matrix* is defined as the number of linearly independent row vectors (or column vectors) in the matrix. For example:

$$\begin{bmatrix} -1 & -1 & 1 \\ 3 & 0 & -2 \\ 4 & 1 & -3 \end{bmatrix}$$

(–2 × column 1) = column 2 + (3 × column 3)

or: row 1 = row 2 – row 3

rank = 2

$$\begin{bmatrix} -2 & 1 & 4 \\ -2 & 1 & 4 \\ -2 & 1 & 4 \end{bmatrix}$$

(–2 × column 1) = (4 × column 2) = column 3

or: row 1 = row 2 = row 3

rank = 1

According to property 5 of determinants (Section 2.6), a matrix whose *rank* is lower than its *order* has a determinant equal to zero. Finding the rank of a matrix may therefore be based on the determinant of the lower-order submatrices it contains. The *rank* of a square matrix is the order of the largest square submatrix with non-zero determinant that it contains; this is also the maximum number of linearly independent vectors found among the rows or the columns.

$$\begin{vmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 10 \end{vmatrix} = -3 \neq 0, \text{ so that the } rank = 3$$

$$\begin{vmatrix} -1 & -1 & 1 \\ 3 & 0 & -2 \\ 4 & 1 & -3 \end{vmatrix} = 0 \qquad \begin{vmatrix} -1 & -1 \\ 3 & 0 \end{vmatrix} = 3$$

$$rank = 2$$

The determinant can be used to diagnose the independence of the vectors forming a matrix **X**. For a *square* matrix **X** (symmetric or not), all row and column vectors are linearly independent if $\det(\mathbf{X}) \neq 0$.

Linear independence of the vectors in a *rectangular* matrix **X** with more rows than columns ($n > p$) can be determined from the covariance matrix **S** computed from **X** (eq. 4.6): if $\det(\mathbf{S}) \neq 0$, all column vectors of **X** are linearly independent. This method of diagnosis of the linear independence of the column vectors requires, however, a matrix **X** with $n > p$; if $n \leq p$, $\det(\mathbf{S}) = 0$.

Rank of a rectangular matrix

Numerical example 1. It is possible to determine the rank of a *rectangular* matrix. Several *square* submatrices may be extracted from a *rectangular* matrix, by eliminating rows or/and columns from the matrix. The *rank* of a rectangular matrix is the highest rank of all the square submatrices that can be extracted from it. A first

example illustrates the case where the rank of a rectangular matrix is equal to the number of rows:

$$\begin{vmatrix} 2 & 0 & 1 & 0 & -1 & -2 & 3 \\ 1 & 2 & 2 & 0 & 0 & 1 & -1 \\ 0 & 1 & 2 & 3 & 1 & -1 & 0 \end{vmatrix} \rightarrow \begin{vmatrix} 2 & 0 & 1 \\ 1 & 2 & 2 \\ 0 & 1 & 2 \end{vmatrix} = 5 \qquad rank = 3$$

Numerical example 2. In this example, the rank is lower than the number of rows:

$$\begin{vmatrix} 2 & 1 & 3 & 4 \\ -1 & 6 & -3 & 0 \\ 1 & 20 & -3 & 8 \end{vmatrix} \rightarrow \begin{vmatrix} 2 & 1 & 3 \\ -1 & 6 & -3 \\ 1 & 20 & -3 \end{vmatrix} = \begin{vmatrix} 2 & 1 & 4 \\ -1 & 6 & 0 \\ 1 & 20 & 8 \end{vmatrix} = \begin{vmatrix} 2 & 3 & 4 \\ -1 & -3 & 0 \\ 1 & -3 & 8 \end{vmatrix} = \begin{vmatrix} 1 & 3 & 4 \\ 6 & -3 & 0 \\ 20 & -3 & 8 \end{vmatrix} = 0$$

$$rank < 3 \rightarrow \begin{vmatrix} 2 & 1 \\ -1 & 6 \end{vmatrix} = 13 \qquad rank = 2$$

In this case, the three rows are clearly linearly dependent: (2 × row 1) + (3 × row 2) = row 3. Since it is possible to find a square matrix of order 2 that has a non-null determinant, the rank of the rectangular matrix is 2.

In practice, singular value decomposition (SVD, Section 2.11) can be used to determine the rank of a square or rectangular matrix: the rank is equal to the number of singular values larger than zero. Numerical example 2 will be analysed again in Application 1 of Section 2.11. For square symmetric matrices like covariance matrices, the number of nonzero eigenvalues can also be used to determine the rank of the matrix; see Section 2.10, Second property.

2.8 Matrix inversion

In algebra, division is expressed as either $c \div b$, or c/b, or $c\ (1/b)$, or $c\ b^{-1}$. In the last two expressions, division as such is replaced by multiplication with a reciprocal or inverse quantity. In matrix algebra, the division operation of **C** by **B** does not exist. The equivalent operation is multiplication of **C** with the *inverse* or *reciprocal* of matrix **B**. The inverse of matrix **B** is denoted $\mathbf{B}^{-1}$; the operation through which it is computed is called the *inversion* of matrix **B**.

To serve its purpose, matrix $\mathbf{B}^{-1}$ must be unique and the relation $\mathbf{B}\mathbf{B}^{-1} = \mathbf{B}^{-1}\mathbf{B} = \mathbf{I}$ must be satisfied. It can be shown that only *square matrices* have unique inverses. It is also only for square matrices that the relation $\mathbf{B}\mathbf{B}^{-1} = \mathbf{B}^{-1}\mathbf{B}$ is satisfied. Indeed, there are *rectangular* matrices **B** for which several matrices **C** can be found, satisfying for example $\mathbf{C}\mathbf{B} = \mathbf{I}$ but not $\mathbf{B}\mathbf{C} = \mathbf{I}$. There are also rectangular matrices for which no

matrix **C** can be found such that **CB** = **I**, whereas an infinite number of matrices **C** may exist that satisfy **BC** = **I**. For example:

$$\mathbf{B} = \begin{bmatrix} 1 & 1 \\ -1 & 0 \\ 3 & -1 \end{bmatrix} \qquad \mathbf{C} = \begin{bmatrix} 1 & 3 & 1 \\ 2 & 5 & 1 \end{bmatrix} \qquad \mathbf{CB} = \mathbf{I} \qquad \mathbf{BC} \neq \mathbf{I}$$

$$\mathbf{C} = \begin{bmatrix} 4 & 15 & 4 \\ 7 & 25 & 6 \end{bmatrix} \qquad \mathbf{CB} = \mathbf{I} \qquad \mathbf{BC} \neq \mathbf{I}$$

Generalized inverses can be computed for rectangular matrices by singular value decomposition (Section 2.11, Application 3). Note that several types of generalized inverses, described in textbooks of advanced linear algebra, are not unique.

Inverse of a square matrix

To calculate the inverse of a square matrix **B**, the *adjugate* or *adjoint matrix* of **B** is first defined. In the *matrix of cofactors* of **B**, each element b_{ij} is replaced by its cofactor (cof b_{ij}; see Section 2.6). The adjugate matrix of **B** is the *transpose* of the matrix of cofactors:

$$\begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ b_{n1} & b_{n2} & \cdots & b_{nn} \end{bmatrix} \rightarrow \begin{bmatrix} \text{cof } b_{11} & \text{cof } b_{21} & \cdots & \text{cof } b_{n1} \\ \text{cof } b_{12} & \text{cof } b_{22} & \cdots & \text{cof } b_{n2} \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \text{cof } b_{1n} & \text{cof } b_{2n} & \cdots & \text{cof } b_{nn} \end{bmatrix} \qquad \textbf{(2.16)}$$

matrix **B** adjugate matrix of **B**

In the case of second order matrices, cofactors are scalar values, e.g. cof $b_{11} = b_{22}$, cof $b_{12} = -b_{21}$, etc.

The *inverse* of matrix **B** is the adjugate matrix of **B** divided by the determinant $|\mathbf{B}|$. The product of the matrix with its inverse gives the unit matrix:

$$\underbrace{\frac{1}{|\mathbf{B}|}\begin{bmatrix} \text{cof } b_{11} & \text{cof } b_{21} & \cdots & \text{cof } b_{n1} \\ \text{cof } b_{12} & \text{cof } b_{22} & \cdots & \text{cof } b_{n2} \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \text{cof } b_{1n} & \text{cof } b_{2n} & \cdots & \text{cof } b_{nn} \end{bmatrix}}_{\mathbf{B}^{-1}} \underbrace{\begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ b_{n1} & b_{n2} & \cdots & b_{nn} \end{bmatrix}}_{\mathbf{B}} = \mathbf{I} \qquad \textbf{(2.17)}$$

All diagonal terms resulting from the multiplication $\mathbf{B}^{-1}\mathbf{B}$ (or $\mathbf{B}\mathbf{B}^{-1}$) are of the form $\sum b_{ij}\text{cof}\, b_{ij}$, which is the expansion by minors of a determinant (not taking into account, at this stage, the division of each element of the matrix by $|\mathbf{B}|$). Each diagonal element consequently has the value of the determinant $|\mathbf{B}|$ (eq. 2.14). All other elements of matrix $\mathbf{B}^{-1}\mathbf{B}$ are sums of the products of the elements of a row with the corresponding cofactors of a different row. According to property 10 of determinants (Section 2.6), each non-diagonal element is therefore null. It follows that:

$$\mathbf{B}^{-1}\mathbf{B} = \frac{1}{|\mathbf{B}|}\begin{bmatrix} |\mathbf{B}| & 0 & \dots & 0 \\ 0 & |\mathbf{B}| & \dots & 0 \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ 0 & 0 & \dots & |\mathbf{B}| \end{bmatrix} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ 0 & 0 & \dots & 1 \end{bmatrix} = \mathbf{I} \qquad \textbf{(2.18)}$$

Singular matrix

An important point is that $\mathbf{B}^{-1}$ exists only if $|\mathbf{B}| \neq 0$. A square matrix with a null determinant is called a *singular* matrix; it has no ordinary inverse (but see *singular value decomposition*, Section 2.11). Matrices that can be inverted are called *nonsingular*.

Numerical example. The numerical example of Sections 2.6 and 2.7 is used again to illustrate the calculations:

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 10 \end{bmatrix}$$

The determinant is already known (Section 2.6); its value is –3. The matrix of cofactors is computed, and its *transpose* (adjugate matrix) is divided by the determinant to give the inverse matrix:

$$\begin{bmatrix} 2 & 2 & -3 \\ 4 & -11 & 6 \\ -3 & 6 & -3 \end{bmatrix} \qquad \begin{bmatrix} 2 & 4 & -3 \\ 2 & -11 & 6 \\ -3 & 6 & -3 \end{bmatrix} \qquad -\frac{1}{3}\begin{bmatrix} 2 & 4 & -3 \\ 2 & -11 & 6 \\ -3 & 6 & -3 \end{bmatrix}$$

matrix of cofactors adjugate matrix inverse of matrix

As for the determinant (Section 2.6), various methods exist for quickly inverting matrices using computers; they are especially useful for matrices of higher ranks. Description of these methods, which are available in computer packages, is beyond the scope of the present book. A popular method is briefly explained here; it is somewhat similar to the pivotal condensation presented above for determinants.

Gauss-Jordan

Inversion of matrix $\mathbf{B}$ may be conducted using the *method of Gauss-Jordan*. To do so, matrix $\mathbf{B}_{(n \times n)}$ is first augmented to the right with a same-size identity matrix $\mathbf{I}$, thus creating a $n \times 2n$ matrix. This is illustrated for $n = 3$:

$$\begin{bmatrix} b_{11} & b_{12} & b_{13} & 1 & 0 & 0 \\ b_{21} & b_{22} & b_{23} & 0 & 1 & 0 \\ b_{31} & b_{32} & b_{33} & 0 & 0 & 1 \end{bmatrix}$$

If the augmented matrix is multiplied by matrix $\mathbf{C}_{(n \times n)}$, and if $\mathbf{C} = \mathbf{B}^{-1}$, then the resulting matrix $(n \times 2n)$ has an identity matrix in its first n columns and matrix $\mathbf{C} = \mathbf{B}^{-1}$ in the last n columns.

$$[\ \mathbf{C} = \mathbf{B}^{-1}]\,[\ \mathbf{B}\quad ,\quad \mathbf{I}\ ] = [\ \mathbf{I}\ , \mathbf{C} = \mathbf{B}^{-1}]$$

$$\begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} & b_{13} & 1 & 0 & 0 \\ b_{21} & b_{22} & b_{23} & 0 & 1 & 0 \\ b_{31} & b_{32} & b_{33} & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & c_{11} & c_{12} & c_{13} \\ 0 & 1 & 0 & c_{21} & c_{22} & c_{23} \\ 0 & 0 & 1 & c_{31} & c_{32} & c_{33} \end{bmatrix}$$

This shows that, if matrix $[\mathbf{B},\mathbf{I}]$ is transformed into an equivalent matrix $[\mathbf{I},\mathbf{C}]$, then $\mathbf{C} = \mathbf{B}^{-1}$.

The Gauss-Jordan transformation proceeds in two steps.

• In the first step, the diagonal terms are used, one after the other and from left to right, as pivots to make all the off-diagonal terms equal to zero. This is done in exactly the same way as for the determinant: a factor is calculated to cancel out the target term, using the pivot, and property 8 of the determinants is applied using this factor as multiplier. The difference with determinants is that the whole row of the augmented matrix is modified, not only the part belonging to matrix $\mathbf{B}$. If an off-diagonal zero value is encountered, then of course it is left as is, no cancellation by a multiple of the pivot being necessary or even possible. If a zero is found on the diagonal, this pivot has to be left aside for the time being (in actual programs, rows and columns are interchanged in a process called pivoting); this zero will be changed to a non-zero value during the next cycle unless the matrix is singular. Pivoting makes programming of this method a bit complex.

• Second step. When all the off-diagonal terms are zero, the diagonal terms of the former matrix $\mathbf{B}$ are brought to 1. This is accomplished by dividing each row of the augmented matrix by the value now present in the diagonal term of the former $\mathbf{B}$ (left) portion. If the changes introduced during the first step have made one of the diagonal elements equal to zero, then of course no division can bring it back to 1 and the matrix is singular (i.e. it cannot be inverted).

A Gauss-Jordan algorithm with pivoting is described in the book *Numerical recipes* (Press *et al.*, 2007).

Numerical example. To illustrate the Gauss-Jordan method, the same square matrix as above is first augmented, then transformed so that its left-hand portion becomes the identity matrix:

$$(a)\ \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 10 \end{bmatrix} \rightarrow \left[\begin{array}{ccc|ccc} 1 & 2 & 3 & 1 & 0 & 0 \\ 4 & 5 & 6 & 0 & 1 & 0 \\ 7 & 8 & 10 & 0 & 0 & 1 \end{array}\right]$$

$$(b)\ \left[\begin{array}{ccc|ccc} 1 & 2 & 3 & 1 & 0 & 0 \\ 0 & -3 & -6 & -4 & 1 & 0 \\ 0 & -6 & -11 & -7 & 0 & 1 \end{array}\right]$$

New row 2 ← row 2 – 4row 1
New row 3 ← row 3 – 7row 1

$$(c)\ \left[\begin{array}{ccc|ccc} 3 & 0 & -3 & -5 & 2 & 0 \\ 0 & -3 & -6 & -4 & 1 & 0 \\ 0 & 0 & 1 & 1 & -2 & 1 \end{array}\right]$$

New row 1 ← 3row 1 + 2row 2
New row 3 ← row 3 – 2row 2

$$(d)\ \left[\begin{array}{ccc|ccc} 3 & 0 & 0 & -2 & -4 & 3 \\ 0 & -3 & 0 & 2 & -11 & 6 \\ 0 & 0 & 1 & 1 & -2 & 1 \end{array}\right]$$

New row 1 ← row 1 + 3row 3
New row 2 ← row 2 + 6row 3

$$(e)\ \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & -2/3 & -4/3 & 1 \\ 0 & 1 & 0 & -2/3 & 11/3 & -2 \\ 0 & 0 & 1 & 1 & -2 & 1 \end{array}\right]$$

New row 1 ← (1/3) row 1
New row 2 ← – (1/3) row 2
New row 3 ← row 3

$$(f)\ -\frac{1}{3}\begin{bmatrix} 2 & 4 & -3 \\ 2 & -11 & 6 \\ -3 & 6 & -3 \end{bmatrix}$$

inverse of matrix **B**

The inverse of matrix **B** is the same as calculated above.

The inverse of a matrix has several interesting properties, including:

(1) $\mathbf{B}^{-1}\mathbf{B} = \mathbf{B}\mathbf{B}^{-1} = \mathbf{I}$.

(2) $\left|\mathbf{B}^{-1}\right| = 1 / \left|\mathbf{B}\right|$.

(3) $\left[\mathbf{B}^{-1}\right]^{-1} = \mathbf{B}$.

(4) $\left[\mathbf{B}'\right]^{-1} = \left[\mathbf{B}^{-1}\right]'$.

(5) If **B** and **C** are nonsingular square matrices, $\left[\mathbf{BC}\right]^{-1} = \mathbf{C}^{-1}\mathbf{B}^{-1}$.

(6) In the case of a symmetric matrix, since $\mathbf{B}' = \mathbf{B}$, then $\left[\mathbf{B}^{-1}\right]' = \mathbf{B}^{-1}$.

Orthonormal matrix

(7) An orthogonal matrix (Section 2.5) whose column vectors are normalized (scaled to length 1: Section 2.4) is called *orthonormal*. A square orthonormal matrix **B** has the property that $\mathbf{B}' = \mathbf{B}^{-1}$. This may be shown as follows: on the one hand, $\mathbf{B}^{-1}\mathbf{B} = \mathbf{I}$ by definition of the inverse of a square matrix. On the other hand, property 8 of matrix products (Section 2.5) shows that $\mathbf{B}'\mathbf{B} = \mathbf{D}(1)$ when the column vectors in **B** are normalized (which is the case for an orthonormal matrix); $\mathbf{D}(1)$ is a diagonal matrix of 1's, which is the identity matrix **I** (eq. 2.5). Given that $\mathbf{B}'\mathbf{B} = \mathbf{B}^{-1}\mathbf{B} = \mathbf{I}$, then

$\mathbf{B}' = \mathbf{B}^{-1}$. Furthermore, combining the properties $\mathbf{B}\mathbf{B}^{-1} = \mathbf{I}$ (which is true for any square matrix) and $\mathbf{B}' = \mathbf{B}^{-1}$ shows that $\mathbf{B}\mathbf{B}' = \mathbf{I}$. For example, the matrix of normalized eigenvectors of a symmetric matrix, which is square and orthonormal (Section 2.9), has these properties.

(8) The inverse of a diagonal matrix is a diagonal matrix whose elements are the reciprocals of the original elements: $[\mathbf{D}(x_i)]^{-1} = \mathbf{D}(1/x_i)$.

Inversion is used in many types of applications, as will be seen in the remainder of this book. Classical examples of the role of inverse matrices are solving systems of linear equations and the calculation of regression coefficients.

System of linear equations

A system of linear equations can be represented in matrix form; for example:

$$\begin{aligned} b_1 + 2b_2 + 3b_3 &= 2 \\ 4b_1 + 5b_2 + 6b_3 &= 2 \\ 7b_1 + 8b_2 + 10b_3 &= 3 \end{aligned} \quad \rightarrow \quad \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 10 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \\ 3 \end{bmatrix}$$

which may be written $\mathbf{A}\mathbf{b} = \mathbf{c}$. To find the values of the unknowns b_1, b_2 and b_3, vector $\mathbf{b}$ must be isolated to the left, which necessitates an inversion of the square matrix $\mathbf{A}$:

$$\begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 10 \end{bmatrix}^{-1} \begin{bmatrix} 2 \\ 2 \\ 3 \end{bmatrix}$$

The inverse of $\mathbf{A}$ has been calculated above. Multiplication with vector $\mathbf{c}$ provides the solution for the three unknowns:

$$\begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = -\frac{1}{3} \begin{bmatrix} 2 & 4 & -3 \\ 2 & -11 & 6 \\ -3 & 6 & -3 \end{bmatrix} \begin{bmatrix} 2 \\ 2 \\ 3 \end{bmatrix} = -\frac{1}{3} \begin{bmatrix} 3 \\ 0 \\ -3 \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} \qquad \begin{aligned} b_1 &= -1 \\ b_2 &= 0 \\ b_3 &= 1 \end{aligned}$$

Systems of linear equations are solved in that way in Subsections 13.2.2 and 13.3.3.

Simple linear regression

Linear regression analysis is reviewed in Section 10.3. *Regression coefficients* are easily calculated for several models using matrix inversion; the approach is briefly discussed here. The mathematical model for *simple linear regression* (model I, Subsection 10.3.1) is:

$$\hat{y} = b_0 + b_1 x$$

The regression coefficients b_0 and b_1 are estimated from the observed data **x** and **y**. This is equivalent to resolving the following system of equations:

$$\begin{matrix} y_1 = b_0 + b_1 x_1 \\ y_2 = b_0 + b_1 x_2 \\ \cdot \qquad\quad \cdot \\ \cdot \qquad\quad \cdot \\ \cdot \qquad\quad \cdot \\ y_n = b_0 + b_1 x_n \end{matrix} \rightarrow \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & x_n \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$$

Least squares

Matrix **X** was augmented with a column of 1's in order to estimate the intercept of the regression equation, b_0. Coefficients b are estimated by the *method of least squares* (Subsection 10.3.1), which minimizes the sum of squares of the differences between observed values y and values $\hat{y}$ calculated using the regression equation. In order to obtain a least-squares best fit, each member (left and right) of matrix equation $\mathbf{y} = \mathbf{Xb}$ is multiplied by the transpose of matrix **X**, i.e. $\mathbf{X'y} = \mathbf{X'Xb}$. By doing so, the rectangular matrix **X** produces a square matrix $\mathbf{X'X}$, which can be inverted. The values of coefficients b_0 and b_1 forming vector **b** are computed directly, after inverting the square matrix $[\mathbf{X'X}]$:

$$\mathbf{b} = [\mathbf{X'X}]^{-1} [\mathbf{X'y}] \qquad \textbf{(2.19)}$$

Multiple linear regression

Using the same approach, it is easy to compute coefficients $b_0, b_1, \ldots, b_m$ of a *multiple linear regression* (Subsection 10.3.3). In this type of regression, variable y is a linear function of several (m) variables x_j, so that one can write:

$$\hat{y} = b_0 + b_1 x_1 + \ldots + b_m x_m$$

Vectors **y** and **b** and matrix **X** are defined as follows:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1m} \\ 1 & x_{21} & \cdots & x_{2m} \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ 1 & x_{n1} & \cdots & x_{nm} \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ \cdot \\ \cdot \\ \cdot \\ b_m \end{bmatrix}$$

Again, matrix **X** was augmented with a column of 1's in order to estimate the intercept of the equation, b_0. The least-squares solution is found by computing eq. 2.19. Readers can consult Section 10.3 for computational and variable selection methods to be used in multiple linear regression when the variables x_j are strongly intercorrelated, as is often the case in ecology.

Polynomial regression

In *polynomial regression* (Subsection 10.3.4), several regression parameters b, corresponding to powers of a single variable x, are fitted to the observed data. The general regression model is:

$$\hat{y} = b_0 + b_1 x + b_2 x^2 + \ldots + b_k x^k$$

The vector of parameters, **b**, is computed in the same way. Vectors **y** and **b**, and matrix **X**, are defined as follows:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^k \\ 1 & x_2 & x_2^2 & \cdots & x_2^k \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ 1 & x_n & x_n^2 & \cdots & x_n^k \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ \cdot \\ \cdot \\ \cdot \\ b_k \end{bmatrix}$$

The least-squares solution is computed using eq. 2.19. Readers should consult Subsection 10.3.4 where practical considerations concerning the calculation of polynomial regression with ecological data are discussed.

2.9 Eigenvalues and eigenvectors

There are other problems, in addition to those examined above, where the determinant and the inverse of a matrix are used to provide simple and elegant solutions. An important one in data analysis is the derivation of an orthogonal form (i.e. a matrix whose vectors are at right angles; Sections 2.5 and 2.8) for a non-orthogonal symmetric matrix. This will provide the algebraic basis for most of the methods studied in Chapters 9 and 11. In ecology, data sets generally include a large number of variables, which are associated to one another (e.g. linearly correlated; Section 4.2). The basic idea underlying several methods of data analysis is to reduce this large number of intercorrelated variables to a smaller number of composite, but linearly independent (Box 1.1) variables, each explaining a different fraction of the observed variation. One of the main goals of numerical data analysis is indeed to generate a small number of variables, each explaining a large portion of the variation, and to ascertain that these new variables explain different aspects of the phenomena under study. The present section only deals with the mathematics of the computation of *eigenvalues* and *eigenvectors*. Applications to the analysis of multidimensional ecological data are discussed in Chapters 4, 9 and 11.

Mathematically, the problem may be formulated as follows. Given a square matrix **A**, one wishes to find a diagonal matrix that is equivalent to **A**. To fix ideas, **A** is a covariance matrix **S** in principal component analysis. Other types of square, symmetric

association matrices (Section 2.2) are used in numerical ecology, hence the use of the symbol **A**:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

In **A**, the terms above and below the diagonal characterize the degree of association of either the objects, or the ecological variables, with one another (Fig. 2.1). In the new matrix **Λ** (capital lambda) being sought, all elements outside the diagonal are null:

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_{11} & 0 & \cdots & 0 \\ 0 & \lambda_{22} & \cdots & 0 \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ 0 & 0 & \cdots & \lambda_{nn} \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \qquad \textbf{(2.20)}$$

This new matrix is called the *matrix of eigenvalues*[*]. It has the same trace and the same determinant as **A**. The new variables (*eigenvectors*; see below) whose association is described by this matrix **Λ** are thus linearly independent of one another. The use of the Greek letter λ (lower-case lambda) to represent eigenvalues stems from the fact that eigenvalues are actually *Lagrangian multipliers* λ, as will be shown in Section 4.4.

Canonical form

Matrix **Λ** is known as the *canonical form* of matrix **A;** for the exact meaning of *canonical* in mathematics, see Section 11.0.

1 — Computation

The eigenvalues and eigenvectors of matrix **A** are found from equation

$$\mathbf{A}\mathbf{u}_i = \lambda_i \mathbf{u}_i \qquad \textbf{(2.21)}$$

which allows one to compute the different eigenvalues λ_i and their associated eigenvectors $\mathbf{u}_i$. First, the validity of eq. 2.21 must be demonstrated.

---

[*] In the literature, the following expressions are synonymous:

eigenvalue	eigenvector
characteristic root	characteristic vector
latent root	latent vector

Eigen is the German word for *characteristic*.

To do so, one uses any pair h and i of eigenvalues and eigenvectors computed from matrix **A**. Equation 2.21 becomes

$$\mathbf{A}\mathbf{u}_h = \lambda_h \mathbf{u}_h \quad \text{and} \quad \mathbf{A}\mathbf{u}_i = \lambda_i \mathbf{u}_i \,, \quad \text{respectively.}$$

Multiplying these equations by row vectors $\mathbf{u}'_i$ and $\mathbf{u}'_h$, respectively, gives:

$$\mathbf{u}'_i \mathbf{A}\mathbf{u}_h = \lambda_h \mathbf{u}'_i \mathbf{u}_h \quad \text{and} \quad \mathbf{u}'_h \mathbf{A}\mathbf{u}_i = \lambda_i \mathbf{u}'_h \mathbf{u}_i$$

It can be shown that, in the case of a symmetric matrix, the left-hand members of these two equations are equal: $\mathbf{u}'_i \mathbf{A}\mathbf{u}_h = \mathbf{u}'_h \mathbf{A}\mathbf{u}_i$; this would not be true for an asymmetric matrix, however. Using a (2×2) matrix **A** like the one of Numerical example 1 below, readers can easily check that the equality holds only when $a_{12} = a_{21}$, i.e. when **A** is symmetric. So, in the case of a symmetric matrix, the right-hand members are also equal:

$$\lambda_h \mathbf{u}'_i \mathbf{u}_h = \lambda_i \mathbf{u}'_h \mathbf{u}_i$$

Since we are talking about two distinct values for λ_h and λ_i, the only possibility for the above equality to be true is that the product of vectors $\mathbf{u}_h$ and $\mathbf{u}_i$ be 0 (i.e. $\mathbf{u}'_i \mathbf{u}_h = \mathbf{u}'_h \mathbf{u}_i = 0$), which is the condition of orthogonality for two vectors (Section 2.5). It is therefore concluded that eq. 2.21

$$\mathbf{A}\mathbf{u}_i = \lambda_i \mathbf{u}_i$$

can be used to compute vectors $\mathbf{u}_i$ that are orthogonal when matrix **A** is symmetric. In the case of a non-symmetric matrix, eigenvectors can also be calculated, but they are not orthogonal.

If the scalars λ_i and their associated vectors $\mathbf{u}_i$ exist, then eq. 2.21 can be transformed as follows:

$$\mathbf{A}\mathbf{u}_i - \lambda_i \mathbf{u}_i = \mathbf{0} \quad \text{(difference between two vectors)}$$

and vector $\mathbf{u}_i$ can be factorized:

$$(\mathbf{A} - \lambda_i \mathbf{I})\mathbf{u}_i = \mathbf{0} \qquad \textbf{(2.22)}$$

Because of the nature of the elements in eq. 2.22, it is necessary to introduce a unit matrix **I** inside the parentheses, where one now finds a difference between two square matrices. According to eq. 2.22, multiplication of the square matrix $(\mathbf{A} - \lambda_i \mathbf{I})$ with the column vector $\mathbf{u}_i$ must result in a null column vector ($\mathbf{0}$).

Besides the trivial solution where $\mathbf{u}_i$ is a null vector, eq. 2.22 has the following solution:

$$\left| \mathbf{A} - \lambda_i \mathbf{I} \right| = 0 \qquad \textbf{(2.23)}$$

That is, the determinant of the difference between matrices **A** and $\lambda_i\mathbf{I}$ must be equal to 0 for each λ_i. Resolving eq. 2.23 provides the eigenvalues λ_i associated with matrix **A**. Equation 2.23 is known as the *characteristic* or *determinantal equation*.

Character-istic equation

Demonstration of eq. 2.23 goes as follows:

1) One solution to $(\mathbf{A} - \lambda_i\mathbf{I})\mathbf{u}_i = \mathbf{0}$ is that $\mathbf{u}_i$ is the null vector: $\mathbf{u} = [0]$. This solution is trivial, since it corresponds to the centroid of the scatter of data points. A non-trivial solution must thus involve $(\mathbf{A} - \lambda_i\mathbf{I})$.

2) Solution $(\mathbf{A} - \lambda_i\mathbf{I}) = [0]$ is not acceptable either, since it implies that $\mathbf{A} = \lambda_i\mathbf{I}$ and thus that **A** be a scalar matrix, which is generally not true.

3) The solution thus requires that λ_i and $\mathbf{u}_i$ be such that the *scalar product* $(\mathbf{A} - \lambda_i\mathbf{I})\mathbf{u}_i$ is a null vector. In other words, vector $\mathbf{u}_i$ must be orthogonal to the space corresponding to **A** after $\lambda_i\mathbf{I}$ has been subtracted from it; orthogonality of two vectors or matrices is obtained when their scalar product is zero (Section 2.5). The solution $|\mathbf{A} - \lambda_i\mathbf{I}| = 0$ (eq. 2.23) means that, for each value λ_i, the rank of $(\mathbf{A} - \lambda_i\mathbf{I})$ is lower than its order, which makes the determinant equal to zero (Section 2.7). Each λ_i is the variance corresponding to one dimension of matrix **A** (Section 4.4). It is then easy to calculate the eigenvector $\mathbf{u}_i$ that is orthogonal to the space $(\mathbf{A} - \lambda_i\mathbf{I})$ of lower dimension than **A**. That eigenvector is the solution to eq. 2.22, which specifies orthogonality of $\mathbf{u}_i$ with respect to $(\mathbf{A} - \lambda_i\mathbf{I})$.

For a matrix **A** of order n, the characteristic equation is a polynomial of degree n, whose solutions are the eigenvalues λ_i. When these values are found, it is easy to use eq. 2.22 to calculate the eigenvector $\mathbf{u}_i$ corresponding to each eigenvalue λ_i. There are therefore as many eigenvectors as there are eigenvalues.

There are methods that enable the quick and efficient calculation of eigenvalues and eigenvectors by computer. Three of these are described in Subsection 9.1.9.

Ecologists, who are more concerned with shedding light on natural phenomena than on mathematical entities, may find unduly technical this discussion of the computation of eigenvalues and eigenvectors. The same subject will be considered again in Section 4.4 in the context of the multidimensional normal distribution. Mastering the bases of this algebraic operation is essential to understand the methods based on *eigenanalysis* (Chapters 9 and 11), which are of prime importance to the analysis of ecological data.

2 — Numerical examples

This subsection contains two examples of eigen-decomposition.

Numerical example 1. The characteristic equation of the symmetric matrix

$$\mathbf{A} = \begin{bmatrix} 2 & 2 \\ 2 & 5 \end{bmatrix}$$

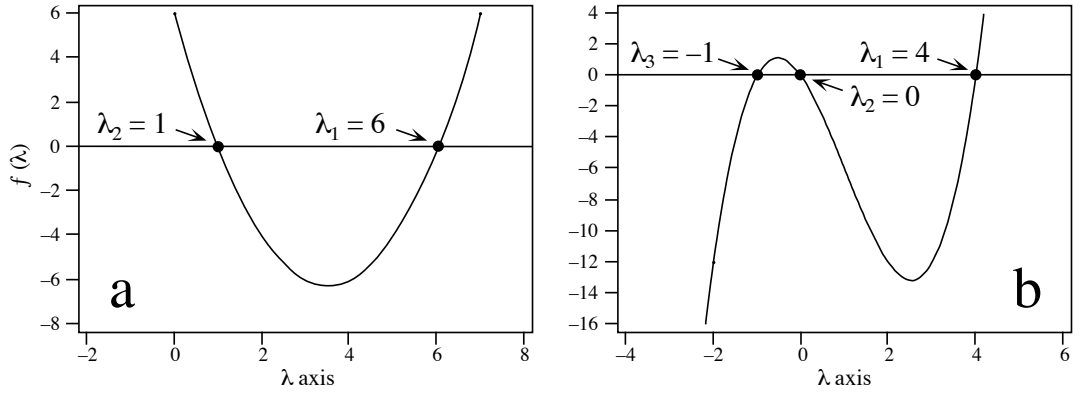


Figure 2.2 (a) The eigenvalues of Numerical example 1 are the values along the λ axis where the function $\lambda^2 - 7\lambda + 6$ is zero. (b) Similarly for Numerical example 2, the eigenvalues are the values along the λ axis where the function $\lambda^3 - 3\lambda^2 - 4\lambda$ is zero.

is (eq. 2.23)
$$\left| \begin{bmatrix} 2 & 2 \\ 2 & 5 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right| = 0$$

therefore
$$\left| \begin{bmatrix} 2 & 2 \\ 2 & 5 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \right| = 0$$

and thus
$$\begin{vmatrix} 2-\lambda & 2 \\ 2 & 5-\lambda \end{vmatrix} = 0$$

The *characteristic polynomial* is found by expanding the determinant (Section 2.6):

$$(2 - \lambda)(5 - \lambda) - 4 = 0$$

which gives
$$\lambda^2 - 7\lambda + 6 = 0$$

from which it is easy to calculate the two values of λ that satisfy the equation (Fig. 2.2a). The two eigenvalues of **A** are:

$$\lambda_1 = 6 \quad \text{and} \quad \lambda_2 = 1$$

The sum of the eigenvalues is equal to the trace (i.e. the sum of the diagonal elements) of **A**.

The ordering of eigenvalues is arbitrary. It would have been equally correct to write that $\lambda_1 = 1$ and $\lambda_2 = 6$, but the convention is to sort the eigenvalues in decreasing order.

Equation 2.22 is used to calculate the eigenvectors $\mathbf{u}_1$ and $\mathbf{u}_2$ corresponding to eigenvalues λ_1 and λ_2:

for $\lambda_1 = 6$

$$\left(\begin{bmatrix} 2 & 2 \\ 2 & 5 \end{bmatrix} - 6 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right) \begin{bmatrix} u_{11} \\ u_{21} \end{bmatrix} = \mathbf{0}$$

$$\begin{bmatrix} -4 & 2 \\ 2 & -1 \end{bmatrix} \begin{bmatrix} u_{11} \\ u_{21} \end{bmatrix} = \mathbf{0}$$

for $\lambda_2 = 1$

$$\left(\begin{bmatrix} 2 & 2 \\ 2 & 5 \end{bmatrix} - 1 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right) \begin{bmatrix} u_{12} \\ u_{22} \end{bmatrix} = \mathbf{0}$$

$$\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} u_{12} \\ u_{22} \end{bmatrix} = \mathbf{0}$$

which is equivalent to the following pairs of linear equations:

$$-4u_{11} + 2u_{21} = 0 \qquad 1u_{12} + 2u_{22} = 0$$

$$2u_{11} - 1u_{21} = 0 \qquad 2u_{12} + 4u_{22} = 0$$

These sets of linear equations are always indeterminate. The solution is given by any point (vector) in the direction of the eigenvector being sought. To remove the indetermination, an arbitrary value is assigned to one of the elements u, which specifies a particular vector. For example, value $u = 1$ may be arbitrarily assigned to the first element u in each set:

given that	$u_{11} = 1$	$u_{12} = 1$
it follows that	$-4u_{11} + 2u_{21} = 0$	$1u_{12} + 2u_{22} = 0$
become	$-4 + 2u_{21} = 0$	$1 + 2u_{22} = 0$
so that	$u_{21} = 2$	$u_{22} = -1/2$

Eigenvectors $\mathbf{u}_1$ and $\mathbf{u}_2$ are therefore:

$$\begin{bmatrix} 1 \\ 2 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 \\ -1/2 \end{bmatrix}$$

Values other than 1 could have been arbitrarily assigned to u_{11} and u_{12} (or, for that matter, to any other term in each vector). For example, the following vectors also satisfy the two pairs of linear equations, since these eigenvectors differ only by a scalar multiplier:

$$\begin{bmatrix} 2 \\ 4 \end{bmatrix} \text{ or } \begin{bmatrix} -3 \\ -6 \end{bmatrix} \qquad \begin{bmatrix} 2 \\ -1 \end{bmatrix} \text{ or } \begin{bmatrix} -4 \\ 2 \end{bmatrix}$$

This is the reason why eigenvectors are generally standardized. One method is to assign value 1 to the *largest element* of each vector, and adjust the other elements accordingly. Another standardization method, used for instance in principal component and principal coordinate analyses (Sections 9.1 and 9.3), is to make the length of each eigenvector $\mathbf{u}_i$ equal to the square root of its eigenvalue (eigenvector *scaled to* $\sqrt{\lambda_i}$).

The most common and practical method is to normalize eigenvectors, i.e. to make their lengths equal to 1. Thus, a *normalized* eigenvector is in fact *scaled to* 1, i.e. $\mathbf{u}'\mathbf{u} = 1$. As explained in Section 2.4, normalization is achieved by dividing each element of a vector by the length of this vector, i.e. the square root of the sum of squares of all elements in the vector. Like most other computer packages, the R function ***eigen()*** outputs normalized eigenvectors.

In the numerical example, the two eigenvectors

$$\begin{bmatrix} 1 \\ 2 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 2 \\ -1 \end{bmatrix}$$

are normalized to

$$\begin{bmatrix} 1/\sqrt{5} \\ 2/\sqrt{5} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 2/\sqrt{5} \\ -1/\sqrt{5} \end{bmatrix}$$

Since the eigenvectors are both orthogonal and normalized, they are *orthonormal* (property 7 in Section 2.8).

Had the eigenvectors been multiplied by a negative scalar, their normalized forms would now be the following:

$$\begin{bmatrix} -1/\sqrt{5} \\ -2/\sqrt{5} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} -2/\sqrt{5} \\ 1/\sqrt{5} \end{bmatrix}$$

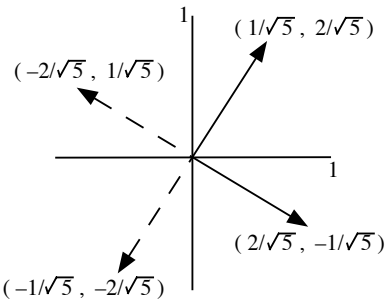
These forms are strictly equivalent to those above.

Since matrix **A** is symmetric, its eigenvectors must be orthogonal. This is easily verified as their product is equal to zero, which is the condition for two vectors to be orthogonal (Section 2.5):

$$\mathbf{u}_1'\mathbf{u}_2 = \begin{bmatrix} 1/\sqrt{5} & 2/\sqrt{5} \end{bmatrix} \begin{bmatrix} 2/\sqrt{5} \\ -1/\sqrt{5} \end{bmatrix} = 2/5 - 2/5 = 0$$

The normalized eigenvectors can be plotted in the original system of coordinates, i.e. the Cartesian plane whose axes are the two original descriptors; the association between these

descriptors is given by matrix **A.** This plot (full arrows) shows that the angle between the eigenvectors is indeed 90° (cos 90° = 0) and that their lengths are 1:



The dashed arrows illustrate the same eigenvectors with inverted signs. The eigenvectors with dashed arrows are equivalent to those with full arrows.

Resolving the system of linear equations used to compute eigenvectors is greatly facilitated by matrix inversion. Defining matrix $\mathbf{C}_{nn} = (\mathbf{A} - \lambda_n \mathbf{I})$ allows eq. 2.22 to be written in a simplified form:

$$\mathbf{C}_{nn}\mathbf{u}_n = \mathbf{0}_n \qquad \textbf{(2.24)}$$

Indices n designate here the dimensions of matrix **C** and vector **u**. Matrix $\mathbf{C}_{nn}$ contains all the coefficients by which a given eigenvector $\mathbf{u}_n$ is multiplied. The system of equations is indeterminate, which prevents the inversion of **C** and calculation of **u**. To remove the indetermination, it is sufficient to determine any one element of vector **u**. For example, one may arbitrarily decide that $u_1 = \alpha$ ($\alpha \neq 0$). Then,

$$\begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{bmatrix} \begin{bmatrix} \alpha \\ u_2 \\ \cdot \\ \cdot \\ \cdot \\ u_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix}$$

can be written

$$\begin{bmatrix} c_{11}\alpha + c_{12}u_2 + \ldots + c_{1n}u_n \\ c_{21}\alpha + c_{22}u_2 + \ldots + c_{2n}u_n \\ \cdot \qquad\qquad\qquad \cdot \\ \cdot \qquad\qquad\qquad \cdot \\ \cdot \qquad\qquad\qquad \cdot \\ c_{n1}\alpha + c_{n2}u_2 + \ldots + c_{nn}u_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix}$$

so that

$$\begin{bmatrix} c_{12}u_2 + \ldots + c_{1n}u_n \\ c_{22}u_2 + \ldots + c_{2n}u_n \\ \cdot \qquad\qquad \cdot \\ \cdot \qquad\qquad \cdot \\ \cdot \qquad\qquad \cdot \\ c_{n2}u_2 + \ldots + c_{nn}u_n \end{bmatrix} = -\alpha \begin{bmatrix} c_{11} \\ c_{21} \\ \cdot \\ \cdot \\ \cdot \\ c_{n1} \end{bmatrix}$$

After setting $u_1 = \alpha$, the first column of matrix **C** is transferred to the right. The last $n - 1$ rows of **C** are then sufficient to define a completely determined system. The first row is removed from **C** in order to obtain a square matrix of order $n - 1$, which can be inverted. The determined system thus obtained is:

$$\begin{bmatrix} c_{22}u_2 + \ldots + c_{2n}u_n \\ \cdot \qquad\qquad \cdot \\ \cdot \qquad\qquad \cdot \\ \cdot \qquad\qquad \cdot \\ c_{n2}u_2 + \ldots + c_{nn}u_n \end{bmatrix} = -\alpha \begin{bmatrix} c_{21} \\ \cdot \\ \cdot \\ \cdot \\ c_{n1} \end{bmatrix}$$

which can be written $$\mathbf{C}_{(n-1)\,(n-1)}\,\mathbf{u}_{(n-1)} = -\alpha\mathbf{c}_{(n-1)} \qquad \textbf{(2.25)}$$

This system can be resolved by inversion of **C**, as in Section 2.8:

$$\mathbf{u}_{(n-1)} = -\alpha\mathbf{C}^{-1}_{(n-1)\,(n-1)}\,\mathbf{c}_{(n-1)} \qquad \textbf{(2.26)}$$

This method of computing the eigenvectors may not work, however, in the case of multiple eigenvalues (see Third property in Section 2.10, below). The following example provides an illustration of the computation through matrix inversion.

Numerical example 2. For the asymmetric matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 3 & -1 \\ 0 & 1 & 2 \\ 1 & 4 & 1 \end{bmatrix},$$

the characteristic polynomial, computed from eq. 2.23, is $\lambda^3 - 3\lambda^2 - 4\lambda = 0$, from which the three eigenvalues 4, 0 and –1 can be calculated (Fig. 2.2b). The sum of the eigenvalues has to be equal to the trace of **A**, which is 3.

The eigenvectors are computed by inserting each eigenvalue, in turn, into eq. 2.22. For $\lambda_1 = 4$:

$$\begin{bmatrix} (1-4) & 3 & -1 \\ 0 & (1-4) & 2 \\ 1 & 4 & (1-4) \end{bmatrix} \begin{bmatrix} u_{11} \\ u_{21} \\ u_{31} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

The above system is determined by setting $u_{11} = 1$. Using eq. 2.25 gives:

$$\begin{bmatrix} (1-4) & 2 \\ 4 & (1-4) \end{bmatrix} \begin{bmatrix} u_{21} \\ u_{31} \end{bmatrix} = -1 \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

from which it follows (eq. 2.26) that

$$\begin{bmatrix} u_{21} \\ u_{31} \end{bmatrix} = \begin{bmatrix} (1-4) & 2 \\ 4 & (1-4) \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ -1 \end{bmatrix}$$

The inverse of matrix $\begin{bmatrix} -3 & 2 \\ 4 & -3 \end{bmatrix}$ is $\begin{bmatrix} -3 & -2 \\ -4 & -3 \end{bmatrix}$ so that

$$\begin{bmatrix} u_{21} \\ u_{31} \end{bmatrix} = \begin{bmatrix} -3 & -2 \\ -4 & -3 \end{bmatrix} \begin{bmatrix} 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

The two other eigenvectors are computed in the same fashion, from eigenvalues $\lambda_2 = 0$ and $\lambda_3 = -1$. The resulting matrix of eigenvectors (columns) is:

$$\mathbf{U} = \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \mathbf{u}_3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 2 & -2/7 & -1/2 \\ 3 & 1/7 & 1/2 \end{bmatrix} \quad \text{or else} \quad \begin{bmatrix} 1 & 7 & 2 \\ 2 & -2 & -1 \\ 3 & 1 & 1 \end{bmatrix}$$

which is normalized to $\mathbf{U} = \begin{bmatrix} 0.27 & 0.95 & 0.82 \\ 0.53 & -0.27 & -0.41 \\ 0.80 & 0.14 & 0.41 \end{bmatrix}$

Readers can easily check that these eigenvectors, which were extracted from a non-symmetric matrix, are indeed not orthogonal; none of the scalar products between pairs of columns is equal to zero. The eigenanalysis of non-symmetric (or *asymmetric*) matrices will be encountered in linear discriminant analysis and canonical correlation analysis, Sections 11.3 and 11.4.

2.10 Some properties of eigenvalues and eigenvectors

First property. — A simple rearrangement of eq. 2.21 shows that matrix **U** of the eigenvectors is a transform matrix, allowing one to go from system **A** to system **Λ**. Indeed, the equation can be rewritten so as to include all eigenvalues and eigenvectors:

$$\mathbf{AU} = \mathbf{U\Lambda} \qquad \textbf{(2.27)}$$

Numerical example. Equation 2.27 can be verified using Numerical example 2 of Section 2.9:

$$\begin{bmatrix} 1 & 3 & -1 \\ 0 & 1 & 2 \\ 1 & 4 & 1 \end{bmatrix} \begin{bmatrix} 1 & 7 & 2 \\ 2 & -2 & -1 \\ 3 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 7 & 2 \\ 2 & -2 & -1 \\ 3 & 1 & 1 \end{bmatrix} \begin{bmatrix} 4 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

The left and right-hand sides of the equation are identical:

$$\begin{bmatrix} 4 & 0 & -2 \\ 8 & 0 & 1 \\ 12 & 0 & -1 \end{bmatrix} = \begin{bmatrix} 4 & 0 & -2 \\ 8 & 0 & 1 \\ 12 & 0 & -1 \end{bmatrix}$$

On the left-hand side of the equation, matrix **A** is postmultiplied by matrix **U** of the eigenvectors whereas, on the right-hand side, the matrix of eigenvalues **Λ** is premultiplied by **U**. It follows that **U** achieves a two-way transformation (rows, columns), from the reference system **A** to the system **Λ**. This transformation can go both ways, as shown by the following equations which are both derived from eq. 2.27:

$$\mathbf{A} = \mathbf{U\Lambda U}^{-1} \qquad \text{and} \qquad \mathbf{\Lambda} = \mathbf{U}^{-1}\mathbf{AU} \qquad \textbf{(2.28)}$$

A simple formula may be derived from $\mathbf{A} = \mathbf{U\Lambda U}^{-1}$, which can be used to raise matrix **A** to any power *x*:

$$\mathbf{A}^{x} = (\mathbf{U\Lambda U}^{-1})\mathbf{U\Lambda} \ldots \mathbf{U}^{-1}(\mathbf{U\Lambda U}^{-1})$$

$$\mathbf{A}^{x} = \mathbf{U\Lambda}(\mathbf{U}^{-1}\mathbf{U})\mathbf{\Lambda} \ldots (\mathbf{U}^{-1}\mathbf{U})\mathbf{\Lambda U}^{-1}$$

$$\mathbf{A}^{\boldsymbol{x}} = \mathbf{U}\boldsymbol{\Lambda}^{\boldsymbol{x}}\mathbf{U}^{-1}, \text{ because } \mathbf{U}^{-1}\mathbf{U} = \mathbf{I}$$

Raising a matrix to some high power is greatly facilitated by the fact that $\boldsymbol{\Lambda}^{\boldsymbol{x}}$ is the matrix of eigenvalues, which is diagonal. Indeed, a diagonal matrix can be raised to any power x by raising each of its diagonal elements to power x. It follows that the last equation can be rewritten as:

$$\mathbf{A}^{x} = \mathbf{U}[\lambda_i^{x}]\mathbf{U}^{-1} \tag{2.29}$$

This may be verified using the above example. Note 1: this calculation cannot be done if there are negative eigenvalues, as in non-symmetric matrices, and the exponent is not an integer. The reason is that a fractional exponent of a negative number is undefined. Note 2: if $\mathbf{U}$ is orthonormal, $\mathbf{U}^{-1} = \mathbf{U}'$, so that $\mathbf{A}^{x} = \mathbf{U}[\lambda_i^{x}]\mathbf{U}^{-1} = \mathbf{U}[\lambda_i^{x}]\mathbf{U}'$ (property of the inverse of an orthonormal matrix, Section 2.8). This equality is true only if $\mathbf{U}$ has been normalized.

Second property. — It was shown in Section 2.7 that, when the rank (r) of matrix $\mathbf{A}_{nn}$ is smaller than its order ($r < n$), the determinant $|\mathbf{A}|$ is 0. It was also shown that, when it is necessary to know the rank of a matrix, as for instance in dimensional analysis (Section 3.3), $|\mathbf{A}| = 0$ indicates that one must check the rank of $\mathbf{A}$. Such a test naturally follows from the calculation of eigenvalues. Indeed, for a square symmetric matrix $\mathbf{A}$, the determinant is equal to the product of its eigenvalues:

$$|\mathbf{A}| = \prod_{i=1}^{n} \lambda_i \tag{2.30}$$

so that $|\mathbf{A}| = 0$ if one or several of the eigenvalues is 0. When the rank of a matrix is smaller than its order ($r < n$), this matrix has ($n - r$) null eigenvalues. Thus, *eigenvalues can be used to determine the rank of a square symmetric matrix*: the rank is equal to *the number of nonzero eigenvalues*. In the case of a covariance or cross-product matrix among variables, the number of nonzero eigenvalues (i.e. the rank of $\mathbf{A}$) is the number of linearly independent dimensions required to account for all the variance (Chapter 9).

Multiple eigenvalues

Third property. — It was implicitly assumed, up to this point, that the eigenvalues were *all different* from one another. It may happen, however, that some (say, m) eigenvalues are equal. These are known as *multiple eigenvalues*. In such a case, the question is whether or not matrix A_{nn} has n *distinct* eigenvectors. In other words, are there m *linearly independent* eigenvectors which correspond to the *same* eigenvalue? In principal component analysis (Section 9.1), a solution corresponding to that situation is called circular.

Values λ_i are chosen in such a way that the determinant $|\mathbf{A} - \lambda_i\mathbf{I}|$ is null (eq. 2.23):

$$|\mathbf{A} - \lambda_i\mathbf{I}| = 0$$

which means that the rank of $(\mathbf{A} - \lambda_i\mathbf{I})$ is smaller than n. In the case of multiple eigenvalues, if there are *m distinct* eigenvectors corresponding to the m identical eigenvalues λ_i, the determinant of $(\mathbf{A} - \lambda_i\mathbf{I})$ must be null for each of these eigenvalues, but in a different way each time. When $m = 1$, the condition for $|\mathbf{A} - \lambda_i\mathbf{I}| = 0$ is for its rank to be $r = n - 1$. Similarly, in a case of *multiplicity*, the condition for $|\mathbf{A} - \lambda_i\mathbf{I}|$ to be null m times, but distinctly, is for its rank to be $r = n - m$. Consequently, for n distinct eigenvectors to exist, the rank of $(\mathbf{A} - \lambda_i\mathbf{I})$ must be $r = n - m$, and this for any eigenvalue λ_i of multiplicity m.

Numerical example. Here is an example of a full-rank asymmetric matrix $\mathbf{A}$ that has two equal eigenvalues corresponding to distinct eigenvectors. The full-rank condition is shown by the fact that $\det(\mathbf{A}) = -1$, which differs from 0. The eigenvalues are $\lambda_1 = \lambda_2 = 1$ and $\lambda_3 = -1$:

$$\mathbf{A} = \begin{bmatrix} -1 & -2 & -2 \\ 1 & 2 & 1 \\ -1 & -1 & 0 \end{bmatrix} \qquad \text{so that, for } \lambda_1 = \lambda_2 = 1, \qquad (\mathbf{A} - 1\mathbf{I}) = \begin{bmatrix} -2 & -2 & -2 \\ 1 & 1 & 1 \\ -1 & -1 & -1 \end{bmatrix}$$

The multiplicity, or number of multiple eigenvalues, is $m = 2$. The rank of $(\mathbf{A} - \lambda_i\mathbf{I})$ is $r = 1$ because all three columns of this matrix are identical. Thus, for $\lambda_1 = \lambda_2 = 1$, $n - m = 3 - 2 = 1$, which shows that $r = n - m$ in this example. It follows that there exist two distinct eigenvectors $\mathbf{u}_1$ and $\mathbf{u}_2$. They can indeed be calculated:

$$\text{for } \lambda_1 = 1,\ \mathbf{u}_1 = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}, \qquad \text{for } \lambda_2 = 1,\ \mathbf{u}_2 = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}, \qquad \text{whereas for } \lambda_3 = -1,\ \mathbf{u}_3 = \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix}$$

Eigenvectors $\mathbf{u}_1$ and $\mathbf{u}_2$ both correspond to the multiple eigenvalue $\lambda = 1$. Any linear combination of such eigenvectors is also an eigenvector of $\mathbf{A}$ corresponding to λ. For example:

$$\mathbf{u}_1 - \mathbf{u}_2 = \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix} \qquad \mathbf{u}_1 + 2\mathbf{u}_2 = \begin{bmatrix} 3 \\ -2 \\ -1 \end{bmatrix}$$

It can easily be verified that the above two eigenvectors, or any other linear combination of $\mathbf{u}_1$ and $\mathbf{u}_2$, are eigenvectors of $\mathbf{A}$ corresponding to $\lambda = 1$. Of course, the new eigenvectors are not linearly independent of $\mathbf{u}_1$ and $\mathbf{u}_2$, so that there are still only two distinct eigenvectors corresponding to the multiple eigenvalue $\lambda = 1$.

Numerical example. Here is an example of a full-rank asymmetric matrix $\mathbf{A}$ that has two indistinguishable eigenvectors. The full-rank condition is shown by the fact that $\det(\mathbf{A}) = 3$, which differs from 0. The eigenvalues are $\lambda_1 = 3$ and $\lambda_2 = \lambda_3 = 1$:

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 1 \\ 3 & 3 & -2 \\ 4 & 1 & 0 \end{bmatrix} \qquad \text{so that, for } \lambda_2 = \lambda_3 = 1, \qquad (\mathbf{A} - 1\mathbf{I}) = \begin{bmatrix} 1 & -1 & 1 \\ 3 & 2 & -2 \\ 4 & 1 & -1 \end{bmatrix}$$

Table 2.2 Types of *symmetric* matrices and corresponding characteristics of their eigenvalues.

Symmetric matrix	Eigenvalues
All elements of matrix **A** are *real* (i.e. non-imaginary)	All eigenvalues are *real* (i.e. non-imaginary)
Matrix **A** is *positive definite*	All eigenvalues are *positive*
Matrix $\mathbf{A}_{nn}$ is *positive semidefinite* and of rank r	There are r positive and $(n - r)$ null eigenvalues
Matrix $\mathbf{A}_{nn}$ is *negative semidefinite* and of rank r	There are r negative and $(n - r)$ null eigenvalues
Matrix $\mathbf{A}_{nn}$ is *indefinite* and of rank r	There are r non-null (positive and negative) and $(n - r)$ null eigenvalues
Matrix **A** is *diagonal*	The *diagonal elements* are the eigenvalues

The multiplicity, or number of multiple eigenvalues, is $m = 2$. The rank of $(\mathbf{A} - \lambda_i\mathbf{I})$ is $r = 2$ because any two of the three rows (or columns) of this matrix are independent of each other. Thus, for $\lambda_2 = \lambda_3 = 1$, $n - m = 3 - 2 = 1$, which shows that $r \neq n - m$ in this example. The conclusion is that there do not exist two independent eigenvectors associated with the eigenvalue of multiplicity $m = 2$. The eigenvectors are the following:

$$\text{for } \lambda_1 = 3, \quad \mathbf{u}_1 = \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix} \qquad \text{whereas for } \lambda_2 = \lambda_3 = 1, \quad \mathbf{u}_1 = \mathbf{u}_2 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$$

In the case of a square *symmetric* matrix, it is always possible to calculate m *orthogonal* eigenvectors corresponding to multiple eigenvalues, when present. This is not necessarily true for *non-symmetric* matrices, where the number of eigenvectors may be smaller than m. Therefore, whatever their multiplicity, eigenvalues of most matrices of interest to ecologists, including association matrices (Section 2.2), have distinct eigenvectors associated with them. In any case, it is unlikely that eigenvalues of matrices computed from field data be exactly equal (i.e. multiple).

Fourth property. — A property of square *symmetric* matrices may be used to predict the nature of their eigenvalues (Table 2.2). A symmetric matrix **A** may be combined with any vector $\mathbf{t} \neq \mathbf{0}$, in a matrix expression of the form $\mathbf{t}'\mathbf{A}\mathbf{t}$ which is

Quadratic form

known as a *quadratic form*. This expression results in a scalar whose value leads to the following definitions:

- if $\mathbf{t}'\mathbf{At}$ is always positive, matrix $\mathbf{A}$ is *positive definite*;
- if $\mathbf{t}'\mathbf{At}$ can be either positive or null, matrix $\mathbf{A}$ is *positive semidefinite*;
- if $\mathbf{t}'\mathbf{At}$ can be either negative or null, matrix $\mathbf{A}$ is *negative semidefinite*;
- if $\mathbf{t}'\mathbf{At}$ can be either negative, null or positive, matrix $\mathbf{A}$ is *indefinite*.

2.11 Singular value decomposition

Another useful method of matrix decomposition is *singular value decomposition* (SVD). The *approximation theorem* of Schmidt (1907), later rediscovered by Eckart & Young (1936), showed that any *rectangular* matrix $\mathbf{Y}$ can be decomposed as follows:

$$\mathbf{Y}(n\times p) = \mathbf{V}(n\times k)\ \mathbf{W}(\text{diagonal}, k\times k)\ \mathbf{U}'(k\times p) \qquad \textbf{(2.31)}$$

Singular value

where both $\mathbf{U}$ and $\mathbf{V}$ are orthonormal matrices (i.e. matrices containing column vectors that are normalized and orthogonal to one another; Section 2.8). $\mathbf{W}$ is a diagonal matrix $\mathbf{D}(w_i)$, of order $k = \min(n, p)$, containing the *singular values*; the illustration hereunder assumes that $n > p$ so that $k = p$. The notation $\mathbf{D}(w_i)$ for the diagonal matrix of singular values will be used in the remainder of this section. The early history of singular value decomposition has been recounted by Stewart (1993). The following illustration shows the shapes of these matrices:

$$\begin{bmatrix} \mathbf{Y}_{(n\times p)} \end{bmatrix} = \begin{bmatrix} \mathbf{V}_{(n\times p)} \end{bmatrix} \begin{bmatrix} w_1 & 0 & 0 & \dots & 0 \\ 0 & w_2 & 0 & \dots & 0 \\ 0 & 0 & w_3 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & w_p \end{bmatrix} \begin{bmatrix} \mathbf{U}'_{(p\times p)} \end{bmatrix}$$

Demonstrating eq. 2.31 is beyond the scope of this book. The diagonal values w_i in $\mathbf{D}(w_i)$ are non-negative; they are the singular values of $\mathbf{Y}$. SVD functions are found in advanced statistical languages such as R, S-PLUS® and MATLAB®. The notation used in different manuals and computer software may, however, differ from the one used here. That is the case of the R language, where function ***svd()*** is said to decomposes $\mathbf{Y}$ into $\mathbf{UD}(w_i)\mathbf{V}'$, instead of the notation $\mathbf{VD}(w_i)\mathbf{U}'$ used here to insure consistency between the results of eigenanalysis and SVD in Subsection 9.1.9.

Application 1: Rank of a rectangular matrix. — The rank of a rectangular matrix is equal to the number of singular values larger than 0. As an illustration, consider the matrix in Numerical example 2 of Section 2.7:

$$\mathbf{Y} = \begin{bmatrix} 2 & 1 & 3 & 4 \\ -1 & 6 & -3 & 0 \\ 1 & 20 & -3 & 8 \end{bmatrix}$$

In this example, (n = 3) < (p = 4), hence $k = n = 3$, and the dimensions of matrices in eq. 2.31 are **V**(3×3), **W**(3×3) and **U**'(3×4). Singular value decomposition of that matrix produces two singular values larger than zero and one null singular value. SVD of the transposed matrix produces the same singular values. **Y** is thus of rank 2. After elimination of the third (null) singular value and the corresponding vector in both **V** and **U'**, the singular value decomposition of **Y** gives:

$$\mathbf{Y} = \begin{bmatrix} -0.08682 & 0.84068 \\ -0.26247 & -0.53689 \\ -0.96103 & 0.07069 \end{bmatrix} \begin{bmatrix} 22.650 & 0 \\ 0 & 6.081 \end{bmatrix} \begin{bmatrix} -0.03851 & -0.92195 & 0.15055 & -0.35477 \\ 0.37642 & -0.15902 & 0.64476 & 0.64600 \end{bmatrix}$$

Application 2: Decomposition of a cross-product matrix. — A covariance matrix is a type of cross-product matrix (Chapter 4). Consider the covariance matrix **S** of the data used to illustrate principal component analysis in Section 9.1. It is decomposed as follows by SVD:

$$\begin{matrix} \mathbf{S} & = & \mathbf{V} & \mathbf{D}(w_i) & \mathbf{U}' \end{matrix}$$

$$\begin{bmatrix} 8.2 & 1.6 \\ 1.6 & 5.8 \end{bmatrix} = \begin{bmatrix} -0.8944 & -0.4472 \\ -0.4472 & 0.8944 \end{bmatrix} \begin{bmatrix} 9 & 0 \\ 0 & 5 \end{bmatrix} \begin{bmatrix} -0.8944 & -0.4472 \\ -0.4472 & 0.8944 \end{bmatrix}$$

The singular values of **S**, found on the diagonal of $\mathbf{D}(w_i)$, are equal to the eigenvalues. This is true for any square symmetric matrix. Matrices **V** and **U** contain vectors identical to the eigenvectors obtained by eigenanalysis; eigenvectors may vary in their signs depending on the program or the computer platform. Negative eigenvalues, which may be found in principal coordinate analysis of symmetric distance matrices (PCoA, Section 9.3), will come out as singular values with positive signs. Example:

matrix $\begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \\ 3 & 4 & 5 \end{bmatrix}$

has the singular values [9.6235, 6.2348, 0.0000] and the following set of eigenvalues: [9.6235, 0.0000, –6.2348]. The singular value 6.2348 with a positive sign corresponds to the negative eigenvalue –6.2348.

Application 3: Generalized matrix inversion. — SVD offers a way of inverting matrices that are singular (Section 2.8) or numerically very close to being singular. SVD may either give users a clear diagnostic of the problem, or solve it. Singularity may be encountered in regression for example: if the matrix of explanatory variables **X** is not of full rank, the cross-product matrix $\mathbf{A} = [\mathbf{X}'\mathbf{X}]$ is singular and it cannot be inverted with the methods described in Section 2.8, although inversion is necessary to solve eq. 2.19.

Inversion of $\mathbf{A} = [\mathbf{X}'\mathbf{X}]$ by SVD involves the following steps. First, **A** is decomposed using eq. 2.31:

$$\mathbf{A} = \mathbf{V}\ \mathbf{D}\ (w_i)\mathbf{U}'$$

Since **A** is symmetric, **V**, $\mathbf{D}(w_i)$, and **U** are all square matrices of the same size as **A**. Using property 5 of matrix inverses (above), the inverse of **A** is easy to compute:

$$\mathbf{A}^{-1} = [\mathbf{V}\ \mathbf{D}(w_i)\mathbf{U}']^{-1} = [\mathbf{U}']^{-1}[\mathbf{D}(w_i)]^{-1}[\mathbf{V}]^{-1}$$

Because **U** and **V** are orthonormal, their inverses are equal to their transposes (property 7 of inverses), whereas the inverse of a diagonal matrix is a diagonal matrix whose elements are the reciprocals of the original elements (property 8). Hence:

$$\mathbf{A}^{-1} = \mathbf{U}\mathbf{D}(1/w_i)\mathbf{V}' \qquad \textbf{(2.32)}$$

Singular matrix

Ill-conditioned matrix

It may happen that one or more of the w_i's are zero, so that their reciprocals are infinite; **A** is then a *singular matrix*. This is what happens in the regression case when **X** is not of full rank. It may also happen that one or more of the w_i's are numerically so small that their values cannot be properly computed because of the machine's precision in floating-point calculation; in that case, **A** is said to be *ill-conditioned*. When **A** is singular, the columns of **U** corresponding to the zero elements in $\mathbf{D}(w_i)$ form an orthonormal basis* for the space where the system of equations has no solution, whereas the columns of **V** corresponding to the non-zero elements in $\mathbf{D}(w_i)$ are an orthonormal basis for the space where the system has a solution. When **A** is singular or ill-conditioned, it is still possible to find its inverse, either exactly or approximately, and use it to compute a regression model. Here is an example:

$$\mathbf{y} = \begin{bmatrix} 1.25 \\ 1.13 \\ 1.60 \\ 2.08 \\ 2.10 \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & 5 \\ 1 & 2 & 2 & 2 \\ 1 & 3 & 3 & 4 \\ 1 & 4 & 4 & 3 \\ 1 & 5 & 5 & 1 \end{bmatrix}$$

* A set of k linearly independent vectors form a basis for a k-dimensional vector space. Any vector in that space can be uniquely written as a linear combination of the base vectors.

The first column of **X** contains 1's to estimate the intercept. Columns 2 and 3 are identical, so **X** is not of full rank. Equation 2.31 produces a decomposition of $\mathbf{A} = [\mathbf{X}'\mathbf{X}]$ that has 3 (not 4) singular values larger than 0. A generalized inverse is obtained by computing eq. 2.32 after removing the last column from **U** and **V** and the last row and column from $\mathbf{D}(w_i)$:

$$\mathbf{A}^{-1} = \mathbf{U}\mathbf{D}(1/w_i)\mathbf{V}'$$

$$= \begin{bmatrix} -0.17891 & 0.07546 & 0.98097 \\ -0.59259 & -0.37762 & -0.07903 \\ -0.59259 & -0.37762 & -0.07903 \\ -0.51544 & 0.84209 & -0.15878 \end{bmatrix} \begin{bmatrix} 0.00678 & 0 & 0 \\ 0 & 0.04492 & 0 \\ 0 & 0 & 6.44243 \end{bmatrix} \begin{bmatrix} -0.17891 & -0.59259 & -0.59259 & -0.51544 \\ 0.07546 & -0.37762 & -0.37762 & 0.84209 \\ 0.98097 & -0.07903 & -0.07903 & -0.15878 \end{bmatrix}$$

$$= \begin{bmatrix} 6.20000 & -0.50000 & -0.50000 & -1.00000 \\ -0.50000 & 0.04902 & 0.04902 & 0.06863 \\ -0.50000 & 0.04902 & 0.04902 & 0.06863 \\ -1.00000 & 0.06863 & 0.06863 & 0.19608 \end{bmatrix}$$

Using the generalized inverse $\mathbf{A}^{-1}$, the regression coefficients can now be computed (eq. 2.19):

$$\mathbf{b} = [\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{y} = \mathbf{A}^{-1}\mathbf{X}'\mathbf{y} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 0.21200 \\ 0.17539 \\ 0.17539 \\ 0.12255 \end{bmatrix}$$

The first value in vector **b** is the intercept. Now remove column 2 from **X** and compute a multiple linear regression equation. The regression coefficients are:

$$\begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 0.21200 \\ 0.35078 \\ 0.12255 \end{bmatrix}$$

The regression coefficient for the second column of **X**, 0.35078, has been split in two equal coefficients of 0.17539 in the SVD solution when the two identical variables were kept in the analysis.

Similar problems may be encountered when solving sets of simultaneous linear equations represented by matrix equation $\mathbf{A}\mathbf{b} = \mathbf{c}$ (Section 2.8). In this book, SVD will also be used in algorithms for principal component analysis (Subsection 9.1.9) and correspondence analysis (Subsection 9.2.1).

2.12 Software

Functions for all matrix operations described in this chapter are available in the R language. Standard matrix operations are available in the BASE package while more specialized operations are found in the MATRIX package.

Among the functions found in BASE are ***det()*** to compute a determinant, ***solve()*** for matrix inversion or solving a system of linear equations, ***eigen()*** for eigenvalue decomposition, and ***svd()*** for singular value decomposition. Other useful decompositions used in later chapters but not discussed in Chapter 2 are the QR decomposition (function ***qr()*** of BASE) and Cholesky factorization (functions ***chol()*** of BASE and MATRIX). Package MASS offers function ***ginv()*** for general inversion.

Functions implementing matrix algebra are also available in S-PLUS®, MATLAB® and SAS®.

Chapter

3

Dimensional analysis in ecology

3.0 Dimensional analysis

Dimensional analysis is generally not part of the curriculum of ecologists, so that relatively few are conversant with this simple but remarkably powerful tool. Yet, applications of dimensional analysis are found in the ecological literature, where results clearly demonstrate the advantage of using this mathematical approach.

"Dimensional analysis treats the *general forms of equations* that describe natural phenomena" (Langhaar, 1951). The basic principles of this discipline were established by physicists (Fourier, 1822; Maxwell, 1871) and later applied by engineers to the very important area of small-scale modelling. Readers interested in the fundamentals and engineering applications of dimensional analysis should refer, for example, to Langhaar (1951), from which are taken several of the topics developed in the present chapter. Other useful references are Ipsen (1960), Huntley (1967), and Schneider (1994).

The use of dimensional analysis in ecology rests on the fact that a growing number of areas in ecological science use *equations*; for example, populations dynamics and ecological modelling. The study of equations is the very basis of dimensional analysis. This powerful approach can easily be used by ecologists, given the facts that it can be reduced to *a single theorem* (the Π theorem) and that many of its applications (Sections 3.1 and 3.2) only require a knowledge of elementary mathematics.

Dimensional analysis can resolve complex ecological problems in a simple and elegant manner. Readers should therefore not be surprised that ecological applications in the present chapter are of a rather high level, since the advantage of dimensional analysis lies precisely in its ability to handle complex problems. It follows that dimensional analysis is mainly useful in those cases where it would be difficult to resolve the ecological problem by conventional approaches.

3.1 Dimensions

All fields of science, including ecology, rest on a number of abstract entities such as the mass, length, time, temperature, speed, acceleration, radioactivity, concentration, energy or volume. These entities, which can be measured, are called *quantities*. Designing a *system of units* requires to: (1) arbitrarily choose a small number of *fundamental quantities*, on which a coherent and practical system can be constructed, and (2) arbitrarily assign, to each of these quantities, *base units* chosen as references for comparing measurements.

International System of Units

Various systems of units have been developed in the past, e.g. the British system and several versions of the metric system. The latter include the CGS metric system used by scientists (based on the centimetre, the gram and the second), the MKS (force) metric system used by engineers (based on the metre, the kilogram and the second, where the kilogram is the unit of *force*), and the MKS (mass) metric system (where the kilogram is the unit of *mass*). Since 1960, there is an internationally accepted version of the metric system, called the *International System of Units* (SI, from the French name *Système international d'unités*; see Plate 3.1, p. 142). The SI is based on seven *quantities*, to which are associated seven *base units* (Table 3.1; the mole was added to the SI in 1971 only). In addition to these seven base units, the SI recognizes two

Table 3.1 Base units of the International System of Units (Sl).

Fundamental quantity	Quantity symbol*	Dimension symbol	Base unit	Unit symbol
mass	m	[M]	kilogram	kg
length	l	[L]	metre†	m
time	t	[T]	second	s
electric current	I	[I]	ampere	A
thermodynamic temperature	T ‡	[θ]	kelvin‡	K
amount of substance	n	[N]	mole	mol
luminous intensity	I_v	[J]	candela	cd

* Quantity symbols are not part of the SI, and they are not unique.

† Spelled meter in the United States of America.

‡ In ecology, temperature is generally measured on the Celsius scale, where the unit is the *degree Celsius* (°C); the quantity symbol for temperatures expressed in °C is usually t. Note that the absolute temperature unit is the kelvin, *not degree kelvin*.

supplementary units, the radian (rad) and the steradian (sr), which measure planar and solid angles, respectively. All other units, called *derived units*, are combinations of the base and supplementary units. Some frequently used derived units have special names, e.g. volt, lux, joule, newton, ohm. It must be noted that: (1) *unit names* are written with small letters only, the sole exception being the degree Celsius; (2) *unit symbols* are written with small letters only, except the symbols of derived units that are surnames, whose first letter is a capital (e.g. Pa for pascal), and the litre (see Table 3.2, footnote). Unit symbols are *not* abbreviations, hence they are *never* followed by a dot.

Dimension

Table 3.2 shows that derived units are not only simple products of the fundamental units, but that they are often *powers* and *combinations of powers* of these units. Maxwell (1871) used symbols such as [M], [L], [T], and [θ] to represent the quantities mass, length, time and temperature (Table 3.1). The *dimensions* of the various quantities are *products of powers* of the symbols of fundamental quantities. Thus, the dimension of an area is $[L^2]$, of a volume $[L^3]$, of a speed $[LT^{-1}]$, and of an acceleration $[LT^{-2}]$. Table 3.2 gives the exponents of the dimensional form of the most frequently encountered quantities.

Since the various quantities are *products of powers,* going from one quantity to another is done simply by *adding* (or *subtracting) exponents* of the dimensions. For example, one calculates the dimensions of *heat conductivity* $W(mK)^{-1}$ by subtracting, from the dimension exponents of *power* W, the sum of the dimension exponents of *length* m and of *temperature* K:

$$[M^1L^2T^{-3}] / ([L^1] \times [\theta^1]) = [M^1L^{(2-1)}T^{-3}\theta^{-(1)}] = [M^1L^1T^{-3}\theta^{-1}]$$

The first three fundamental quantities (Table 3.1), mass [M], length [L], and time [T], are enough to describe any Newtonian mechanical system. Ecologists may require, in addition, temperature [θ], amount of substance [N], and luminous intensity [J]. Research in electromagnetism calls for electric current [I] and, in quantum mechanics, one uses the quantum state of the system [Ψ].

Four types of entities are recognized:

(1) *dimensional variables*, e.g. most of the quantities listed in Table 3.2;

(2) *dimensional constants*, for instance: the speed of light in vacuum $[LT^{-1}]$, $c = 2.998 \times 10^8$ m s^{-1}; the acceleration due to Earth's gravity at sea level $[LT^{-2}]$, $g = 9.807$ m s^{-2}; the number of elementary entities in a mole $N_A = 6.022 \times 10^{23}$ mol^{-1}, where N_A is the Avogadro number (note that the nature of the elementary entities in a mole must always be specified, e.g. mol C, mol photons);

(3) *dimensionless variables,* such as angles, relative density (Table 3.2), or dimensionless products which will be studied in following sections;

(4) *dimensionless constants,* e.g. π, e, 2, 7; it must be noted that exponents are, by definition, dimensionless constants.

Table 3.2 Dimensions, units, and names of quantities. Units follow the standards of the International System of Units (SI).

Quantity	[M]	[L]	[T]	[I]	[θ]	[N]	[J]	Units	Name*
mass	1	0	0	0	0	0	0	kg	kilogram
length	0	1	0	0	0	0	0	m	metre
time	0	0	1	0	0	0	0	s	second
electric current	0	0	0	1	0	0	0	A	ampere
temperature	0	0	0	0	1	0	0	K	kelvin
amount of substance	0	0	0	0	0	1	0	mol	mole
luminous intensity	0	0	0	0	0	0	1	cd	candela
absorbed dose	0	2	–2	0	0	0	0	$J\,kg^{-1} = Gy$	gray
acceleration (angular)	0	0	–2	0	0	0	0	$rad\,s^{-2}$	
acceleration (linear)	0	1	–2	0	0	0	0	$m\,s^{-2}$	
activity of radioactive source	0	0	–1	0	0	0	0	$s^{-1} = Bq$	becquerel
angle (planar)	0	0	0	0	0	0	0	rad	radian
angle (solid)	0	0	0	0	0	0	0	sr	steradian
angular momentum	1	2	–1	0	0	0	0	$kg\,m^2\,s^{-1}$	
angular velocity	0	0	–1	0	0	0	0	$rad\,s^{-1}$	
area	0	2	0	0	0	0	0	m^2	
compressibility	–1	1	2	0	0	0	0	Pa^{-1}	
concentration (molarity)	0	–3	0	0	0	1	0	$mol\,m^{-3}$	
current density	0	–2	0	1	0	0	0	$A\,m^{-2}$	
density (mass density)	1	–3	0	0	0	0	0	$kg\,m^{-3}$	
electric capacitance	–1	–2	4	2	0	0	0	$C\,V^{-1} = F$	farad
electric charge	0	0	1	1	0	0	0	$A\,s = C$	coulomb
electric conductance	–1	–2	3	2	0	0	0	$\Omega^{-1} = S$	siemens
electric field strength	1	1	–3	–1	0	0	0	$V\,m^{-1}$	
electric resistance	1	2	–3	–2	0	0	0	$V\,A^{-1} = \Omega$	ohm
electric potential	1	2	–3	–1	0	0	0	$W\,A^{-1} = V$	volt
energy	1	2	–2	0	0	0	0	$N\,m = J$	joule
force	1	1	–2	0	0	0	0	$kg\,m\,s^{-2} = N$	newton
frequency	0	0	–1	0	0	0	0	$s^{-1} = Hz$	hertz
heat capacity	1	2	–2	0	–1	0	0	$J\,K^{-1}$	
heat conductivity	1	1	–3	0	–1	0	0	$W(m\,K)^{-1}$	
heat flux density	1	0	–3	0	0	0	0	$W\,m^{-2}$	
illuminance	0	–2	0	0	0	0	1	$lm\,m^{-2} = lx$	lux
inductance	1	2	–2	–2	0	0	0	$Wb\,A^{-1} = H$	henry
light exposure	0	–2	1	0	0	0	1	lx s	
luminance	0	–2	0	0	0	0	1	$cd\,m^{-2}$	
luminous flux	0	0	0	0	0	0	1	$cd\,sr = lm$	lumen
magnetic field strength	0	–1	0	1	0	0	0	$A\,m^{-1}$	
magnetic flux	1	2	–2	–1	0	0	0	$V\,s = Wb$	weber
magnetic flux density	1	0	–2	–1	0	0	0	$Wb\,m^{-2} = T$	tesla
magnetic induction	1	0	–2	–1	0	0	0	$Wb\,m^{-2} = T$	tesla

* Only base units and special names of derived units are listed.

Table 3.2 Dimensions, units, and names of quantities (continued).

Quantity	[M]	[L]	[T]	[I]	[θ]	[N]	[J]	Units	Name
magnetic permeability	1	1	–2	–2	0	0	0	Ω s m^{-1}	
mass flow rate	1	0	–1	0	0	0	0	kg s^{-1}	
molality	–1	0	0	0	0	1	0	mol kg^{-1}	
molarity	0	–3	0	0	0	1	0	mol m^{-3}	
molar internal energy	1	2	–2	0	0	–1	0	J mol^{-1}	
molar mass	1	0	0	0	0	–1	0	kg mol^{-1}	
molar volume	0	3	0	0	0	–1	0	m^3 mol^{-1}	
moment of force	1	2	–2	0	0	0	0	N m	
moment of inertia	1	2	0	0	0	0	0	kg m^2	
momentum	1	1	–1	0	0	0	0	kg m s^{-1}	
period	0	0	1	0	0	0	0	s	
permittivity	–1	–3	4	2	0	0	0	F m^{-1}	
power	1	2	–3	0	0	0	0	J s^{-1} = W	watt
pressure	1	–1	–2	0	0	0	0	N m^{-2} = Pa	pascal
quantity of light	0	0	1	0	0	0	1	lm s	
radiant intensity	1	2	–3	0	0	0	0	W sr^{-1}	
relative density	0	0	0	0	0	0	0	(no unit)	
rotational frequency	0	0	–1	0	0	0	0	s^{-1}	
second moment of area	0	4	0	0	0	0	0	m^4	
specific heat capacity	0	2	–2	0	–1	0	0	J(kg K)$^{-1}$	
specific latent heat	0	2	–2	0	0	0	0	J kg^{-1}	
specific volume	–1	3	0	0	0	0	0	m^3 kg^{-1}	
speed	0	1	–1	0	0	0	0	m s^{-1}	
stress	1	–1	–2	0	0	0	0	N m^{-2} = Pa	pascal
surface tension	1	0	–2	0	0	0	0	N m^{-1}	
torque	1	2	–2	0	0	0	0	N m	
viscosity (dynamic)	1	–1	–1	0	0	0	0	Pa s	
viscosity (kinetic)	0	2	–1	0	0	0	0	m^2 s^{-1}	
volume [†]	0	3	0	0	0	0	0	m^3	
volume flow rate	0	3	–1	0	0	0	0	m^3 s^{-1}	
wavelength	0	1	0	0	0	0	0	m	
wave number	0	–1	0	0	0	0	0	m^{-1}	
work	1	2	–2	0	0	0	0	N m = J	joule

[†] The litre (spelt liter in the United States of America) is the *capacity* (vs. *cubic*) unit of volume. Its symbol (letter l) may be confused with digit one (1) in printed texts so that it was decided in 1979 that capital L could be used as well; 1 m^3 = 1000 L.

The very concept of *dimension* leads to immediate applications in physics and ecology. In physics, for example, one can easily demonstrate that the first derivative of distance with respect to time is a speed:

$$\text{dimensions of } \frac{dl}{dt}: \quad \left[\frac{\mathrm{L}}{\mathrm{T}}\right] = [\mathrm{LT}^{-1}], \text{ i.e. speed.}$$

Similarly, it can be shown that the second derivative is an acceleration:

$$\text{dimensions of } \frac{d^2 l}{dt^2} = \frac{d}{dt}\left(\frac{dl}{dt}\right): \quad \left[\frac{\mathrm{L}}{\mathrm{TT}}\right] = [\mathrm{LT}^{-2}], \text{ i.e. acceleration.}$$

Note that *italic*s are used for *quantity symbols* such as length (l), mass (m), time (t), area (A), and so on. This distinguishes them from *unit symbols* (roman type; Tables 3.1 and 3.2), and *dimension symbols* (roman capitals in brackets; Table 3.1).

Ecological application 3.1

Platt (1969) studied the efficiency of *primary (phytoplankton) production* in the *aquatic environment*. Primary production is generally determined at different depths in the water column, so that it is difficult to compare values observed under different conditions. The solution to this problem consists in finding a method to standardize the values, for example by transforming field estimates of *primary production* into values of *energy efficiency*. Such a transformation would eliminate the effect on production of solar irradiance at different locations and different depths. Primary production at a given depth $P(z)$ may be expressed in J m^{-3} s^{-1} [ML^{-1} T^{-3}], while irradiance at the same depth $E(z)$ is in J m^{-2} s^{-1} [MT^{-3}] (energy units).

The dimension of the ratio $P(z)/E(z)$, which defines the energy efficiency of primary production, is thus [L^{-1}]. Another property determined in the water column, which also has dimension [L^{-1}], is the *attenuation* of diffuse light as a function of depth. The *coefficient of diffuse light attenuation* (α) is defined as:

$$E(z_2) = E(z_1)\, e^{-\alpha(z_2 - z_1)}$$

where $E(z_2)$ and $E(z_1)$ are irradiances at depths z_2 and z_1, respectively. Given the fact that an exponent is, by definition, dimensionless, the dimension of α must be [L^{-1}] since that of depth z is [L].

Based on the dimensional similarity between efficiency and attenuation, and considering the physical aspects of light attenuation in the water column, Platt partitioned the attenuation coefficient (α) into physical (k_p) and biological (k_b) components, i.e. $\alpha = k_p + k_b$. The *biological attenuation coefficient* k_p may be used to estimate the attenuation of light caused by photosynthetic processes. In the same paper and in further publications by Platt & Subba Rao (1970) and Legendre (1971), it was shown that there exists a correlation in the marine environment between k_b and the concentration of chlorophyll *a*. The above papers used the calorie as unit of energy but, according to the SI standard, this unit should no longer be used. Coherency requires here that primary production be expressed in J m^{-3} s^{-1} and irradiance in J m^{-2} s^{-1} (or W m^{-2}).

This example illustrates how a simple reflection, based on dimensions, led to an interesting development in the field of ecology.

It is therefore useful to think in terms of *dimensions* when dealing with ecological equations that contain physical *quantities*. Even if this habit is worth cultivating, it would not however, in and of itself, justify an entire chapter in the present book. So, let us move forward in the study of dimensional analysis.

3.2 Fundamental principles and the Pi theorem

It was shown in the previous section that going from one quantity to another is generally done by multiplying or dividing quantities characterized by *different dimensions*. In contrast, additions and subtractions can only be performed on quantities having the *same dimensions* — hence the fundamental principle of *dimensional homogeneity*. Any equation of the general form

Dimensional homogeneity

$$a + b + c + \ldots = g + h + \ldots$$

is dimensionally homogeneous if and only if all quantities $a, b, c, \ldots g, h, \ldots$ have the *same dimensions*. This property applies to all equations of a *theoretical* nature, but it does not necessarily apply to those derived *empirically*. Readers must be aware that dimensional analysis only deals with dimensionally homogeneous equations. In animal ecology, for example, the basic equation for energy budgets is:

$$dW/dt = R - T \qquad \textbf{(3.1)}$$

where W is the mass of an animal, R its food ration, and T its metabolic expenditure rate (oxygen consumption). This equation, which describes growth dW/dt as a function of ration R and metabolic rate T, is dimensionally homogeneous. The rate of oxygen consumption T is expressed as mass per unit time, its dimensions thus being $[MT^{-1}]$, as those of food ration R. The dimensions of dW/dt are also clearly $[MT^{-1}]$. This same equation will be used in Ecological applications 3.2e and 3.3b, together with other ecological equations — all of which are dimensionally homogeneous.

In dimensional analysis, the correct identification of quantities to be included in a given equation is much more important than the exact form of the equation. Researchers using dimensional analysis must therefore have prior knowledge of the phenomenon under study, in order to identify the pertinent *dimensional variables* and *constants*. On the one hand, missing key quantities could lead to incomplete or incorrect results, or even to a deadlock. On the other hand, including unnecessary terms could overburden the solution needlessly. Hence, dimensional analysis cannot be conducted without first considering the ecological bases of the problem. A simple example, taken from hydrodynamics, will illustrate the dimensional method.

The question considered here relates to the work of many ecologists in aquatic environments, i.e. estimating the drag experienced by an object immersed in a current. Ecologists who moor current meters or other probes must consider the drag, lest the equipment might be carried away. To simplify the problem, one assumes that the immersed object is a smooth sphere and that the *velocity* of the current V is constant. The drag *force* F is then a function of: the *velocity* (V), the *diameter* of the sphere (D), the *density* of water (ρ), and its *dynamic viscosity* (η). The simplest equation relating these five quantities is:

$$F = f(V, D, \rho, \eta) \tag{3.2}$$

At first sight, nothing specifies the nature of the dependency of F on V, D, ρ, and η, except that such a dependency exists. Dimensional analysis allows one to find the form of the equation that relates F to the variables identified as governing the drag.

Dimensionless product

A number of variables are regularly encountered in hydrodynamics problems, i.e. F, V, L, ρ, η, to which one must also add g, the acceleration due to gravity. Some of these variables may be combined to form *dimensionless products*. Specialists of hydrodynamics have given names to some often-used dimensionless products:

$$\textit{Reynolds number}: Re = \frac{VL\rho}{\eta} = \frac{[\mathrm{LT^{-1}}]\,[\mathrm{L}]\,[\mathrm{ML^{-3}}]}{[\mathrm{ML^{-1}T^{-1}}]} = \frac{[\mathrm{ML^{-1}T^{-1}}]}{[\mathrm{ML^{-1}T^{-1}}]} = [1] \tag{3.3}$$

$$\textit{Newton number}: Ne = \frac{F}{\rho L^2 V^2} = \frac{[\mathrm{MLT^{-2}}]}{[\mathrm{ML^{-3}}]\,[\mathrm{L^2}]\,[\mathrm{L^2T^{-2}}]} = \frac{[\mathrm{MLT^{-2}}]}{[\mathrm{MLT^{-2}}]} = [1] \tag{3.4}$$

$$\textit{Froude number}: Fr = \frac{V^2}{Lg} = \frac{[\mathrm{L^2T^{-2}}]}{[\mathrm{L}]\,[\mathrm{T^{-2}}]} = \frac{[\mathrm{L^2T^{-2}}]}{[\mathrm{L^2T^{-2}}]} = [1] \tag{3.5}$$

Each of the above *products* is clearly *dimensionless*. It should also be noted that each product of this set is *independent* of the others, since each contains one exclusive variable, i.e. η for *Re*, F for *Ne*, and g for *Fr*. Finally, any other dimensionless product of these same variables would *inevitably* be a product of powers of dimensionless products from the above set. The three dimensionless products thus form a *complete set* of dimensionless products for variables F, V, L, ρ, η and g. It would obviously be possible to form other complete sets of dimensionless products using these same variables, by combining them differently.

The first important concept to remember is that of *dimensionless product*. This concept leads to the *sole* theorem of dimensional analysis, the Π theorem, which is also known as the Buckingham theorem.

Given the fundamental principle of dimensional homogeneity (see above), it follows that any equation that combines dimensionless products is dimensionally

homogeneous. Thus, a *sufficient* condition for an equation to be dimensionally homogeneous is that it could be reduced to an equation combining dimensionless products. Indeed, any equation that can be reduced to an equation made of dimensionless products is dimensionally homogeneous. Buckingham (1914) did show that this condition is not only *sufficient* but also *necessary*. This leads to the Π *(Pi) theorem* (the capital Greek letter Π is the mathematical symbol for product):

Π theorem

If an equation is dimensionally homogeneous, it can be reduced to a relationship among the members of a complete set of dimensionless products.

This theorem alone summarizes the whole theory of dimensional analysis.

The power of the Π theorem is illustrated by the solution of the drag problem, introduced above. Equation 3.2 is, by definition, dimensionally homogeneous:

$$F = f(V, D, \rho, \eta)$$

It may be rewritten as:

$$f(F, V, D, \rho, \eta) = 0 \qquad \textbf{(3.6)}$$

The complete set of dimensionless products of the five variables F, V, D, ρ, η contains two products, i.e. the Reynolds (*Re*) and Newton (*Ne*) numbers (D being a length, it is a quantity of type L). Hence, eq. 3.6 may be rewritten as a relation between the members of this complete set of dimensionless products (Π theorem):

$$Ne = f(Re)$$

$$\frac{F}{\rho V^2 D^2} = f(Re) \qquad \textbf{(3.7)}$$

In this equation, the function f is, for the time being, unknown, except that it depends on the sole dimensionless variable *Re*.

The projected area (A) of a sphere is:

$$A = \pi\,(D/2)^2 = (1/4)\,\pi\,D^2, \text{ so that } D^2 = 4A/\pi$$

which allows one to rewrite eq. 3.7 as:

$$\frac{F}{\rho V^2 \dfrac{4A}{\pi}} = f(Re)$$

$$\frac{F}{\rho V^2 A} = \frac{1}{2}\left(\frac{8}{\pi}\right) f(Re)$$

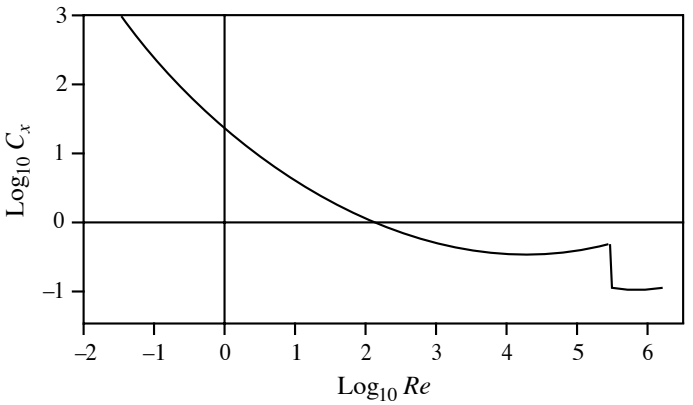


Figure 3.1 Drag coefficient on smooth spheres. Adapted from Eisner (1931).

In hydrodynamics, the term $(8/\pi)\,f\,(Re)$ is called the *drag coefficient* and is represented by $C_{x,}$ so that the drag exerted on a sphere is:

$$F = (1/2)\ C_x \rho V^2 A, \text{ where } C_x = (8/\pi)\,f\,(Re) \qquad \textbf{(3.8)}$$

Since C_x is a function of the sole dimensionless coefficient *Re*, the problem is resolved by determining, in the laboratory, the experimental curve of C_x as a function of *Re*. This curve will be valid for any density (ρ) or dynamic viscosity (η) of any fluid under consideration (the same curve can thus be used for water, air, etc.) and for objects of any size, or any flow speed. The curve may thus be determined by researchers under the most suitable conditions, i.e. choosing fluids and flow speeds that are most convenient for laboratory work. As a matter of fact, this curve is already known (Fig. 3.1).

Two important properties follow from the above example.

(1) First, data to build a *dimensionless graph* should be obtained under the most convenient conditions. For example, determining C_x for a sphere of diameter 3.48 m immersed in air at 14.4°C with a velocity of 15.24 m s^{-1} would be difficult and costly. In contrast, it would be much easier, in most laboratories, to determine C_x by using a sphere of diameter 0.61 m in water at 14.4°C with a speed of 5.79 m s^{-1}. In both cases, *Re* is the same so that the measured value of C_x is the same. This first property is the basis for *model testing* in engineering (Section 3.4), the sphere in air being here the *prototype* and that in water, the *model*.

Chart

(2) The dimensionless graph of Fig. 3.1 contains much more information than a set of *charts* depicting the function of the 4 variables. In a chart (Fig. 3.2), *a function of*

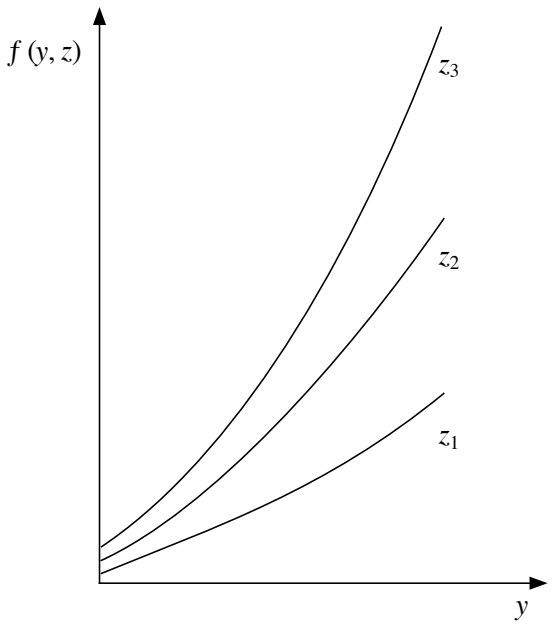


Figure 3.2 Chart representing a function of two variables. One curve is required for each value of the second variable ($z_1, z_2, z_3, \ldots$)

two variables is represented by a *family of curves*, one curve being required for each value of the second variable. A *function of three variables* would be represented by a *set of sets of charts*. Hence, for four variables and assuming that there were only five values measured per variable, a total of 625 experimental points would be required, i.e. five sets of five charts each. With 25 times fewer experimental points, one can easily obtain a dimensionless graph (e.g. Fig. 3.1) which is both more exact and much more convenient.

The above physical example illustrated the great simplicity and remarkable power of dimensional analysis. Let us now examine examples from ecology.

Ecological application 3.2a

This first example belongs to the disciplines of ecology and physiology, since it concerns the dimensions of animals and their muscular dynamics. Hill (1950) compared different cetaceans, as a set of similar animals which differ in size. All these cetaceans (porpoises, dolphins, and whales), with a 5000-fold mass range, travel at high speed (ca. 7.5 m s^{-1}) which they can maintain for a long time. Table 3.3 compares the two extreme cases of the common dolphin (*Delphinus delphis*) and the blue whale *(Balaenoptera musculus*).

Table 3.3 Body characteristics of two cetaceans.

	Common dolphin	Blue whale
Maximum length (m)	2.4	30
Maximum mass (10^3 kg)	0.14	150
Mass/length^3	0.01	0.006
Area/length^2	0.45	0.40

Since these two animals can maintain a cruising speed of ca. 7.5 m s^{-1} for long periods, one may assume that they are then in a physiological steady state. The question is: how is it possible for two species with such different sizes to cruise at the same speed?

To answer this question, one must first consider the drag (F) on a streamlined body moving in a fluid. The equation is similar to eq. 3.8, except that the drag coefficient C_x is replaced here by the *friction coefficient* C_f:

$$F = 0.5\, C_f \rho V^2 A$$

where ρ is the *density* of the fluid, V the *velocity* of the body, and A its *total surface area*. For *laminar* flow, $C_f \approx 1.33\, Re^{-1/2}$ whereas, for *turbulent* flow, $C_f \approx 0.455\, (\log_{10} Re)^{-2.58}$, Re being the *Reynolds number*. Low values of Re correspond to laminar flow, where resistance to motion is relatively weak, whereas high values of Re are associated with turbulent flow, which creates stronger resistance to motion. Normally, for a streamlined body, the flow is laminar over the front portion only and is turbulent towards the back.

The *power* developed by the muscles of moving cetaceans is calculated in three steps.

• Calculation of Re, for the animal under study:

$$Re \approx 7 \times 10^5\ (\text{s m}^{-2})\ VL, \text{ in sea water at } 5°\text{C}$$

• Calculation of drag (F):

$$F = 0.5\, C_f \rho V^2 A$$

C_f being computed from Re, using the equation for either laminar or turbulent flow.

• Calculation of power (P) developed during motion:

$$P = FV$$

For the purpose of the calculation, consider (1) a dolphin with a length of 2 m, weighing 80 kg, whose surface area is 1.75 m^2 and (2) a whale 25 m long, with a mass of 100 t and surface area of 250 m^2.

(1) The value of *Re* for a dolphin moving at 7.5 m s^{-1} is of the order of 10^7, which seems to indicate highly turbulent flow. In the case of *laminar* flow,

$$C_f = 1.33 \times (10^7)^{-1/2} = 4.2 \times 10^{-4}$$

and, for *turbulent* flow,

$$C_f = 0.455\ (\log_{10} 10^7)^{-2.58} = 3 \times 10^{-3}$$

The drag (F) corresponding to these two flow regimes is:

$$F \text{ (laminar)} = 0.5\ (4.2 \times 10^{-4})\ (1028 \text{ kg m}^{-3})\ (7.5 \text{ m s}^{-1})^2\ (1.75 \text{ m}^2) = 22 \text{ N}$$

$$F \text{ (turbulent)} = 0.5\ (3 \times 10^{-3})\ (1028 \text{ kg m}^{-3})\ (7.5 \text{ m s}^{-1})^2\ (1.75 \text{ m}^2) = 155 \text{ N}$$

The *power* ($P = F \times 7.5$ m s^{-1}) that a dolphin should develop, if its motion resulted in perfectly *laminar* flow, would be 165 W and, for *turbulent* flow, 1165 W. Since the size of a dolphin is of the same order as that of a man, it is reasonable to assume that the power it can develop under normal conditions is not higher than that of an athlete, i.e. a *maximum power* of 260 W. It follows that the flow must be laminar for the 9/10 front portion of the dolphin's body, with the rear 1/10 being perhaps turbulent. This conclusion is consistent with observations made in nature on dolphins. It is assumed that the absence of turbulence along the front part of the dolphin's body comes from the fact that the animal only uses its rear section for propulsion.

(2) The blue whale also swims at 7.5 m s^{-1}, its *Re* being ca. 12.5×10^7 which corresponds to a turbulent flow regime. A *laminar* flow would lead to a value

$$C_f = 1.33 \times (12.5 \text{ x } 10^7)^{-1/2} = 1.2 \times 10^{-4}$$

and a *turbulent* flow to

$$C_f = 0.455\ (\log_{10} 12.5 \times 10^7)^{-2.58} = 2.1 \text{ x } 10^{-3}$$

The corresponding drag (F) would be:

$$F \text{ (laminar)} = 0.5\ (1.2 \times 10^{-4})\ (1028 \text{ kg m}^{-3})\ (7.5 \text{ m s}^{-1})^2\ (250 \text{ m}^2) = 745 \text{ N}$$

$$F \text{ (turbulent)} = 0.5\ (2.1 \times 10^{-3})\ (1028 \text{ kg m}^{-3})\ (7.5 \text{ m s}^{-1})^2\ (250 \text{ m}^2) = 13 \text{ kN}.$$

The *power* a whale should develop, if its motion at 7.5 m s^{-1} was accompanied by *laminar* flow, would be 5.6 kW and, in the case of *turbulent* flow, 100 kW. The maximum power developed by a 80 kg dolphin was estimated to be 260 W so that, if the maximum power of an animal was proportional to its mass, a 10^5 kg whale should be able to develop 325 kW. One should, however, take into account the fact that the available energy depends on blood flow. Since cardiac rate is proportional to $(\text{mass})^{-0.27}$, the heart of a whale beats at a rate $(100/0.08)^{-0.27} \approx 1/7$ that of a dolphin. The *maximum power* of a whale is thus ca. 1/7 of 325 kW, i.e. 46.5 kW. This leads to the conclusion that laminar flow takes place along the 2/3 front portion of the animal and that only the 1/3 rear part can sustain turbulent flow.

Ecological application 3.2b

A second study, taken from the same paper as the previous application (Hill, 1950), deals with land animals. It has been observed that several terrestrial mammals run more or less at the same speed and jump approximately the same height, even if their sizes are very different. Table 3.4 gives some approximate maximal values. The question is to explain the similarities observed between the performances of animals with such different sizes.

Table 3.4 Performances (maximal values) of five mammals.

	Running speed (m s^{-1})	Height of jump (m)
Man	12	2
Horse	20	2
Greyhound (25 kg)	18	—
Hare	20	1.5
Deer	15	2.5

One of the explanations proposed by the author involves a relatively simple dimensional argument. The strength of tissues in the bodies of animals cannot be exceeded, during athletic performances, without great risk. For two differently sized animals, consider a pair of systems with lengths l_1 and l_2, respectively, carrying out similar movements within times t_1 and t_2, respectively. The stress at any point in these systems has dimensions $[ML^{-1}T^{-2}]$, which corresponds to the product of density $[ML^{-3}]$ with the square of speed $[L^2T^{-2}]$.

Assuming that the densities of systems are the same for the two species (i.e. $m_1 l_1^{-3} = m_2 l_2^{-3}$, which is reasonable, since the densities of bones, muscles, etc. are similar for all mammals), the stresses at corresponding points of the systems are in the ratio $(l_1^2 t_1^{-2}) : (l_2^2 t_2^{-2})$. If the two systems operate at speeds such that the stresses are the same at corresponding points, it follows that $(l_1 t_1^{-1}) = (l_2 t_2^{-1})$. In other words, the speed is the same at corresponding points of the two systems. It is therefore the strength of their tissues which would explain why athletic animals of very different sizes have the same upper limits for running speeds and jumping heights.

It is interesting to note that, over the years, the topic of maximal running speed of terrestrial mammals has been the subject of many papers, which considered at least four competing theories. These include the theory of geometric similarity, briefly explained in this example, and theories that predict an increase of maximum running speed with body mass. These are summarized in the introduction of a paper by Garland (1983), where maximum running speeds for 106 species of terrestrial mammals are analysed. The study led to several interesting conclusions, including that, even if maximal running speed is mass-independent within some mammalian orders, this is not the case when species from different orders are put together; there is then a tendency for running speed to increase with mass, up to an optimal mass of ca. 120 kg. This is quite paradoxical since, when considering mammals in general, limb bone proportions do scale consistently with geometric similarity. The author refers to Günther's (1975, p. 672) conclusion that "no single similarity criterion can provide a satisfactory quantitative explanation for every single function of an organism that can be submitted to dimensional analysis".

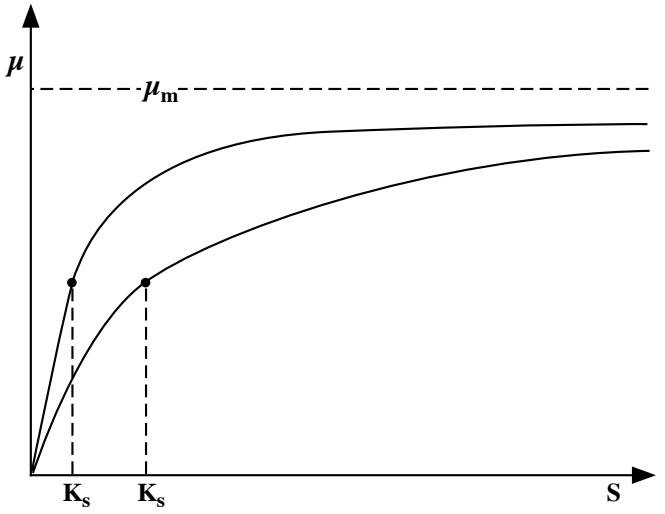


Figure 3.3 Illustration of the Michaelis-Menten equation, showing the role of parameter K_s. In the curve with higher K_s, μ approaches the asymptote μ_m more slowly than in the other curve.

Ecological application 3.2c

An example from aquatic ecology (Platt & Subba Rao, 1973) illustrates the use of dimensionless graphs. The dependence of phytoplankton growth on a given nutrient is often described by means of the Michaelis-Menten equation, borrowed from enzymology. In this equation, the *growth rate* (μ), with dimension $[T^{-1}]$, is a function of the *maximum specific* growth rate (μ_m), the *concentration* (S) of the nutrient, and the *concentration* (K_s) of nutrient at which the growth rate $\mu = 1/2\ \mu_m$:

$$\mu = \frac{1}{B}\frac{dB}{dt} = \frac{\mu_m S}{K_s + S}$$

$$[T^{-1}] = \frac{[1]}{[ML^{-3}]}\frac{[ML^{-3}]}{[T]} = \frac{[T^{-1}]\,[ML^{-3}]}{[ML^{-3}] + [ML^{-3}]}$$

where B is the concentration of phytoplankton *biomass*. This equation is that of a rectangular hyperbola, where K_s determines how fast the asymptote μ_m is approached. When K_s is high, μ approaches the asymptote μ_m slowly, which indicates a weak dependence of μ on S in the unsaturated part of the curve (Fig. 3.3).

In order to compare the effects of two different variables on phytoplankton growth, the authors defined a new entity $S_* = S/K_s$. Since this entity is dimensionless, the abscissa of the

graph $\mu(S_*)$ as a function of S_* is dimensionless; $\mu(S_*)$ stands for the specific growth rate, normalized to S_*. The Michaelis-Menten equation is thus rewritten as:

$$\mu(S_*) = \frac{\mu_m S_*}{(1 + S_*)}$$

Hence, the strength of the dependence of μ on S_* is:

$$\frac{d\mu(S_*)}{dS_*} = \frac{d}{dS_*}\left(\frac{\mu_m S_*}{1 + S_*}\right) = \frac{\mu_m}{(1 + S_*)^2}$$

Using this expression, it is possible to determine the relative strength of the dependence of μ on two different variables (i and j):

$$\xi(i,j) = \frac{d\mu(S^i_*)/dS^i_*}{d\mu(S^j_*)/dS^j_*} = \frac{\mu^i_m}{\mu^j_m}\left[\frac{(1 + S^j_*)^2}{(1 + S^i_*)^2}\right]$$

Under conditions that do not limit phytoplankton growth, the maximum specific growth rate is the same for the two variables, i.e. $\mu^i_m = \mu^j_m$. In such a case, the dependence of μ on the two variables becomes:

$$\xi(i,j) = (1 + S^j_*)^2 / (1 + S^i_*)^2$$

This dimensionless approach makes it possible to compare the effects of different variables on phytoplankton growth, regardless of the dimensions of these variables. Using the above equation, one could assess, for example, the relative importance of irradiance (µmol photons $m^{-2}s^{-1}$, also denoted µEinstein $m^{-2}s^{-1}$) [$NL^{-2}T^{-1}$] and of a nutrient [ML^{-3}] for phytoplankton growth.

Characteristic value

The method described here is actually of general interest in ecology, since it shows how to approach a problem involving several variables with no common measure. In all cases, it is recommended to transform the *dimensional* variables into *dimensionless ones*. The most obvious transformation, proposed by Platt & Subba Rao (1973), consists in dividing each variable by a *characteristic value*, which has the same dimensions as the variable itself. In the case of the Michaelis-Menten equation, the characteristic value is K_s, which has the same dimensions as S. This elegant and efficient approach is also used in parametric statistics, where variables are transformed through division by their standard deviations. For this and other transformations, see Section 1.5. The approach which consists in dividing an ecologically interesting variable by another variable with the same dimensions, so as to create a dimensionless variable, is known as "scaling" (e.g. in Schneider, 1994). Scaling analysis has been used, for example, in coral reef studies (Hatcher and Firth, 1985; Hatcher *et al*., 1987) and by Murray & Jumars (2002) to model steady-state diffusive uptake of nutrients by a spherical attached bacterium (study summarized by Legendre, 2004: 81-83).

The following example illustrates some basic characteristics of dimensional analysis. It also stresses a major weakness of the method, of which ecologists should be aware.

Ecological application 3.2d

The study discussed here (Kierstead & Slobodkin, 1953) did not use dimensional analysis, but it provides material to which the method may usefully be applied. The authors did develop their theory for phytoplankton, but it is general enough to be used with several other types of organisms. Given a water mass containing a growing population, which loses individuals (e.g. phytoplankton cells) by diffusion and regenerates itself by multiplication, the problem is to define the minimum size of the water mass below which the growth of the population is no longer possible.

The problem is simplified by assuming that: (1) the *diffusion* (D) of organisms remains constant within the water mass, but is very large outside where the population cannot maintain itself, and (2) the water mass is *one-dimensional* (long and narrow), so that the *concentration* (c) of organisms is a function of the *position* (x) along the axis of the water mass. The equation describing the growth of the population is thus:

$$\frac{\partial c}{\partial t} = D\frac{\partial^2 c}{\partial x^2} + Kc$$

where K is the growth rate. On the right-hand side of the equation, the first term accounts for diffusion, while the second represents linear growth. A complicated algebraic solution led the authors to define a critical *length* (L_c) for the water mass, under which the population would decrease and above which it could increase:

$$L_c = \pi\sqrt{D/K}$$

It must be noted that this equation is analogous to that of the critical mass in a nuclear reactor. Associated with this critical length is a *characteristic time* (t) of the process, after which the critical length L_c becomes operative:

$$t = L_c^2/(8\pi^2 D)$$

The above results are those given in the paper of Kierstead and Slobodkin. The same problem is now approached by means of dimensional analysis, which will allow one to compare the *dimensional solution* of Platt (1981) to the algebraic solution of Kierstead and Slobodkin. In order to approach the question from a dimensional point of view, the dimensions of variables in the problem must first be specified:

x:	[L]	K:	$[T^{-1}]$
t:	[T]	D:	$[L^2T^{-1}]$

The only dimensions that are not immediately evident are those of D, but these can easily be found using the principle of dimensional homogeneity of theoretical equations.

The equation of Kierstead & Slobodkin involves three variables (c, t, x) and two constants (D, K). According to the general method developed in the previous ecological application, the variables are first *transformed* to dimensionless forms, through division by suitable *characteristic values*. *Dimensionless variables* C, T and X are defined using *characteristic values* c_*, t_* and x_*:

$$C = c/c_* \qquad T = t/t_* \qquad X = x/x_*$$

hence

$$c = C\,c_* \qquad t = T\,t_* \qquad x = X\,x_*$$

Substitution of these values in the equation gives:

$$\frac{c_* \partial C}{t_* \partial T} = D \frac{c_* \partial^2 C}{x_*^2 \partial X^2} + K c_* C$$

The next step is to make all terms in the equation dimensionless, by multiplying each one by x_*^2 and dividing it by D, after eliminating from all terms the common constant c_* :

$$\left[\frac{x_*^2}{D t_*}\right] \frac{\partial C}{\partial T} = \frac{\partial^2 C}{\partial X^2} + \left[\frac{K x_*^2}{D}\right] C$$

The resulting equation thus contains three *dimensionless variables* (C, T and X) and two *dimensionless products* (in brackets).

Since the dimensions of the two products are [1], these may be transformed to isolate the characteristic values x_* and t_* :

$$\text{since } \left[\frac{x_*^2}{D t_*}\right] = [1], \text{ it follows that } [\, t_* \,] = \left[\frac{x_*^2}{D}\right]$$

$$\text{since } \left[\frac{K x_*^2}{D}\right] = [1], \text{ it follows that } [\, x_*^2 \,] = \left[\frac{D}{K}\right] \text{ and thus } [\, x_* \,] = \left[\frac{D}{K}\right]^{1/2}$$

Using these relationships, the following proportionalities are obtained:

$$x_* \propto \sqrt{D/K} \text{ and } t_* \propto x_*^2 / D$$

Dimensional analysis thus easily led to the same results as those obtained by Kierstead and Slobodkin (1953), reported above, except for the constant factors π and $8\pi^2$. This same example will be reconsidered in the next section (Ecological application 3.3a), where the two dimensionless products will be calculated directly.

The above example illustrates the fact that *dimensional analysis cannot generate dimensionless constants*, which is a limit of the method that must be kept in mind. Thus, in order to take advantage of the power of dimensional analysis, one must give up some precision. It is obvious that such a simple method as dimensional analysis cannot produce the same detailed results as complex algebraic developments. As mentioned above (Section 3.0), dimensional analysis deals with *general forms* of equations. Yet, starting from simple concepts, one can progress quite far into complex problems, but the final solution is only partial. As noted by Langhaar (1951): "The generality of the method is both its strength and its weakness. With little effort, a partial solution to nearly any problem is obtained. On the other hand, a complete solution is not obtained."

Ecological application 3.2e

It often happens that ecologists must synthesize published data on a given subject, either as a starting point for new research, or to resolve a problem using existing knowledge, or else as a

basis for a new theoretical perspective. This is nowadays more necessary than ever, because of the explosion of ecological information. However, such syntheses are confronted to a real difficulty, which is the fact that available data are often very diversified, and must thus be unified before being used. Paloheimo & Dickie (1965) met this problem when they synthesized the mass of information available in the literature on the growth of fish as a function of food intake. As in the previous application, the authors did not themselves use dimensional analysis in their work. The dimensional solution discussed here is modified from Platt (1981).

The metabolism of fish may be described using the following relationship:

$$T = \alpha W^{\gamma}$$

where T is the rate of *oxygen consumption*, α specifies the level of *metabolic expenditure* per unit time, W is the *mass* of the fish, and γ specifies the rate of *change of metabolism* with body mass. Growth is expressed as a function of food ration (R), by means of the following equation:

$$\frac{dW}{dt} = R\left[e^{-(a+bR)}\right]$$

which shows that growth efficiency decreases by a constant fraction e^{-b} for each unit increase in the amount of food consumed per unit time. The value of R at maximum growth is determined, as usual, by setting the partial derivative equal to 0:

$$\frac{\partial}{\partial R}\left(\frac{dW}{dt}\right) = (1-bR)\,e^{-(a+bR)} = 0$$

Growth is thus maximum when $bR = 1$.

The basic equation for the energy budget (eq. 3.1) is:

$$\frac{dW}{dt} = R - T$$

so that

$$T = R - \frac{dW}{dt}$$

Replacing, in this last equation, dW/dt by its expression in the second equation, above, and isolating R, one obtains:

$$T = R\left[1 - e^{-(a+br)}\right]$$

Then, replacing T by its expression in the first equation leads to:

$$\alpha W^{\gamma} = R\left[1 - e^{-(a+br)}\right]$$

which is a general equation for energy budgets. This equation may be used to calculate, for any fish of mass W, the ration R required to maintain a given metabolic level. Furthermore, with an increase in ration, the term $[1 - e^{-(a+bR)}]$ tends towards 1, which indicates that the metabolism then approaches R. In other words, growth decreases at high values of R.

Values for coefficient b and food intake found in the literature are quite variable. It was shown above that the product bR determines growth. Paloheimo & Dickie therefore suggested to standardize the relationship between growth and ration in terms of bR.

Since growth is maximum when $bR = 1$, the *ration* can be brought to a common measure by expressing it in units of $1/b$. On this new *scale*, the ration (r) is defined as:

$$r = bR$$

When growth is maximum, $bR = 1$, so that $R = 1/b$. Replacing, in the general equation for the energy budget, R by $1/b$ (and bR by 1) yields:

$$\alpha W^{\gamma} = 1/b\,[1 - e^{-(a+1)}]$$

so that

$$W = \left[\frac{1 - e^{-(a+1)}}{\alpha b}\right]^{1/\gamma}$$

from which it is concluded that the *mass* should be expressed in units of $(1/\alpha b)^{1/\gamma}$ in order to bring data from the literature to a common measure. On this new *scale*, the mass (w) is defined as:

$$w = (\alpha b)^{1/\gamma}\, W$$

so that

$$\frac{w^{\gamma}}{b} = \alpha W^{\gamma} = T$$

Using the scaled ration (r) and mass (w), the general equation for energy budgets may be rewritten as:

$$\frac{w^{\gamma}}{b} = \frac{r}{b}\,[1 - e^{-(a+r)}]$$

and finally

$$w^{\gamma} = r\,[1 - e^{-(a+r)}]$$

In this last equation, the use of r and w brings to a common measure the highly variable values of R and W, which are available in the literature for different species or for different groups within a given fish species.

These same results could have been obtained much more easily using dimensional analysis. As with all problems of the kind, it is essential, first of all, to identify the dimensions of variables involved in the problem. The first two equations are used to identify the dimensions of all variables in the study:

$$T = \alpha W^{\gamma}$$

$$[\mathrm{MT}^{-1}] = [\mathrm{M}^{(1-\gamma)}\mathrm{T}^{-1}]\,[\mathrm{M}^{\gamma}]$$

$$\frac{dW}{dt} = R\,[e^{-(a+bR)}]$$

$$[\mathrm{MT}^{-1}] = [\mathrm{MT}^{-1}]\,[1]^{\,[1]\,+\,[\mathrm{M}^{-1}\mathrm{T}]\,[\mathrm{MT}^{-1}]}$$

The dimensions of α, which were not immediately obvious, are determined using the principle of dimensional homogeneity (i.e. same dimensions on the two sides of the equation). The dimensions of a and b are also found by applying the principle of dimensional homogeneity, taking into account the fact that an exponent is by definition dimensionless.

The problem is then to define *characteristic values* (or, more appropriately, *scale factors*) so as to obtain dimensionless ration (r), mass (w), and time (τ). Obviously, these scale factors must contain the two dimensional parameters of the above equations, α and b.

Because the product bR is dimensionless, the scale factor r for ration is:

$$r = bR$$

The cases of w and τ require the calculation of unknown exponents. These are easily found by dimensional analysis. In order to do so, unknown exponents y and z are assigned to α and b, and these unknowns are solved using the principle of dimensional homogeneity:

Calculation of w:

$$[w] = [1] = [\alpha]^y\,[b]^z\,[W]$$

$$[W]^{-1} = [\alpha]^y\,[b]^z$$

$$[M^{-1}T^0] = [M^{(1-\gamma)}\,T^{-1}]^y\,[M^{-1}T]^z = [M^{y(1-\gamma)-z}\,T^{-y+z}]$$

so that $y(1-\gamma) - z = -1$

and $-y + z = 0$

hence $y = 1/\gamma = z$

Consequently, the scale factor w for the mass is:

$$w = (\alpha b)^{1/\gamma} W$$

Calculation of τ:

$$[\tau] = [1] = [\alpha]^y\,[b]^z\,[t]$$

$$[t]^{-1} = [\alpha]^y\,[b]^z$$

$$[M^0T^{-1}] = [M^{y(1-\gamma)-z}\,T^{-y+z}]$$

so that $y(1-\gamma) - z = 0$

and $-y + z = -1$

hence $y = 1/\gamma$ and $z = 1/\gamma - 1$

It follows that the scale factor τ for time is:

$$\tau = \alpha^{1/\gamma}\, b^{(1/\gamma - 1)} t$$

$$\tau = [(\alpha b)^{1/\gamma}/b]t$$

These scale factors can be used to compare highly diversified data. *Ration* is then expressed in units of $(1/b)$, *mass* in units of $(\alpha b)^{-1/\gamma}$, and *time* in units of $b/(\alpha b)^{-1/\gamma}$. With this approach, it is possible to conduct generalized studies on the food intake and growth of fish as a function of time.

Other applications of dimensionless products in ecology are found, for example, in Tranter & Smith (1968), Rubenstein & Koehl (1977), and Okubo (1987). The first application analyses the performance of plankton nets, the second explores the mechanisms of filter feeding by aquatic organisms, and the third examines various aspects of biofluid mechanics, including a general relationship between the Reynolds number (*Re*) and the sizes and swimming speeds of aquatic organisms from bacteria to whales. Platt (1981) provides other examples of application of dimensional analysis in

the field of biological oceanography. Legendre (2004, pp. 87-91) explains how the dimensional approach provided the main guideline to derive operational equations from a conceptual model on the fate of biogenic carbon in oceans. These equations were used by Beaugrand *et al*. (2010) to compute the effects of long-term changes in copepod biodiversity on carbon flows in the extratropical North Atlantic Ocean.

Ecological applications 3.2d and 3.2e showed that dimensional analysis may be a powerful tool in ecology. They do, however, leave potential users somewhat uncertain as to how personally apply this approach to new problems. The next section outlines a general method for solving problems of dimensional analysis, which will lead to more straightforward use of the method. It will be shown that it is not even necessary to know the basic *equations* pertaining to a problem, provided that all the pertinent *variables* are identified. The above last two examples will then be reconsidered as applications of the *systematic calculation of dimensionless products*.

3.3 The complete set of dimensionless products

As shown in the previous section, the resolution of problems using dimensional analysis involves two distinct steps: (1) the identification of variables pertinent to the phenomenon under study — these are derived from fundamental principles, for example of ecological nature — and (2) the computation of a complete set of dimensionless products. When the number of variables involved is small, complete sets of dimensionless products can be formed quite easily, as seen above. However, as the number of variables increases, this soon becomes unwieldy, so that one must proceed to a *systematic calculation of the complete set of dimensionless products*.

The physical example of the *drag on smooth spheres* (Section 3.2) will first be used to illustrate the principles of the calculation. The problem involved five variables (F, V, L, ρ, and η; see eq. 3.2), whose dimensions are written here in a *dimensional matrix*:

$$\begin{array}{c} \\ \mathrm{M} \\ \mathrm{L} \\ \mathrm{T} \end{array} \begin{array}{c} \begin{array}{ccccc} F & \eta & \rho & L & V \end{array} \\ \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & -1 & -3 & 1 & 1 \\ -2 & -1 & 0 & 0 & -1 \end{bmatrix} \end{array} \qquad \textbf{(3.9)}$$

It must be kept in mind that the numbers in matrix 3.9 (i.e. dimensions) are *exponents*. The *dimensionless products* being sought are *products* of *powers* of variables in the matrix (columns). In each product, the exponents given to the variables must be such that the result is *dimensionless*.

In other words, the systematic calculation of dimensionless products consists in finding exponents x_1, x_2, x_3, x_4 and x_5 for variables F, η, ρ, L, and V, such that a product Π, of the general form

$$\Pi = F^{x_1}\eta^{x_2}\rho^{x_3}L^{x_4}V^{x_5}$$

be dimensionless. Taking into account the respective dimensions of the five variables, the general dimensions of Π are:

$$\Pi = [\mathrm{MLT^{-2}}]^{x_1}[\mathrm{ML^{-1}T^{-1}}]^{x_2}[\mathrm{ML^{-3}}]^{x_3}[\mathrm{L}]^{x_4}[\mathrm{LT^{-1}}]^{x_5}$$

$$\Pi = [\mathrm{M}^{(x_1+x_2+x_3)}\mathrm{L}^{(x_1-x_2-3x_3+x_4+x_5)}\mathrm{T}^{(-2x_1-x_2-x_5)}]$$

The exponents of dimensions [M], [L], and [T] carry exactly the same information as the dimensional matrix (eq. 3.9). These exponents could therefore have been written directly, using matrix notation:

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & -1 & -3 & 1 & 1 \\ -2 & -1 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} \qquad \textbf{(3.10)}$$

where the dimensional matrix is on the left-hand side.

Since the products Π are dimensionless, the exponent of each dimension [M], [L], and [T], respectively, must be *zero*. In follows that:

$$x_1 + x_2 + x_3 = 0$$

$$x_1 - x_2 - 3x_3 + x_4 + x_5 = 0$$

$$-2x_1 - x_2 - x_5 = 0$$

or, in matrix notation:

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & -1 & -3 & 1 & 1 \\ -2 & -1 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \mathbf{0} \qquad \textbf{(3.11)}$$

Calculation of dimensionless products Π is thus achieved by simultaneously solving three equations. However, the above system of equations is *indeterminate*, since there are only three equations for five unknowns. Arbitrary values must thus be assigned to two of the unknowns, for example x_1 and x_2. The general solution is then given in terms of x_1 and x_2. The steps are as follows:

(1) Matrix equation 3.11 is rewritten so as to isolate x_1 and x_2 together with the associated first two columns of the matrix. This operation simply involves transferring all terms in x_3, x_4 and x_5 to the right-hand side of the equation:

$$\begin{bmatrix} 1 & 1 \\ 1 & -1 \\ -2 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = - \begin{bmatrix} 1 & 0 & 0 \\ -3 & 1 & 1 \\ 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} x_3 \\ x_4 \\ x_5 \end{bmatrix} \tag{3.12}$$

Note that there is now a negative sign in front of the matrix on the right-hand side. Matrix eq. 3.12 is identical to the algebraic form:

$$x_1 + x_2 = -x_3$$

$$x_1 - x_2 = 3x_3 - x_4 - x_5$$

$$-2x_1 - x_2 = x_5$$

(2) One then solves for the unknowns x_3, x_4 and x_5, using the general method of matrix inversion (Section 2.8):

$$-\begin{bmatrix} 1 & 0 & 0 \\ -3 & 1 & 1 \\ 0 & 0 & -1 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 1 \\ 1 & -1 \\ -2 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_3 \\ x_4 \\ x_5 \end{bmatrix}$$

$$-\begin{bmatrix} 1 & 0 & 0 \\ 3 & 1 & 1 \\ 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & -1 \\ -2 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_3 \\ x_4 \\ x_5 \end{bmatrix}$$

$$\begin{bmatrix} -1 & -1 \\ -2 & -1 \\ -2 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_3 \\ x_4 \\ x_5 \end{bmatrix} \tag{3.13}$$

(3) The simplest approach consists in successively assigning the value 1 to each unknown while setting the other equal to 0, i.e. (1) $x_1 = 1$ and $x_2 = 0$ and (2) $x_1 = 0$ and $x_2 = 1$. It follows that the first two columns of the solution matrix are a *unit matrix*:

$$\begin{array}{c} F \;\; \eta \;\; \rho \;\; L \;\; V \\ x_1 \;\; x_2 \;\; x_3 \;\; x_4 \;\; x_5 \\ \begin{matrix} \Pi_1 \\ \Pi_2 \end{matrix} \begin{bmatrix} 1 & 0 & -1 & -2 & -2 \\ 0 & 1 & -1 & -1 & -1 \end{bmatrix} \end{array} \qquad \textbf{(3.14)}$$

The dimensionless products of the *complete set* are therefore (as in Section 3.2):

$$\Pi_1 = \frac{F}{\rho L^2 V^2}$$, the *Newton number* (*Ne*; eq. 3.4)

$$\Pi_2 = \frac{\eta}{\rho L V}$$, the inverse of the *Reynolds number* ($1/Re$; eq. 3.3)

This example clearly shows that the systematic calculation of dimensionless products rests *solely* on recognizing the *variables* involved in the problem under consideration, without necessarily knowing the corresponding *equations*. The above solution, which was developed using a simple example, can be applied to all problems of dimensional analysis, since it has the following characteristics:

(1) Because the left-hand part of the solution matrix is an *identity matrix* (**I**), the dimensionless products Π are *independent* of one another. Indeed, given **I**, each product contains one variable which is not included in any other product, i.e. the first variable is only in Π_1, the second is only in Π_2, and so on.

(2) When partitioning the dimensional matrix, one must isolate *on the right-hand side* a matrix that can be *inverted*, i.e. a matrix whose determinant is non-zero.

(3) The *rank* (r) of the dimensional matrix is the order of the largest non-zero determinant it contains (Section 2.7). Therefore, it is always possible to isolate, *on the right-hand side*, a matrix of order r whose determinant is non-zero. The order r may however be lower than the number of rows in the dimensional matrix, as seen later.

(4) The *number* of *dimensionless products* in the *complete set* is equal to the number of variables isolated *on the left-hand side* of the dimensional matrix. It follows from item (3) that the number of dimensionless products is equal to the *total number of variables* minus the *rank of the dimensional matrix*. In the preceding example, the number of dimensionless products in the complete set was equal to the number of variables (5) minus the rank of the dimensional matrix (3), i.e. $5 - 3 = 2$ dimensionless products.

(5) When the last r columns of a dimensional matrix of order r do not lead to a non-zero determinant, the columns of the matrix must be rearranged so as to obtain a non-zero determinant.

Numerical example 1. An example will help understand the consequences of the above five characteristics on the general method for the systematic calculation of the complete set of dimensionless products. The dimensional matrix is as follows:

$$\begin{array}{c} \\ \mathrm{M} \\ \mathrm{L} \\ \mathrm{T} \end{array} \begin{array}{c} \begin{array}{ccccccc} V_1 & V_2 & V_3 & V_4 & V_5 & V_6 & V_7 \end{array} \\ \begin{bmatrix} 2 & 0 & 1 & 0 & -1 & -2 & 3 \\ 1 & 2 & 2 & 0 & 0 & 1 & -1 \\ 0 & 1 & 2 & 3 & 1 & -1 & 0 \end{bmatrix} \end{array}$$

The rank (r) of this matrix is 3 (numerical example in Section 2.7), so that the number of dimensionless products of the complete set is equal to $7 - 3 = 4$. However, the determinant of the $r = 3$ last columns is zero:

$$\begin{vmatrix} -1 & -2 & 3 \\ 0 & 1 & -1 \\ 1 & -1 & 0 \end{vmatrix} = 0$$

Calculating the complete set of dimensionless products thus requires a reorganization of the dimensional matrix by rearranging, for example, the columns as follows:

$$\begin{array}{c} \\ \mathrm{M} \\ \mathrm{L} \\ \mathrm{T} \end{array} \begin{array}{c} \begin{array}{ccccccc} V_1 & V_5 & V_7 & V_4 & V_2 & V_6 & V_3 \end{array} \\ \begin{bmatrix} 2 & -1 & 3 & 0 & 0 & -2 & 1 \\ 1 & 0 & -1 & 0 & 2 & 1 & 2 \\ 0 & 1 & 0 & 3 & 1 & -1 & 2 \end{bmatrix} \end{array}$$

The solution then follows from the general method described above:

$$\begin{bmatrix} x_2 \\ x_6 \\ x_3 \end{bmatrix} = -\begin{bmatrix} 0 & -2 & 1 \\ 2 & 1 & 2 \\ 1 & -1 & 2 \end{bmatrix}^{-1} \begin{bmatrix} 2 & -1 & 3 & 0 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_5 \\ x_7 \\ x_4 \end{bmatrix}$$

$$\begin{bmatrix} x_2 \\ x_6 \\ x_3 \end{bmatrix} = -\begin{bmatrix} 4 & 3 & -5 \\ -2 & -1 & 2 \\ -3 & -2 & 4 \end{bmatrix} \begin{bmatrix} 2 & -1 & 3 & 0 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_5 \\ x_7 \\ x_4 \end{bmatrix}$$

$$\begin{bmatrix} x_2 \\ x_6 \\ x_3 \end{bmatrix} = \begin{bmatrix} -11 & 9 & -9 & 15 \\ 5 & -4 & 5 & -6 \\ 8 & -7 & 7 & -12 \end{bmatrix} \begin{bmatrix} x_1 \\ x_5 \\ x_7 \\ x_4 \end{bmatrix}$$

$$\begin{array}{c} \\ \Pi_1 \\ \Pi_2 \\ \Pi_3 \\ \Pi_4 \end{array} \begin{array}{c} \begin{array}{ccccccc} V_1 & V_5 & V_7 & V_4 & V_2 & V_6 & V_3 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & -11 & 5 & 8 \\ 0 & 1 & 0 & 0 & 9 & -4 & -7 \\ 0 & 0 & 1 & 0 & -9 & 5 & 7 \\ 0 & 0 & 0 & 1 & 15 & -6 & -12 \end{bmatrix} \end{array}$$

Numerical example 2. This example illustrates the case of a dimensional matrix whose *rank* is less than its number of rows. This matrix has already been considered in Section 2.7:

$$\begin{array}{c} \\ \mathrm{M} \\ \mathrm{L} \\ \mathrm{T} \end{array} \begin{array}{c} \begin{array}{cccc} V_1 & V_2 & V_3 & V_4 \end{array} \\ \begin{bmatrix} 2 & 1 & 3 & 4 \\ -1 & 6 & -3 & 0 \\ 1 & 20 & -3 & 8 \end{bmatrix} \end{array}$$

It was shown (Section 2.7) that the *rank* of this matrix is $r = 2$, so that it is not possible to find a combination of three columns that could be inverted. Any 3×3 submatrix would be *singular* (Section 2.8).

The solution consists in making the *number of rows* equal to the *rank*. This is done by eliminating any one row of the dimensional matrix, since the matrix has only two independent rows (Section 2.7). The number of dimensionless products in the complete set is thus equal to $4 - 2 = 2$.

$$\begin{array}{c} \\ \mathrm{M} \\ \mathrm{L} \end{array} \begin{array}{c} \begin{array}{cccc} V_1 & V_2 & V_3 & V_4 \end{array} \\ \begin{bmatrix} 2 & 1 & 3 & 4 \\ -1 & 6 & -3 & 0 \end{bmatrix} \end{array}$$

$$\begin{array}{c} \\ \Pi_1 \\ \Pi_2 \end{array} \begin{array}{c} \\ \begin{bmatrix} 1 & 0 & -1/3 & -1/4 \\ 0 & 1 & 2 & -7/4 \end{bmatrix} \end{array}$$

It is possible to eliminate fractional exponents by multiplying each row of the solution matrix by its lowest common denominator:

$$\begin{array}{c} \\ \Pi_1 \\ \Pi_2 \end{array} \begin{array}{c} \\ \begin{bmatrix} 12 & 0 & -4 & -3 \\ 0 & 4 & 8 & -7 \end{bmatrix} \end{array}$$

Identical results would have been obtained if any other row of the dimensional matrix had been eliminated instead of row 3, since each of the three rows is a linear combination of the other two. This can easily be checked as exercise.

There now remains to discuss how to choose the ordering of variables in a dimensional matrix. This order determines the complete set of dimensionless products obtained from the calculation. The rules are as follows:

(1) The *dependent variable* is, of necessity, in the first column of the dimensional matrix, since it must be present in only one Π (the first dimensionless product is thus

called the *dependent dimensionless variable*). As a consequence, this first variable can be expressed as a function of all the others, which is the goal here. For example, in eq. 3.9, the *drag F* is in the first column of the dimensional matrix since it is clearly the *dependent variable*.

(2) The other variables are then arranged in decreasing order, based on their potential for experimental variation. Indeed, a maximum amount of information will result from experimentation if those variables with a wide range of experimental variability occur in a single Π.

(3) The initial ordering of variables must obviously be changed when the last *r columns* of the dimensional matrix have a zero determinant. However, one must then still comply as well as possible with the first two rules.

Two ecological applications, already discussed in Section 3.2, will now be treated using the systematic calculation of complete sets of dimensionless products.

Ecological application 3.3a

The first example reconsiders Ecological application 3.2d, devoted to the model of Kierstead & Slobodkin (1953). This model provided equations for the *critical size* of a growing phytoplankton patch and the *characteristic time* after which this critical size becomes operative.

The dimensional matrix of variables involved in the problem includes: *length x*, *time t*, *diffusion of cells D*, and *growth rate k*. The dependent variables being *x* and *t*, they are in the first two columns of the dimensional matrix:

$$\begin{array}{c} \\ \mathrm{L} \\ \mathrm{T} \end{array} \begin{array}{c} \begin{array}{cccc} x & t & D & k \end{array} \\ \begin{bmatrix} 1 & 0 & 2 & 0 \\ 0 & 1 & -1 & -1 \end{bmatrix} \end{array}$$

The rank of the dimensional matrix being 2, the number of dimensionless products is 4 – 2 = 2. These two products are found using the general method for calculating the complete set:

$$-\begin{bmatrix} 2 & 0 \\ -1 & -1 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} -1/2 & 0 \\ 1/2 & 1 \end{bmatrix}$$

$$\begin{array}{c} \begin{array}{cccc} x & t & D & k \end{array} \\ \begin{array}{c} \Pi_1 \\ \Pi_2 \end{array} \begin{bmatrix} 1 & 0 & -1/2 & 1/2 \\ 0 & 1 & 0 & 1 \end{bmatrix} \end{array} = \begin{array}{c} \begin{array}{cccc} x & t & D & k \end{array} \\ \begin{bmatrix} 2 & 0 & -1 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix} \end{array}$$

$$\Pi_1 = kx^2/D \text{ and } \Pi_2 = tk$$

These two dimensionless products describe, as in Ecological application 3.2d, the *critical length x* and the *characteristic time t* as:

$$x \propto \sqrt{D/k} \text{ and } t \propto 1/k \propto x^2/D$$

Ecological application 3.3b

A second example provides an easy solution to the problem that confronted Paloheimo & Dickie (1965) concerning the synthesis of data on the growth of fish with respect to food intake. The question was discussed at length in Ecological application 3.2e, which led to three scale factors, for *food ration, mass,* and *time*. These scale factors were used by the authors to compare heterogeneous data from the ecological literature.

The solution is found directly, here, using the dimensional matrix of the six variables involved in the problem: *time t, mass W,* food *ration R,* rate of *oxygen consumption T,* rate of *metabolic expenditure* α, and coefficient b. The variables to be isolated being t, W, and R, they are in the first three columns of the dimensional matrix:

$$\begin{array}{c} \\ \mathrm{M} \\ \mathrm{T} \end{array} \begin{array}{c} \begin{array}{cccccc} t & W & R & T & \alpha & b \end{array} \\ \begin{bmatrix} 0 & 1 & 1 & 1 & (1-\gamma) & -1 \\ 1 & 0 & -1 & -1 & -1 & 1 \end{bmatrix} \end{array}$$

Since the *rank* of the dimensional matrix is $r = 2$, the number of dimensionless products is $6 - 2 = 4$. The four products are calculated by the method of the complete set:

$$-\begin{bmatrix} (1-\gamma) & -1 \\ -1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & -1 & -1 \end{bmatrix} = \begin{bmatrix} 1/\gamma & 1/\gamma & 0 & 0 \\ [(1/\gamma)-1] & 1/\gamma & 1 & 1 \end{bmatrix}$$

$$\begin{array}{c} \\ \Pi_1 \\ \Pi_2 \\ \Pi_3 \\ \Pi_4 \end{array} \begin{array}{c} \begin{array}{cccccc} t & W & R & T & \alpha & b \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 1/\gamma & (1/\gamma)-1 \\ 0 & 1 & 0 & 0 & 1/\gamma & 1/\gamma \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \end{bmatrix} \end{array}$$

$$\Pi_1 = t\alpha^{1/\gamma} b^{(1/\gamma - 1)} = [(\alpha b)^{1/\gamma}/b]t$$

$$\Pi_2 = W\alpha^{1/\gamma} b^{1/\gamma} = (\alpha b)^{1/\gamma} W$$

$$\Pi_3 = Rb = bR$$

$$\Pi_4 = Tb = bT$$

The first three dimensionless products define the three scale factors already found in Ecological application 3.2e, i.e. II_1 for *time,* II_2 for *mass*, and II_3 for *ration*. II_4 defines a scale factor for *oxygen consumption*.

Direct calculations of complete sets of dimensionless products thus led to the same results as obtained before, but operations here were more straightforward than in Section 3.2.

It should not be necessary, after these examples, to dwell on the advantage of systematically calculating the complete set of dimensionless products. In addition to providing a rapid and elegant solution to problems of dimensional analysis, the above matrix method sets researchers on the right track when tackling a problem to be investigated using the dimensional tool. The success of a dimensional study depends

on: (1) adequate knowledge of the problem under study, so that *all* the pertinent variables are considered; and (2) clear ideas about which variables are functions of the others. It should be noted, as explained above, that the systematic calculation of the complete set of dimensionless products does not require prior knowledge of the fundamental equations. These, however, may be necessary to derive the dimensions of some complex variables. Dimensional analysis may be a powerful tool, provided that the ecological bases of the problem under consideration are thoroughly understood and that the objectives of the research are clearly stated.

3.4 Scale factors and models

Given the increased awareness in society for environmental problems, major engineering projects cannot be undertaken, in most countries, before their environmental impacts have been assessed. As a consequence, an increasing number of ecologists now work within multidisciplinary teams of consultants. At the planning stage, a powerful tool available to engineers, although very costly, is the small-scale *model*. *Tests* performed with such models help choose the most appropriate engineering solution. Actually, ecologists may encounter two types of model, i.e. mathematical and physical. *Mathematical models* are defined at the beginning of Section 10.3. *Physical models* are small-scale replica of the natural environment, to which changes can be made that reproduce those planned for the real situation. Tests with physical models (e.g. in wind tunnels or hydraulic flumes) are generally more costly to perform than mathematical simulations, so that the latter are becoming increasingly more popular than the former. *Physical models* are often based on dimensional analysis, so that it is this type of model that is considered here. It should be noted that physical models may originate from the empirical approach of engineers, which is distinct from the dimensional approach.

Physical model

In order to communicate with engineers conducting tests on small-scale models, ecologists must have some basic understanding of the principles governing model testing. In some cases, ecologists may even play a role in the study, when it is possible to integrate in the model variables of ecological significance (e.g. in a model of a harbour or estuary, such variables as salinity, sediment transport, etc.). Since small-scale models are based in part on dimensional analysis, their basic theory is thus relatively easy to understand. The actual testing, however, requires the specific knowledge and experience of model engineers. In addition to their possible involvement in applications of modelling to environmental impact studies, ecologists may at times use small-scale models to resolve problems of their own (e.g. studying the interactions between benthic organisms and sediment in a hydraulic flume). These various aspects are introduced here very briefly.

Prototype

In the vocabulary of physical modelling, the full-size system is called *prototype* and the small-size replica is called *model*. A model may be geometrically similar to the

Geometric similarity

prototype, or it may be distorted. In the case of *geometric similarity,* all parts of the model have the same shapes as the corresponding parts of the prototype. In certain cases, geometric similarity would lead to errors, so that one must use a *distorted model*. In such models, one or several scales may be distorted. For example, a geometrically similar model of an estuary could result in some excessively small water depths. With such depths, the flow in the model could become subject to surface tension, which would clearly be incorrect with respect to the real flow. In the model, the depth must therefore be relatively greater than in nature, hence a distorted model.

The physical example of the *drag on smooth spheres*, already discussed in Sections 3.2 and 3.3, is now used to introduce the basic principles of scaling and small-scale modelling. Equation 3.7 describes the *drag* (F) acting on a smooth sphere of *diameter* D, immersed in a stream with *velocity* V of a fluid with density ρ and dynamic viscosity η:

$$F = \rho V^2 D^2 f(Re) \tag{3.7}$$

$$F = \rho V^2 D^2 f\left(\frac{VD\rho}{\eta}\right)$$

In order to experimentally determine the drag, under convenient laboratory conditions (e.g. wind tunnel or hydraulic flume), it may be appropriate to use a geometrically similar model of the sphere. Quantities pertaining to the *model* are assigned *prime indices*. If the curve of the drag coefficient for smooth spheres was not known (Fig. 3.1), the estimation of F in the laboratory would require that the value of the *unknown function f* be the same for both the model and the prototype. In order to do so, the test engineer should make sure that the *Reynolds numbers* for the two systems are equal:

$$Re = Re'$$

$$\frac{VD\rho}{\eta} = \frac{V'D'\rho'}{\eta'} \tag{3.15}$$

Scale factor

A *scale factor* is defined as the ratio of the size of the model to that of the prototype. Scale factors are therefore *dimensionless numbers*. The scale factors (K) corresponding to eq. 3.15 are:

$$K_V = V'/V \qquad K_D = D'/D \qquad K_\rho = \rho'/\rho \qquad K_\eta = \eta'/\eta$$

These scales are used to rewrite eq. 3.15 as:

$$K_V K_D K_\rho = K_\eta \tag{3.16}$$

Because $Re = Re'$, the *scale factor of the unknown function f* is equal to unity:

$$K_{f(Re)} = 1 \qquad \textbf{(3.17)}$$

The ratio between the drag measured for the model and the real drag on the prototype is computed by combining eq. 3.7 with the above scale factors:

$$K_F = K_\rho K_V^2 K_D^2 K_{f(Re)}$$

Because of eq. 3.17, it follows that:

$$K_F = K_\rho K_V^2 K_D^2 \qquad \textbf{(3.18)}$$

Equation 3.16 is used to find the value of K_F:

$$K_V K_D K_\rho = K_\eta \qquad \textbf{(3.16)}$$

is squared $\quad K_V^2 K_D^2 K_\rho^2 = K_\eta^2$

from which $\quad K_V^2 K_D^2 K_\rho = K_\eta^2 / K_\rho$

and, given eq. 3.18 $\quad K_F = K_\eta^2 / K_\rho \qquad \textbf{(3.19)}$

Equation 3.19 leads to the following practical conclusions, for determining the drag on smooth spheres in the laboratory:

(1) If the model is tested using the *same fluid* as for the prototype, the *drag* measured during the test is the same as for the prototype. This follows from the fact that, if $K_\eta = 1$ and $K_\rho = 1$ (same fluid), K_F is equal to unity (eq. 3.19), hence $F' = F$.

(2) If testing is conducted using the *same fluid* as for the prototype, conservation of *Re* requires that the *velocity* for the model be greater than for the prototype (i.e. the model is smaller than the prototype). This follows from the fact that, when $K_\eta = 1$ and $K_\rho = 1$ (same fluid), $K_V K_D = 1$ (eq. 3.16); consequently any decrease in K_D must be compensated by a proportional increase in K_V.

(3) When it is more convenient to use *different fluids*, testing may be conducted while conserving *Re*. It has already been shown (Section 3.2) that, for example, going from a large-size prototype, in air, to a model 6 times smaller, in water, allows a reduction of the flow speed during the test by a factor of 3. The drag measured for the model would not, however, be necessarily the same as that of the prototype, since that force varies as a function of the ratio between the squares of the dynamic viscosities

(K_η^2) and the densities (K_ρ) of the two fluids (eq. 3.19). Knowing this ratio (K_F), it is easy to derive the drag for the model (F) from that measured during the test (F') since:

$$F = F'/K_F$$

In more complex cases, it is sometimes necessary to simultaneously conserve two or more dimensionless products that are incompatible. In such a situation, where a choice must be made between contradictory constraints, it rests on the test engineer to justify discrepancies in similarity and to apply theoretical corrections to compensate for them. Hence modelling, although derived from scientific concepts, becomes an art based on the experience of the researcher.

Similarity

A *general concept of similarity* follows from the previous discussion. In a Cartesian space, the *model* and the *prototype* are described by coordinates (x' y' z') and (x y z), respectively. Correspondence between the two systems is established by means of *scale factors* (K), which define *homologous* times as well as *homologous* points in the three dimensions of space:

$$t' = K_t t \qquad x' = K_x x \qquad y' = K_y y \qquad z' = K_z z$$

The *time scale factor* (K_t) would be used, for example, in the case of a flow where Δ'_t and Δ_t are the time intervals during which two homologous particles go through homologous parts of their respective trajectories. It would then be defined as

$$K_t = \Delta'_t/\Delta_t$$

Geometric similarity is defined as: $K_x = K_y = K_z = K_L$. In *distorted models*, a single length scale is usually modified, so that $K_x = K_y \neq K_z$. The ratio K_z/K_x is the *distortion factor*. It would be possible, using this same approach, to define characteristics of *kinematic similarity,* for similar motions, and of *dynamic similarity*, for systems subjected to homologous forces.

There are several types of similarity in addition to the geometric, dynamic and kinematic similarities. These include the *hydrodynamic*, *transport*, and *thermal similarities*. Readers interested in applications of dimensional analysis to the theory of biological similarity may refer to the review of Günther (1975), where the various types of physical similarity are briefly described.

Plate 3.1 The metre is the basis of the metric system, which was established during the French Revolution and became the *Système International d'Unités* (*SI*, International System of Units) in 1960. The metre was originally set as 10^{-7} of a quarter of the Earth meridional perimeter. In order to define the metre precisely, French astronomers Delambre and Méchain measured the meridian between Dunkerque and Barcelona between 1792 and 1799. The story of their work can be found in Guedj (1999, 2001). Copies of the standard metre engraved in marble were displayed at 16 locations in Paris to make the new measurement unit known and used by the people. The picture shows the last of these marble metres that is still at the site where it was originally placed, under the arcades of 36 rue de Vaugirard in Paris, across the street from the Palais du Luxembourg (seat of the French Senate), where it can be seen nowadays. Photo P. Legendre, 2002.

Chapter

4 *Multidimensional quantitative data*

4.0 Multidimensional statistics

Basic statistics are now part of the training of most ecologists. However, statistical techniques based on simple distributions such as the unidimensional normal distribution are not really appropriate for analysing complex ecological data sets. Nevertheless, researchers sometimes perform series of simple analyses on the various descriptors in their data set, expecting to obtain results that are pertinent to the problem under study. This type of approach is incorrect because it does not take into account the covariances among descriptors; see also Box 1.3 where the statistical problem created by multiple testing is explained. In addition, such an approach only extracts minimum information from data that have often been collected at great cost and it usually generates a mass of results from which it may be difficult to draw synthetic conclusions. Finally, in studies involving species assemblages, it is usually more interesting to describe the variability of the structure of the assemblage as a whole (i.e. *mensurative* variation observed through space or time, or *manipulative* variation resulting from experimental manipulation; Hurlbert, 1984) than to analyse each species independently.

Fortunately, methods derived from *multidimensional statistics*, which are used throughout this book, are designed for analysing complex data sets. These methods take into account the co-varying nature of ecological data and can evidence the structures that underlie the data. The present chapter discusses the basic theory and characteristics of multidimensional data analysis. Mathematics are kept to a minimum, so that readers can easily reach a high level of understanding. Many approaches of practical interest are discussed, including several types of linear correlation with their statistical tests. It must be noted that this chapter is limited to linear statistics.

A number of excellent textbooks deal with detailed aspects of multidimensional statistics, for example Mardia *et al*. (1979), Muirhead (1982), Anderson (2003), and Hair *et al*. (2010). There are also several titles on specialized topics such as linear

Table 4.1 Numerical example of two species observed at four sampling sites. Figure 4.1 shows that each row of the data matrix may be construed as a vector, as defined in Section 2.4.

Sampling sites (objects)	Species (descriptors) 1	2	($p = 2$)
1	5	1	
2	3	2	
3	8	3	
4	6	4	
($n = 4$)			

models, linear regression, and time series analysis. None of these books specifically deals with ecological data, however.

Multidimensional Multivariate

Several authors use the term *multivariate* as an abbreviation for *multidimensional variate* (the latter term meaning *random variable*; Section 1.0). As an adjective, *multivariate* is interchangeable with *multidimensional*.

4.1 Multidimensional variables and dispersion matrix

As stated in Section 1.0, the present textbook deals with the analysis of *random variables*. Ecological data matrices have n rows and p columns (Section 2.1). Each row is a *vector* (Section 2.4) which is, statistically speaking, one realization of a p-dimensional random variable. When, for example, p species are observed at n sampling sites, the species are the p dimensions of a random variable "species" and each site provides one realization of this p-dimensional random variable.

To illustrate this concept, four sampling units with two species (Table 4.1) are plotted in a two-dimensional Euclidean space (Fig. 4.1). Vector "site 1" is the doublet (5,1). It is plotted in the same two-dimensional space as the three other vectors "site i". Each row of the data matrix is a two-dimensional vector, which is one realization of the (bivariate) random variable "species". The random variable "species" is said to be two-dimensional because the sampling units (objects) contain two species (descriptors), the two dimensions being species 1 and 2, respectively. The species descriptors of this example are the axes of the attribute space, or A-space (Fig. 7.2).

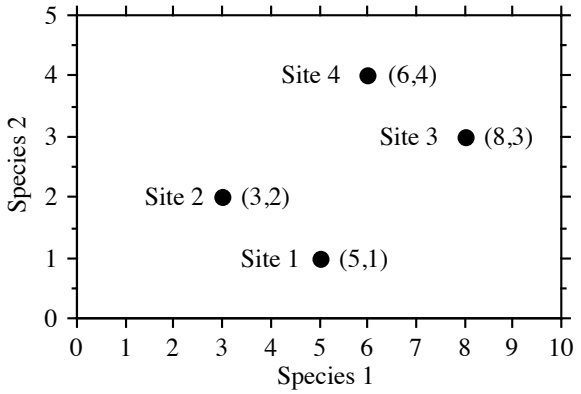


Figure 4.1 Four realizations (sampling sites from Table 4.1) of the two-dimensional random variable "species" are plotted in a two-dimensional Euclidean space.

As the number of descriptors (e.g. species) increases, the number of dimensions of the random variable "species" similarly increases, so that more axes are necessary to construct the space in which the objects are plotted. Thus, the p descriptors make up a p-dimensional random variable and the n vectors of observations (objects) are as many realizations of the p-dimensional vector "descriptors". The present chapter does not deal with *samples* of observations, which result from field or laboratory work (for a brief discussion on sampling, see Section 1.0). It focuses instead on *populations*, which are investigated by means of samples.

Before approaching the multidimensional normal distribution, it is necessary to define a p-dimensional random variable "descriptors":

$$\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots \mathbf{y}_j, \ldots \mathbf{y}_p] \qquad \textbf{(4.1)}$$

Each element $\mathbf{y}_j$ of the multidimensional variable $\mathbf{Y}$ is a one-dimensional random variable. Every descriptor $\mathbf{y}_j$ is observed in each of the n vectors "object", each sampling unit i providing a realization of the p-dimensional random variable.

In ecology, the structure of *dependence* among descriptors is, in many instances, the matter being investigated. Researchers who study multidimensional data using univariate statistics assume that the p unidimensional $\mathbf{y}_j$ variables in $\mathbf{Y}$ are *linearly independent* of one another (third meaning of *independence* in Box 1.1). This is the reason why univariate statistical methods are inappropriate with most ecological data and why methods that take into account the *dependence* among descriptors must be used to analyse multidimensional data sets. Only these methods will generate proper results when there is dependence among descriptors; it is never acceptable to replace a multidimensional analysis by a series of unidimensional treatments.

Table 4.2 Symbols used to identify (population) parameters and (sample) statistics.

	Parameter		Statistic	
	Matrix or vector	Elements	Matrix or vector	Elements
Covariance	$\boldsymbol{\Sigma}$ (sigma)	σ_{jk} (sigma)	$\mathbf{S}$	s_{jk}
Correlation	$\mathbf{P}$ (rho)	ρ_{jk} (rho)	$\mathbf{R}$	r_{jk}
Mean	$\boldsymbol{\mu}$ (mu)	μ_j (mu)	$\bar{\mathbf{y}}$	$\bar{y}_j$

The symbols for covariance matrix $\boldsymbol{\Sigma}$ and summation $\sum$ should not be confused.

The usual tests of significance require, however, "that successive sample observation vectors from the multidimensional population have been drawn in such a way that they can be construed as realizations of independent random vectors" (Morrison, 1990, p. 80). Subsection 1.1.1 has shown that this assumption of independence among observations is most often not realistic in ecology. Lack of independence among the observations (data rows) does not really matter when statistical models are used for descriptive purposes only, as it is often the case in the present book. For statistical testing, however, corrected tests of significance have to be used when the observations are spatially or temporally correlated (Subsection 1.1.2).

To sum up: (1) the p descriptors in ecological data matrices are the p dimensions of a random variable "descriptors"; (2) in general, the p descriptors are *not linearly independent* of one another; methods of multidimensional analysis are designed to bring out the structure of linear dependence among descriptors; (3) each of the n sampling units is a realization of the p-dimensional vector "descriptors"; (4) the usual tests of significance assume that the n sampling units are realizations of *independent* random vectors. The latter condition is generally not met in ecology, with consequences that were discussed in the previous paragraph and in Subsection 1.1.1. For the various meanings of the term *independence* in statistics, see Box 1.1.

Parameter
Statistic

Greek and roman letters are used here and in the remainder of the book (Table 4.2). The properties of a *population* (called *parameters*) are denoted by *greek* letters. Their *estimates* (called *statistics*), computed from *samples,* are symbolized by the corresponding *roman* letters. These conventions are complemented by those pertaining to matrix notation (Section 2.1).

The dependence among quantitative variables $\mathbf{y}_j$ brings up the concept of *covariance*. Covariance is the extension, to two descriptors, of the concept of *variance*. The variance is a measure of the *dispersion* of a random variable $\mathbf{y}_j$ around its mean; it is denoted σ_j^2. Covariance measures the *joint dispersion* of two random variables $\mathbf{y}_j$

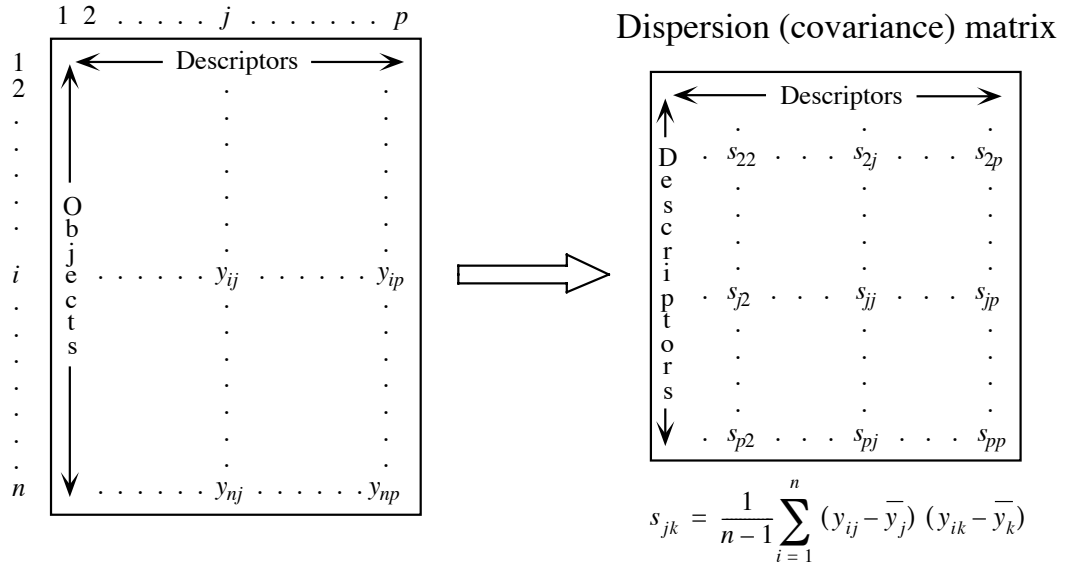


Figure 4.2 Structure of ecological data. Given their nature, ecological descriptors are, in most cases, *linearly dependent* on one another (Box 1.1).

Dispersion matrix

and $\mathbf{y}_k$ around their means; it is denoted σ_{jk}. The *dispersion matrix* of $\mathbf{Y}$, called matrix $\mathbf{\Sigma}$ (sigma), contains the variances and covariances of the p descriptors (Fig. 4.2):

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix} \quad \textbf{(4.2)}$$

Matrix $\mathbf{\Sigma}$ is an *association matrix* [descriptors × descriptors] (Section 2.2). The elements σ_{jk} of matrix $\mathbf{\Sigma}$ are the covariances between all pairs of the p random variables. The matrix is symmetric because the covariance of $\mathbf{y}_j$ and $\mathbf{y}_k$ is identical to that of $\mathbf{y}_k$ and $\mathbf{y}_j$. Each diagonal element of $\mathbf{\Sigma}$ is the covariance of a descriptor $\mathbf{y}_j$ with itself, which is the variance of $\mathbf{y}_j$, so that $\sigma_{jj} = \sigma_j^2$.

Variance

The *estimate* of the *variance* of $\mathbf{y}_j$, denoted s_j^2, is computed on the *centred variable* $(y_{ij} - \bar{y}_j)$. Variable $\mathbf{y}_j$ is centred by subtracting the mean $\bar{y}_j$ from each of the n

observations y_{ij}. As a result, the mean of the centred variable is zero. The unbiased estimator of the population variance s_j^2 is computed using the well-known formula:

$$\text{var}(\mathbf{y}_j) = s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n} (y_{ij} - \bar{y}_j)^2 \qquad \textbf{(4.3)}$$

where the sum of *squares* of the *centred data*, for descriptor j, is divided by the number of objects minus one $(n - 1)$. The summation is over the n observations of descriptor j. The variance of $\mathbf{y}_j$ is expressed in the squared physical dimension of $\mathbf{y}_j$. In the same way, the estimate (s_{jk}) of the *covariance* (σ_{jk}) of $\mathbf{y}_j$ and $\mathbf{y}_k$ is computed on the centred variables $(y_{ij} - \bar{y}_j)$ and $(y_{ik} - \bar{y}_k)$, using the formula of a "bivariate variance". The *covariance* s_{jk} is calculated as:

Covariance

$$\text{cov}(\mathbf{y}_j,\mathbf{y}_k) = s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n} (y_{ij} - \bar{y}_j)\,(y_{ik} - \bar{y}_k) \qquad \textbf{(4.4)}$$

When $k = j$, eq. 4.4 is identical to eq. 4.3. The positive square root of the variance is called the *standard deviation* (σ_j); it has the same dimension as $\mathbf{y}_j$. Its estimate s_j is:

Standard deviation

$$s_j = \sqrt{s_j^2} \qquad \textbf{(4.5)}$$

The coefficient of variation is a dimensionless measure of variation. *CV* is used to compare the variation of variables expressed in different physical units. It is obtained by dividing the standard deviation s_j by the mean $\bar{x}_j$ of variable j:

Coefficient of variation

$$CV_j = s_j / \bar{x}_j$$

Since the standard deviation and the mean of a variable have the same physical units, CV_j is dimensionless. CV_j is only defined for quantitative variables that have non-zero means and it does not make sense for interval-scale variables (Subsection 1.4.1), for which the value of the mean is arbitrary. The coefficient of variation may be rescaled to percentages by multiplying its value by 100. For small n, an estimate with reduced bias is obtained by multiplying *CV* by $(1 + 1/(4n))$.

Contrary to the variance, which is always positive, the covariance may take positive or negative values. To understand the meaning of the covariance, imagine that the object points are plotted in a scatter diagram where the axes are descriptors $\mathbf{y}_j$ and $\mathbf{y}_k$. The data are centred by drawing new axes, whose origin is at the centroid $(\bar{y}_j, \bar{y}_k)$ of the cloud of points (centred plots of that kind with positive and negative correlations are shown in Fig. 4.7). A positive covariance (e.g. Fig. 4.7, right) means that most of the points are in quadrants I and III of the centred plot, where the centred values $(y_{ij} - \bar{y}_j)$ and $(y_{ik} - \bar{y}_k)$ have the same signs. This corresponds to a positive relationship between the two descriptors. The converse is true for a negative covariance (e.g. Fig. 4.7, left), for which most of the points are in quadrants II and IV

of the centred plot. When the covariance is null (e.g. Fig. 4.8, left) or small, the points are equally distributed among the four quadrants of the centred plot.

The covariance or dispersion matrix* **S** can be computed directly by multiplying the *matrix of centred data* $[y - \bar{y}]$ with its transpose $[y - \bar{y}]'$:

$$\text{cov}(\mathbf{Y}) = \mathbf{S} = \frac{1}{n-1}[y - \bar{y}]'[y - \bar{y}] \qquad \textbf{(4.6)}$$

$$\mathbf{S} = \frac{1}{n-1}\begin{bmatrix} (y_{11}-\bar{y_1}) & (y_{21}-\bar{y_1}) & \cdots & (y_{n1}-\bar{y_1}) \\ (y_{12}-\bar{y_2}) & (y_{22}-\bar{y_2}) & \cdots & (y_{n2}-\bar{y_2}) \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ (y_{1p}-\bar{y_p}) & (y_{2p}-\bar{y_p}) & \cdots & (y_{np}-\bar{y_p}) \end{bmatrix} \begin{bmatrix} (y_{11}-\bar{y_1}) & (y_{12}-\bar{y_2}) & \cdots & (y_{1p}-\bar{y_p}) \\ (y_{21}-\bar{y_1}) & (y_{22}-\bar{y_2}) & \cdots & (y_{2p}-\bar{y_p}) \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ (y_{n1}-\bar{y_1}) & (y_{n2}-\bar{y_2}) & \cdots & (y_{np}-\bar{y_p}) \end{bmatrix}$$

$$\mathbf{S} = \frac{1}{n-1}\begin{bmatrix} \sum_{i=1}^{n}(y_{i1}-\bar{y_1})^2 & \sum_{i=1}^{n}(y_{i1}-\bar{y_1})(y_{i2}-\bar{y_2}) & \cdots & \sum_{i=1}^{n}(y_{i1}-\bar{y_1})(y_{ip}-\bar{y_p}) \\ \sum_{i=1}^{n}(y_{i2}-\bar{y_2})(y_{i1}-\bar{y_1}) & \sum_{i=1}^{n}(y_{i2}-\bar{y_2})^2 & \cdots & \sum_{i=1}^{n}(y_{i2}-\bar{y_2})(y_{ip}-\bar{y_p}) \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \sum_{i=1}^{n}(y_{ip}-\bar{y_p})(y_{i1}-\bar{y_1}) & \sum_{i=1}^{n}(y_{ip}-\bar{y_p})(y_{i2}-\bar{y_2}) & \cdots & \sum_{i=1}^{n}(y_{ip}-\bar{y_p})^2 \end{bmatrix}$$

This elegant and rapid procedure shows once again the advantage of matrix algebra in numerical ecology, where the data sets are generally large.

Numerical example. Four species ($p = 4$) were observed at five stations ($n = 5$). The estimated population parameters, for the species, are the means ($\bar{y}_j$), the variances (s_j^2), and the covariances (s_{jk}). The original and centred data are shown in Table 4.3. Because $s_{jk} = s_{kj}$, the dispersion matrix is symmetric. The mean of each *centred variable* is zero.

In this numerical example, the covariance between species 2 and the other three species is zero. This does not necessarily mean that species 2 is independent of the other three, but simply that the joint *linear* dispersion of species 2 with any one of the other three is zero. This example will be revisited in Section 4.2.

* Some authors call $[y - \bar{y}]'[y - \bar{y}]$ a *dispersion matrix* and **S** a *covariance matrix*. For these authors, a covariance matrix is then a dispersion matrix divided by $(n - 1)$.

Table 4.3 Numerical example. Calculation of centred data and covariances.

Sites	Original data	Centred data
1 2 3 4 5	$\mathbf{Y} = \begin{bmatrix} 1 & 5 & 2 & 6 \\ 2 & 2 & 1 & 8 \\ 3 & 1 & 3 & 4 \\ 4 & 2 & 5 & 0 \\ 5 & 5 & 4 & 2 \end{bmatrix}$	$[y - \bar{y}] = \begin{bmatrix} -2 & 2 & -1 & 2 \\ -1 & -1 & -2 & 4 \\ 0 & -2 & 0 & 0 \\ 1 & -1 & 2 & -4 \\ 2 & 2 & 1 & -2 \end{bmatrix}$
Means	$\bar{\mathbf{y}}' = \begin{bmatrix} 3 & 3 & 3 & 4 \end{bmatrix}$	$[\overline{y - \bar{y}}]' = \begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix}$
$n - 1 = 4$	$\mathbf{S} = \frac{1}{n-1}[y - \bar{y}]'[y - \bar{y}] = \begin{bmatrix} 2.5 & 0 & 2 & -4 \\ 0 & 3.5 & 0 & 0 \\ 2 & 0 & 2.5 & -5 \\ -4 & 0 & -5 & 10 \end{bmatrix}$	

The square root of the determinant of the dispersion matrix $|\mathbf{S}|^{1/2}$ is known as the *generalized variance*. It is also equal to the square root of the product of the eigenvalues of **S**.

Any dispersion matrix **S** is *positive semidefinite* (Table 2.2). Indeed, the quadratic form of **S** ($p \times p$) with any real and non-null vector **t** (of size p) is:

$$\mathbf{t}'\mathbf{S}\mathbf{t}$$

This expression can be expanded using eq. 4.6:

$$\mathbf{t}'\mathbf{S}\mathbf{t} = \mathbf{t}'\frac{1}{n-1}[y - \bar{y}]'[y - \bar{y}]\mathbf{t}$$

$$\mathbf{t}'\mathbf{S}\mathbf{t} = \frac{1}{n-1}[[y - \bar{y}]\mathbf{t}]'[[y - \bar{y}]\mathbf{t}] = \text{a scalar}$$

This scalar is the variance of the variable resulting from the product **Yt**. Since a variance, which is a sum of squared values, can only be positive or null, it follows that:

$$\mathbf{t}'\mathbf{S}\mathbf{t} \geq 0$$

so that **S** is positive semidefinite. This means that **S** cannot have negative eigenvalues.

This important property can be derived by computing the quadratic form of the dispersion matrix **S** using eq. 2.28 (right), $\mathbf{\Lambda} = \mathbf{U}^{-1}\mathbf{A}\mathbf{U}$. Because **S** is symmetric, its eigenvectors found in matrix **U** are orthogonal. Since they are also normalized, **U** is an orthonormal matrix, hence $\mathbf{U}^{-1} = \mathbf{U}'$ (property #7 of inverses, Section 2.8), and eq. 2.28 (right) can be written:

$$\mathbf{U}'\mathbf{S}\mathbf{U} = \mathbf{\Lambda}$$

In the quadratic form, vector **t** is replaced by each successive eigenvector $\mathbf{u}_j$ in turn, i.e. each column of matrix **U**. For each vector $\mathbf{u}_j$, the development above shows that

$$\mathbf{u}_j'\mathbf{S}\mathbf{u}_j \geq 0$$

Since $\mathbf{u}_j'\mathbf{S}\mathbf{u}_j = \lambda_j$, this demonstrates that *all eigenvalues* λ_j *of* **S** *are positive or null*. This property of dispersion matrices is fundamental in numerical ecology: it allows one to partition the variance of a matrix **Y** among real (i.e. non-imaginary) *principal axes* (Sections 4.4 and 9.1).

Another property of the dispersion matrix is that the sum of all values in **S** is equal to the variance of a synthetic variable **y** computed as the sum by rows (objects) of all descriptors in **Y**. For example, if **Y** contains species abundance data, the sum by rows (sites) of all species abundances is a new variable **y** corresponding to the total number of individuals at the sites, which can in some cases be interpreted as the total yield or the support capacity of the sites. If **Y** consists of species presence-absence data, **y** is the species richness of the sites. The variance of the synthetic variable **y** can be obtained by summing all values in **S** instead of computing **y** and then its variance. This property will be used in Subsection 13.1.4.

Ideally, matrix **S** (of order p) should be estimated from a number of observations n larger than the number of descriptors p. When $n \leq p$, the rank of matrix **S** is $n - 1$ and, consequently, only $n - 1$ of its rows or columns are independent of one another, so that $p - (n - 1)$ null eigenvalues are produced. The only practical consequence of $n \leq p$ is thus the presence of null eigenvalues in the principal component solution (Section 9.1). The first few eigenvalues of **S**, which are generally those of interest, have positive eigenvalues.

4.2 Correlation matrix

The previous section has shown that the covariance provides information on the orientation of the cloud of data points in the space defined by the descriptors. That statistic, however, does not provide any information on the intensity of the relationship between variables $\mathbf{y}_j$ and $\mathbf{y}_k$. Indeed, the covariance may increase or decrease without changing the relationship between $\mathbf{y}_j$ and $\mathbf{y}_k$. For example, in Fig. 4.3, the two clouds of points correspond to different covariance values (factor two in size, and thus in

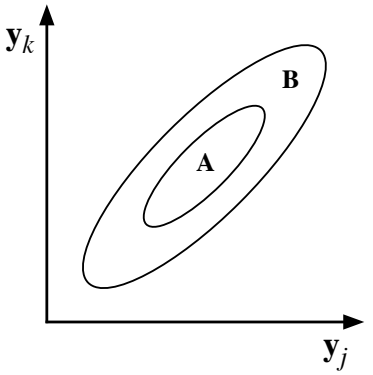


Figure 4.3 Several observations (objects), with descriptors $\mathbf{y}_j$ and $\mathbf{y}_k$, were made under two different sets of conditions (A and B). The two ellipses delineate clouds of point-objects corresponding to A and B, respectively. The covariance of $\mathbf{y}_j$ and $\mathbf{y}_k$ is twice as large for B as it is for A (larger ellipse), but the correlation between the two descriptors is the same in these two cases (i.e. the ellipses have the same shape).

covariance), but the relationship between the variables is identical (same shape). Since the covariance depends on the dispersion of the points around the mean of each variable (i.e. their variances), determining the intensity of the relationship between variables requires to control for the variances.

The *covariance* measures the joint *dispersion* of two random variables around the bivariate mean. The *correlation* is defined as a measure of the *dependence* between two random variables $\mathbf{y}_j$ and $\mathbf{y}_k$. As explained in Section 1.5, it often happens that matrices of ecological data contain descriptors with scales that are not commensurate, e.g. when some species have larger biomass than others by orders of magnitude, or when the descriptors have different physical dimensions (Chapter 3). Calculating covariances on such variables obviously does not make sense, except if the descriptors are first reduced to a common scale. The standardization procedure consists in centring all descriptors on a zero mean and reducing them to unit standard deviation (eq. 1.12). By using *standardized descriptors*, it is possible to calculate meaningful covariances because the new variables have the same scale (i.e. unit standard deviation) and are dimensionless (see Chapter 3).

Linear correlation

The covariance of two standardized descriptors is called the coefficient of linear correlation (Pearson *r*). This statistic has been proposed by the statistician Karl Pearson and is named after him. Given two standardized descriptors (eq. 1.12)

$$z_{ij} = \frac{y_{ij} - \overline{y}_j}{s_j} \qquad \text{and} \qquad z_{ik} = \frac{y_{ik} - \overline{y}_k}{s_k}$$

calculating their covariance (eq. 4.4) gives

$$s(z_j,z_k) = \frac{1}{n-1}\sum_{i=1}^{n} (z_{ij} - 0)\,(z_{ik} - 0) \qquad \text{because} \qquad \overline{z}_j = \overline{z}_k = 0$$

$$s(z_j,z_k) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{y_{ij}-\bar{y}_j}{s_j}\right)\left(\frac{y_{ik}-\bar{y}_k}{s_k}\right)$$

$$s(z_j,z_k) = \left(\frac{1}{s_j s_k}\right)\frac{1}{n-1}\sum_{i=1}^{n}(y_{ij}-\bar{y}_j)\,(y_{ik}-\bar{y}_k)$$

$$s(z_j,z_k) = \left(\frac{1}{s_j s_k}\right)s_{jk} = r_{jk}\text{, the } \textit{coefficient of linear correlation} \text{ between } \mathbf{y}_j \text{ and } \mathbf{y}_k.$$

The developed formula is:

$$\text{cor}(\mathbf{y}_j,\mathbf{y}_k) = r_{jk} = \frac{s_{jk}}{s_j s_k} = \frac{\sum_{i=1}^{n}(y_{ij}-\bar{y}_j)\,(y_{ik}-\bar{y}_k)}{\sqrt{\sum_{i=1}^{n}(y_{ij}-\bar{y}_j)^2\sum_{i=1}^{n}(y_{ik}-\bar{y}_k)^2}} \qquad \textbf{(4.7)}$$

Correlation matrix

As in the case of dispersion (Section 4.1), it is possible to construct the *correlation matrix* of $\mathbf{Y}$, i.e. the $\mathbf{P}$ (rho) matrix, whose elements are the coefficients of linear correlation ρ_{jk}:

$$\mathbf{P} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix} \qquad \textbf{(4.8)}$$

The *correlation matrix is the dispersion matrix of the standardized variables*. This concept will play a fundamental role in principal component analysis (Section 9.1). It should be noted that the diagonal elements of $\mathbf{P}$ are all equal to 1. This is because the comparison of any descriptor with itself is a case of complete dependence, which leads to a correlation $\rho = 1$. When $\mathbf{y}_j$ and $\mathbf{y}_k$ are independent of each other, $\rho_j = 0$. However, a correlation equal to zero does not necessarily imply that $\mathbf{y}_j$ and $\mathbf{y}_k$ are independent of each other, as shown by the following numerical example. A correlation $\rho_{jk} = -1$ is indicative of a complete, but inverse dependence of the two variables.

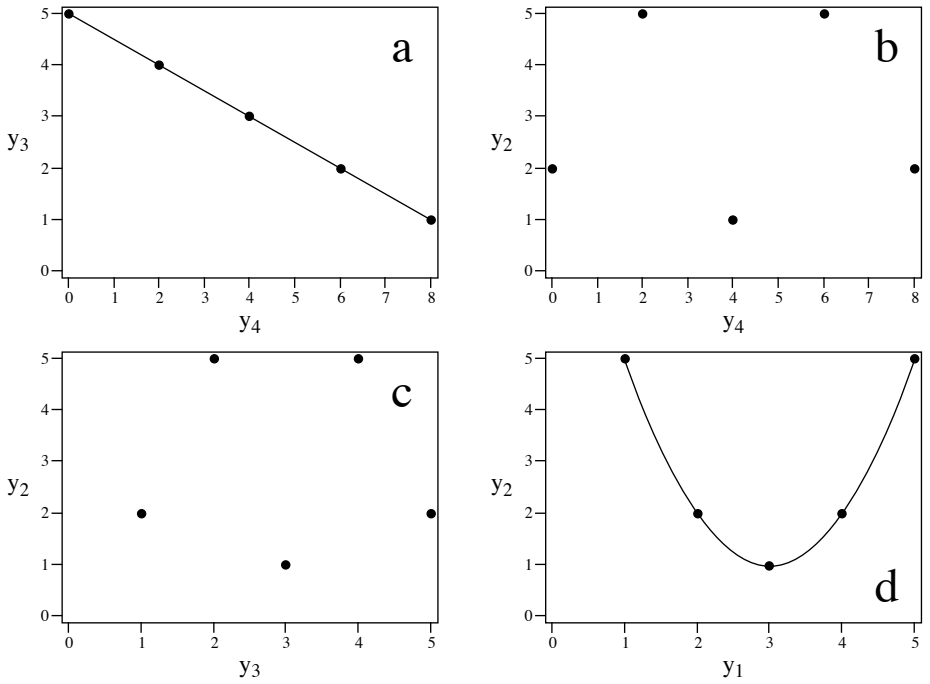


Figure 4.4 Numerical example. Relationships between species (a) 3 and 4, (b) 2 and 4, (c) 2 and 3, and (d) 2 and 1.

Numerical example. Using the values in Table 4.3, matrix **R** can easily be computed. According to eq. 4.7, each element r_{jk} combines the covariance s_{jk} with the variances s_j and s_k:

$$\mathbf{R} = \begin{bmatrix} 1 & 0 & 0.8 & -0.8 \\ 0 & 1 & 0 & 0 \\ 0.8 & 0 & 1 & -1 \\ -0.8 & 0 & -1 & 1 \end{bmatrix}$$

Matrix **R** is symmetric, as was matrix **S**. The correlation $r = -1$ between species 3 and 4 means that these species are fully, but inversely, dependent (Fig. 4.4a). Correlations $r = 0.8$ and -0.8 are interpreted as indications of strong dependence between species 1 and 3 (direct) and species 1 and 4 (inverse), respectively. The *zero* correlation between species 2 and the other three species must be interpreted with caution. Figure 4.4d clearly shows that species 1 and 2 are completely *dependent* on each other since they are related by equation $y_2 = 1 + (3 - y_1)^2$; the zero correlation is, in this case, a consequence of the *linear* model underlying statistic *r*. Therefore, only the correlations that are *significantly* different from zero should be considered, since a null correlation has no unique interpretation.

Table 4.4 Numerical example. Calculation of standardized data and correlations.

Sites	Original data	Standardized data
1	$\mathbf{Y} = \begin{bmatrix} 1 & 5 & 2 & 6 \\ 2 & 2 & 1 & 8 \\ 3 & 1 & 3 & 4 \\ 4 & 2 & 5 & 0 \\ 5 & 5 & 4 & 2 \end{bmatrix}$	$\mathbf{Z} = \begin{bmatrix} -1.27 & 1.07 & -0.63 & 0.63 \\ -0.63 & -0.53 & -1.27 & 1.27 \\ 0 & -1.07 & 0 & 0 \\ 0.63 & -0.53 & 1.27 & -1.27 \\ 1.27 & 1.07 & 0.63 & -0.63 \end{bmatrix}$
2		
3		
4		
5		
Means	$\bar{\mathbf{y}}' = \begin{bmatrix} 3 & 3 & 3 & 4 \end{bmatrix}$	$\bar{\mathbf{z}}' = \begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix}$
$n - 1 = 4$	$\mathbf{R}(y) = \mathbf{S}(z) = \frac{1}{n-1}\mathbf{Z}'\mathbf{Z} = \begin{bmatrix} 1 & 0 & 0.8 & -0.8 \\ 0 & 1 & 0 & 0 \\ 0.8 & 0 & 1 & -1 \\ -0.8 & 0 & -1 & 1 \end{bmatrix}$	

Since the correlation matrix is the dispersion matrix of standardized variables, it is possible, as in the case of matrix **S** (eq. 4.6), to compute **R** directly by multiplying the *matrix of standardized data* with its transpose:

$$\text{cor}(\mathbf{Y}) = \mathbf{R} = \frac{1}{n-1}\left[(y - \bar{y}) / s_y\right]'\left[(y - \bar{y}) / s_y\right] = \frac{1}{n-1}\mathbf{Z}'\mathbf{Z} \qquad \textbf{(4.9)}$$

Table 4.4 shows how to calculate correlations r_{jk} of the example as in Table 4.3, using this time the *standardized data*. The mean of each *standardized variable is* zero and its standard deviation is equal to unity. The *dispersion* matrix of **Z** is identical to the *correlation* matrix of **Y**, which was calculated above using the covariances and variances.

Matrices $\mathbf{\Sigma}$ and $\mathbf{P}$ are related to each other by the diagonal matrix of standard deviations of $\mathbf{Y}$. This new matrix, which was specifically designed here to relate $\mathbf{\Sigma}$ and $\mathbf{P}$, is symbolized by $\mathbf{D}(\sigma)$ and its inverse by $\mathbf{D}(\sigma)^{-1}$:

$$\mathbf{D}(\sigma) = \begin{bmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ 0 & \cdot & \dots & \sigma_p \end{bmatrix} \quad \text{and} \quad \mathbf{D}(\sigma)^{-1} = \begin{bmatrix} 1/\sigma_1 & 0 & \dots & 0 \\ 0 & 1/\sigma_2 & \dots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ 0 & \cdot & \dots & 1/\sigma_p \end{bmatrix}$$

Using these two matrices, one can write:

$$\mathbf{P} = \mathbf{D}(\sigma^2)^{-1/2}\ \mathbf{\Sigma}\ \mathbf{D}(\sigma^2)^{-1/2} = \mathbf{D}(\sigma)^{-1}\ \mathbf{\Sigma}\ \mathbf{D}(\sigma)^{-1} \qquad \textbf{(4.10)}$$

where $\mathbf{D}(\sigma^2)$ is the matrix of the diagonal elements of $\mathbf{\Sigma}$. It follows from eq. 4.10 that:

$$\mathbf{\Sigma} = \mathbf{D}(\sigma)\ \mathbf{P}\ \mathbf{D}(\sigma) \qquad \textbf{(4.11)}$$

Significance of *r*

The theory underlying tests of significance is summarized in Section 1.2. In the case of *r*, inference about the statistical population is in most instances through the null hypothesis H_0: $\rho = 0$. H_0 may also state that ρ has some other value than zero, which would be derived from ecological hypotheses. The general formula for testing correlation coefficients is given in Section 4.5 (eq. 4.39). The Pearson correlation coefficient r_{jk} involves two descriptors $\mathbf{y}_j$ and $\mathbf{y}_k$ (hence $m = 2$ when testing a coefficient of simple linear correlation using eq. 4.39), so that $\nu_1 = 2 - 1 = 1$ and $\nu_2 = n - 2 = \nu$. The general formula then becomes:

$$F = \frac{r_{jk}^2 / 1}{(1 - r_{jk}^2)/\nu} = \nu \frac{r_{jk}^2}{1 - r_{jk}^2} \qquad \textbf{(4.12)}$$

where $\nu = n - 2$. Statistic F is tested against $F_{\alpha[1,\nu]}$. Since the square root of a statistic $F_{[\nu_1,\nu_2]}$ is a statistic $t_{[\nu = \nu_2]}$ when $\nu_1 = 1$, r may also be tested using:

$$t = \frac{r_{jk}\sqrt{\nu}}{\sqrt{1 - r_{jk}^2}} \qquad \textbf{(4.13)}$$

The t-statistic is tested against the value $t_{\alpha[\nu]}$. In other words, H_0 is tested by comparing the F (or t) statistic to the value found in a table of critical values of F (or t). Equations 4.12 and 4.13 produce identical tests. The number of degrees of freedom is $\nu = (n - 2)$ because calculating a correlation coefficient requires prior estimation of two parameters, i.e. the means of the two populations (eq. 4.7). H_0 is rejected when the probability corresponding to F (or t) is smaller than or equal to a predetermined *level*

of significance (α for a two-tailed test, and $\alpha/2$ for a one-tailed test; the difference between the two types of tests is explained in Section 1.2). In principle, this test requires that the sample of observations be drawn from a population having a *bivariate normal distributio*n (Section 4.3). Testing for normality and multinormality is discussed in Section 4.6, and normalizing transformations in Section 1.5. When the data do not satisfy the condition of normality, t can be tested by permutation, as explained in Section 1.2.

Test of independence of variables

It is also possible to test the *independence of all variables* in a data matrix by considering the set of all correlation coefficients found in matrix **R**. The null hypothesis here is that the $p(p-1)/2$ coefficients are all equal to zero, H_0: $\mathbf{R} = \mathbf{I}$ (unit matrix). According to Bartlett (1954), the determinant of **R**, $|\mathbf{R}|$, can be transformed into a X^2 (chi-square) test statistic:

$$X^2 = -[n - (2p + 5)/6] \log_e |\mathbf{R}| \qquad \textbf{(4.14)}$$

where $\log_e |\mathbf{R}|$ is the natural logarithm of the determinant of **R**. This statistic is approximately distributed as χ^2 with $\nu = p(p-1)/2$ degrees of freedom. When the probability associated with X^2 is significantly low, the null hypothesis of complete independence of the p descriptors is rejected. In principle, this test requires the observations to be drawn from a population with a *multivariate normal distributio*n (Section 4.3). If the null hypothesis of independence of all variables is rejected, the $p(p-1)/2$ correlation coefficients in matrix **R** may be tested individually. Box 1.3 describes how to correct individual p-values in situations of multiple testing.

Other correlation coefficients are described in Sections 4.5 and 5.3. When the coefficient of linear correlation must be distinguished from other coefficients, it is referred to as *Pearson r*. Elsewhere, r is called the *coefficient of linear correlation* or *correlation coefficient*. Table 4.5 summarizes the main properties of this coefficient.

4.3 Multinormal distribution

In general, the mathematics of the normal distribution is of little concern to ecologists using unidimensional statistical methods. In the best case, data are normalized (Section 1.5) before being subjected to tests that are based on *parametric* hypotheses. It must be remembered that all *parametric tests* require the data to follow a specific *distribution,* most often the normal distribution. When the data do not obey this condition, the results of parametric tests may be *invalid*.

There also exist nonparametric tests (Chapter 5), in which no reference is made to any theoretical distribution of the population, hence no use of parameters. That is also the case with permutation tests based on the usual parametric statistics, e.g. the Pearson correlation coefficient r (Subsection 1.2.2). Another advantage of nonparametric and permutational tests is that they remain valid for samples of very

Table 4.5 Main properties of the coefficient of linear correlation. Some of these properties are discussed in a later sections.

	Properties	Sections
1.	The coefficient of linear correlation measures the *intensity of the linear relationship* between two random variables.	4.2
2.	The coefficient of linear correlation between two variables can be calculated using their respective *variances* and their *covariance*.	4.2
3.	The correlation matrix is the *dispersion* matrix of *standardized variables*.	4.2
4.	The square of the coefficient of linear correlation is the *coefficient of determination*. It measures how much of the variance of each variable is explained by the other.	10.3
5.	The coefficient of linear correlation is a *parameter* of a multinormal distribution.	4.3
6.	The absolute value of the coefficient of linear correlation is the *geometric mean* of the *coefficients of linear regression* of each variable on the other.	10.3

small sizes, which are often encountered in ecological research. These tests are of great interest to ecologists. Researchers may nevertheless attempt to normalize their data to have access to the powerful toolbox of parametric statistics or because some of the methods of multivariate analysis, e.g. principal component analysis (Section 9.1), perform better when the response data distributions are not strongly asymmetric.

Multidimensional statistics require careful examination of the main characteristics of the *multinormal* (or *multivariate normal*) *distribution*. Several of the methods described in the present chapter, and also in Chapters 9, 10 and 11, are founded on principles derived from the multinormal distribution. This is true even in cases where no test of significance is performed, which is often the case in numerical ecology (i.e. descriptive versus inferential statistics, Sections 1.2).

The logic of an approach centred on the multinormal distribution is based upon a theorem which is undoubtedly one of the most important of statistics. According to the *central limit theorem*, when a random variable results from several independent and additive effects, of which none has a dominant variance, then this variable tends towards a normal distribution even if the effects are not themselves normally distributed. Since ecological variables, and species abundances in particular, are often influenced by several independent random factors, the above theorem explains why the normal distribution is frequently invoked to describe ecological phenomena. This justifies a careful examination of the properties of the multinormal distribution before studying the methods for analysing multidimensional quantitative data.

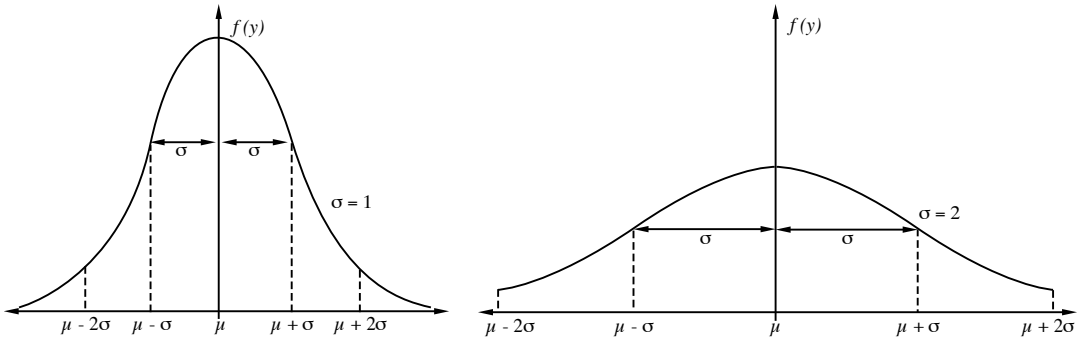


Figure 4.5 Role of the standard deviation σ in the normal distribution function. The abscissa is variable *y*.

Normal

The probability density of a *normal* random variable *y* is:

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right] \qquad \textbf{(4.15)}$$

(Laplace-Gauss equation) where exp […] reads "*e* to the power […]", *e* being the Napierian base (*e* = 2.71828…). Calculation of $f(y)$, for a given value *y*, only requires μ and σ. The mean (μ) and standard deviation (σ) of the theoretical population completely determine the shape of the probability distribution. This is why they are called the *parameters* of the normal distribution. The curve is symmetric on both sides of μ and its exact shape depends on σ (Fig. 4.5).

The value σ determines the positions of the inflexion points along the normal curve. These points are located on both sides of μ, at a distance σ, whereas μ positions the curve on the abscissa. In Fig. 4.5, the surface under each of the two curves is identical for the same number of σ units on either side of μ. The height of the curve is the probability density corresponding to the *y* value; for a continuous function such as that of the normal distribution, the probability of finding a value between $y = a$ and $y = b$ ($a < b$) is given by the surface under the curve between *a* and *b*. For example, the probability of finding a value between $\mu - 1.96\sigma$ and $\mu + 1.96\sigma$ is 0.95.

In view of examining the properties of the multinormal distribution, it is convenient to first consider the joint probability density of *p* *independent* unidimensional normal variables. For *each* of these *p* variables y_j, the probability density is given by eq. 4.15, with mean μ_j and standard deviation σ_j:

$$f(y_j) = \frac{1}{\sqrt{2\pi}\sigma_j}\exp\left[-\frac{1}{2}\left(\frac{y_j-\mu_j}{\sigma_j}\right)^2\right] \qquad \textbf{(4.16)}$$

A basic law of probabilities states that the joint probability density of several independent variables is the product of their individual densities. It follows that the joint probability density for p independent variables is:

$$f(y_1, y_2, \ldots, y_p) = f(y_1) \times f(y_2) \times \ldots \times f(y_p)$$

$$f(y_1, y_2, \ldots, y_p) = \frac{1}{(2\pi)^{p/2}\sigma_1\sigma_2\ldots\sigma_p} \exp\left[-\frac{1}{2} \sum_{j=1}^{p} \left(\frac{y_j - \mu_j}{\sigma_j} \right)^2 \right] \qquad \textbf{(4.17)}$$

Using the conventions of Table 4.2, one defines the following matrices:

$$\mathbf{y} = \begin{bmatrix} y_1 & y_2 & \ldots & y_p \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & \ldots & 0 \\ 0 & \sigma_2^2 & \ldots & 0 \\ \cdot & \cdot & \ldots & \cdot \\ \cdot & \cdot & \ldots & \cdot \\ \cdot & \cdot & \ldots & \cdot \\ 0 & 0 & \ldots & \sigma_p^2 \end{bmatrix} \qquad \textbf{(4.18)}$$

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 & \mu_2 & \ldots & \mu_p \end{bmatrix}$$

where $\mathbf{y} = [y_1\ y_2\ \ldots\ y_p]$ is the p-dimensional vector of coordinates of a point for which the probability density, i.e. the height (ordinate) of the p-dimensional normal curve, is sought; $\boldsymbol{\mu}$ is the vector of means, and $\boldsymbol{\Sigma}$ is the dispersion matrix among the p independent variables. The determinant of $\boldsymbol{\Sigma}$ is the *generalized variance* of the multivariate distribution. The determinant of a diagonal matrix being equal to the product of the diagonal elements (Section 2.6), it follows that:

Generalized variance

$$\left| \boldsymbol{\Sigma} \right|^{1/2} = (\sigma_1\ \sigma_2 \ldots \sigma_p)$$

From definitions (4.18), and for a single row vector $[y - \mu]$ of p-dimensional centred data, one can write:

$$[y - \mu]\ \boldsymbol{\Sigma}^{-1}\ [y - \mu]' = \sum_{j=1}^{p} \left(\frac{y_j - \mu_j}{\sigma_j} \right)^2$$

which is a scalar. Do not confuse, here, the summation symbol $\sum$ with dispersion matrix $\boldsymbol{\Sigma}$. Using these relationships, eq. 4.17 is rewritten as:

$$f(\mathbf{y}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}}\ \exp\ \{-(1/2)\ [y - \mu]\ \boldsymbol{\Sigma}^{-1}\ [y - \mu]'\} \qquad \textbf{(4.19)}$$

The above equations are for the joint probability density of p *independent* unidimensional normal variables y_j. It is easy to go from there to the *multinormal distribution,* where $\mathbf{y}$ is a p-dimensional random variable whose p dimensions are *not*

Multi-normal

independent. In order to do so, one simply replaces the above matrix $\mathbf{\Sigma}$ by a dispersion matrix containing variances and non-zero covariances, i.e. (eq. 4.2):

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}$$

Using this dispersion matrix $\mathbf{\Sigma}$, eq. 4.19 now describes the probability density $f(\mathbf{y})$ for a p-dimensional multinormal distribution.

Given eq. 4.11, eq. 4.19 for point $\mathbf{y}$ may be rewritten as:

$$f(\mathbf{y}) = \frac{1}{(2\pi)^{p/2}|\mathbf{D}(\sigma)||\mathbf{P}|^{1/2}} \exp\left\{-(1/2)\left[y-\mu\right]\mathbf{D}(\sigma)^{-1}\,\mathbf{P}^{-1}\,\mathbf{D}(\sigma)^{-1}\left[y-\mu\right]'\right\} \quad \textbf{(4.20)}$$

Replacing, in eq. 4.20, vector $\mathbf{y}$ from the p-dimensional matrix $\mathbf{Y}$ by vector $\mathbf{z}$ from the p-dimensional standardized matrix $\mathbf{Z}$ (eq. 1.12) gives:

$$f(\mathbf{z}) = \frac{1}{(2\pi)^{p/2}|\mathbf{P}|^{1/2}} \exp\left\{-(1/2)\,\mathbf{z}\,\mathbf{P}^{-1}\,\mathbf{z}'\right\} \quad \textbf{(4.21)}$$

because $\left[y-\mu\right]\mathbf{D}(\sigma)^{-1} = \mathbf{z}$ and, for a standardized variable $\mathbf{z}$, $\mathbf{D}(\sigma) = \mathbf{I}$.

Equation 4.21 stresses a fundamental point, which was already clear in eq. 4.20: *the correlations* ρ *are parameters of the multinormal distribution,* together with the means μ and standard deviations σ. This property of ρ is shown in Table 4.5.

Three sets of parameters are therefore necessary to specify a multidimensional normal distribution, i.e. the vector of *means* $\boldsymbol{\mu}$, the diagonal matrix of *standard deviations* $\mathbf{D}(\sigma)$, and the *correlation matrix* $\mathbf{P}$. In the unidimensional normal distribution (eq. 4.15), μ and σ were the only parameters because there is no correlation ρ for a single variable.

Bivariate normal

It is not possible to represent, in a plane, more than three dimensions. Thus, for the purpose of illustration, only the simplest case of multinormal distribution will be considered, i.e. the *bivariate normal distribution*, where:

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 & \mu_2 \end{bmatrix} \qquad \mathbf{D}(\sigma) = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \qquad \mathbf{P} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

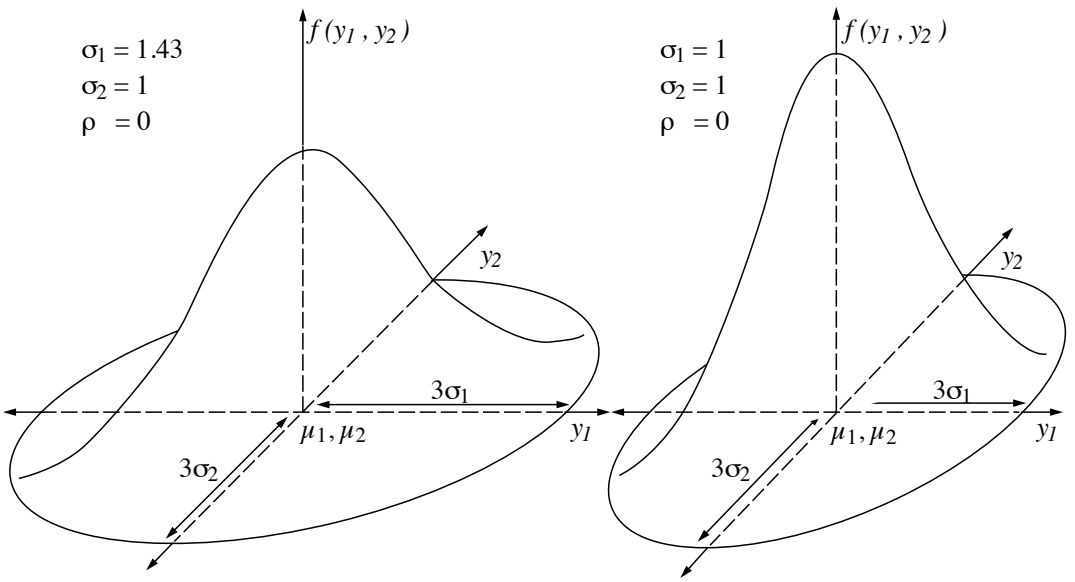


Figure 4.6 Roles of σ_1 and σ_2 in the bivariate normal distribution.

Since $|\mathbf{D}(\sigma)| = \sigma_1\sigma_2$ and $|\mathbf{P}| = (1 - \rho^2)$ in this case, eq. 4.20 becomes:

$$f(y_1,y_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\{-(1/2)\,[y-\mu]\,\mathbf{D}(1/\sigma)\,(1-\rho^2)^{-1}\begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix}\,\mathbf{D}(1/\sigma)\,[y-\mu]'\}$$

$$= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2}\frac{1}{(1-\rho^2)}\left[\left(\frac{y_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{y_1-\mu_1}{\sigma_1}\right)\left(\frac{y_2-\mu_2}{\sigma_2}\right) + \left(\frac{y_2-\mu_2}{\sigma_2}\right)^2\right]\right\}$$

Figure 4.6 shows bivariate normal distributions with their typical "bell" shapes. The two examples illustrate the roles of σ_1 and σ_2. Further examination of the multinormal mathematics is required to specify the role of ρ.

Coming back to the probability density of the multidimensional distribution and neglecting the constant –1/2, the remainder of the exponent in eq. 4.19 is:

$$[y-\mu]\,\mathbf{\Sigma}^{-1}\,[y-\mu]'$$

When it is made equal to a positive constant (α), this algebraic form specifies the equation of any of the points $[y-\mu]$ on a *p*-dimensional *ellipse*:

$$[y-\mu]\,\mathbf{\Sigma}^{-1}\,[y-\mu]' = \alpha \qquad \textbf{(4.22)}$$

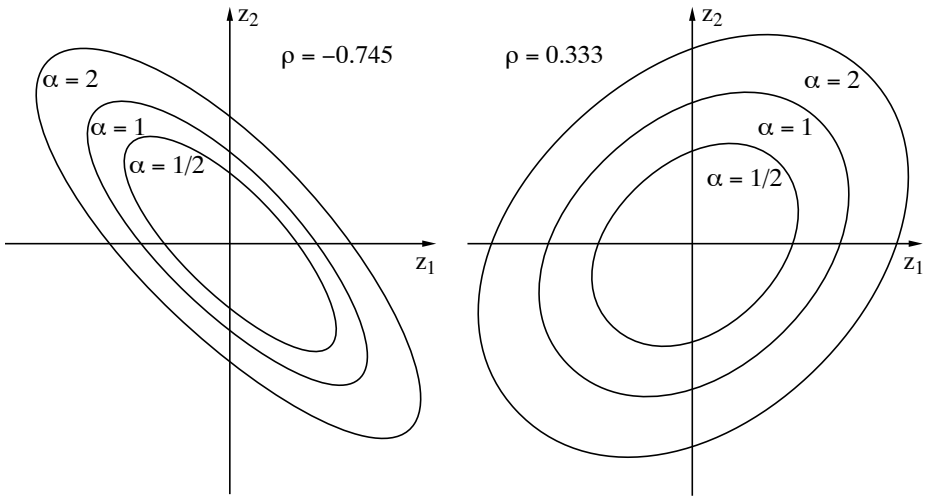


Figure 4.7 Concentration ellipses of a standardized bivariate normal distribution. Role of the correlation ρ.

A family of such multidimensional ellipses may be generated by varying the constant α. All these ellipses have the multidimensional point $\boldsymbol{\mu}$ as their common centre.

It is easy to understand the meaning of eq. 4.22 by examining the two-dimensional case. Without loss of generality, it is convenient to use the standardized variable (z_1,z_2) instead of (y_1,y_2). In that case, the family of *ellipses* (i.e. two-dimensional ellipsoids) is centred on the origin $\boldsymbol{\mu} = [0\ 0]$. For each point with coordinates $[z_1\ z_2]$, the exponent of the *standardized bivariate normal density* is (from expression on the previous page):

$$\frac{1}{1-\rho^2}[z_1^2 - 2\rho z_1 z_2 + z_2^2]$$

This exponent specifies, in two-dimensional space, the equation of a family of ellipses:

$$\frac{1}{1-\rho^2}[z_1^2 - 2\rho z_1 z_2 + z_2^2] = \alpha$$

$$z_1^2 - 2\rho z_1 z_2 + z_2^2 = \alpha(1-\rho^2)$$

Figure 4.7 illustrates the role played by ρ in determining the general shape of the family of ellipses. As ρ approaches zero, the ellipses tend to become circular. In contrast, as ρ approaches +1 or –1, the ellipses tend to elongate. The sign of ρ determines the orientation of the ellipses relative to the axes.

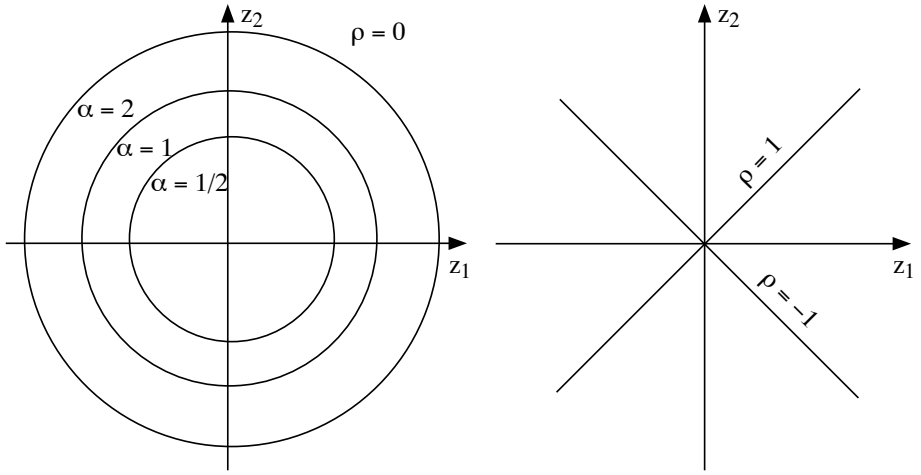


Figure 4.8 Concentration ellipses of a standardized bivariate normal distribution. Extreme values of the correlation ρ.

Actually, when $\rho = 0$ (Fig. 4.8, left), the equation for the family of ellipses becomes:

$$z_1^2 - [2 \times 0 \times z_1 z_2] + z_2^2 = \alpha(1 - 0)$$

or $z_1^2 + z_2^2 = \alpha$, which is the equation of a *circle*.

In contrast, when $\rho = \pm 1$, the equation becomes:

$$z_1^2 - [2 \times (\pm 1) \times z_1 z_2] + z_2^2 = \alpha[1 - (\pm 1)^2]$$

$$z_1^2 \mp 2 z_1 z_2 + z_2^2 = 0$$

hence $[z_1 \mp z_2]^2 = 0$, so that $z_1 \mp z_2 = 0$, and thus $z_1 = \pm z_2$,

which is the equation of a *straight line* with a positive or negative slope of 1 (±45° angle).

Such a family of ellipses, called *concentration ellipses*, is comparable to a series of contour lines on the two-dimensional normal distribution (Fig. 4.6). Increasing the value of α corresponds to moving down along the sides of the distribution. The concentration ellipses pass through points of equal probabilities around the bivariate normal distribution. The role of ρ then becomes clear: when $\rho = 0$, the "bell" of probability densities is perfectly circular (in overhead view); as ρ increases in absolute

value, the "bell" of the probability densities flattens out, until it becomes unidimensional when $\rho = \pm 1$. Indeed, when there is a perfect correlation between two dimensions (i.e. $\rho = \pm 1$), a single dimension, at an angle of 45° with respect to the two original variables, is sufficient to specify the joint distribution of probability densities.

When the number of dimensions is $p = 3$, the family of concentration ellipses becomes a family of concentration *ellipsoids* and, when $p > 3$, a family of *hyperellipsoids*. The meaning of these ellipsoids and hyperellipsoids is the same as in the two-dimensional case although it is not possible to draw them on a sheet of paper.

4.4 Principal axes

Various aspects of the multinormal distribution have been examined in the previous section. One of these, namely the *concentration ellipses* (Fig. 4.7), is the gateway to a topic of great importance for ecologists. In the present section, a method will be developed for determining the *principal axes* of the concentration hyperellipsoids; for simplicity, the term ellipsoid will be used in the following discussion. The *first principal axis* is the line that passes through the dimension of greatest variance of the ellipsoid. The next *principal axes* go through the next dimensions of greatest variance, smaller and smaller, of the p-dimensional ellipsoid. Hence, p consecutive principal axes are determined. These principal axes will be used, in Section 9.1, as the basis for *principal component analysis*.

In the two-dimensional case, the *first principal axis* corresponds to the *major axis* of the concentration ellipse and the *second principal axis* to the *minor axis*. These two axes are perpendicular to each other. Similarly in the p-dimensional case, there are p consecutive axes, which are all perpendicular to one another in the hyperspace.

The first principal axis goes through the p-dimensional centre $\boldsymbol{\mu} = [\mu_1\ \mu_2\ \ldots\ \mu_p]$ of the ellipsoid, and it crosses the surface of the ellipsoid at a point designated here by $\mathbf{y} = [y_1\ y_2\ \ldots\ y_p]$. The values of $\boldsymbol{\mu}$ and $\mathbf{y}$ specify a vector in the p-dimensional space (Section 2.4). The length of the axis, from $\boldsymbol{\mu}$ to the surface of the ellipsoid, is calculated using Pythagoras' formula:

$$[(y_1 - \mu_1)^2 + (y_2 - \mu_2)^2 + \ldots + (y_p - \mu_p)^2]^{1/2} = ([y - \mu][y - \mu]')^{1/2}$$

Actually, this is only half the length of the axis, which extends equally on both sides of $\boldsymbol{\mu}$. The coordinates of the *first* principal axis must be chosen in such a way as to maximize the length of the axis. This can be achieved by maximizing the square of the half-length:

$$[y - \mu][y - \mu]'$$

Calculating coordinates corresponding to the axis with the greatest length is subjected to the constraint that the end point **y** be on the surface of the ellipsoid. This constraint is made explicit using eq. 4.22, which specifies the ellipsoid:

$$[y - \mu]\ \boldsymbol{\Sigma}^{-1}\ [y - \mu]' = \alpha$$

$$[y - \mu]\ \boldsymbol{\Sigma}^{-1}\ [y - \mu]' - \alpha = 0$$

Principal axis

Lagrangian multipliers are used to compute the maximum and minimum values of a function of several variables when the relationships among the variables are known. In the present case, the above two equations, which describe the square of the half-length of the first principal axis and the constraint, are combined into a single function:

$$f(\mathbf{y}) = [y - \mu][y - \mu]' - \lambda\ \{[y - \mu]\ \boldsymbol{\Sigma}^{-1}\ [y - \mu]' - \alpha\}$$

Lagrangian multiplier

Scalar λ is called a *Lagrangian multiplier*[*]. The values that maximize this function are found by the usual method of setting the equation's partial derivative equal to 0:

$$\frac{\partial}{\partial \mathbf{y}} f(\mathbf{y}) = \mathbf{0}$$

$$\frac{\partial}{\partial \mathbf{y}} [y - \mu][y - \mu]' - \lambda\ \frac{\partial}{\partial \mathbf{y}} \{[y - \mu]\ \boldsymbol{\Sigma}^{-1}\ [y - \mu]' - \alpha\} = \mathbf{0}$$

It is important to remember here that **y** is a *p*-dimensional *vector* $(y_1, y_2, \ldots, y_p)$, which means that the above equation is successively derived with respect to y_1, y_2, … and y_p. Therefore, derivation with respect to **y** represents in fact calculating a series of *p* partial derivatives (∂y_j). The results of the derivation may be rewritten as a (column) vector with *p* elements:

$$2\ [y - \mu] - 2\ \lambda\ \boldsymbol{\Sigma}^{-1}\ [y - \mu] = \mathbf{0}$$

One may factor out $[y - \mu]$ and eliminate the constant 2:

$$(\mathbf{I} - \lambda\ \boldsymbol{\Sigma}^{-1})\ [y - \mu] = \mathbf{0}$$

Multiplying both sides of the equation by $\boldsymbol{\Sigma}$ gives:

$$(\boldsymbol{\Sigma} - \lambda \mathbf{I})\ [y - \mu] = \mathbf{0} \qquad \textbf{(4.23)}$$

The general equation defining eigenvectors (eq. 2.22) is $(\mathbf{A} - \lambda \mathbf{I})\ \mathbf{u} = \mathbf{0}$. Replacing, in that equation, **A** by $\boldsymbol{\Sigma}$ and **u** by $[y - \mu]$ produces eq. 4.23. This leads to the conclusion that the *vector of coordinates that specifies the first principal axis is one of the eigenvectors* $[y - \mu]$ *of matrix* $\boldsymbol{\Sigma}$.

[*] After Joseph-Louis Lagrange (1736-1813), mathematician and astronomer.

In order to find out which of the p eigenvectors of $\mathbf{\Sigma}$ is the vector of coordinates of the *first principal axis,* go back to the equation resulting from the partial derivation (above) and transfer the second term to the right, after eliminating the constant 2:

$$[y - \mu] = \lambda\, \mathbf{\Sigma}^{-1}\, [y - \mu]$$

The two sides are then premultiplied by $[y - \mu]'$:

$$[y - \mu]'\, [y - \mu] = \lambda\, [y - \mu]'\, \mathbf{\Sigma}^{-1}\, [y - \mu]$$

Since $[y - \mu]'\, \mathbf{\Sigma}^{-1}\, [y - \mu] = \alpha$ (eq. 4.22), it follows that:

$$[y - \mu]'\, [y - \mu] = \lambda\alpha$$

Considering the first eigenvalue λ_1, the term on the left-hand side of the equation is the square of the half-length of the first principal axis (see above). Thus, for a given value α, the length of the *first principal axis* is maximized by taking the largest possible value for λ or, in other words, the *largest eigenvalue,* λ_1, *of matrix* $\mathbf{\Sigma}$. The vector of coordinates of the *first principal axis* is therefore the eigenvector corresponding to the largest eigenvalue of $\mathbf{\Sigma}$.

Eigenvalue

Numerical example. The above equations are illustrated using the bivariate data matrix from Section 9.1 (principal component analysis). The sample covariance matrix is:

$$\mathbf{S} = \begin{bmatrix} 8.2 & 1.6 \\ 1.6 & 5.8 \end{bmatrix}$$

There are two eigenvalues, $\lambda_1 = 9$ and $\lambda_2 = 5$, computed using eq. 2.23. To normalize the eigenvectors (written as column vectors), put $[y - \mu]'\, [y - \mu] = \lambda\alpha = 1$ for each of them; in other words, $\alpha_1 = 1/9$ and $\alpha_2 = 1/5$. The normalized eigenvectors were called $\mathbf{y}_1$ and $\mathbf{y}_2$ until now in this section; they will be denoted $\mathbf{u}_j$ from now on, as in Sections 2.9 and 2.10. They form matrix $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2]$:

$$\mathbf{y}_1 = \mathbf{u}_1 = \begin{bmatrix} 0.8944 \\ 0.4472 \end{bmatrix} \text{ and } \mathbf{y}_2 = \mathbf{u}_2 = \begin{bmatrix} -0.4472 \\ 0.8944 \end{bmatrix}$$

These eigenvectors are of length 1 since they have been normalized. They determine the *directions* of the major and minor axes of the bivariate distribution. The matrix of eigenvectors $\mathbf{U}$ must be multiplied by the diagonal matrix containing the square roots of the eigenvalues ($\mathbf{U\Lambda}^{1/2}$, eq. 9.10) to provide a new matrix whose columns give the coordinates where the two principal axes cross an ellipsoid with size $\alpha = 1$. This example is further developed in Chapter 9.

To find the vectors of coordinates specifying the p successive principal axes,

• rank the p eigenvalues of matrix $\mathbf{\Sigma}$ in decreasing order:

$$\lambda_1 > \lambda_2 > \ldots > \lambda_p \geq 0$$

Note that the eigenvalues of a matrix $\mathbf{\Sigma}$ are all positive (end of Section 4.1);

• associate the p eigenvectors to their corresponding eigenvalues. The orientation of the p successive principal axes is given by the eigenvectors, which are associated with the p eigenvalues ranked in decreasing order. The eigenvectors of a covariance matrix $\mathbf{\Sigma}$ are orthogonal to one another because $\mathbf{\Sigma}$ is symmetric (Section 2.9). In the case of multiplicity (Section 2.10, Third property), the orthogonal axes may be rotated to an infinity of "principal" directions, i.e. two equal λ's result in a circle and several determine a hypersphere (multidimensional sphere) where no orientation prevails.

The next step consists in calculating a new p-dimensional set of variables, forming matrix $\mathbf{V}$, that position the dispersion ellipses with respect to the principal axes instead of the original Cartesian system. $\mathbf{V}$ is related to the original data matrix $\mathbf{Y}$ (eq. 4.1) through the following transformation:

$$\mathbf{V} = [y - \mu]\,\mathbf{U} \qquad \textbf{(4.24)}$$

where each of the p columns in matrix $\mathbf{U}$ is the normalized eigenvector $\mathbf{u}_k$ corresponding to the k-th principal axis. Because vectors $\mathbf{u}_k$ are both orthogonal and normalized, matrix $\mathbf{U}$ is said to be *orthonormal* (Section 2.8). This transformation results in shifting the origin of the system of axes to the p-dimensional point $\boldsymbol{\mu}$ followed by a rigid rotation of the translated axes into the principal axes (Fig. 4.9), which form matrix $\mathbf{V}$.

The dispersion matrix of $\mathbf{V}$ is:

$$\mathbf{\Sigma}_\mathbf{V} = \frac{1}{(n-1)}\,(\mathbf{V}'\mathbf{V}) \;=\; \frac{1}{(n-1)}\mathbf{U}'\;[y-\mu]'\,[y-\mu]\,\mathbf{U} = \mathbf{U}'\mathbf{\Sigma}\mathbf{U}$$

where $\mathbf{\Sigma}$ is the dispersion matrix of the original matrix $\mathbf{Y}$. So, the *variance* of the k-th dimension $\mathbf{v}_k$ (i.e. the k-th principal axis) is:

$$s^2(\mathbf{v}_k) = \mathbf{u}_k'\,\mathbf{\Sigma}\mathbf{u}_k$$

Since, by definition, $\mathbf{\Sigma}\mathbf{u}_k = \lambda_k\mathbf{u}_k$ (eq. 2.21) and $\mathbf{u}_k'\,\mathbf{u}_k = 1$, it follows that:

$$s^2(\mathbf{v}_k) = \mathbf{u}_k'\,\mathbf{\Sigma}\mathbf{u}_k = \mathbf{u}_k'\,\lambda_k\mathbf{u}_k = \lambda_k\;\mathbf{u}_k'\,\mathbf{u}_k = \lambda_k\,(1) = \lambda_k \qquad \textbf{(4.25)}$$

with $\lambda_k \geq 0$ in all cases since $\mathbf{\Sigma}$ is positive semi-definite. The *covariance* of any two vectors of matrix $\mathbf{V}$ is zero because the product of two orthogonal vectors $\mathbf{u}_k$ and $\mathbf{u}_h$ is zero (Section 2.5):

$$s(\mathbf{v}_k,\mathbf{v}_h) = \mathbf{u}_k'\,\mathbf{\Sigma}\mathbf{u}_h = \mathbf{u}_k'\,\lambda_h\mathbf{u}_h = \lambda_h\;\mathbf{u}_k'\,\mathbf{u}_h = \lambda_k\,(0) = 0 \qquad \textbf{(4.26)}$$

The last two points are of utmost importance, since they are the basis for using the principal axes (and thus principal component analysis; Section 9.1) in ecology: (1) *the variance of a principal axis is equal to the eigenvalue associated with that axis*

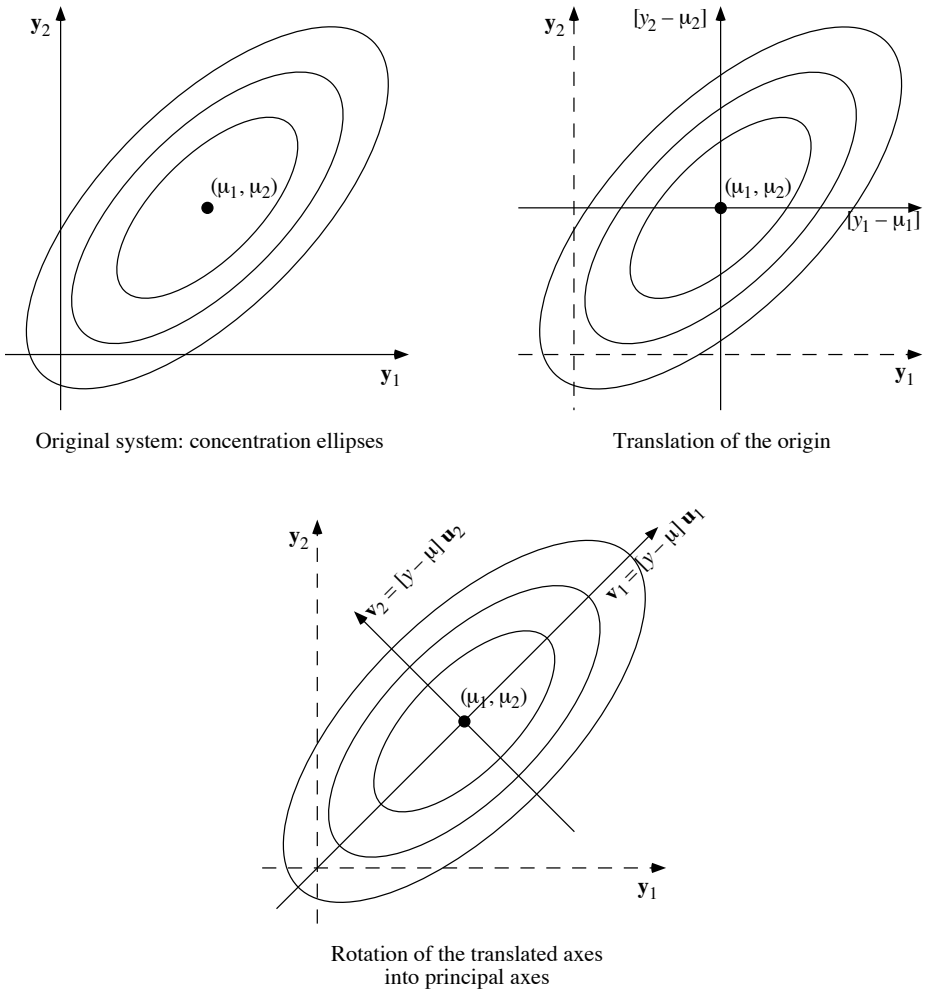


Figure 4.9 Result of the transformation $\mathbf{V} = [y - \mu]\,\mathbf{U}$ (eq. 4.24)**.**

(eq. 4.25) and (2) *the p dimensions of the transformed variable are linearly independent* since their covariances are zero (eq. 4.26).

A last point concerns the meaning of the p elements u_{jk} of the normalized eigenvectors $\mathbf{u}_k$. The values of these elements determine the rotation of the system of axes, so that they correspond to angles. Figure 4.10 illustrates, for the two-dimensional case, how the elements of the eigenvectors are related to the rotation angles. Using the

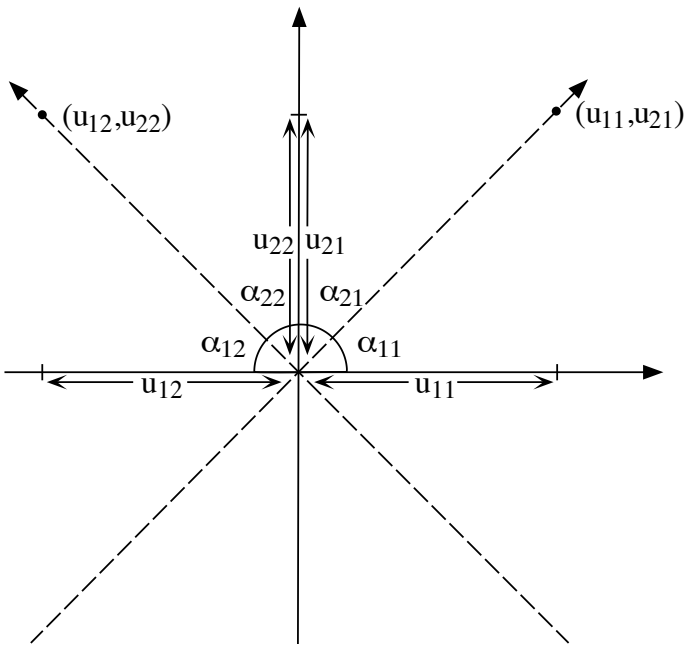


Figure 4.10 Geometrical meaning of the principal axes.

trigonometric functions for right-angled triangles, the angular relationships in Fig. 4.10 may be rewritten as cosines:

$$\cos \alpha_{11} = \text{length } u_{11} / \text{length of vector } (u_{11}, u_{21}) = u_{11}$$

$$\cos \alpha_{21} = \text{length } u_{21} / \text{length of vector } (u_{11}, u_{21}) = u_{21}$$

$$\cos \alpha_{12} = \text{length } u_{12} / \text{length of vector } (u_{12}, u_{22}) = u_{12}$$

$$\cos \alpha_{22} = \text{length } u_{22} / \text{length of vector } (u_{12}, u_{22}) = u_{22}$$

because the lengths of the *normalized* vectors (u_{11}, u_{21}) and (u_{12}, u_{22}) are 1 (Section 2.4). Eigenvector $\mathbf{u}_k$ determines the direction of the k-th main axis; it follows from the above trigonometric relationships that elements u_{jk} of the normalized eigenvectors are *direction cosines*. Each direction cosine specifies the angle between an original Cartesian axis j and a principal axis k.

Direction cosine

The two-dimensional case, illustrated in Figs. 4.9 and 4.10, is the simplest to compute. The standardized dispersion matrix is of the general form:

$$\mathbf{P} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

When ρ is positive, the eigenvalues of $\mathbf{P}$ are $\lambda_1 = (1 + \rho)$ and $\lambda_2 = (1 - \rho)$. The *normalized* eigenvectors are:

$$\mathbf{u}_1 = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} \qquad \mathbf{u}_2 = \begin{bmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$$

Therefore, the first principal axis goes through the point $(1/\sqrt{2}\ ,\ 1/\sqrt{2}\ )$, so that it cuts the first and third quadrants at a 45° angle. Its direction cosines are $\cos \alpha_{11} = 1/\sqrt{2}$ and $\cos \alpha_{12} = 1/\sqrt{2}$, which indeed specify 45° angles with respect to the two axes of the first quadrant. The second principal axis goes through $(-1/\sqrt{2}\ ,\ 1/\sqrt{2}\ )$, so that it cuts the second and fourth quadrants at 45°. Its direction cosines are $\cos \alpha_{21} = -1/\sqrt{2}$ and $\cos \alpha_{22} = 1/\sqrt{2}$, which determine 45°angles with respect to the two axes of the second quadrant.

When ρ is *negative,* the eigenvalues of $\mathbf{P}$ are $\lambda_1 = (1 - \rho)$ and $\lambda_2 = (1 + \rho)$. Consequently the first principal axis goes through $(-1/\sqrt{2}\ ,\ 1/\sqrt{2}\ )$ in the second quadrant, while the second principal axis with coordinates $(1/\sqrt{2}\ ,\ 1/\sqrt{2}\ )$ cuts the first quadrant. A value $\rho = 0$ entails a case of multiplicity since $\lambda_1 = \lambda_2 = 1$. This results in an infinite number of "principal" axes, i.e. any two perpendicular diameters would fit the circular concentration ellipse (Fig. 4.8, left).

These concepts, so far quite abstract, will find direct applications to ecology in Section 9.1, where principal component analysis is described.

4.5 Multiple and partial correlations

Section 4.2 considered, in a multidimensional context, the correlation of pairs of variables, which represent two dimensions of a *p*-dimensional random variable. However, the multidimensional nature of ecological data allows other approaches to correlation analysis. These statistics are examined in the present section and compared graphically in Box 4.1

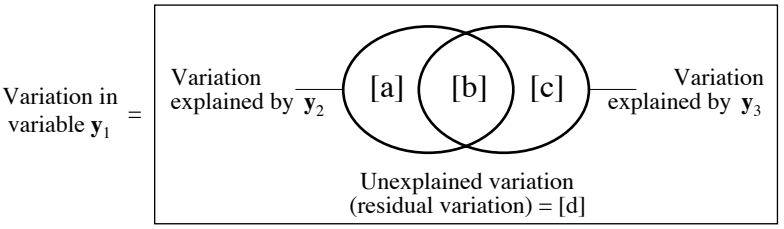
The following developments will require that the *p*-dimensional correlation matrix $\mathbf{R}$ be partitioned into four submatrices. Indices assigned to the submatrices follow the general convention on matrix indices (Section 2.1):

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{21} & \mathbf{R}_{22} \end{bmatrix} \qquad \textbf{(4.27)}$$

Variation partitioning — Box 4.1

Variation partitioning, which is described in detail in Subsection 10.3.5, provides a general framework to illustrate the similarities and differences between the coefficient of multiple determination and the partial and semipartial correlation coefficients, as well as the corresponding F-statistics.

Three variables only, $\mathbf{y}_1$, $\mathbf{y}_2$, and $\mathbf{y}_3$, are considered in this example. In the following Venn diagram, the rectangle represents the total sum of squares of variable $\mathbf{y}_1$:



In the multiple regression of $\mathbf{y}_1$ on $\mathbf{y}_2$ and $\mathbf{y}_3$, $\hat{y}_1 = b_0 + b_2 y_2 + b_3 y_3$ (this is an application of eq. 10.15), the coefficient of multiple determination, which is the square of the coefficient of multiple correlation, is:

$$R^2_{1.23} = \frac{[a + b + c]}{[a + b + c + d]} \qquad \text{with} \quad F = \frac{[a + b + c]/2}{[d]/(n-3)}$$

The partial correlation of $\mathbf{y}_1$ with $\mathbf{y}_2$ while controlling for the effect of $\mathbf{y}_3$ is:

$$r_{12.3} = \sqrt{\frac{[a]}{[a + d]}} \qquad \text{with} \quad F = \frac{[a]/1}{[d]/(n-3)}$$

The semipartial correlation of $\mathbf{y}_1$ with $\mathbf{y}_2$ in the presence of $\mathbf{y}_3$ is:

$$r_{1(2.3)} = \sqrt{\frac{[a]}{[a+b+c+d]}} \qquad \text{with} \quad F = \frac{[a]/1}{[d]/(n-3)}$$

The coefficients of partial and semipartial correlation receive the same sign as the corresponding coefficient of partial regression.

The test of a partial regression coefficient, b_2 or b_3, is the same (i.e. it has the same F-statistic) as the test of the corresponding partial correlation coefficient, $r_{12.3}$ or $r_{13.2}$. The F-statistic is always the ratio of two *independent* portions of the variation of $\mathbf{y}_1$, each one divided by its degrees of freedom; see eqs. 4.39 and 4.40.

There are two possible approaches to linear correlation involving several variables or several dimensions of a multidimensional variable. The first one, which is called *multiple* (*linear*) *correlation*, measures the intensity of the relationship between a *response* variable and a linear combination of several *explanatory* variables. The second approach, called *partial* (*linear*) *correlation*, measures the intensity of the linear relationship between two variables, while taking into account their relationships with other variables.

1 — Multiple linear correlation

Multiple correlation applies to cases where there is one *response* variable and several *explanatory* variables. This situation is further studied in Section 10.3, within the context of *multiple regression*. The *coefficient of multiple determination* (R^2; eq. 10.20) measures the fraction of the variance of $\mathbf{y}_k$ that is explained by a linear combination of $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_j, \ldots$ and $\mathbf{y}_p$:

$$R^2_{k.12..j..p} = \frac{b_1 s_{1k} + b_2 s_{2k} + \ldots + b_j s_{jk} + \ldots + b_p s_{pk}}{s^2_k} \qquad \textbf{(4.28)}$$

where p is here the number of explanatory variables.The concept is illustrated in Box 4.1. In eq. 4.28, coefficients b are the coefficients of the multiple regression (Subsection 10.3.3) of $\mathbf{y}_k$ on the explanatory variables. A coefficient $R^2_{k.12..j..p} = 0.73$, for example, means that the linear relationships of variables $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_j, \ldots$ and $\mathbf{y}_p$ with $\mathbf{y}_k$ explain 73% of the variation of $\mathbf{y}_k$ around its mean. The *multiple correlation coefficient* (R) is the square root of the coefficient of multiple determination:

$$R_{k.12..j..p} = \sqrt{R^2_{k.12..j..p}} \qquad \textbf{(4.29)}$$

To calculate R^2 using matrix algebra, a correlation matrix $\mathbf{R}$ is written for variables $\mathbf{y}_k$ and $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_j, \ldots, \mathbf{y}_p\}$, with $\mathbf{y}_k$ in the first position. Partitioning this matrix following eq. 4.27 to compute a multiple correlation coefficient gives:

$$\mathbf{R} = \left[\begin{array}{c|ccccc} 1 & r_{k1} & r_{k2} & \cdots & r_{kp} \\ \hline r_{1k} & 1 & r_{12} & \cdots & r_{1p} \\ r_{2k} & r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ r_{pk} & r_{p1} & r_{p2} & \cdots & 1 \end{array}\right] = \left[\begin{array}{cc} 1 & \mathbf{r}_{12} \\ \mathbf{r}_{21} & \mathbf{R}_{22} \end{array}\right] \qquad \textbf{(4.30)}$$

where $\mathbf{r}_{12} = \mathbf{r}'_{21}$ is a vector containing the correlation coefficients $r_{k1}, r_{k2}, \ldots, r_{kp}$. Using $\mathbf{r}_{12}, \mathbf{r}_{21}$ and $\mathbf{R}_{22}$ as defined in eq. 4.30, R^2 is calculated as:

$$R^2 = \mathbf{r}_{12}\mathbf{R}_{22}^{-1}\mathbf{r}_{21} = \mathbf{r}'_{21}\mathbf{R}_{22}^{-1}\mathbf{r}_{21} \qquad \textbf{(4.31)}$$

Equation 4.31 is expanded using eq. 2.17:

$$R^2 = \mathbf{r}'_{21}\mathbf{R}_{22}^{-1}\mathbf{r}_{21} = \mathbf{r}'_{21}\frac{1}{|\mathbf{R}_{22}|}\begin{bmatrix} \text{cof}(r_{11}) & \text{cof}(r_{21}) & \ldots & \text{cof}(r_{p1}) \\ \text{cof}(r_{12}) & \text{cof}(r_{22}) & \ldots & \text{cof}(r_{p2}) \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \text{cof}(r_{1p}) & \text{cof}(r_{2p}) & \ldots & \text{cof}(r_{pp}) \end{bmatrix}\mathbf{r}_{21}$$

$$R^2 = \frac{1}{|\mathbf{R}_{22}|}\left(|\mathbf{R}_{22}| - |\mathbf{R}|\right) = 1 - \frac{|\mathbf{R}|}{|\mathbf{R}_{22}|} \qquad \textbf{(4.32)}$$

As an exercise, it is easy to check that

$$|\mathbf{R}_{22}| - |\mathbf{R}| = \mathbf{r}'_{21}\ [\text{adjugate matrix of } \mathbf{R}_{22}]\ \mathbf{r}_{21}$$

Multiple correlation

The *coefficient of multiple correlation* is calculated from eqs. 4.31 or 4.32:

$$R_{k.12\ldots j\ldots p} = \sqrt{\mathbf{r}'_{21}\mathbf{R}_{22}^{-1}\mathbf{r}_{21}} \quad \text{or} \quad R_{k.12\ldots j\ldots p} = \sqrt{1 - \frac{|\mathbf{R}|}{|\mathbf{R}_{22}|}} \qquad \textbf{(4.33)}$$

A third way of calculating R^2 is described in eq. 4.38, near the end of Subsection 4.5.2 on partial correlation.

When two or more variables in matrix $\mathbf{R}_{22}$ are perfectly correlated (i.e. $r = 1$ or $r = -1$), the rank of $\mathbf{R}_{22}$ is smaller than its order (Section 2.7), hence $|\mathbf{R}_{22}| = 0$. Calculation of R thus requires the elimination of redundant variables from matrix $\mathbf{R}$.

Numerical example. A simple example, with three variables ($\mathbf{y}_1$, $\mathbf{y}_2$ and $\mathbf{y}_3$), illustrates the above equations. Matrix $\mathbf{R}$ is:

$$\mathbf{R} = \begin{bmatrix} 1 & 0.4 & 0.8 \\ 0.4 & 1 & 0.5 \\ 0.8 & 0.5 & 1 \end{bmatrix}$$

The *coefficient of multiple determination* $R^2_{1.23}$ is first calculated using eq. 4.31:

$$R^2_{1.23} = \begin{bmatrix}0.4 & 0.8\end{bmatrix}\begin{bmatrix}1 & 0.5\\ 0.5 & 1\end{bmatrix}^{-1}\begin{bmatrix}0.4\\ 0.8\end{bmatrix}$$

$$R^2_{1.23} = \begin{bmatrix}0.4 & 0.8\end{bmatrix}\begin{bmatrix}1.33 & -0.67\\ -0.67 & 1.33\end{bmatrix}\begin{bmatrix}0.4\\ 0.8\end{bmatrix}$$

$$R^2_{1.23} = 0.64$$

Equation 4.32 leads to an identical result:

$$R^2_{1.23} = 1 - \frac{\begin{vmatrix}1 & 0.4 & 0.8\\ 0.4 & 1 & 0.5\\ 0.8 & 0.5 & 1\end{vmatrix}}{\begin{vmatrix}1 & 0.5\\ 0.5 & 1\end{vmatrix}}$$

$$R^2_{1.23} = 1 - \frac{0.27}{0.75} = 0.64$$

The linear combination of variables $\mathbf{y}_2$ and $\mathbf{y}_3$ explains 64% of the variance of $\mathbf{y}_1$. The *multiple correlation coefficient* is $R_{1.23} = 0.8$.

2 — Partial correlation

The second approach to correlation, in the multidimensional context, applies to situations where the relationship between two variables is influenced by their relationships with other variables. Two coefficients are described in Box 4.1: the *partial* and *semipartial correlation coefficients*.

The *partial correlation coefficient* is related to partial multiple regression (Subsection 10.3.5). It measures what the correlation between y_j and y_k would be if other variables y_1, y_2, … and y_p, hypothesized to influence both y_j and y_k, were held constant at their means. The partial correlation between variables y_j and y_k, when controlling for their relationships with y_1, y_2, … and y_p, is written $r_{jk.12\ldots p}$.

In order to calculate the partial correlation coefficients, the set of variables is divided into two subsets. The *first subset* contains the variables between which the partial correlation is to be computed while controlling for the influence of the variables

in the second subset. The *second subset* thus contains the variables whose influence is to be taken into account. Matrix $\mathbf{R}$ is partitioned as follows (eq. 4.27):

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{21} & \mathbf{R}_{22} \end{bmatrix}$$

$\mathbf{R}_{11}$ (of order 2 × 2 for partial correlations) and $\mathbf{R}_{22}$ contain the correlations among variables in the first and the second subsets, respectively, whereas $\mathbf{R}_{12}$ and $\mathbf{R}_{21}$ both contain the correlations between variables of the two subsets; $\mathbf{R}_{12} = \mathbf{R}'_{21}$.

The number of variables in the second subset determines the *order* of the partial correlation coefficient. This order is the number of variables whose effects are eliminated from the correlation between $\mathbf{y}_j$ and $\mathbf{y}_k$. For example $r_{12.345}$ (third-order partial correlation coefficient) means that the correlation between variables $\mathbf{y}_1$ and $\mathbf{y}_2$ is calculated while controlling for the linear effects of $\mathbf{y}_3$, $\mathbf{y}_4$, and $\mathbf{y}_5$.

The computation consists in subtracting from $\mathbf{R}_{11}$ (correlation matrix among the variables in the first subset) a second matrix containing the coefficients of multiple determination of the variables in the second subset on those in the first subset. These coefficients measure the fraction of the variance and covariance of the variables in the first subset that is explained by linear combinations of the variables in the second subset. They are computed by replacing in eq. 4.31 vector $\mathbf{r}_{21}$ by submatrix $\mathbf{R}_{21}$:

$$\mathbf{R}_{12}\mathbf{R}_{22}^{-1}\mathbf{R}_{21} = \mathbf{R}'_{21}\mathbf{R}_{22}^{-1}\mathbf{R}_{21}$$

Subtracting this expression from $\mathbf{R}_{11}$ gives the *matrix of conditional correlations*:

$$\text{Matrix of conditional correlations} = \mathbf{R}_{11} - \mathbf{R}_{12}\mathbf{R}_{22}^{-1}\mathbf{R}_{21} \qquad \textbf{(4.34)}$$

It can be shown that the maximum likelihood estimate ($\mathbf{R}_{1.2}$) of the partial correlation matrix $\mathbf{P}_{1.2}$ is:

$$\mathbf{R}_{1.2} = \mathbf{D}\,(r_{1.2})^{-1/2}\,(\mathbf{R}_{11} - \mathbf{R}_{12}\mathbf{R}_{22}^{-1}\mathbf{R}_{21})\,\mathbf{D}\,(r_{1.2})^{-1/2} \qquad \textbf{(4.35)}$$

where $\mathbf{D}(r_{1.2})$ is the matrix of diagonal elements of the matrix of conditional correlation (eq. 4.34).

The calculation is illustrated for the three-dimensional case, in which there is a single controlled variable $\mathbf{y}_3$:

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{bmatrix}$$

This development will provide the algebraic formula for the partial correlation coefficients of order 1. Coefficients pertaining to variables of the first subset ($\mathbf{y}_1$ and $\mathbf{y}_2$) are in the first two rows and columns. Using eq. 4.35 gives:

$$\mathbf{R}_{12}\mathbf{R}_{22}^{-1}\mathbf{R}_{21} = \begin{bmatrix} r_{13} \\ r_{23} \end{bmatrix} \begin{bmatrix} 1 \end{bmatrix}^{-1} \begin{bmatrix} r_{31} & r_{32} \end{bmatrix} = \begin{bmatrix} r_{13}^2 & r_{13}r_{23} \\ r_{13}r_{23} & r_{23}^2 \end{bmatrix}$$

$$\mathbf{R}_{11} - \mathbf{R}_{12}\mathbf{R}_{22}^{-1}\mathbf{R}_{21} = \begin{bmatrix} 1 & r_{12} \\ r_{21} & 1 \end{bmatrix} - \begin{bmatrix} r_{13}^2 & r_{13}r_{23} \\ r_{13}r_{23} & r_{23}^2 \end{bmatrix} = \begin{bmatrix} (1-r_{13}^2) & (r_{12}-r_{13}r_{23}) \\ (r_{12}-r_{13}r_{23}) & (1-r_{23}^2) \end{bmatrix}$$

$$\mathbf{R}_{1.2} = \begin{bmatrix} 1/\sqrt{1-r_{13}^2} & 0 \\ 0 & 1/\sqrt{1-r_{23}^2} \end{bmatrix} \begin{bmatrix} (1-r_{13}^2) & (r_{12}-r_{13}r_{23}) \\ (r_{12}-r_{13}r_{23}) & (1-r_{23}^2) \end{bmatrix} \begin{bmatrix} 1/\sqrt{1-r_{13}^2} & 0 \\ 0 & 1/\sqrt{1-r_{23}^2} \end{bmatrix}$$

$$\mathbf{R}_{1.2} = \begin{bmatrix} 1 & \dfrac{r_{12}-r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}} \\ \dfrac{r_{12}-r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}} & 1 \end{bmatrix} = \begin{bmatrix} 1 & r_{12.3} \\ r_{12.3} & 1 \end{bmatrix}$$

The previous matrix equation provides the formula for the first-order partial correlation coefficient:

$$r_{12.3} = \frac{r_{12}-r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}} \qquad \textbf{(4.36)}$$

The general formula, for coefficients of order p, is:

$$r_{jk.1\ldots p} = \frac{r_{jk.1\ldots(p-1)} - r_{jp.1\ldots(p-1)}r_{kp.1\ldots(p-1)}}{\sqrt{1-r_{jp.1\ldots(p-1)}^2}\sqrt{1-r_{kp.1\ldots(p-1)}^2}} \qquad \textbf{(4.37)}$$

When there are four variables, it is possible to calculate 12 first-order and 6 second-order partial correlation coefficients. Computing a second-order coefficient necessitates the calculation of 3 first-order coefficients. For example:

$$r_{12.34} = \frac{r_{12.3}-r_{14.3}r_{24.3}}{\sqrt{1-r_{14.3}^2}\sqrt{1-r_{24.3}^2}} = r_{12.43} = \frac{r_{12.4}-r_{13.4}r_{23.4}}{\sqrt{1-r_{13.4}^2}\sqrt{1-r_{23.4}^2}}$$

It is thus possible, as the number of variables increases, to calculate higher-order coefficients. Computing a coefficient of a given order requires the calculation of three

coefficients of the previous order, each of these requiring the calculation of coefficients of the previous order, and so on depending on the number of variables involved. Such a cascade of calculations is advantageously replaced by the direct matrix approach of eq. 4.35.

Numerical example. Partial correlations are calculated on the simple example already used for multiple correlation. Matrix **R** is:

$$\mathbf{R} = \begin{bmatrix} 1 & 0.4 & 0.8 \\ 0.4 & 1 & 0.5 \\ 0.8 & 0.5 & 1 \end{bmatrix}$$

Two subsets are formed, the first one containing descriptors $\mathbf{y}_1$ and $\mathbf{y}_2$ (between which the partial correlation is computed) and the second one $\mathbf{y}_3$ (whose influence on r_{12} is controlled for). Computations follow eqs. 4.34 and 4.35:

eq. 4.34 Matrix of conditional correlations $= \begin{bmatrix} 1 & 0.4 \\ 0.4 & 1 \end{bmatrix} - \begin{bmatrix} 0.8 \\ 0.5 \end{bmatrix} \begin{bmatrix} 1 \end{bmatrix}^{-1} \begin{bmatrix} 0.8 & 0.5 \end{bmatrix}$

$$= \begin{bmatrix} 1 & 0.4 \\ 0.4 & 1 \end{bmatrix} - \begin{bmatrix} 0.64 & 0.40 \\ 0.40 & 0.25 \end{bmatrix} = \begin{bmatrix} 0.36 & 0 \\ 0 & 0.75 \end{bmatrix}$$

eq. 4.35 $$\mathbf{R}_{1.2} = \begin{bmatrix} 1.67 & 0 \\ 0 & 1.15 \end{bmatrix} \begin{bmatrix} 0.36 & 0 \\ 0 & 0.75 \end{bmatrix} \begin{bmatrix} 1.67 & 0 \\ 0 & 1.15 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Thus, the partial correlation $r_{12.3} = 0$; this was unexpected given that $r_{12} = 0.4$. In other words, fraction [a] displayed in Box 4.1 is 0. The conclusion is that, when their (linear) relationships with $\mathbf{y}_3$ are taken into account, descriptors $\mathbf{y}_1$ and $\mathbf{y}_2$ are (linearly) independent. Similar calculations for the other two pairs of descriptors give: $r_{13.2} = 0.76$ and $r_{23.1} = 0.33$. The interpretation of these correlation coefficients will be further discussed in Subsection 4.5.4.

There is a relationship between the coefficients of *multiple* and *partial* correlation. The equation linking the two types of coefficients can be easily derived; in the multiple correlation equation, p is the number of variables other than $\mathbf{y}_k$:

when $p = 1$, the fraction of the variance of $\mathbf{y}_k$ not explained by $\mathbf{y}_1$ is the complement of the coefficient of determination $(1 - r_{k1}^2)$; this expression is sometimes called the *coefficient of nondetermination*;

Nondetermination

when $p = 2$, the fraction of the variance of $\mathbf{y}_k$ not explained by $\mathbf{y}_2$, while controlling for the linear influence of $\mathbf{y}_1$, is $(1 - r_{k2.1}^2)$, so that the fraction of the variance of $\mathbf{y}_k$ not explained by $\mathbf{y}_1$ and $\mathbf{y}_2$ is $(1 - r_{k1}^2)\,(1 - r_{k2.1}^2)$.

This leads to a general expression for the fraction of the variance of $\mathbf{y}_k$ that is not explained by $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_j, \ldots$ and $\mathbf{y}_p$:

$$(1 - r_{k1}^2)\ (1 - r_{k2.1}^2) \ldots (1 - r_{kj.12\ldots}^2) \ldots (1 - r_{kp.12\ldots j\ldots(p-1)}^2)$$

Multiple determination

The fraction of the variance of $\mathbf{y}_k$ that is explained by $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_j, \ldots$ and $\mathbf{y}_p$, i.e. the *coefficient of multiple determination* (square of the *multiple correlation*), is thus:

$$R_{k.12\ldots p}^2 = 1 - [\,(1 - r_{k1}^2)\ (1 - r_{k2.1}^2) \ldots (1 - r_{kp.12\ldots p-1}^2)\,] \qquad \textbf{(4.38)}$$

Numerical example. The same example as above is used to illustrate the calculation of the multiple correlation coefficient, using eq. 4.38:

$$R_{1.23}^2 = 1 - [\,(1 - r_{12}^2)\ (1 - r_{13.2}^2)\,]$$

$$R_{1.23}^2 = 1 - [\,1 - (0.4)^2\,]\,[\,1 - (0.76)^2\,] = 0.64$$

which is identical to the result obtained in Subsection 4.5.1 using eqs. 4.31 and 4.32.

Like the partial correlation, the *semipartial correlation coefficient* measures the correlation between $\mathbf{y}_j$ and $\mathbf{y}_k$ while controlling for the linear effect of other variables $\mathbf{y}_1, \mathbf{y}_2, \ldots$ and $\mathbf{y}_p$. The difference is in the denominator, which is the total variation in the response variable, i.e. the quantity [a+b+c+d] in Box 4.1. The formula for the first-order semipartial correlation coefficient is:

$$r_{1(2.3)} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{23}^2}}$$

The value of $r_{1(2.3)}$ is 0 for the numerical example because [a] = 0. The *semipartial correlation* can also be calculated as the square root of the difference between two multiple determination coefficients:

$$r_{1(2.3)} = \sqrt{R_{1.23}^2 - R_{1.3}^2}$$

Because the latter equation does not specify the sign of the semipartial correlation coefficient, the previous equation must be used to obtain that sign, which is the same as the sign of the partial regression coefficient. In the Venn diagram of Box 4.1, $R_{1.23}^2$ is the union of the two ellipses or the quantity [a+b+c], whereas $R_{1.3}^2$ is the right-hand ellipse or the quantity [b+c], each of these quantities being divided by the total variation (total sum of squares) in the response variable, [a+b+c+d]. Hence $R_{1\,(2.3)}^2$ is ([a+b+c]−[b+c])/[a+b+c+d], or [a]/[a+b+c+d]. The semipartial correlation coefficient is especially useful in variation partitioning (Subsection 10.3.5) because it expresses all fractions of variation with respect to the same common denominator, which is the total sum of squares in the response variable [a+b+c+d].

Table 4.6 Main properties of the multiple (linear) correlation coefficient.

	Properties	Sections
1.	The multiple correlation coefficient measures the *intensity of the relationship* between a *response* variable and a *linear* combination of several *explanatory* variables.	4.5
2.	The square of the multiple correlation coefficient, called *coefficient of multiple determination*, measures the fraction of the variance of the response variable that is explained by a linear combination of the explanatory variables.	4.5
3.	The coefficient of multiple determination is the extension, to the multidimensional case, of the *coefficient of determination between two variables*.	4.5 and 10.3
4.	The multiple correlation coefficient can be computed from the matrix of correlations among *explanatory* variables and the vector of correlations between the *explanatory* and *response* variables.	4.5
5.	The multiple correlation coefficient can be computed from the determinant of the matrix of correlations among the *explanatory* variables and that of the matrix of correlations among all variables involved.	4.5
6.	The multiple correlation coefficient can be computed from the product of a series of *complements of coefficients of partial determination*.	4.5

Tables 4.6 and 4.7 summarize the main conclusions relative to the coefficients of multiple and partial correlation, respectively.

3 — Tests of statistical significance

In correlation analysis, the null hypothesis H_0 is usually that the correlation coefficient is equal to zero (i.e. independence of the descriptors). One can also test the hypothesis that ρ has some particular value other than zero. The general formula for testing correlation coefficients (for H_0: $\rho = 0$) is:

$$F = \frac{r_{jk}^2 / \nu_1}{(1 - r_{jk}^2) / \nu_2} \qquad \textbf{(4.39)}$$

with $\nu_1 = m$ and $\nu_2 = n - m - 1$, where m is the number of variables correlated to j. This F-statistic is compared to the critical value $F_{\alpha[\nu_1,\nu_2]}$. In the case of the simple correlation coefficient, where $m = 1$ (there is a single variable correlated to j), eq. 4.39 becomes eq. 4.12.

Table 4.7 Main properties of the partial (linear) correlation coefficient. One of these properties is discussed in a later chapter.

	Properties	Sections
1.	The partial and semipartial correlation coefficients measure the *intensity of the linear relationship* between two random variables while taking into account their relationships with other variables.	4.5
2.	The difference between the partial and semipartial correlation coefficients is in the denominator, which excludes the variation of the controlled variables in the partial correlation but not in the semipartial correlation.	4.5
3.	The partial correlation coefficient can be computed from the submatrix of correlations among the variables *in partial relationship* (first subset), the submatrix of variables that *influence* the first subset, and the submatrix of correlations between the *two subsets* of variables.	4.5
4.	The partial and semipartial correlation coefficients can be computed from the *coefficients of simple correlation* between all pairs of variables involved.	4.5
5.	The square of the partial correlation coefficient (*coefficient of partial determination*; name seldom used) measures the fraction of the total variance of each variable that is mutually explained by the other, the influence of some other variables being taken into account.	10.3

In regression analysis, the null hypothesis is that the coefficient of multiple determination (R^2) is zero. To test the *coefficient of multiple determination* R^2 and the *multiple correlation coefficient* R, the F-statistic is:

$$F = \frac{R^2_{1.2\ldots p} / \nu_1}{(1 - R^2_{1.2\ldots p}) / \nu_2} \qquad \textbf{(4.40)}$$

with $\nu_1 = m$ and $\nu_2 = n - m - 1$, where m is the number of explanatory variables; $m = p - 1$ in the notation of eq. 4.40.

Partial correlation coefficients are tested in the same way as coefficients of simple correlation (eq. 4.12 for the F-test and eq. 4.13 for the t-test, where $\nu = n - 2$), except that one additional degree of freedom is lost for each successive *order* of the coefficient, or each *covariable* in the model. For example, the number of degrees of freedom for $r_{jk.123}$ (third-order partial correlation coefficient) is $\nu = (n - 2) - 3 = n - 5$.

This is the same as counting $\nu = n - m - 1$, where m is the number of variables in the model besides j. For partial correlations, eqs. 4.12 and 4.13 become respectively:

$$F = \nu \frac{r^2_{jk.1\ldots p}}{1 - r^2_{jk.1\ldots p}} \quad \textbf{(4.12)} \qquad \text{and} \qquad t = \sqrt{\nu} \frac{r_{jk.1\ldots p}}{\sqrt{1 - r^2_{jk.1\ldots p}}} \quad \textbf{(4.13)}$$

The number of covariables will be called q in Subsections 10.3.5 and 11.1.7 which describe, respectively, the tests of significance in partial regression and partial canonical analysis. *Semipartial correlation coefficients* are tested using the same F-statistic as for partial correlations, as shown in Box 4.1. As usual (Sections 1.2 and 4.2), H_0 is tested either by comparing the computed statistic (F or t) to a critical value found in a table for a predetermined significance level α, or by computing the probability associated with the computed statistic.

4 — Causal modelling using correlations

In the ecological literature, correlation coefficients are often interpreted in terms of causal relationships among descriptors. That should never be done when the only information available is that provided by the correlation coefficients themselves.

Causality

In statistics, "causality" refers to the hypothesis that changes occurring in one variable cause changes in another variable; *causality resides in the hypotheses only.* Within the framework of a given sampling design (i.e. spatial, temporal, or experimental) where variation is controlled, data are said to support the causality hypothesis if a significant portion of the variation in **b** is explained by changes taking place in **a**. If the relationship is assumed to be linear, a significant linear correlation coefficient is interpreted as supporting the hypothesis of linear causation.

Let us consider the simple case of three linearly related variables $\mathbf{y}_1$, $\mathbf{y}_2$, and $\mathbf{y}_3$. In the following paragraphs, these variables will be noted **a**, **b**, and **c** for simplicity. A simple form of causal modelling is obtained by looking at the simple and partial correlation coefficients between these variables, following the pioneering work of De Neufville & Stafford (1971). One basic condition must be fulfilled for such a model to encompass the three variables; it is that at least two of the simple correlations be significantly different from zero. Under the assumption of linear relationships among variables, these two coefficients provide statistical support for two "causal arrows".

Causal model

There are four elementary models describing the possible interactions among three variables (Fig. 4.11), each with possible permutations of **a**, **b** and **c**, for a total of 18 distinguishable models. These four elementary *causal models* show how difficult it is to interpret correlation matrices, especially when several ecological descriptors are interacting in complex ways. Partial correlations may be used to elucidate the relationships among descriptors. However, the choice of a causal model always requires hypotheses, or else the input of external ecological information. When it is possible, from a priori information or ecological hypotheses, to specify the causal

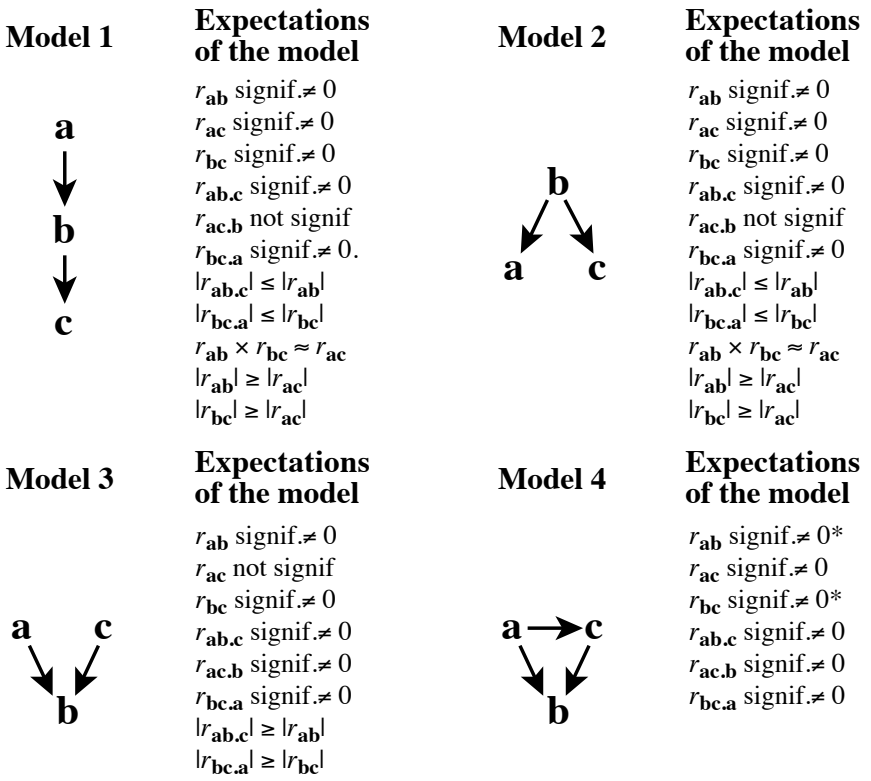


Figure 4.11 Predictions of the four possible models of causal relationships involving three variables, in terms of the expected values for the simple and partial linear correlation coefficients. 'r_{ab} signif.≠ 0' means that, under the model, the correlation must be significantly different from zero. 'r_{ab} not signif.' means that the correlation is not necessarily significantly different from zero at the pre-selected significance level. * Model 4 holds even if one, *but only one*, of these two simple correlation coefficients is not significant. Adapted from Legendre (1993).

ordering among descriptors, path analysis (Section 10.4) may be used to assess the correspondence between the data (i.e. correlations) and causal models. Note again that a causal model may never be derived from a correlation matrix, whereas a causal model is required to interpret a correlation matrix in terms of causality.

In Fig. 4.11, model 1 describes a *causal chain*, with six possible permutations of **a**, **b** and **c**, and model 2 is the *double effect* model with three distinguishable permutations: each of the three variables may be at the origin of the two arrows. Model 3 is the *double cause* model, with three distinguishable permutations. Model 4 describes a *triangular relationship* with six possible permutations; it may be seen as a combination of models 1 and 2 or 1 and 3. The direct and indirect effects implied in

model 4 may be further analysed using path analysis (Section 10.4), which requires precise hypotheses about arrow directions.

In Fig. 4.11, the predictions of the four models were obtained by numerical simulations. Examining model 1 in some detail illustrates how the "expectations of the model" can also be derived analytically.

• Significance of the simple correlations. Obviously, $r_{\mathbf{ab}}$ and $r_{\mathbf{bc}}$ must be significantly different from zero for the model to hold. The model can accommodate $r_{\mathbf{ac}}$ being significant or not, although the value of $r_{\mathbf{ac}}$ should always be different from zero since $r_{\mathbf{ac}} = r_{\mathbf{ab}} r_{\mathbf{bc}}$.

• Significance of the partial correlations. The condition $r_{\mathbf{ac}} = r_{\mathbf{ab}} r_{\mathbf{bc}}$ implies that $r_{\mathbf{ac}} - r_{\mathbf{ab}} r_{\mathbf{bc}} = 0$ or, in other words (eq. 4.36), $r_{\mathbf{ac.b}} = 0$. For the model to hold, partial correlations $r_{\mathbf{ab.c}}$ and $r_{\mathbf{bc.a}}$ must be significantly different from 0. Indeed, $r_{\mathbf{ab.c}}$ being equal to zero would mean that $r_{\mathbf{ab}} = r_{\mathbf{ac}} r_{\mathbf{bc}}$, which would imply that **c** is in the centre of the sequence; this is not the case in the model as specified, where **b** is in the centre. The same reasoning explains the relationship $r_{\mathbf{bc.a}} \neq 0$.

• Comparison of simple correlation values. Since correlation coefficients are smaller than or equal to 1 in absolute value, the relationship $r_{\mathbf{ac}} = r_{\mathbf{ab}} r_{\mathbf{bc}}$ implies that $|r_{\mathbf{ab}}| \geq |r_{\mathbf{ac}}|$ and $|r_{\mathbf{bc}}| \geq |r_{\mathbf{ac}}|$.

• Comparison of partial correlation values. Consider the partial correlation formula for $r_{\mathbf{ab.c}}$ (eq. 4.36). Is it true that $|r_{\mathbf{ab.c}}| \leq |r_{\mathbf{ab}}|$? The relationship $r_{\mathbf{ac}} = r_{\mathbf{ab}} r_{\mathbf{bc}}$ allows one to replace $r_{\mathbf{ac}}$ by $r_{\mathbf{ab}} r_{\mathbf{bc}}$ in that equation. After a few lines of algebra, the following inequality

$$|r_{\mathbf{ab.c}}| = \frac{|r_{\mathbf{ab}}| \, [1 - r_{\mathbf{bc}}^2]}{\sqrt{[1 - r_{\mathbf{ab}}^2 r_{\mathbf{bc}}^2] \, [1 - r_{\mathbf{bc}}^2]}} \leq |r_{\mathbf{ab}}|$$

leads to the relationship $r^2_{\mathbf{bc}} (1 - r^2_{\mathbf{ab}}) \geq 0$, which is true in all cases because $r_{\mathbf{bc}} \neq 0$ and $|r_{\mathbf{ab}}| \leq 1$. This also shows that $r_{\mathbf{ab.c}} = r_{\mathbf{ab}}$ only when $r_{\mathbf{ab}} = 1$. The same method can be used to demonstrate that $|r_{\mathbf{bc.a}}| \leq |r_{\mathbf{bc}}|$.

The model predictions in Fig. 4.11 show that it is not possible to distinguish between models 1 and 2 from the correlation coefficients alone: these two models differ only in their hypotheses (arrow directions). Their key common characteristic is the non-significance of the partial correlation $r_{\mathbf{ac.b}}$. Model 3 is distinct in the fact that $r_{\mathbf{ac}}$ is not significant and that the partial correlations are, in absolute values, larger than or equal to the corresponding simple correlations, whereas they are smaller in models 1 and 2. For model 4, some of the predictions depend on the signs of the effects depicted by the arrows; for example, the three partial correlations may be larger or smaller, in absolute values, than the simple correlations. Model 4 may apply even if one, but only one, of the two simple correlations, $r_{\mathbf{ab}}$ or $r_{\mathbf{bc}}$, is not significant. When n

is small, the tests may not have enough power to evidence the significance of the relationships and, as a consequence, evidence may be lacking to support a model.

Numerical example. The simple example already used for multiple and partial correlations illustrates here the problem inherent to all correlation matrices, i.e. that it is never possible to interpret correlations *per se* in terms of causal relationships. In the following matrix, the upper triangle contains the coefficients of simple correlation whereas the lower triangle contains the partial correlation coefficients:

$$\begin{bmatrix} 1 & 0.4 & 0.8 \\ 0 & 1 & 0.5 \\ 0.76 & 0.33 & 1 \end{bmatrix}$$

It may have seemed that descriptors $\mathbf{y}_1$ and $\mathbf{y}_2$ were correlated ($r_{12} = 0.4$), but the first-order partial correlation coefficient $r_{12.3} = 0$ shows that this is not the case. The predictions of models 1 and 2 in Fig. 4.11, with $\mathbf{a} = \mathbf{y}_1$, $\mathbf{b} = \mathbf{y}_3$ and $\mathbf{c} = \mathbf{y}_2$, are in agreement with these results. In the absence of external information or ecological hypotheses, there is no way of determining which pattern of causal relationships, model 1 or model 2, actually fits this correlation matrix.

Ecological application 4.5

Bach *et al*. (1992) analysed a 28-month long time series (weekly sampling, $n = 122$) of eel catches (*Anguilla anguilla*) in the Thau marine lagoon in southern France. Fixed gears called 'capêchades', composed of three funnel nets (6-mm mesh) and an enclosure, were used near the shore in less than 1.5 m of water. In the deeper parts of the lagoon, other types of gears were used: heavier assemblages of funnel nets with larger mesh sizes, called 'brandines', 'triangles' and 'gangui', as well as longlines. Various hypotheses were stated by the authors and tested using partial correlation analysis and path analysis. These concerned the influence of environmental variables, including air temperature as a proxy for seasons, on the behaviour of fish and fishermen, and their effects on landings. Coefficients of linear correlation reported in the paper are used here to study the relationships among air temperature, fishing effort, and landings, for the catches by 'capêchade' (Fig 4.12). The analysis in the paper was more complex; it also considered the effects of wind and lunar phases. Linearity of the relationships was checked. The correlation coefficients are consistent with a type-4 model stating that both effort and temperature affect the landings (temperature increases eel metabolism and thus their activity and catchability) and that the effort, represented by the number of active 'capêchade' fishermen, is affected by seasonality (lower effort at high temperature, 'capêchades' being not much used from August to October). Interesting is the non-significant simple linear correlation between temperature and catches. The partial correlations indicate that this simple correlation corresponds to two effects of temperature on catches that are both significant but of opposite signs: a positive partial correlation of temperature on catches and a negative one of temperature on effort. In the paper of Bach *et al*., partial correlation analysis was used as a first screen to eliminate variables that clearly did not influence catches. Path analysis (Section 10.4) was then used to study the direct and indirect effects of the potentially explanatory variables on catches.

Partial correlations do not provide the same information as path analysis (Section 10.4). On the one hand, partial correlations, like partial regression coefficients (Subsection 10.3.3), indicate whether a given variable has some unique (linear)

	Eel catches	Fishing effort	Air temperature
Eel catches	----	$r = 0.730$***	$r = 0.096$ (N.S.)
Fishing effort	partial r = 0.763***	----	$r = -0.180$*
Air temperature	partial r = 0.338***	partial r = –0.368***	----

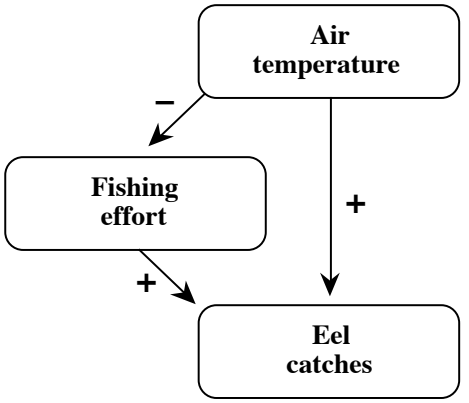


Figure 4.12 Left: simple and partial correlations among temperature, fishing effort, and eel catches using the 'capêchade' fishing gear, from Bach *et al*. (1992). Right: causal model supported by the data. The signs of the partial correlation coefficients are shown along the arrows. *: $0.05 \geq p > 0.01$; ***: $p \leq 0.001$; N.S.: non-significant correlation at significance level $\alpha = 0.05$.

relationship with some other variable, after the linear effects of all the other variables in the model have been taken into account. In path analysis on the other hand, one is mostly interested in partitioning the relationship between predictor (explanatory, independent) and criterion (response, dependent) variables into direct and indirect components.

The above discussion was based on linear correlation coefficients. Advantages of the linear model include ease of computation and simplicity of interpretation. However, environmental processes are not necessarily linear. This is why linearity must be checked, not only assumed, before embarking in this type of computation. When the variables are not linearly related, two choices are open: either proceed with non-linear statistics (nonparametric simple and partial correlation coefficients, in particular, are available and may be used in this type of calculation), or linearize the relationships that seem promising. Monotonic relationships, identified in scatter diagrams, can often be linearized by applying the transformations of Section 1.5 to one or both variables. There is no 'cheating' involved in doing that; either a monotonic relationship exists, and linearizing transformations allow one to apply linear statistics to the data; or there is no monotonic relationship, and no amount of transformation will ever create one.

Simple causal modelling, as presented in this subsection, may be used in two different types of circumstances. A first, common application is exploratory analysis, which is performed when 'weak' ecological hypotheses only can be formulated. What this means is the following: in many studies, a large number of causal hypotheses may

be formulated *a priori*, some being contradictory, because the processes at work in ecosystems are too numerous for ecologists to decide which ones are dominant under given circumstances. So, insofar as each of the models derived from ecological theory can be translated into hypothesized correlation coefficients, partial correlation analysis may be used to clear away those hypotheses that are not consistent with the data and to keep only those that look promising for further analysis. Considering three variables, for instance, one may look at the set of simple and partial correlation coefficients and decide which of the four models of Fig. 4.11 are not consistent with the data. Alternatively, when the ecosystem is better understood, one may wish to test a single set of hypotheses (i.e. a single model), to the exclusion of all others. With three variables, this would mean testing only one of the models of Fig. 4.11, to the exclusion of all others, and deciding whether the data are consistent, or not, with that model.

Several correlation coefficients are tested in each panel of Fig. 4.11. Three simultaneous tests are performed for the simple correlation coefficients and three for the partial correlation coefficients. In order to determine whether such results could have been obtained by chance alone, some kind of global test of significance, or correction, must be performed (eq. 4.14; Box 1.3).

The simple form of modelling described here may be extended beyond the frame of linear modelling, as long as formulas exist for computing partial relationships. Examples are the partial nonparametric correlation coefficient (partial Kendall τ, eq. 5.9) and the partial Mantel statistic (Subsection 10.5.2).

4.6 Tests of normality and multinormality

Testing the normality of empirical distributions is an important concern for ecologists who want to use linear models for analysing their data. Tests of normality are carried out in two types of circumstances. On the one hand, many tests of statistical significance, including those described in the present chapter, require the empirical data to be drawn from normally distributed populations. On the other hand, the linear methods of multivariate data analysis discussed in Chapters 9, 10, and 11 do summarize data in more informative ways if their underlying distributions are multinormal — or at least are not markedly skewed, as discussed below. Estimating the skewness and testing the normality of empirical variables is thus an important initial step in the analysis of a data set. Variables that are not normally distributed may be subjected to normalizing transformations (Section 1.5). The historical development of the tests of normality has been reviewed by D'Agostino (1982) and by Dutilleul & Legendre (1992).

The problem may first be approached by plotting frequency histograms of empirical variables. Looking at these plots immediately identifies distributions that have several modes, for instance, or are obviously too skewed or too 'flat' or 'peaked' to have possibly been drawn from normally distributed populations.

Next, for unimodal distributions, one may examine the skewness and kurtosis parameters. The first centred moment of a distribution is $m_1 = 0$ and the second is the variance, $m_2 = s_x^2$ (unbiased estimator: eq. 4.3). The unbiased estimator of the third centred moment is:

$$m_3 = \frac{n \sum (x_i - \bar{x})^3}{(n-1)\,(n-2)}$$

Skewness

Skewness (α_3) is a measure of asymmetry. It is estimated as the third moment of the distribution divided by the cube of the standard deviation:

$$\alpha_3 = m_3 / s_x^3 \qquad \textbf{(4.41)}$$

Skewness is 0 for a symmetric distribution like the normal distribution. Positive skewness corresponds to a frequency distribution with a longer tail to the right than to the left, whereas a distribution with a longer tail to the left shows negative skewness. The unbiased estimator of the fourth moment of a distribution is:

$$m_4 = \frac{n\,(n+1) \sum (x_i - \bar{x})^4 - 3\,(n-1)\,(\sum (x_i - \bar{x})^2)^2}{(n-1)\,(n-2)\,(n-3)}$$

Kurtosis

Kurtosis (α_4) is a measure of flatness or peakedness of a distribution. It is estimated as the fourth moment divided by the standard deviation to the power 4:

$$\alpha_4 = m_4 / s_x^4 \qquad \textbf{(4.42)}$$

The kurtosis of a normal distribution is $\alpha_4 = 0$. Distributions flatter than the normal distribution ('platycurtic') have negative values for α_4 whereas distributions that have more observations around the mean than the normal distribution have positive values for α_4, indicating that they are 'leptokurtic' which means more 'peaked'. The value of α_4 for a uniform (flat, rectangular) distribution is -1.2.

Although tests of significance have been developed for skewness and kurtosis, they are not used any longer because more powerful tests of normality are now available. For the same reason, testing the goodness-of-fit of an empirical frequency distribution to a normal distribution with same mean and variance (as in Fig 4.13a) using a chi-square test is no longer in fashion because it is not very sensitive to departures from normality (Stephens, 1974; D'Agostino, 1982), even though it may still be presented in some texts of biological statistics as an acceptable procedure. The main problem is that it does not take into account the ordering of classes of the two frequency distributions that are being compared. This explains why the main statistical packages do not use it, but propose instead one or the other (or both) procedure described below.

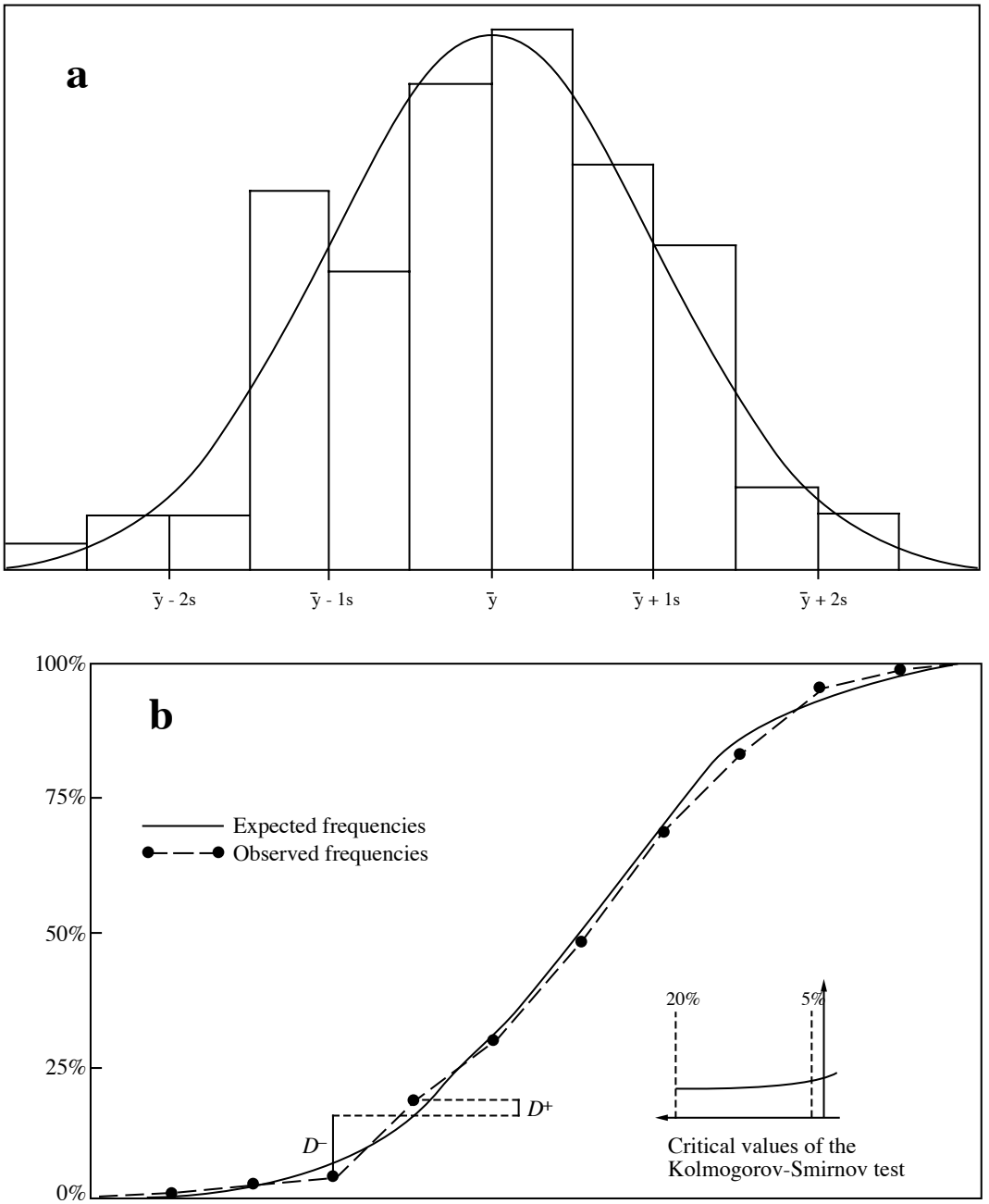


Figure 4.13 Numerical example with $n = 100$. (a) Frequency distribution and fitted theoretical normal curve, (b) relative cumulative frequencies and Kolmogorov-Smirnov test of goodness-of-fit, showing that the maximum deviation $D = 0.032$ is too small in this case to reject the hypothesis of normality.

One of the widely used tests of normality is the Kolmogorov-Smirnov test of goodness-of-fit. In Fig. 4.13b, the same data as in Fig. 4.13a are plotted as a cumulative frequency distribution. The cumulative theoretical normal distribution is also plotted on the same graph; it can easily be obtained from a published table, or by requesting in a statistical package the normal probability values corresponding to the relative cumulative frequencies (function ***pnorm()*** in R). One looks for the largest deviation D between the cumulative empirical relative frequency distribution and the cumulative theoretical normal distribution. If D is larger than or equal to the critical value in the table, for a given number of observations n and significance level α, the hypothesis of normality is rejected.

K-S test

The Kolmogorov-Smirnov (K-S) test of goodness-of-fit is especially interesting for small sample sizes because it does not require to lump the data into classes. When they are divided into classes, the empirical data are discontinuous and their cumulative distribution is a step-function, whereas the theoretical normal distribution to which they are compared is a continuous function. D is then formally defined as the maximum of D^- and D^+, where D^- is the maximum difference computed just before a data value and D^+ is the maximum difference computed at the data value (i.e. at the top of each step of the cumulative empirical step-function). A detailed numerical example of the procedure is presented by Sokal & Rohlf (1995).

Standard Kolmogorov-Smirnov tables for the comparison of two samples, where the distribution functions are completely specified (i.e. the mean and standard deviation are stated by hypothesis), are not appropriate for testing the normality of *empirical data* since the mean and standard deviation of the reference normal distribution must then be estimated from the observed data; critical values given in these tables are systematically too large, and thus lead too often to not rejecting the null hypothesis of normality. Corrected critical values for testing whether a set of observations is drawn from a normal population, that are valid for stated probabilities of type I error, have been computed by Lilliefors (1967) and, with additional corrections based on larger Monte Carlo simulations, by Stephens (1974). The same paper by Stephens evaluates other statistics to perform tests of normality, such as Cramér-von Mises W^2 and Anderson-Darling A^2 which, like D, are based on the empirical cumulative distribution function (only the statistics differ); it proposes corrections where needed for the situation where the mean and variance of the reference normal distribution are unknown and are thus estimated from the data.

Normal probability plot

The second widely used test of normality is due to Shapiro & Wilk (1965). It is based on an older graphical technique that will be described first. This technique, called *normal probability plotting*, was developed as an informal way of assessing deviations from normality. The objective is to plot the data in such a way that, if they come from a normally distributed population, they will fall along a straight line. Deviations from a straight line may be used as indication of the type of non-normality. In these plots, the values along the abscissa are either the observed or the standardized data (in which case the values are transformed to standard deviation units), while the ordinate is the percent cumulative frequency value of each point plotted on a normal

probability scale. Sokal & Rohlf (1995) give computation details. Figure 4.14 shows the same data as in Fig 4.13a, which are divided into classes, plotted on normal probability paper. The same type of plot could also be produced for the raw data, not grouped into classes. For each point, the *upper limit* of a class is used as the abscissa, while the ordinate is the percent cumulative frequency (or the cumulative percentage) of that class. Perfectly normal data would fall on a straight line passing through the point ($\bar{y}$, 50%). A straight line is fitted trough the points, using reference points based on the mean and variance of the empirical data (see the legend of Fig. 4.14); deviations from that line indicate non-normality. Alternatively, a straight line may be fitted through the points, either by eye or by regression; the mean of the distribution may be estimated as the abscissa value that has an ordinate value of 50% on that line. D'Agostino (1982) gives examples illustrating how deviations from linearity in such plots indicate the degree and type of non-normality of the data.

Shapiro-Wilk test

Shapiro & Wilk (1965) proposed to quantify the information in normal probability plots using a statistic called 'analysis of variance *W*', which they defined as the *F*-ratio of the estimated variance obtained from the weighted least-squares of the slope of the straight line (numerator) to the variance of the sample data (denominator). The statistic is used to assess the goodness of the linear fit:

$$W = \left(\sum_{i=1}^{n} w_i x_i \right)^2 \Big/ \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad \textbf{(4.43)}$$

where the x_i are the ordered observations ($x_1 \le x_2 \le \ldots \le x_n$) and coefficients w_i are optimal weights for a population assumed to be normally distributed. Statistic *W* may be viewed as the square of the correlation coefficient (i.e. the coefficient of determination) between the abscissa and ordinate of the normal probability plot described above. Large values of *W* indicate normality (points lying along a straight line give r^2 close to 1), whereas small values indicate lack of normality. Shapiro & Wilk did provide critical values of *W* for sample sizes up to 50. D'Agostino (1971, 1972) and Royston (1982a, b, c) proposed modifications to the *W* formula (better estimates of the weights w_i), which extend its application to much larger sample sizes. Extensive simulation studies have shown that *W* is a sensitive *omnibus* test statistic, meaning that it has good power properties over a wide range of non-normal distribution types and sample sizes.

The Shapiro-Wilk test is available in the ***shapiro.test()*** function of the R STATS package. Five other functions are available in the NORTEST package to carry out tests of normality, including function ***lillie.test()*** for the Lilliefors (1967) K-S test using Stephens' (1974) corrections. Which of these tests is best? Reviewing the studies on the power of tests of normality published during the previous 25 years, D'Agostino (1982) concluded that the best *omnibus* tests are the Shapiro-Wilk *W*-test and a modification by Stephens (1974) of the Anderson-Darling A^2-test mentioned above (***ad.test()*** function in NORTEST). In a Monte Carlo study involving autocorrelated data (Section 1.1), however, Dutilleul & Legendre (1992) showed (1) that, for moderate

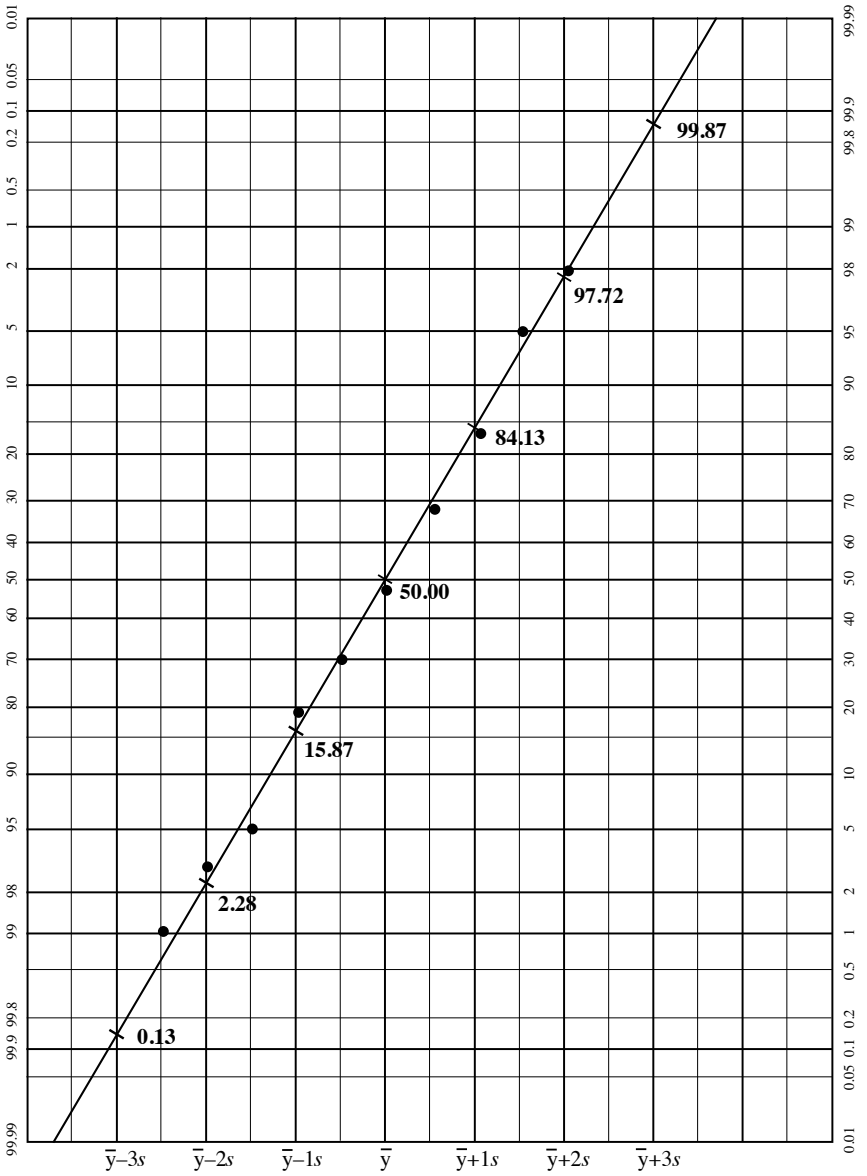


Figure 4.14 The cumulative percentages of data in Fig. 4.13a are plotted here on normal probability paper (probit transformation) as a function of the upper limits of classes. Cumulative percentiles are indicated on the right-hand side of the graph. The last data value cannot be plotted on this graph because its cumulated percentage value is 100. The diagonal line represents the theoretical cumulative normal distribution with same mean and variance as the data. This line is positioned on the graph using reference values of the cumulative normal distribution, for example 0.13% at $\bar{y} - 3s$ and 99.87% at $\bar{y} + 3s$, and it passes through the point ($\bar{y}$, 50%). This graph contains exactly the same information as Fig. 4.13b; the difference lies in the scale of the ordinate.

sample sizes, both the *D*-test and the *W*-test were too liberal (in an asymmetric way) for high positive ($\rho > 0.4$) and very high negative ($\rho < -0.8$) values of autocorrelation along time series and for high positive values of spatial autocorrelation ($\rho > 0.2$), and (2) that, overall, the Kolmogorov-Smirnov *D*-test was more robust against autocorrelation than the Shapiro-Wilk *W*-test, whatever the sign of the first-order autocorrelation.

As stated at the beginning of this section, ecologists must absolutely check the normality of data only when they are planning to use parametric statistical tests that assume normality of the distributions; permutation tests (Section 1.2) can be used with non-normal data. Most methods presented in this book, including clustering and ordination techniques, do not require statistical testing and hence may be applied to non-normal data. With many of these methods, however, ecological structures emerge more clearly when the data do not present strong asymmetry; this is the case, for example, with principal component analysis. Since normal data are not skewed (coefficient $\alpha_3 = 0$), testing the normality of data is also testing for asymmetry; normalizing transformations, applied to data with unimodal distributions, reduce or eliminate asymmetry. So, with multidimensional data, it is recommended to check at least the skewness of the variables one by one.

Some tests of significance require that the data be multinormal (Section 4.3). Normality of the *p* individual variables can easily be tested as described above. In a multivariate situation, however, showing that each variable does not significantly depart from normality does not demonstrate that the multivariate data set is multinormal although, in many instances, this is the best that researchers can do.

Test of multinormality

Dagnelie (1975) proposed an elegant way of testing the multivariate normality of a set of multivariate observations. The method is based on the *Mahalanobis generalized distance* (D_5; Section 7.4, eq. 7.38) described in Chapter 7. Generalized distances are computed, in the multidimensional space, between each object and the multidimensional mean of all objects. The distance between object $\mathbf{x}_i$ and the mean point $\bar{\mathbf{x}}$ is computed as:

$$D(\mathbf{x}_i, \bar{\mathbf{x}}) = \sqrt{[y - \bar{y}]_i \mathbf{S}^{-1} [y - \bar{y}]_i'} \qquad \textbf{(4.44)}$$

where $[y - \bar{y}]_i$ is the vector corresponding to object $\mathbf{x}_i$ in the matrix of centred data and **S** is the multivariate dispersion matrix (Section 4.1). For standardized variables $z_{ij} = (y_{ij} - \bar{y}_j) / s_j$, eq. 4.44 becomes:

$$D(\mathbf{x}_i, \bar{\mathbf{x}}) = \sqrt{\mathbf{z}_i \mathbf{R}^{-1} \mathbf{z}_i'} \qquad \textbf{(4.45)}$$

where **R** is the correlation matrix. Dagnelie's approach is that, for multinormal data, the generalized distances should be normally distributed. He suggested to do a visual examination of the cumulative frequency distribution as in Fig. 4.14. Actually, the generalized distances can be subjected to a Shapiro-Wilk test of normality, whose conclusions are applied to the multinormality of the original multivariate data.

The Dagnelie test of multivariate normality based on the Shapiro-Wilk test of normality of Mahalanobis generalized distances is invalid for univariate data (type I error rate too high). Numerical simulations by D. Borcard (personal communication) showed that the test had correct levels of type I error for values of n between $3p$ and $7.5p$, where p is the number of variables in the data table (simulations with $1 \leq p \leq 50$). Outside that range of n values, the results were too liberal, meaning that the test rejected too often the null hypothesis of normality. For $p = 2$, the simulations showed the test to be valid for $6 \leq n \leq 11$. If H_0 is not rejected in a situation where the test is too liberal, the result is trustworthy.

4.7 Software

Functions for all operations described in this chapter are available in the R language.

1. Covariance matrices are computed using functions ***var()*** and ***cov()*** of the STATS package; correlation matrices are computed by ***cor()***. The F-test comparing two variances is carried out by ***var.test()*** and correlation coefficients are tested using ***cor.test()*** in STATS.

2. Eigenanalysis is computed by ***eigen()*** in STATS.

3. Partial correlations are computed by function ***partial.cor()*** of the RCMDR package.

4. Tests of normality are computed using ***shapiro.test()*** in STATS, ***lillie.test()*** in NORTEST, and ***ad.test()*** in NORTEST. Function ***qqnorm()*** of STATS produces normal quantile-quantile plots like Fig. 4.14.

5. Function ***pnorm()*** in STATS computes p-values for the normal distribution, ***pf()*** for the F-distribution, ***pt()*** for the Student t-distribution, ***pchisq()*** for the chi-square distribution, and so on for other statistical distributions.

Commercial statistical packages, as well as S-PLUS® and MATLAB®, also provide functions for these calculations.

Chapter

5

Multidimensional semiquantitative data

5.0 Nonparametric statistics

Statistical testing often refers to the concepts of *parameter* and *reference population*, as explained in Section 1.2. Section 4.3 showed that the mean, standard deviation and correlation are *parameters* of the multinormal distribution, so that this statistical distribution and others play a key role in testing *quantitative* data. When the data are *semiquantitative*, however, it does not make sense to compute statistics such as the mean or the standard deviation. In that case, hypothesis testing must be conducted with *nonparametric statistics*. This expression cover all statistical methods developed for analysing either *semiquantitative* (rank statistics; Section 5.2) or *qualitative* (Chapter 6) data.

Nonparametric tests are *distribution-free*, i.e. they do not assume that the samples were drawn from a population with a specified distribution (e.g. multinormal). Because of that, nonparametric statistics are useful not only when descriptors are semiquantitative, but also when quantitative descriptors do not conform to the multinormal distribution and researchers do not wish, or succeed, to normalize them. Many of the nonparametric tests for semiquantitative data are called *ranking tests* because they are based on ranked observations instead of quantitative values. Another advantage of nonparametric statistics is computational simplicity. Last but not least, nonparametric tests may be used with small samples, a situation that frequently occurs with ecological data; permutation tests based upon parametric statistics (Section 1.2) share this last advantage of nonparametric tests. For semiquantitative data, the nonparametric statistics corresponding to the *mean* and *variance* (Section 4.1) are the *median* and *range*, respectively.

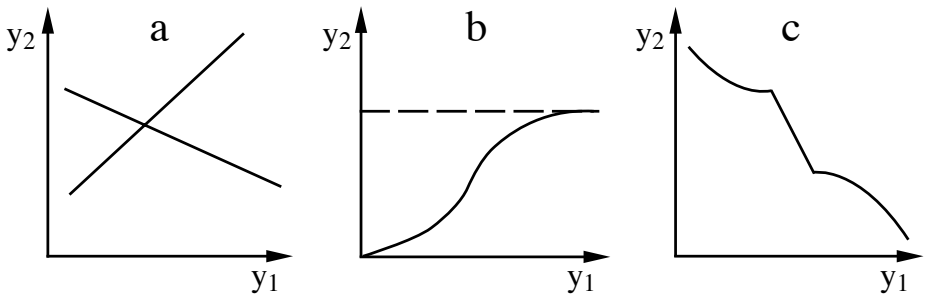


Figure 5.1 Three types of monotonic relationships between two descriptors: (a) linear (increasing and decreasing); (b) logistic (increasing monotonic); (c) atypical (decreasing monotonic).

Ranking tests should be used in the following situations:

- One or several descriptors among those to be compared are semiquantitative.

Monotonic

- The purpose of the study is not to evidence linear, but *monotonic* relationships between quantitative descriptors. In a bivariate monotonic relationship, one of the two descriptors keeps increasing or decreasing as the other increases (Fig. 5.1); the increase (or decrease) is not necessarily *linear* nor *smoothly curvilinear*.

Ranking tests or *permutation tests* (Section 1.2) can be used in the following cases:

- One or several quantitative descriptors are not normally distributed (tests of normality and multinormality are described in Section 4.6) and researchers do not wish to normalize them or do not succeed in doing so. Normalizing transformations are described in Subsection 1.5.6.

- The number of observations is small.

The present chapter first summarizes the methods available in the nonparametric approach, with reference to the corresponding parametric methods (Section 5.1). Tests for differences among groups using quantitative, semiquantitative or qualitative descriptors are compared in Section 5.2. Rank correlation coefficients are presented in Section 5.3. Section 5.4 is devoted to the Kendall coefficient of concordance, which is a generalization of the Spearman correlation coefficient to several descriptors. Most statistical computer packages, including R, offer nonparametric testing procedures.

5.1 Quantitative, semiquantitative, and qualitative multivariates

As discussed in Section 1.4, ecological descriptors may be of different levels of precision (Table 1.2). Ecologists generally observe several descriptors on the same objects, so that multidimensional ecological variates may be either quantitative, semiquantitative, or qualitative, or mixed, i.e. consisting of descriptors with different precision levels. For many years, quantitative ecology has been based almost exclusively on quantitative descriptors and on parametric tests, even though there exist a large number of methods that can efficiently analyse semiquantitative or qualitative multivariates as well as multivariates of mixed precision levels. These methods have become increasingly popular in ecology, not only because non-quantitative descriptors often provide unique information, but also because parametric statistics cannot be tested for significance when quantitative data do not conform to a number of conditions, including normality. This section briefly reviews numerical methods for analysing multivariates with various levels of precision.

Table 5.1 summarizes and compares methods described elsewhere in the present book. In the same row are found corresponding methods, listed under four column headings. The applicability of methods increases from left to right. Methods in the first (left-hand) column are restricted to *quantitative* multivariates, which must also, in most cases, be linearly related or/and multinormally distributed. Methods in the second column have been developed for *semiquantitative* descriptors exhibiting *monotonic* relationships. These methods may also be used (a) with quantitative descriptors, in particular when they do not follow the conditions underlying methods in the first column, and (b) for the combined analysis of quantitative and semiquantitative descriptors. Methods in the third column were developed for the numerical analysis of *qualitative* descriptors. They may also be used for analysing quantitative or semiquantitative descriptors exhibiting nonmonotonic relationships, after dividing these continuous descriptors into classes. Methods for qualitative descriptors thus represent a first type of techniques for multivariates of mixed precision, since they can be used for analysing together quantitative, semiquantitative, and qualitative descriptors, after the former have been divided into classes. An alternative is to recode multiclass qualitative descriptors into dummy variables (Subsection 1.5.7) and use parametric methods (first column of Table 5.1) on the resulting assemblage of quantitative and binary descriptors; this approach is often used in regression and canonical analyses (Chapters 10 and 11).

The methods listed in the right-hand column can be used for analysing data tables containing mixtures of quantitative, semiquantitative and qualitative descriptors. Of special interest are the distance-based methods (dbRDA, PCoA, nMDS, clustering), which can be applied after computing an association coefficient for mixed-level data. These methods are very general, since they may replace equivalent methods in the other three columns; the cost is sometimes greater mathematical and/or computational complexity.

Table 5.1 Methods for analysing *multidimensional* ecological data sets, classified here according to the levels of precision of descriptors (columns). For methods concerning data series, see Table 12.2. The Subject index at the end of the book shows where each method is described.

Quantitative descriptors	Semiquantitative descriptors	Qualitative descriptors	Descriptors of mixed precision
Difference between two samples:			
Hotelling T^2	---	Log-linear models	---
RDA, tbRDA, dbRDA	dbRDA	tbRDA, dbRDA	dbRDA
CCA		CCA	
Difference among several samples:			
MANOVA	---	Log-linear models	---
RDA, tbRDA, dbRDA	db-RDA	tbRDA, db-RDA	db-RDA
CCA		CCA	
Scatter diagram	Rank diagram	Multiway contingency table	Quantitative-rank diagram
Association coefficients R:			
Covariance	---	Information, X^2	---
Pearson *r*	Spearman *r*	Contingency	---
	Kendall τ		
Partial *r*	Partial τ		
Multiple *R*	Kendall *W*		
Species diversity:			
Diversity measures	Diversity measures	Number of species	---
Association coeff. Q	Association coeff. Q	Association coeff. Q	Association coeff. Q
Clustering	Clustering	Clustering	Clustering
Ordination:			
PCA, tbPCA, CA		tbPCA, CA	
PCoA	PCoA	PCoA	PCoA
nMDS	nMDS	nMDS	nMDS
Regression	Regression	Correspondence	Regression
simple linear (I and II)	nonparametric		logistic
multiple linear			dummy
polynomial			
partial linear			
nonlinear, logistic			
smoothing (splines, LOWESS)			
multivariate; see also canonical a.			
Path analysis	---	Log-linear models	
		Logit models	
Canonical analysis:			
RDA, tbRDA, dbRDA	dbRDA	tbRDA, dbRDA	db-RDA
CCA		CCA	
CCorA, CoIA		tbCCorA, tbCoIA	
LDA	---	Discrete discriminant a.	
		Log-linear models	
		Logistic regression	

There are many types of methods for multidimensional analysis (rows of Table 5.1). One interesting aspect of the table is that there is always at least one, and often several methods for descriptors with low precision levels. Thus, ecologists should not hesitate to collect information in semiquantitative or qualitative form since there exist numerical methods for processing descriptors with all levels of precision. However, it is always important to consider, at the stage of the sampling design (Fig. 1.3), how the data will be analysed, so as to avoid problems at the later stage of analysis. These problems may include the lack of human resources to efficiently use advanced numerical methods. Researchers could use the period devoted to sampling to improve their knowledge of methods and become familiar with computer programs and functions.

Coming back to Table 5.1, it is possible to compare groups of objects, described by quantitative multivariate data, using multidimensional analysis of variance (MANOVA). When there are only two groups, another approach is Hotelling's T^2 (Section 7.4). In the case of qualitative multivariate data, the comparison may be done by adjusting log-linear models (Section 6.3) to a multiway contingency table; the relationship between contingency table analysis and analysis of variance is explained in Section 6.0. Multivariate analysis of variance of species presence-absence or abundance tables may be carried out using either transformation-based redundancy analysis (tbRDA) or distance-based redundancy analysis (db-RDA) (Subsections 11.1.5 and 11.1.10), or else canonical correspondence analysis (CCA, Section 11.2).

The simplest approach to investigate the relationships among descriptors considered two at a time (Fig. 5.2) is to plot the data as a scatter diagram, whose semiquantitative and qualitative equivalent are the rank-rank diagram and the contingency table, respectively. Quantitative-rank diagrams may be used to compare a quantitative to a semiquantitative descriptor (Legendre & Legendre, 1982).

Two families of methods follow from these diagrams, for either *measuring* the dependence among descriptors, or *forecasting* one or several descriptors using other ones. The R-mode coefficients of dependence, described in Chapter 4 for quantitative descriptors, in Chapter 5 for semiquantitative descriptors, and in Chapter 6 for qualitative descriptors, measure the dependence between descriptors. These coefficients are summarized in Subsection 7.5.1. It is interesting to note that measures of information and X^2 (chi-square) calculated on contingency tables (Chapter 6) are the equivalent, for qualitative descriptors, of the covariance computed between quantitative descriptors. Methods in the second family belong to regression analysis (Section 10.3), which has nonparametric forms (e.g. the monotone regression method used in Section 9.4), and whose qualitative equivalent is the analysis of correspondence in contingency tables (Section 6.4).

Various measures of species diversity are reviewed in Section 6.5. They are usually computed on quantitative species counts, but Dévaux & Millerioux (1977) have shown that this may be done just as well on semiquantitative counts. When there are no counts, the number of species present may be used to assess diversity; this is indeed

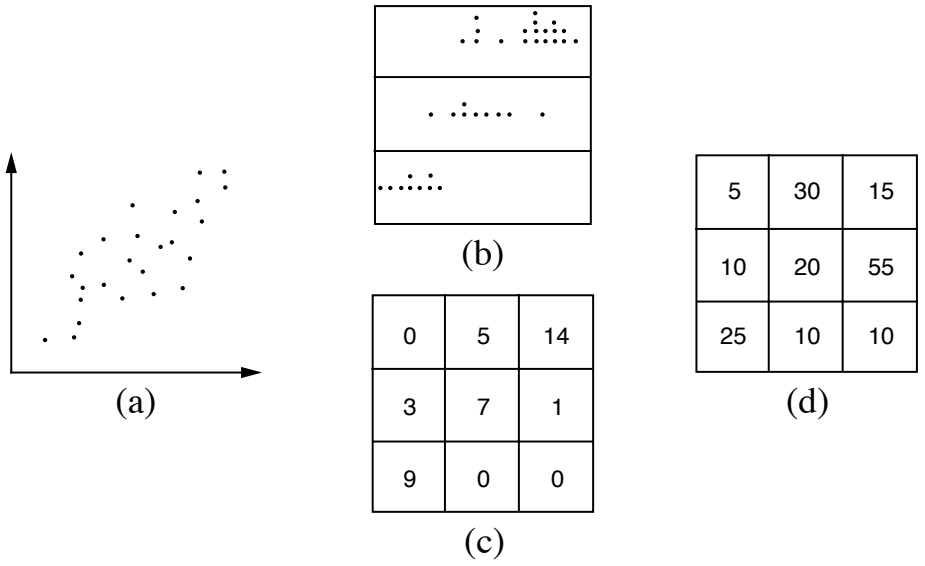


Figure 5.2 Comparison of two descriptors. (a) Scatter diagram (quantitative descriptors on both axes). (b) Quantitative-rank diagram (quantitative descriptor on the abscissa, ranked classes of a semiquantitative descriptor on the ordinate). (c) Rank-rank diagram (ranked classes of semiquantitative descriptors on both axes). (d) Two-way contingency table (nonordered classes of qualitative descriptors on both axes). From Legendre & Legendre (1982).

the first diversity index to have been described in the ecological literature (Patrick, 1949; Subsection 6.5.1).

There are Q-mode association coefficients (Sections 7.3 and 7.4) adapted to descriptors of all levels of precision (see Tables 7.4 and 7.5). Some of the similarity coefficients (Chapter 7: S_{15}, S_{16}) are yet another way of combining quantitative and qualitative descriptors in multivariate data analysis. Concerning clustering algorithms (Chapter 8), most of them are indifferent to the precision of descriptors, since clustering is in general conducted on an association matrix, most often of type Q.

Among the ordination methods in reduced space, principal component analysis (PCA, Section 9.1) is the main method to use with quantitative descriptors, although it can also be applied to semiquantitative data (Subsection 9.1.7). Species abundance or presence-absence data, as well as other types of frequency data, can be analysed by correspondence analysis (CA, Section 9.2) or by transformation-based PCA (tbPCA). Principal coordinate analysis (PCoA, Section 9.3) and nonmetric multidimensional scaling (nMDS, Section 9.4) are indifferent to the precision of descriptors since they are computed on an association matrix (generally Q-type).

For the interpretation of ecological structures, regression, which was briefly discussed a few paragraphs above, is the chief technique when the dependent variable is a single quantitative variable. Logistic regression is used when the response is presence-absence data. Various forms of canonical analysis are available to interpret the structure of quantitative data using one or several tables of explanatory variables: redundancy analysis (RDA, Section 11.1), canonical correspondence analysis (CCA, Section 11.2), linear discriminant analysis (LDA, Section 11.3), canonical correlation analysis (CCorA, Section 11.4), and co-inertia analysis (CoIA, Section 11.5). Canonical correspondence analysis, as well as tbRDA, allow the interpretation of the structure of species abundance or presence-absence data using explanatory variables. For non-quantitative data, distance-based RDA (dbRDA) can be used after computing a distance matrix using an appropriate distance function. There are also methods equivalent to discriminant and path analyses for qualitative descriptors.

Table 5.1 shows that ecological data can efficiently be analysed irrespective of their levels of precision. Researchers should use ecological criteria, such as allowable effort in the field and biological meaningfulness of the decimal places to be recorded, to decide about the level of precision of their data. The strictly numerical aspects should play a secondary role in that decision.

5.2 One-dimensional nonparametric statistics

The present book is devoted to numerical methods for analysing sets of *multidimensional* ecological data. Methods for one-dimensional variables are not discussed in depth since they are the subject of many excellent textbooks. Nonparametric tests for one-dimensional descriptors are explained, among others, in the books of Siegel (1956), Hájek (1969), Siegel & Castellan (1988), and Sokal & Rohlf (1995). Because ecologists are often not fully conversant with these tests, the correspondence between approaches for quantitative, semiquantitative, and qualitative descriptors is not always clearly understood. This is why the one-dimensional methods to carry out tests of differences among groups of objects are summarized in Table 5.2.

Independent samples
Related samples

Methods in the table are divided in two main families: those for independent samples, which are the most generally applicable, and those for related samples. Related samples are often called *matched* or *paired* samples (Box 1.1). With such samples of observations, the analysis may focus either on the differences between the matched observation units, or on the differences among the classes of another factor while controlling for the differences between the matched observations. Matching may be achieved, for example, by repeating observations at the same sampling sites at different times, or by making observations at points representing corresponding conditions, e.g. in several lakes with sampling units taken from the same depths in the water columns. Sampling units observed before and after a treatment also form matched pairs. When related samples are analysed using the methods for independent

Table 5.2 Methods to carry out tests of differences among groups of objects (*one-dimensional* data) are classified here according to the levels of precision of the descriptors (columns). Most of these methods are not discussed elsewhere in the present book. Table modified from Siegel (1956) and Legendre & Legendre (1982).

Number of groups (k)	Quantitative descriptors*	Semiquantitative descriptors	Qualitative descriptors
Independent samples:			
$k = 2$	Student t (unpaired)	Mann-Whitney U-test Median test Kolmogorov-Smirnov test etc.	X^2 (2 × no. states) Fisher's exact probability test Logistic regression
$k \geq 2$ (one-way)	One-way ANOVA and F-test	Kruskal-Wallis' H Extension of the median test	X^2 (k × no. states) Discriminant a.
Related samples:			
$k = 2$	Student t (paired)	Sign test Wilcoxon signed-ranks test	McNemar test (binary descriptors)
$k \geq 2$ (two-way)	Two-way ANOVA and F-tests	Friedman's two-way ANOVA by ranks	Cochran Q (binary descriptors)
$k \geq 2$ (multiway)	Multiway ANOVA and F-tests	---	---

* When quantitative data do not meet the distributional assumptions underlying parametric tests, they must be analysed using ranking tests (for semiquantitative descriptors). Another way would be to test the parametric statistics by permutation (Section 1.2).

samples, the matching information is not taken into account and this results in a less powerful statistical test. Within each of the two families, methods in Table 5.2 are classified according to the number of groups (k) that are compared.

Univariate comparison of *two independent samples* ($k = 2$), when the data are *quantitative*, is generally done by using the Student t-statistic to test the hypothesis (H_0) of equality of the group means (i.e. that the two groups of objects were drawn from the same statistical population, or perhaps from populations with equal means, assuming equal standard deviations). When the data are *semiquantitative*, computing means and standard deviations would not make sense, so that the approach must be nonparametric. The Mann-Whitney U-statistic first combines and ranks all objects in a single series and then tests the hypothesis (H_0) that the ranked observations come from the same statistical population or from populations that have the same median. The median test, which is not as powerful as the previous one (except in cases when there are ties), is used for testing the hypothesis (H_0) that the two groups of objects have similar medians. Other nonparametric tests consider not only the positions of the two

groups along the abscissa but also the differences in dispersion and shape (e.g. skewness) of their distributions. The best-known is the Kolmogorov-Smirnov test; this is not the same test as the one described in Section 4.6 for comparing an empirical to a theoretical distribution. The Kolmogorov-Smirnov method discussed here allows one to test the hypothesis (H_0) that the largest difference between the cumulative distributions of the two groups is so small that they may come from the same or identical populations. Finally, when the data are *qualitative*, the significance of differences between two groups of objects may be tested using a X^2-statistic calculated on a two-way contingency table. Section 6.2 describes contingency table analysis for the comparison of two descriptors. In the present case, the contingency table has two rows (i.e. two groups of objects) and as many columns as there are states in the quantitative descriptor. The hypothesis tested (H_0) is that the frequency distributions in the two rows are similar; this is the same as stating the more usual hypothesis of independence between rows and columns of the contingency table (Section 6.0). When the descriptor is *binary* (e.g. presence or absence) and the number of observations in the two groups is small, it is possible to test the hypothesis (H_0) that the two groups exhibit similar proportions for the two states, using Fisher's powerful exact probability test. Logistic regression (Subsection 10.3.7) may also be used in this context; in the regression, the two groups are represented by a binary response variable while the qualitative explanatory descriptors are recoded as a series of dummy variables, coded as shown in Subsection 1.5.7.

The standard parametric technique for testing that the means of *several independent samples* ($k \geq 2$) are equal, when the data are *quantitative*, is one-way analysis of variance (ANOVA). It is a k-group generalization of the Student t-test. In one-way ANOVA, the overall variance is partitioned between two orthogonal (i.e. linearly independent; see Box 1.1) components, the first one reflecting differences among the k groups and the second one accounting for the variability among objects within the groups. The hypothesis (H_0) of equal means is rejected (F-test) when the among-group variability is significantly larger than the within-group component. For *semiquantitative* data, the Kruskal-Wallis' H-test (also called Kruskal-Wallis' one-way ANOVA by ranks) first ranks all objects from the k groups into a single series, and then tests (H_0) that the sums of ranks calculated for the various groups are so similar that the objects are likely to have been drawn from the same or identical populations. When applied to quantitative data that are meeting all the assumptions of parametric ANOVA, Kruskal-Wallis' H is almost as powerful as the F-test. Another possibility is to extend to $k \geq 2$ groups the median test, described in the previous paragraph for $k = 2$. The latter is less powerful than Kruskal-Wallis' H because it uses less of the information in the data. As in the above case where $k = 2$, *qualitative* data can be analysed using a contingency table, but this time with $k \geq 2$ rows.

Multinomial logistic regression

To model multistate qualitative response data, *multinomial logistic regression* is available in R (see Section 5.5) as well as in procedure CATMOD of SAS. Discriminant analysis could be used in the same spirit. See the discussion on discriminant analysis *versus* logistic regression in Section 11.6 (point 2).

Comparing *two related samples* ($k = 2$) is usually done, for *quantitative* data, by testing (H_0) that the mean of the differences between matched pairs of observations is null (Student *t*-test; the differences are assumed to be normally and independently distributed). When the data are *semiquantitative*, one can use the sign test, which first codes pairs of observations (y_i, y_k) as either (+) when $y_i > y_k$ or (–) when $y_i < y_k$, and then tests the hypothesis (H_0) that the numbers of pairs with each sign are equal; an equivalent formulation is that the proportion of pairs with either sign is equal to 0.5. This test uses information about the direction of the differences between pairs. When the relative magnitude of the differences between pairs is also known, it becomes possible to use the more powerful Wilcoxon matched-pairs signed-ranks test. Differences between pairs are first ranked according to their magnitude (absolute values), after which the sign of the difference is affixed to each rank. The null hypothesis of the test (H_0) is that the sum of the ranks having a (+) sign is equal to that of the ranks with a (–) sign. The McNemar test provides a means of comparing paired samples of *binary* data. For example, using binary observations (e.g. presence or absence) made at the same sites, before and after some event, one could test (H_0) that no overall change has occurred.

When there are *several related samples* ($k \geq 2$) and the data are *quantitative*, the parametric approach for testing (H_0) that the means of the k groups are equal is two-way analysis of variance, with or without replication. One classification criterion of the two-way ANOVA accounts for the variability among the k groups (as in one-way ANOVA above, for $k \geq 2$ independent samples) and the other for that among the related samples. Consider, as example, 16 sites (i.e. k groups) that have been sampled at 5 depths in the water column (or at 5 different times, or using 5 different methods, etc.). The nonparametric equivalent, for *semiquantitative* data, is Friedman's *two-way analysis of variance by ranks without replication*, which is based on a two-way table like Table 5.7. In the two-way table, the k groups (e.g. 16 sites) are in rows and the corresponding samples (e.g. 5 depths) are in columns. Values within each column are ranked separately, and the Friedman X^2-statistic (eq. 5.15) is used to test (H_0) that the rank totals of the various rows (e.g. 16 sites) are equal. For *binary* data, the Cochran Q test is an extension to $k \geq 2$ groups of the McNemar test, described above for $k = 2$.

Finally, when there are *several samples ($k \geq 2$), related across several classification criteria* (e.g. 16 sites all sampled at 8 different times, using each time 5 different methods), multiway ANOVA is the standard parametric method for testing the null hypothesis (H_0) that the means of the k groups are equal (F-test). In that case, there are no obvious equivalent approaches for semiquantitative or qualitative data.

How to analyse multivariate data representing related samples is described in Subsection 11.1.10, point 3.

Table 5.3 Numerical example. Perfect rank correlation between descriptors $\mathbf{y}_1$ and $\mathbf{y}_2$.

Objects (observation units)	Ranks of objects on the two descriptors	
	$\mathbf{y}_1$	$\mathbf{y}_2$
$\mathbf{x}_1$	5	5
$\mathbf{x}_2$	1	1
$\mathbf{x}_3$	4	4
$\mathbf{x}_4$	2	2
$\mathbf{x}_5$	3	3

5.3 Rank correlations

Textbooks of nonparametric statistics propose a few methods only for the analysis of bi- or multivariate semiquantitative data. Section 5.1 has shown that there actually exist many numerical approaches for analysing multidimensional data, corresponding to all levels of precision (Table 5.1). These methods, which include most of those described in this book, belong to *nonparametric statistics* in a general sense, because they do not focus on the parameters of the data distributions. Within the specific realm of *ranking tests*, however, the statistical techniques available for multidimensional semiquantitative data are two *rank correlation coefficients* (Spearman r and Kendall τ), which both quantify the relationship between two descriptors, and the *coefficient of concordance* (Kendall W), which assesses the relationship among several descriptors. The two correlation coefficients are described in the present section and coefficient W in Section 5.4.

1 — Spearman r

Spearman corr. coeff.

The Spearman r coefficient, also called ρ (rho), is based on the idea that two descriptors $\mathbf{y}_1$ and $\mathbf{y}_2$ carry the same information if the object with the largest rank on $\mathbf{y}_1$ also has the highest rank on $\mathbf{y}_2$, and so on for all other objects. Two descriptors are said to be in perfect correlation when the ranks of all objects are the same on both descriptors, as in the numerical example shown in Table 5.3. If, however, object $\mathbf{x}_1$ which has rank 5 on $\mathbf{y}_1$ had rank 2 on $\mathbf{y}_2$, it would be natural to use the difference between these ranks, $d_1 = (y_{11} - y_{12}) = (5 - 2) = 3$, as a measure of the difference between the rankings given to this object by the two descriptors. For the whole set of objects, differences d_i are squared before summing them, to prevent differences with opposite signs from cancelling each other out.

The expression for the Spearman r correlation coefficient may be derived from the general formula of correlation coefficients (Kendall, 1948):

$$r_{jk} = \frac{\sum_{i=1}^{n} (y_{ij} - \bar{y}_j)(y_{ik} - \bar{y}_k)}{\sqrt{\sum_{i=1}^{n} (y_{ij} - \bar{y}_j)^2 \sum_{i=1}^{n} (y_{ik} - \bar{y}_k)^2}} \quad \textbf{(5.1)}$$

For quantitative data, this equation is used to compute the Pearson linear correlation coefficient (eq. 4.7).

For ranked data, the average ranks $\bar{y}_j$ and $\bar{y}_k$ are equal, so that $(y_{ij} - \bar{y}_j) - (y_{ik} - \bar{y}_k) = (y_{ij} - y_{ik})$. One can write the difference between the ranks of object i on the two descriptors as $d_i = (y_{ij} - y_{ik}) = (y_{ij} - \bar{y}_j) - (y_{ik} - \bar{y}_k)$, which leads to:

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (y_{ij} - \bar{y}_j)^2 + \sum_{i=1}^{n} (y_{ik} - \bar{y}_k)^2 - 2 \sum_{i=1}^{n} (y_{ij} - \bar{y}_j)(y_{ik} - \bar{y}_k)$$

Isolating the right-hand sum gives:

$$\sum_{i=1}^{n} (y_{ij} - \bar{y}_j)(y_{ik} - \bar{y}_k) = \frac{1}{2}\left[\sum_{i=1}^{n} (y_{ij} - \bar{y}_j)^2 + \sum_{i=1}^{n} (y_{ik} - \bar{y}_k)^2 - \sum_{i=1}^{n} d_i^2\right]$$

Using this result, eq. 5.1 is rewritten as:

$$r_{jk} = \frac{\frac{1}{2}\left[\sum_{i=1}^{n} (y_{ij} - \bar{y}_j)^2 + \sum_{i=1}^{n} (y_{ik} - \bar{y}_k)^2 - \sum_{i=1}^{n} d_i^2\right]}{\sqrt{\sum_{i=1}^{n} (y_{ij} - \bar{y}_j)^2 \sum_{i=1}^{n} (y_{ik} - \bar{y}_k)^2}} \quad \textbf{(5.2)}$$

The sum of ranks for each descriptor, which is the sum of the first n integers, is equal to $\sum_{i=1}^{n} y_{ij} = n(n+1)/2$ and the sum of their squares is $\sum_{i=1}^{n} y_{ij}^2 = n(n+1)(2n+1)/6$. Since the sum of deviations from the mean rank is

$$\sum_{i=1}^{n} (y_{ij} - \bar{y}_j)^2 = \sum_{i=1}^{n} y_{ij}^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_{ij}\right)^2$$

one can write:

$$\sum_{i=1}^{n} (y_{ij} - \bar{y}_j)^2 = \frac{n(n+1)(2n+1)}{6} - \frac{1}{n}\left[\frac{n^2(n+1)^2}{4}\right] = \frac{n^3 - n}{12}$$

It follows that, when using ranks, the numerator of eq. 5.2 becomes:

$$\frac{1}{2}\left[\sum_{i=1}^{n}(y_{ij}-\bar{y}_j)^2+\sum_{i=1}^{n}(y_{ik}-\bar{y}_k)^2-\sum_{i=1}^{n}d_i^2\right]=\frac{1}{2}\left[\frac{n^3-n}{12}+\frac{n^3-n}{12}-\sum_{i=1}^{n}d_i^2\right]$$

while its denominator reduces to:

$$\sqrt{\sum_{i=1}^{n}(y_{ij}-\bar{y}_j)^2\sum_{i=1}^{n}(y_{ik}-\bar{y}_k)^2}=\sqrt{\left(\frac{n^3-n}{12}\right)\left(\frac{n^3-n}{12}\right)}=\frac{n^3-n}{12}$$

The final formula is obtained by replacing the above two expressions in eq. 5.2. This development shows that, when using ranks, eq. 5.1 simplifies to the following formula for Spearman *r*:

$$r_{jk}=\frac{\frac{1}{2}\left[\frac{n^3-n}{12}+\frac{n^3-n}{12}-\sum_{i=1}^{n}d_i^2\right]}{\frac{n^3-n}{12}}=1-\frac{6\sum_{i=1}^{n}d_i^2}{n^3-n} \qquad \textbf{(5.3)}$$

Alternatively, the Spearman rank correlation coefficient can be obtained in two steps: (1) replace all observations by ranks (columnwise) and (2) compute the Pearson correlation coefficient (eq. 4.7, formula identical to eq. 5.1) between the ranked variables. The result is the same as obtained from eq. 5.3.

The Spearman *r* coefficient varies between +1 and –1, just like the Pearson *r*. Descriptors that are perfectly matched, in terms of ranks, exhibit values $r = +1$ (direct relationship) or $r = -1$ (inverse relationship), whereas $r = 0$ indicates the absence of a monotonic relationship between the two descriptors. (Relationships that are not monotonic, e.g. Fig. 4.4d, can be quantified using polynomial or nonlinear regression, or else contingency coefficients; see Section 6.2 and Subsection 10.3.4.)

Numerical example. A small example (ranked data, Table 5.4) illustrates the equivalence between eq. 5.1 computed on ranks and eq. 5.3. Using eq. 5.1 gives:

$$r_{12}=\frac{-2}{\sqrt{5\times 5}}=\frac{-2}{5}=-0.4$$

The same result is obtained from eq. 5.3:

$$r_{12}=1-\frac{6\times 14}{4^3-4}=1-\frac{84}{60}=1-1.4=-0.4$$

Two or more objects may have the same rank on a given descriptor. This is often the case with descriptors used in ecology, which may have a small number of states or ordered classes. Such observations are said to be *tied*. Each of them is assigned the

Table 5.4 Numerical example. Ranks of four objects on two descriptors, $\mathbf{y}_1$ and $\mathbf{y}_2$.

Objects (observation units)	Ranks of objects on the two descriptors	
	$\mathbf{y}_1$	$\mathbf{y}_2$
$\mathbf{x}_1$	3	3
$\mathbf{x}_2$	4	1
$\mathbf{x}_3$	2	4
$\mathbf{x}_4$	1	2

average of the ranks that would have been assigned had no ties occurred. If the proportion of tied observations is large, correction factors must be introduced into the sums of squared deviations of eq. 5.2, which become:

$$\sum_{i=1}^{n} (y_{ij} - \bar{y}_j)^2 = \frac{1}{12}\left[(n^3 - n) - \sum_{r=1}^{q} (t_{rj}^3 - t_{rj}) \right]$$

and

$$\sum_{i=1}^{n} (y_{ik} - \bar{y}_k)^2 = \frac{1}{12}\left[(n^3 - n) - \sum_{r=1}^{s} (t_{rk}^3 - t_{rk}) \right]$$

where t_{rj} and t_{rk} are the numbers of observations in descriptors $\mathbf{y}_j$ and $\mathbf{y}_k$ that are tied at ranks r, these values being summed over the q sets of tied observations in descriptor j and the s sets in descriptor k.

Significance of the Spearman coefficient is usually tested against the null hypothesis H_0: $r = 0$. When $n \geq 10$, the test statistic is the same as for Pearson r (eq. 4.13):

$$t = \sqrt{\nu} \frac{r_{ki}}{\sqrt{1 - r_{ki}^2}} \qquad \textbf{(5.4)}$$

H_0 is tested by comparing statistic t to the value found in a table of critical values of t with $\nu = n - 2$ degrees of freedom. H_0 is rejected when the probability corresponding to t is smaller than or equal to a predetermined level of significance (α, for a two-tailed test). The rules for one-tailed and two-tailed tests are the same as for the Pearson r (Section 4.2). When $n < 10$, which is not often the case in ecology, one must refer to a

Table 5.5 Numerical example. The order of the four objects (rows) of Table 5.4 has been rearranged in such a way that the ranks on $\mathbf{y}_1$ are now in increasing order

Objects (observation units)	Ranks of objects on the two descriptors	
	$\mathbf{y}_1$	$\mathbf{y}_2$
$\mathbf{x}_4$	1	2
$\mathbf{x}_3$	2	4
$\mathbf{x}_1$	3	3
$\mathbf{x}_2$	4	1

special table of critical values of the Spearman rank correlation coefficient, found in textbooks of nonparametric statistics.

2 — Kendall τ

Kendall corr. coeff.

Kendall τ (tau) is another rank correlation coefficient, which can be used for the same types of descriptors as Spearman *r*. One major advantage of τ over Spearman *r* is that the former can be generalized to a partial correlation coefficient (below), which is not the case for the latter. While Spearman *r* was based on the differences between the ranks of objects on the two descriptors being compared, Kendall τ refers to a somewhat different concept, which is best explained using an example.

Numerical example. Kendall τ is calculated on the example of Table 5.4, already used for computing Spearman *r*. In Table 5.5, the order of the objects was rearranged so as to obtain increasing ranks on one of the two descriptors (here $\mathbf{y}_1$). The table is used to determine the degree of dependence between the two descriptors. Since the ranks are now in increasing order on $\mathbf{y}_1$, it is sufficient to determine how many *pairs of ranks* are also in increasing order on $\mathbf{y}_2$ to obtain a measure of the association between the two descriptors. Considering the object in first rank (i.e. $\mathbf{x}_4$), at the top of the right-hand column, the first pair of ranks (2 and 4, belonging to objects $\mathbf{x}_4$ and $\mathbf{x}_3$) is in increasing order; a score of +1 is assigned to it. The same goes for the second pair (2 and 3, belonging to objects $\mathbf{x}_4$ and $\mathbf{x}_1$). The third pair of ranks (2 and 1, belonging to objects $\mathbf{x}_4$ and $\mathbf{x}_2$) is in decreasing order, however, so that it earns a negative score –1. The same operation is repeated for every object in successive ranks along $\mathbf{y}_1$, i.e. for the object in second rank ($\mathbf{x}_3$): first pair of ranks (4 and 3, belonging to objects $\mathbf{x}_3$ and $\mathbf{x}_1$), etc. The sum *S* of scores assigned to each of the $n(n-1)/2$ different pairs of ranks is then computed.

Kendall's rank correlation coefficient is defined as follows:

$$\tau_a = \frac{S}{n(n-1)/2} = \frac{2S}{n(n-1)} \qquad \textbf{(5.5)}$$

Table 5.6 Numerical example. Contingency table giving the distribution of 80 objects among the states of two semiquantitative descriptors, **a** and **b**. Numbers in the table are frequencies (*f*).

	b_1	b_2	b_3	b_4	t_j
a_1	20	10	10	0	40
a_2	0	10	0	10	20
a_3	0	0	10	0	10
a_4	0	0	0	10	10
t_k	20	20	20	20	80

where *S* stands for "sum of scores". Kendall τ_a is thus the sum of scores for pairs in increasing and decreasing order divided by the total number of pairs ($n(n-1)/2$). For the example of Tables 5.4 and 5.5, τ_a is:

$$\tau_a = \frac{2\,(1+1-1-1-1-1)}{4 \times 3} = \frac{2\,(-2)}{12} = -0.33$$

Clearly, in the case of perfect agreement between two descriptors, all pairs receive a positive score, so that $S = n(n-1)/2$ and thus $\tau_a = +1$. When there is complete disagreement, $S = -n(n-1)/2$ and thus $\tau_a = -1$. When the descriptors are totally unrelated, the positive and negative scores cancel out, so that *S* as well as τ_a are 0 or near 0.

Equation 5.5 cannot be used for computing τ when there are tied observations. This is often the case with ecological *semiquantitative* descriptors, which may have a small number of states. The Kendall rank correlation is then computed on a contingency table crossing two semiquantitative descriptors.

Table 5.6 is a contingency (or frequency) table crossing two ordered descriptors. For example, descriptor **a** could represent the relative abundances of arthropods in soil enumerated on a semiquantitative scale (e.g. absent, present, abundant, and very abundant), while descriptor **b** could be the concentration of organic matter in the soil, divided into 4 classes. For simplicity, descriptors are called **a** and **b** here, as in Chapter 6. The states of **a** vary from 1 to *r* (number of rows) while the states of **b** go from 1 to *c* (number of columns).

To compute τ with tied observations, *S* is calculated as the difference between the numbers of positive (*P*) and negative (*Q*) scores, $S = P - Q$. *P* is the sum of all

frequencies f in the contingency table, each one multiplied by the sum of all frequencies located *lower* and on its *right*:

$$P = \sum_{j=1}^{r} \sum_{k=1}^{c} \left[f_{jk} \times \sum_{l=j+1}^{r} \sum_{m=k+1}^{c} f_{lm} \right]$$

Likewise, Q is the sum of all frequencies f in the table, each one multiplied by the sum of all frequencies *lower* and on its *left*:

$$Q = \sum_{j=1}^{r} \sum_{k=1}^{c} \left[f_{jk} \times \sum_{l=j+1}^{r} \sum_{m=1}^{k-1} f_{lm} \right]$$

Numerical example. For the data in Table 5.6:

$$P = (20 \times 40) + (10 \times 30) + (10 \times 20) + (10 \times 20) + (10 \times 10) = 1600$$
$$Q = (10 \times 10) + (10 \times 10) = 200$$
$$S = P - Q = 1600 - 200 = 1400$$

Using this value S, there are two approaches for calculating τ, depending on the numbers of states in the two descriptors. When **a** and **b** have the same numbers of states ($r = c$), τ_b is computed using a formula that includes the total number of pairs $n(n-1)/2$, as in the case of τ_a (eq. 5.5). The difference with eq. 5.5 is that τ_b includes corrections for the number of pairs L_1 tied in **a** and the number of pairs L_2 tied in **b**, where

$$L_1 = \sum_{j=1}^{r} \frac{1}{2} t_j (t_j - 1) \quad \text{in which } t_j \text{ is the marginal total for row } j$$

$$L_2 = \sum_{k=1}^{c} \frac{1}{2} t_k (t_k - 1) \quad \text{in which } t_k \text{ is the marginal total for column } k.$$

The formula for τ_b is:

$$\tau_b = \frac{S}{\sqrt{\frac{1}{2} n(n-1) - L_1}\sqrt{\frac{1}{2} n(n-1) - L_2}} \qquad \textbf{(5.6)}$$

When there are no tied observations, $L_1 = L_2 = 0$ and eq. 5.6 is identical to eq. 5.5.

Numerical example. For the data in Table 5.6:

$$L_1 = \frac{40 \times 39}{2} + \frac{20 \times 19}{2} + \frac{10 \times 9}{2} + \frac{10 \times 9}{2} = 1060$$

$$L_2 = \frac{20 \times 19}{2} + \frac{20 \times 19}{2} + \frac{20 \times 19}{2} + \frac{20 \times 19}{2} = 760$$

$$\tau_b = \frac{1400}{\sqrt{\frac{1}{2}(80 \times 79) - 1060}\sqrt{\frac{1}{2}(80 \times 79) - 760}} = 0.62$$

Without correction for ties, the calculated value (eq. 5.5) would have been

$$\tau_a = (2 \times 1400) / (80 \times 79) = 0.44$$

The second approach for calculating τ with tied observations is used when **a** and **b** do not have the same number of states ($r \neq c$). The formula for τ_c uses the minimum number of states in either **a** or **b**, $min(r, c)$:

$$\tau_c = \frac{S}{\frac{1}{2}n^2\left(\frac{min-1}{min}\right)} \quad \textbf{(5.7)}$$

The significance of Kendall τ is tested against the null hypothesis H_0: $r = 0$ (i.e. independence of the two descriptors). Kendall (1948) has shown that the distribution of τ approximates the normal distribution with mean $\mu_\tau = 0$ and standard deviation $\sqrt{2(2n+5)(9n(n-1))}$. Hence a z-test statistic can be obtained by transforming τ into a standard normal variate z using the formula:

$$z = \left[|\tau|\sqrt{\frac{9n(n-1)}{2(2n+5)}}\right] - \sqrt{\frac{18}{n(n-1)(2n+5)}} \quad \textbf{(5.8)}$$

With this statistic, H_0 can be tested using a table of z (or t_∞). Since z tables are one-tailed, the z-statistic of eq. 5.8 may be used directly for one-tailed tests by comparing it to the value z_α read in the table. For two-tailed tests, the statistic is compared to the value $z_{\alpha/2}$ from the z-table. When n is large, the second term of eq. 5.8 (correction for small n) becomes small: for $n = 30$, its value is 0.0178, and it is 0.0084 for $n = 50$.

Spearman r provides a better approximation of Pearson r when the data are almost quantitative and there are but a few tied observations, whereas Kendall τ does better when there are many ties. Computing both Spearman r and Kendall τ_a on the same numerical example, above, produced different numerical values (i.e. $r = -0.40$ versus $\tau_a = -0.33$). This is because the two coefficients have different underlying scales, so that their numerical values cannot be directly compared. However, given their different sampling distributions, they both reject H_0 at the same level of significance. If applied to quantitative data that are meeting all the requirements of Pearson r, both Spearman r and Kendall τ have power nearly as high (about 91%; Hotelling & Pabst, 1936) as their parametric equivalent. In all other cases, they are more powerful than Pearson r. This

Power

refers to the notion of *power* of statistical tests: a test is more powerful than another if it is more likely to detect small deviations from H_0 (i.e. smaller type II error), for constant type I error.

The chief advantage of Kendall τ over Spearman *r*, as already mentioned, is that it can be generalized to a partial correlation coefficient, which cannot be done with Spearman (Siegel, 1956: 214). The formula for a partial τ is:

$$\tau_{12.3} = \frac{\tau_{12} - \tau_{13}\tau_{23}}{\sqrt{1-\tau_{13}^2}\sqrt{1-\tau_{23}^2}} \qquad \textbf{(5.9)}$$

This formula is algebraically the same as that of first-order partial Pearson *r* (eq. 4.36) although, according to Kendall (1948: 103), this would be merely coincidental because the two formulae are derived using entirely different approaches. The three τ coefficients on the right-hand side of eq. 5.9 may themselves be partial τ's, thus allowing one to control for more than one descriptor (i.e. high order partial correlation coefficients). Siegel & Castellan (1988) give tables for testing the significance of the Kendall partial correlation coefficient.

Rank correlation coefficients should not be computed in the Q mode, i.e. for comparing objects instead of descriptors; see Box 7.1, Chapter 7.

5.4 Coefficient of concordance

Kendall coeff. of concordance

The rank correlation coefficients described in the previous section measure the correlation between two descriptors for *n* objects. Kendall's coefficient of concordance *W* (Kendall & Babington Smith, 1939) measures the agreement among several (*p*) quantitative or semiquantitative variables over a set of *n* objects. In community ecology, the *p* variables may be species whose abundances are used to assess habitat quality at *n* study sites. In taxonomy, they may be *p* characters measured over *n* different species, biological populations, or individuals. In the social sciences, the variables are often *p* "judges" assessing *n* different subjects or situations.

Friedman's two-way ANOVA

There is a close relationship between Friedman's two-way analysis of variance without replication by ranks (Section 5.2) and Kendall's coefficient of concordance. Indeed, they both address hypotheses concerning the same data table and use the same statistic for testing. They only differ in the formulation of their respective null hypothesis. Consider Table 5.7, which contains illustrative data. In Friedman's test, the null hypothesis is that there is no real difference among the $n = 6$ objects because they pertain to the same statistical population. Under H_0, they should have received random ranks along the $p = 3$ variables, so that their sums of ranks should be approximately equal. Kendall's test focuses on the relationships among the $p = 3$ variables. If the null hypothesis of Friedman's test is true, this means that the variables have produced rankings of the objects that are independent of one another. This is the null hypothesis of Kendall's test of *W*.

Table 5.7 Numerical example. Ranks of six objects on three descriptors, $\mathbf{y}_1$, $\mathbf{y}_2$, and $\mathbf{y}_3$.

Objects (observation units)	Ranks of objects on the three descriptors			Row sums
	$\mathbf{y}_1$	$\mathbf{y}_2$	$\mathbf{y}_3$	R_i
$\mathbf{x}_1$	1	1	6	8
$\mathbf{x}_2$	6	5	3	14
$\mathbf{x}_3$	3	6	2	11
$\mathbf{x}_4$	2	4	5	11
$\mathbf{x}_5$	5	2	4	11
$\mathbf{x}_6$	4	3	1	8

1 — Computing Kendall W

The Kendall W coefficient is an estimate of the variance of the row sums of ranks R_i divided by the maximum possible value the variance can take; this occurs when all variables are in total agreement. Hence $0 \leq W \leq 1$, the value 1 representing perfect concordance. There are two ways of computing the Kendall W coefficient (i.e. either form of eq. 5.11); they lead to the same result. The computation proceeds in two steps.

Firstly, S or S' is computed from the row-marginal sums of ranks R_i received by the objects:

$$S = \sum_{i=1}^{n} (R_i - \bar{R})^2 \qquad \text{or} \qquad S' = SSR = \sum_{i=1}^{n} R_i^2 \qquad \textbf{(5.10)}$$

where S is a sum of squared deviations statistic over the row sums of ranks R_i and $\bar{R}$ is the mean of the R_i values. SSR designates the Sum of Squared R_i values.

Secondly, the Kendall W coefficient is obtained using either of the following formulas:

$$W = \frac{12S}{p^2 (n^3 - n) - pT} \qquad \text{or} \qquad W = \frac{12S' - 3p^2 n (n+1)^2}{p^2 (n^3 - n) - pT} \qquad \textbf{(5.11)}$$

where n is the number of objects and p the number of variables. To derive these formulas, one has to know that the sum of all ranks in the data table is $pn(n + 1)/2$ and

that the sum of squares of all ranks is $p^2 n(n+1)(2n+1)/6$. T is a correction factor for tied ranks (Siegel, 1956; Siegel & Castellan, 1988; Zar, 1999):

$$T = \sum_{k=1}^{g} (t_k^3 - t_k) \qquad \textbf{(5.12)}$$

in which t_k is the number of tied ranks in each (k) of g groups of ties. The sum is computed over all groups of ties found in all p variables of the data table. $T = 0$ when there are no tied values.

There is a close relationship between the Spearman r_S correlation coefficient and the Kendall W coefficient: W can be directly calculated from the mean ($\bar{r}_S$) of the pairwise Spearman correlations r_S using the following relationship (Siegel and Castellan, 1988; Zar, 1999):

$$W = \frac{(p-1)\,\bar{r}_S + 1}{p} \qquad \textbf{(5.13)}$$

where p is the number of variables among which the pairwise Spearman correlations are computed. Equation 5.13 is strictly true for untied observations only; for tied observations, ties are handled in a bivariate way in each Spearman r_S coefficient whereas in Kendall W the correction for ties is computed over all variables (eq. 5.12). For two variables only, W is simply a linear transformation of r_S: $W = (r_S + 1)/2$. In that case, a permutation test of W for two variables is the exact equivalent of a permutation test of r_S for the same variables.

The relationship described by eq. 5.13 clearly shows that W will consider p variables to be concordant only if their Spearman correlations are positive. Two variables that give perfectly opposite ranks to a set of objects have a Spearman correlation of –1, hence $W = 0$ for these two variables (eq. 5.13); this is the lower bound of the coefficient of concordance. For two variables only, $r_S = 0$ gives $W = 0.5$; for a group of p uncorrelated variables, $W = 1/p$. So coefficient W applies well to rankings given by a panel of "judges" called in to assess overall performance in sports, quality of wines, or food in restaurants, to rankings obtained from criteria used in quality tests of appliances or services by consumer organizations, or to the study of species associations in multi-species communities. It does not apply to variables used in multivariate analysis where negative as well as positive relationships are informative. Zar (1999), for example, used wing length, tail length and bill length of birds to illustrate the use of the coefficient of concordance. These data are appropriate for W because they are all related to the same common property, the size of the birds.

Numerical example. The calculation of Kendall's coefficient of concordance is illustrated using the numerical example of Table 5.7. The data could be semiquantitative rank scores, or quantitative descriptors coded into ranks. It is important to note that the $n = 6$ objects are ranked on each descriptor (column) separately. The last column gives, for each object i, the sum R_i of its

ranks on the $p = 3$ descriptors. The sum of squared deviations from the mean, $\sum (R_i - \bar{R})^2$ (eq. 5.10 left), is equal to 25.5 for this example. The W-statistic is calculated with eq. 5.11 (left):

$$W = \frac{12 \times 25.5}{9\,(216 - 6)} = 0.1619$$

There are no tied ranks in this example. The F and X^2 (chi-square) statistics are computed as follows (eqs. 5.14 and 5.15, next subsection):

$$F = \frac{(3 - 1) \times 0.1619}{(1 - 0.1619)} = 0.386$$

$$X^2 = 3 \times (6 - 1) \times 0.1619$$

The p-value associated with the F-statistic, found using the F-distribution, is 0.825. The permutational p-value after 999 random permutations within the variables is 0.835. The hypothesis (H_0) that the row sums R_i of Table 5.7 are equal cannot be rejected. The conclusion is that the 3 descriptors differ in the way they rank the 6 objects.

2 — Testing the significance of W

The recommended method for testing the significance of W is to compute the following F-statistic:

$$F = \frac{(p - 1)\,W}{(1 - W)} \qquad \textbf{(5.14)}$$

which is asymptotically distributed like F with $\nu_1 = n - 1 - (2/p)$ and $\nu_2 = \nu_1 (p - 1)$ degrees of freedom (Kendall & Babington Smith, 1939). Numerical simulations showed that this F-statistic had correct levels of type I error for any value of n and p (Legendre, 2010). It is unfortunate that this statistic has been overlooked by authors of recent textbooks on nonparametric statistics who recommend testing the significance of W with Friedman's (1937) X^2-statistic, which is obtained from W as follows:

$$X^2 = p(n - 1)\,W \qquad \textbf{(5.15)}$$

This X^2 (chi-square) statistic is asymptotically distributed like χ^2 with $\nu = (n - 1)$ degrees of freedom. Kendall & Babington Smith (1939) considered this test of W to be satisfactory for moderately large values of p and n only, not for small p. This was confirmed by simulations reported by Legendre (2005), who recommended not to use the theoretical χ^2-distribution to test X^2 when $p < 20$. The X^2-statistic can, however, be tested by permutation.

Permutation test

Permutation tests can be used with all combinations of values of p and n (Legendre, 2005). For the global test of significance, the rank values in all variables are permuted at random, independently over each variable, because the null hypothesis is the independence of the rankings produced by the p variables. The alternative hypothesis (H_1) is that at least one of the variables is concordant with one or more of

the other variables; so when H_0 is rejected, one cannot conclude that all variables are concordant with one another, but only that at least one variable is concordant with one or more of the others. Actually, for permutation testing, the four statistics *SSR* (eq. 5.10), *W* (eq. 5.11), *F* (eq. 5.14), and X^2 (eq. 5.15) are monotonic to one another since *n*, *p* and *T* are constant within a given permutation test; they are thus equivalent statistics for testing since they produce the same permutational probabilities. The test is one-tailed because it only recognizes positive associations between the ranked variables.

Many of the problems subjected to Kendall's concordance analysis involve fewer than 20 variables: the parametric χ^2-test should be avoided in these cases. The *F*-test (eq. 5.14) and the permutation test can be safely used with all values of *p* and *n*.

3 — Contributions of individual variables to Kendall's concordance

The contribution of individual variables (e.g. the *p* species) to the *W*-statistic can be assessed by a permutation test proposed by Legendre (2005). The null hypothesis is the monotonic independence of the variable subjected to the test with respect to all other variables in the group under study. The alternative hypothesis is that this variable is positively correlated with one or several other variables in the set under study (one-tailed test). The statistic *W* can be used directly in *a posteriori* permutation tests; alternatively, one can use two other statistics described in Legendre (2005) that are equivalent to *W* for *a posteriori* tests. Contrary to the global test, only the variable under test (e.g. one of the *p* species) is permuted here. If that variable has values that are monotonically independent of the other variables, permuting its values at random should have little influence on the *W*-statistic. If on the contrary it is concordant with one or several other variables, permuting its values at random should break the concordance and induce a noticeable decrease of *W*.

Concordance analysis is applied to the identification of species associations in Subsection 8.9.2, where an ecological application (mite data) is presented. Another example (fish associations) is found in Section 4.10.2 of Borcard *et al*. (2011).

Concordance analysis is also useful in phylogenetic analysis: prior to phylogenetic reconstruction, the degree of congruence among distance matrices (CADM) corresponding to different types of data or different genes can be tested using a test of significance proposed by Legendre & Lapointe (2004). The distance matrices under comparison are strung out like the descriptors in Table 5.7. The coefficient of concordance (*W*, eq. 5.11) is computed, then tested using the same permutation procedure as in the Mantel test (Subsection 10.5.1). The CADM test is actually a generalization of the Mantel test of correspondence between two distance matrices to any number of distance matrices. It can be used to compare distance matrices computed from evolutionary data (genetic congruence), the topologies of phylogenetic trees derived from these data (topological congruence), or the full phylogenetic trees

including topologies and branch lengths (phylogenetic congruence) (Campbell *et al*., 2011). Applications of this method are found in Campbell *et al*. (2009, 2011).

5.5 Software

All major commercial statistical packages allow the calculation of rank correlation coefficients, as well a choice of the methods listed in Table 5.2. In the R language,

1. Methods listed in Table 5.2 are available in the following functions of the STATS package: ***t.test()*** (*t*-test for independent and related samples), ***aov()*** (different forms of ANOVA), ***wilcox.test()*** (Mann-Whitney and Wilcoxon tests), ***kruskal.test()*** (Kruskal-Wallis test), ***friedman.test()*** (Friedman test), ***chisq.test()*** (chi-square test), ***fisher.test()*** (Fisher exact probability test), and ***mcnemar.test()*** (McNemar test). ***chisq.test()*** and ***fisher.test()*** offer permutation tests among their options. Rank correlation coefficients are available as options in function ***cor()*** of the STATS package, which can also be used to compute correlation matrices among several descriptors.

2. Logistic regression can be computed using the ***glm()*** function of the STATS package. Multinomial logistic regression is computed by function ***mlogit()*** of the MLOGIT package.

3. The global coefficient of concordance and *a posteriori* tests are available in functions ***kendall.global()*** and ***kendall.post()*** of VEGAN. Congruence among distance matrices is available in functions ***CADM.global()*** and ***CADM.post()*** of APE.

Chapter 6 *Multidimensional qualitative data*

6.0 General principles

Ecologists often use variables that are neither quantitative nor ordered (Table 1.2). Variables of this type may be of physical or biological nature. Examples of qualitative physical descriptors are the colour, locality, geological substrate, or nature of surface deposits. Qualitative biological descriptors include the captured or observed species, where the different states of the nonordered descriptor are the different possible species. Likewise, the presence or absence of a species cannot, in most cases, be analysed as a quantitative variable; it must be treated as a semiquantitative or qualitative descriptor. A third group of qualitative descriptors includes the results of classifications — for example, the biological associations to which the zooplankton of various lakes belong, or the chemical groups describing soil cores. Such classifications, obtained or not by clustering (Chapter 8), define qualitative descriptors and, as such, they are amenable to numerical interpretation (see Chapter 10).

The present chapter discusses the analysis of *qualitative* descriptors; methods appropriate for bivariate and multivariate analysis are presented. Because information theory is an intuitively appealing way of introducing these methods of analysis, Section 6.1 shows how to measure the amount of information in a qualitative descriptor. This paradigm is then used in the following sections.

Contingency table

The comparison of qualitative descriptors is based on *contingency tables*. In order to compare two qualitative descriptors, the objects are first allocated to the cells of a two-way contingency table whose rows and columns respectively correspond to the two descriptors. In such a table, the number of rows is equal to the number of states of the first descriptor and the number of columns to that of the second descriptor. Any cell in the table, at the intersection of a row and a column, corresponds to one state of each descriptor. The number of objects with these two states is recorded in the cell, hence the values in contingency tables are *frequencies*. The analysis of *two-way contingency tables* is described in Section 6.2. When there are more than two descriptors, *multiway*

(or *multidimensional*) *contingency tables* are constructed as extensions of two-way tables. Their analysis is discussed in Section 6.3. Finally, Section 6.4 analyses the *correspondence* between descriptors in a contingency table.

Contingency table analysis is the qualitative equivalent of both *correlation analysis* and *analysis of variance*; in the particular case of a two-way contingency table, the analysis is the equivalent of a one-way ANOVA. It involves the computation of X^2 (chi-square) statistics or related measures, instead of correlation or F-statistics. Two types of null hypotheses (H_0) may be tested. The first one is the independence of the two descriptors, which is the usual null hypothesis in correlation analysis (H_0: the correlation coefficient $\rho = 0$ in the statistical population). The second type of hypothesis is similar to that of the analysis of variance. In a two-way contingency table, one of the descriptors (called "first descriptor" in the next sentence) corresponds to the classification criterion of the analysis of variance, and the other descriptor (called "second descriptor") corresponds to the dependent variable. The analysis compares, among the states of the first descriptor, the distribution of frequencies among the states of the second descriptors. The null hypothesis says that the frequency distributions are the same, i.e. that the observations form a homogeneous group. For example, if the groups (classification criterion) form the columns whereas the dependent variable is in the rows, H_0 states that the frequency distributions of the row frequencies are the same in all columns. These two types of hypotheses require the calculation of the same expected values and the same test statistics. The examples in the present chapter will be formulated as correlation hypotheses. In multiway tables, the hypotheses tested are often quite complex because they take into account interactions among the descriptors (Section 6.3).

Correlation hypothesis

ANOVA hypothesis

Considering species data, the names of the various species observed at a sampling site are the states of a qualitative multi-state descriptor. Section 6.5 will discuss *species diversity* as a measure of dispersion of this qualitative descriptor.

The mathematics used throughout this chapter are quite simple and require no prior knowledge other than the intuitive notion of probability. Readers interested in applications only may skip Section 6.1 and come back to it when necessary. To simplify the notation, the following conventions are followed throughout the chapter. When a single descriptor is considered, this descriptor is called **a** and its states have subscripts i going from 1 to q, as in Fig. 1.1. In two-way contingency tables, the descriptors are called **a** and **b**. The states of **a** are denoted a_i with subscripts i varying from 1 to r (number of rows), while the states of **b** are denoted b_j with subscripts j varying from 1 to c (number of columns).

6.1 Information and entropy

Chapters 1 and 2 have shown that the ecological information available about the objects under study is usually (or may be reformulated as) a set of biological and/or

environmental characteristics, which correspond to as many descriptors. Searching for groups of descriptors that behave similarly across the set of objects, or that may be used to forecast one from the other(s) (R analysis, Section 7.1), requires measuring the *amount of information* that these descriptors have in common. In the simplest case of two descriptors **a** and **b** (called $\mathbf{y}_1$ and $\mathbf{y}_2$ in previous chapters), one must assess how much *information* is provided by the distribution of the objects among the states of **a**, that could be used to forecast their distribution among the states of **b**. This approach is central to the analysis of relationships among ecological phenomena.

In 1968, Ludwig von Bertalanffy wrote, in his *General System Theory* (p. 32): "Thus, there exist models, principles, and laws that apply to generalized systems or their subclasses, irrespective of their particular kind, the nature of their component elements, and the relations or 'forces' between them". This is the case with information, which can be viewed and measured in the same manner for all systems. Some authors, including Pielou (1975), think that the concepts derived from information theory are, in ecology, a model and not a homology. Notwithstanding this opinion, the following sections will discuss how to measure information for biological descriptors in terms of information to be acquired, because such a presentation provides a better understanding of the nature of information in ecological systems.

The approach consists in measuring the amount of information contained in each descriptor and, further, the amount of information that two (or several) descriptors have in common. If, for example, two descriptors share 100% of their information, then they obviously carry the same information. Since descriptors are constructed so as to partition the objects under study into a number of states, two descriptors have 100% of their information in common when they partition a set of objects in exactly the same way, i.e. into equal and corresponding sets of states. When descriptors are qualitative, this correspondence does not need to follow any ordering of the states of the two descriptors. For ordered descriptors, the ordering of the correspondence between states is important and the techniques for analysing the information in common belong to correlation analysis (Chapters 4 and 5).

Entropy

The mathematical theory of information is based on the concept of *entropy*. Its mathematical formulation was developed by Shannon (Bell Laboratories) who proposed, in 1948, the well-known equation*:

$$H = -\sum_{i=1}^{q} p_i \log p_i \qquad \textbf{(6.1)}$$

* This equation is sometimes referred to as the Shannon-Weaver or the Shannon-Wiener equation. Norbert Wiener had developed elements of probability theory that were used by Claude E. Shannon in his 1948 paper. In 1963, Warren Weaver co-authored with Shannon a book where Shannon's 1948 article was reprinted.

Table 6.1 Contingency table (numerical example). Distribution of 120 objects on descriptors **a** and **b**.

	b_1	b_2	b_3	b_4
	30	30	30	30
$a_1 = 60$	30	10	15	5
$a_2 = 30$	0	20	0	10
$a_3 = 15$	0	0	0	15
$a_4 = 15$	0	0	15	0

where H is a measure of the uncertainty or choice associated with a frequency distribution (vector) **p**; p_i is the probability that an observation belongs to state i of the descriptor (Fig. 1.1). In practice, p_i is the proportion (or relative frequency, on a 0-1 scale) of observations in state i. Shannon recognized that his equation was similar to the equation of entropy, published in 1898 by physicist Boltzmann as a quantitative formulation of the second law of thermodynamics, which concerns the degree of disorganization in closed physical systems. He thus concluded that H corresponds to the entropy of information systems.

Negative entropy

The entropy of information theory is actually the *negative entropy* of physicists. In thermodynamics, an increase in entropy corresponds to an *increase in disorder*, which is accompanied by a *decrease of information*. Strictly speaking, information is negative entropy and it is only for convenience that it is simply called entropy. *In information theory, entropy and information are taken as synonymous*.

Information

Numerical example. In order to facilitate the understanding of the presentation up to Section 6.4, a small numerical example will be used in which 120 objects are described by two descriptors (**a** and **b**) with 4 states each. The question is to determine to what extent one descriptor can be used to forecast the other. The data in the numerical example could result from the survey of 120 sites of an estuary, or the trees observed in 120 vegetation quadrats. Descriptor **a** could be the dominant species at each sampling site, assuming there are 4 possible species, and descriptor **b**, some environmental variable with 4 states. The following discussion is valid for any type of qualitative descriptor as well as for ordered descriptors divided into classes.

Assume that the 120 observations are distributed as 60, 30, 15 and 15 among the 4 states of descriptor **a** and that there are 30 observations in each of the 4 states of descriptor **b**. The frequencies in the combined states of the descriptors (i.e. the table cells) are shown in Table 6.1.

For each descriptor, the probability of a state is estimated by the relative frequency with which the state is found in the set of observations. Thus, the probability distributions associated with descriptors **a** and **b** are:

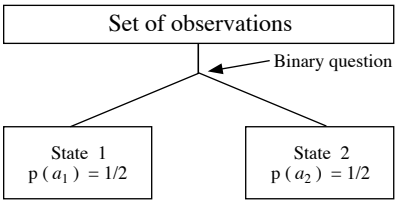
a_1:	60 p(a_1) = 1/2	b_1:	30 p(b_1) = 1/4
a_2:	30 p(a_2) = 1/4	b_2:	30 p(b_2) = 1/4
a_3:	15 p(a_3) = 1/8	b_3:	30 p(b_3) = 1/4
a_4:	15 p(a_4) = 1/8	b_4:	30 p(b_4) = 1/4
	120		120

The relative frequency of a given state is the probability of observing that state when taking an object at random.

Within the framework of information theory, the entropy of a probability distribution is measured, not in kilograms, metres per second, or other such units, but in terms of decisions. The measurement of entropy must reflect how difficult it is to find, among the objects under study, one that has a given state of the descriptor. An approximate measure of entropy is the average minimum number of binary questions that are required for assigning each object to its correct state. Hence, the *amount of information* gained by asking binary questions, and answering them after observing the objects, is equal to the *degree of disorder* or *uncertainty* initially displayed by the frequency distribution. Given that context, the terms *entropy* and *information* are used synonymously. A few numerical examples will help understand this measure.

1. When all the objects exhibit the same state for a descriptor, everything is known *a priori* about the distribution of observations among the different states of the descriptor. There is a single state in this case; hence, the number of binary questions required to assign a state to an object is zero ($H = 0$), which is the minimum value of entropy.

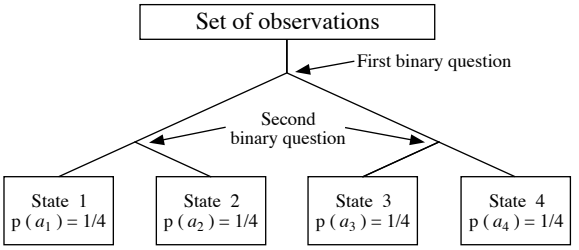
2. The simplest case of a descriptor with non-null entropy is when there are two states among which the objects are distributed equally:



Binary question

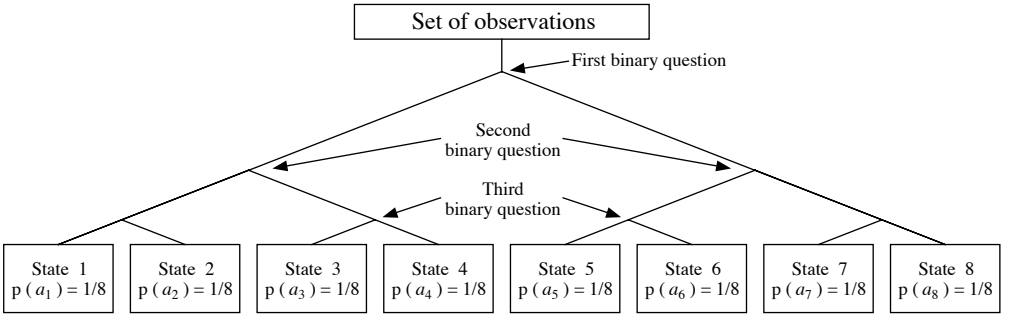
In order to assign a state to any given object, a single binary question is necessary, of the type "Does this object belong to state 1?" If it does, state 1 is assigned to the object; if it does not, the object belongs to state 2. The entropy associated with the descriptor is thus $H = 1$.

3. Applying the above approach to a descriptor with four states among which the objects are distributed equally, one gets an entropy $H = 2$ since exactly two binary questions are required to assign a state to each object:



This would be the case of descriptor **b** in the numerical example of Table 6.1.

4. For an eight-state descriptor with the objects equally distributed among the states, the binary questions are as follows:



The total entropy of the descriptor is thus:

$$[3 \text{ questions} \times 8 \text{ (1/8 of the objects)}] = 3$$

and, in general, the entropy H associated with a descriptor in which the objects are equally distributed among states is equal to the base 2 logarithm (if the questions are binary) of the number of states:

$\log_2 1 = 0$	$\log_2 8 = 3$
$\log_2 2 = 1$	$\log_2 16 = 4$
$\log_2 4 = 2$	etc.

Hence the general formula in that case is $H = \log_2(\text{number of states})$.

Measuring the entropy from the number of binary questions is strictly equal to the logarithmic measure only when the number of states is an integer power of 2, or when the number of observations in the various states is such that binary questions divide them into equal groups (see the numerical example, below). In all other cases, the number of binary questions required is slightly larger than $\log_2(\text{number of states})$,

Table 6.2 The average minimum number of binary questions required to remove the uncertainty about the position of an object in the state-vector is equal to $\log_2$(number of states) when the number of states is an integer power of 2 (in boldface) and the objects are equally distributed among the states. In all other cases, the number of binary questions is slightly larger than the entropy $H = \log_2$(number of states). For example, for a three-state descriptor with equal frequencies, the minimum number of binary questions is (2 questions × 2/3 of the objects) + (1 question × 1/3 of the objects) = 1.66666 binary questions.

Number of states	$\log_2$(number of states)	Average minimum number of binary questions
1	**0.00000**	**0.00000**
2	**1.00000**	**1.00000**
3	1.58496	1.66666
4	**2.00000**	**2.00000**
5	2.32193	2.40000
6	2.58496	2.66666
7	2.80735	2.85714
8	**3.00000**	**3.00000**
9	3.16993	3.22222
10	3.32193	3.40000
11	3.45943	3.54545
12	3.58496	3.66666
13	3.70044	3.76154
14	3.80735	3.85714
15	3.90689	3.93333
16	**4.00000**	**4.00000**

because binary questions are then a little less efficient than in the previous case (Table 6.2). Binary questions have been used in the above discussion only to provide readers with a better understanding of entropy, the true measure being the logarithmic one. One may refer to Shannon (1948), or a textbook on information theory, for a more formal discussion of the measure of entropy.

The following example illustrates the relationship between probability and information. If an ecologist states that water in the Loch Ness is fresh, this is trivial since the probability of the event is 1 (information content null). If, however, he/she announces that she/he has captured a specimen of the famous monster, this statement contains much information because of its low probability (the dynamic aspects of Loch

Ness Monster populations have been discussed by Sheldon & Kerr, 1972, Scheider & Wallis, 1973, and Rigler, 1982; see also Lehn, 1979, and Lehn & Schroeder, 1981, for a physical explanation of the Loch Ness and other aquatic monsters). Thus, information theory deals with a specific technical definition of information, which may not correspond to the intuitive concept. A nontechnical example is that a book should contain the same amount of information before and after one has read it. From the information theory point of view, however, after one has read the book once, there is no information to be gained the next time he/she reads it — unless she/he has forgotten part of it after the first reading.

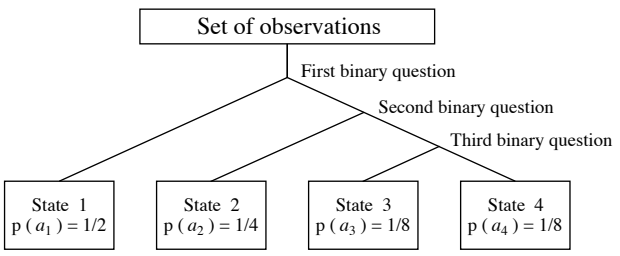
It should be clear, at this point of the discussion, that the entropy of a descriptor depends, among other characteristics, on the number of its states among which the entropy is partitioned. In the case of the above four-state descriptor, for example, 1/4 of the entropy of the descriptor is attributed to each state, i.e. $[1/4 \log_2 4]$, which is equal to $[1/4 \log_2(1/4)^{-1}]$. The total entropy of the descriptor is thus:

$$H = \sum_{4 \text{ states}} (1/4) \log_2 (1/4)^{-1} = \log_2 4 = 2$$

The same holds for the example of the eight-state descriptor. The entropy of each state is $[1/8 \log_2 8] = [1/8 \log_2(1/8)^{-1}]$, so that the total entropy of the descriptor is

$$H = \sum_{8 \text{ states}} (1/8) \log_2 (1/8)^{-1} = \log_2 8 = 3$$

5. Descriptor **a** in the numerical example (Table 6.1) illustrates the case of a descriptor for which the objects are not equally distributed among states. The probability distribution is [1/2, 1/4, 1/8, 1/8], which corresponds to the following scheme of optimal binary questions:



When the objects are not distributed evenly among the states, the amount of information one has *a priori* is higher than in the case of an even distribution, so that the information to be acquired by actual observation of the objects (i.e. the entropy) decreases. It follows that the entropy of the above descriptor should be $H < 2$, which is the maximum entropy for a four-state descriptor. Using binary questions, it is more economical to isolate half of the objects with the first question, then half of the remaining objects with the second question, and use a third question for the last two groups of 1/8 of the objects (see above). Since half of the objects require one question, 1/4 require 2, and the two groups of 1/8 require 3, the total entropy of this descriptor is:

$$H(\mathbf{a}) = (1/2 \times 1) + (1/4 \times 2) + (1/8 \times 3) + (1/8 \times 3) = 1.75$$

As in the previous examples, this is equal to:

$$H(\mathbf{a}) = 1/2\ 1\log_2 2 + 1/4\ 1\log_2 4 + 1/8\ 1\log_2 8 + 1/8\ 1\log_2 8$$

$$H(\mathbf{a}) = 1/2\ 1\log_2 (1/2)^{-1} + 1/4\ 1\log_2 (1/4)^{-1} + 1/8\ 1\log_2 (1/8)^{-1} + 1/8\ 1\log_2 (1/8)^{-1}$$

$$H(\mathbf{a}) = \sum_{\text{all states}} \mathrm{p}(i)\ \log_2 [\mathrm{p}(i)]^{-1}$$

Following the law of exponents for logarithms, exponent –1 is eliminated by writing the equation as:

$$H(\mathbf{a}) = -\sum_{\text{all states}} \mathrm{p}(i) \log_2 \mathrm{p}(i)$$

Bit
Hartley
Decit
Nat

This is Shannon's formula for entropy (eq. 6.1). When the base for the logarithms is 2, the model is that of binary questions and the unit of entropy is the *bit* (contraction of *binary digit*) or *hartley* (Pinty & Gaultier, 1971). The model may be reformulated using questions with 10 answers, in which case the base of the logarithms is 10 and the unit is the *decit*. For natural logarithms, the unit is the *nat* (Pielou, 1975). These units are dimensionless, as are angles for example (Chapter 3).

Communication

Equation 6.1 may be applied to *human communications*, to calculate the information content of strings of symbols. For example, in a system of numbers with base n, there are n^N possible numbers containing N digits (in a base-10 system, there are $10^2 = 100$ numbers containing 2 digits, i.e. the numbers 00 to 99). It follows that the information content of a number with N digits is:

$$H = \log_2 n^N = N \log_2 n$$

The information per symbol (digit) is thus:

$$H/N = \log_2 n \qquad \textbf{(6.2)}$$

In the case of a binary (base 2) number, the information per symbol is $\log_2 2 = 1$ bit; for a decimal (base 10) number, it is $\log_2 10 = 3.32$ bits. A decimal digit contains 3.32 bits of information so that, consequently, a *binary* representation requires on average 3.32 times more digits than a *decimal* representation of the same number.

Alphabet

For an alphabet possessing 27 symbols (26 letters and the blank space), the information per symbol is $\log_2 27 = 4.76$ bits, assuming that all symbols have the same frequency. In languages such as English and French, each letter has a frequency of its own, so that the information per symbol is less than 4.76 bits. The information per letter is 4.03 bits in English and 3.95 bits in French. Hence, the translation from French to English should entail shorter text, which is generally the case.

English
French

Each language is characterized by a number of properties, such as the frequencies of letters, groups of letters, etc. These statistical properties, together with a defined syntax, determine a particular structure. For a given alphabet, the specific constraints

Table 6.3 Redundancy in the French language. Number of lexical elements with 4 to 6 letters (from Bourbeau *et al*., 1984).

Number of letters	Possible number of lexical elements	Actual number of lexical elements in French
4	$26^4 \approx$ 457 000	3 558
5	$26^5 \approx$ 12 000 000	11 351
6	$26^6 \approx$ 300 000 000	24 800

of a language limit the number of messages that can actually be formulated. Thus, the number of lexical elements with 4, 5 or 6 letters is much smaller than the theoretical possible number (Table 6.3). This difference arises from the fact that every language contains a certain amount of information that is inherently embodied in its structure, which is termed *redundancy*. Without redundancy, it would be impossible to detect errors slipping into communications, since any possible group of symbols would have meaning.

In a language with n different symbols, each having a characteristic frequency ($N_1, N_2 \ldots N_n$), the total number of possible messages (P) made up of N symbols is equal to the number of *combinations*:

$$P = N! / (N_1!\, N_2! \ldots N_n!)$$

The information content of a message with N symbols is:

$$H = \log_2 P = \log_2 [N! / (N_1!\, N_2! \ldots N_n!)]$$

Hence, the information per symbol is:

$$H/N = 1/N \log_2 [N! / (N_1!\, N_2! \ldots N_n!)] \qquad \textbf{(6.3)}$$

which is the formula of Brillouin (1956). It will be used later (Section 6.5) to calculate the species diversity of a sample, considered to be representing a "message".

6.2 Two-way contingency tables

In order to compare two qualitative descriptors, the objects are allocated to the cells of a table with two criteria, i.e. the rows and columns. Each cell of the *two-way contingency table* (e.g. Tables 6.1 and 6.4) contains the number of observations

Table 6.4 Contingency table giving the observed (from Table 6.1) and expected (in parentheses) frequencies in each cell; $n = 120$. The observed frequencies that exceed the corresponding expected frequencies are in boldface. Wilks' chi-square statistic: $X^2_W = 150.7$ ($\nu = 9, p < 0.001$).

	b_1	b_2	b_3	b_4
	30	30	30	30
$a_1 = 60$	**30** (15)	10 (15)	15 (15)	5 (15)
$a_2 = 30$	0 (7.5)	**20** (7.5)	0 (7.5)	**10** (7.5)
$a_3 = 15$	0 (3.75)	0 (3.75)	0 (3.75)	**15** (3.75)
$a_4 = 15$	0 (3.75)	0 (3.75)	**15** (3.75)	0 (3.75)

described by that pair of states of the qualitative descriptors. Numbers in the cells of a contingency table are absolute frequencies, i.e. *not* relative frequencies. The number of cells in the table is equal to the product of the number of states in the two descriptors. The first question relative to a contingency table concerns the relationship between the two descriptors: given the bivariate distribution of observations in the table, are the two descriptors related to each other, or not? This question is answered by calculating the expected frequency E for each cell of the table, according to a null hypothesis H_0, and performing a chi-square (X^2) test of the null hypothesis.

Null hypothesis

The simplest null hypothesis is the independence of the two descriptors. E_{ij} is the number of observations that is expected in each cell (i, j) under H_0. Under this null hypothesis, E_{ij} is computed as the product of the marginal totals (i.e. the product of the sum of row i with the sum of column j), divided by n which is the total number of observations in the table:

Expected frequency

$$E_{ij} = [(\text{row sum})_i \times (\text{column sum})_j] / n \qquad \textbf{(6.4)}$$

This equation generates expected frequencies whose relative distribution across the states of descriptor **a**, *within* each state of descriptor **b**, is the same as the distribution of all observed data across the states of **a**, and conversely (Table 6.4). The null hypothesis is tested using a X^2-statistic that compares the observed (O_{ij}) to the expected frequencies (E_{ij}).

In basic statistics textbooks, the significance of relationships in two-way contingency tables is often tested using the *Pearson chi-square statistic* (Pearson, 1900):

Pearson chi-square

$$X_P^2 = \sum_{\text{all cells}} \frac{(O-E)^2}{E} \qquad \textbf{(6.5)}$$

where $(O - E)$ measures the contingency of each cell. Instead of X_P^2, one can compute Wilks' likelihood ratio (1935), also known as the *G* or *2I-statistic* (Sokal & Rohlf, 1995) or G^2 (Bishop *et al*., 1975; Dixon, 1981):

Wilks chi-square

$$X_W^2 = 2 \sum_{\text{all cells}} O \log_e\left(\frac{O}{E}\right) \qquad \textbf{(6.6)}$$

where $\log_e$ is the natural logarithm. For null frequencies, $\lim_{O \to 0} [O \log_e (O/E)] = 0$.

Degrees of freedom

For a contingency table with *r* rows and *c* columns, the number of degrees of freedom used to determine the probability (p-value) of the data under H_0 is:

$$\nu = (r-1)(c-1) \qquad \textbf{(6.7)}$$

When the p-value is smaller than or equal to a predetermined significance level, e.g. $\alpha = 0.05$, the null hypothesis (H_0) of independence of the descriptors is rejected.

When the number of observations (*n*) is large (i.e. larger than ten times the number of cells, *rc*, in the table), the asymptotic distributions of X_P^2 and X_W^2 are both χ^2. In other words, the two statistics are equivalent when H_0 is true. There is however a problem when the number of observations is small, i.e. less than five times the number of cells. Small numbers of observations often lead to several null observed values (O_{ij}) in the contingency table, with correspondingly very low expected frequencies (E_{ij}). According to Cochran (1954) and Siegel (1956), when there is at least *one* value of E_{ij} smaller than 1, or when 20% or more of the expected values E_{ij} are smaller than 5, some states (rows or columns) must be grouped to increase the expected frequencies, provided that there is a logical basis to do so. It now appears that only the first part of this empirical rule should be kept. Indeed Fienberg (1980, p.172) cites results of simulations indicating that, for $\alpha = 0.05$, the computed statistic is distributed like χ^2 if H_0 is true, as long as all E_{ij} values are larger than 1.

Concerning the choice of X_P^2 or X_W^2, there is no difference when the number of observations *n* is large (see the previous paragraph). When *n* is small, Larntz (1978) is of the opinion that X_P^2 is better than X_W^2. Sokal & Rohlf (1995) still recommend using X_W^2 but suggest to correct it as proposed by Williams (1976a) to obtain a better approximation of χ^2. This correction consists in dividing X_W^2 by a correction factor q_{min}. The correction factor, which is based on ν (eq. 6.7), is computed as:

Williams' correction

$$q_{min} = 1 + [(r^2-1)(c^2-1)/6\nu n] \qquad \textbf{(6.8)}$$

When n is large relative to the number of cells in the contingency table, it is not necessary to apply a correction to X^2_W since $q_{min} \approx 1$ in that case. William's correction is especially interesting when one must use X^2_W, as in the study of multiway contingency tables; the general formula for q_{min} is given in Subsection 6.3. Several computer programs allow users to compute both X^2_P and X^2_W.

Another correction, available in some computer programs, consists in adding a small value (e.g. 0.5) to *each* observed value O_{ij} in the contingency table when some of the O_{ij}'s are small. As indicated by Dixon (1981) and Sokal & Rohlf (1995), the effect of this correction is to lower the X^2-statistic, which makes the test more conservative. H_0 may then be rejected in a proportion of cases smaller than α when the null hypothesis is true.

Another measure of interest to ecologists, which is related to the Wilks statistic (see below), refers to the concept of entropy (or information) discussed above. In the numerical example with four rows and columns (Tables 6.1 and 6.4), if the correspondence between the states of descriptors **a** and **b** was perfect (i.e. descriptors completely dependent of each other), the contingency table would only have four non-zero cells — one in each row and each column. These non-zero cells could be anywhere in the table, not necessarily on the diagonal, because the states of the two descriptors are not ordered. It would then be possible, using **a**, to perfectly predict the distribution of observations among the states of **b**, and vice versa. In other words, given one state of the first descriptor, one would immediately know the state of the other descriptor. Thus, there would be no uncertainty (or entropy) concerning the distribution of the objects on **b** after observing **a**, hence the entropy remaining in **b** after observing **a** would be null, i.e. $H(\mathbf{b}|\mathbf{a}) = 0$. On the contrary, if the descriptors were completely independent of each other, the distribution of observations in each row of descriptor **a** would be in the same proportions as their overall distribution in **b** (found at top of Tables 6.1 and 6.4); the same would be true for the columns. $H(\mathbf{b}|\mathbf{a}) = H(\mathbf{b})$ would indicate that all the entropy contained in the distribution of **b** remains after observing **a**.

The two conditional entropies $H(\mathbf{a}|\mathbf{b})$ and $H(\mathbf{b}|\mathbf{a})$, as well as the entropy shared by the two descriptors, can be computed using the total information in the contingency table, $H(\mathbf{a},\mathbf{b})$, and the information of each descriptor, $H(\mathbf{a})$ and $H(\mathbf{b})$, already computed in Section 6.1. $H(\mathbf{a},\mathbf{b})$ is computed on all observed frequencies in the contingency table using Shannon's formula (eq. 6.1):

$$H(\mathbf{a},\mathbf{b}) = -\sum_{\text{states of } \mathbf{a}} \sum_{\text{states of } \mathbf{b}} p(i,j) \log p(i,j) \qquad \textbf{(6.9)}$$

where $p(i,j)$ is the observed frequency in each cell (i,j) of the contingency table, divided by the total number of observations n. For the example (Tables 6.1 or 6.4):

$$H(\mathbf{a},\mathbf{b}) = -\{1/4 \log_2 (1/4) + 1/6 \log_2 (1/6) + 3\,[1/8 \log_2 (1/8)] + 2\,[1/12 \log_2 (1/12)] + 1/24 \log_2 (1/24)\} = 2.84$$

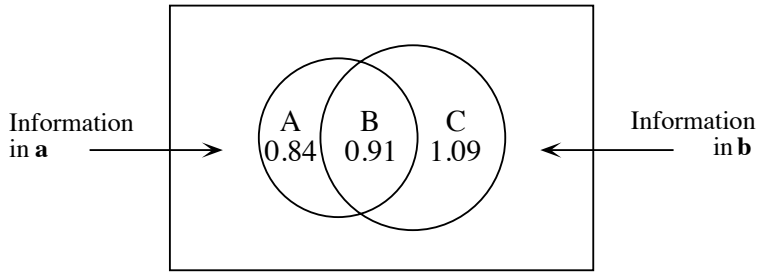


Figure 6.1 Venn diagram partitioning the information of two qualitative descriptors, denoted **a** and **b**. B is the information the two descriptors have in common.

The values of $H(\mathbf{a}) = A + B = 1.75$ and $H(\mathbf{b}) = B + C = 2.00$, represented by circles in the Venn diagram of Fig. 6.1, have been computed in Section 6.1. $H(\mathbf{a},\mathbf{b}) = 2.84$ is the total information in the union of the two descriptors, represented by A + B + C. The information (B) shared by the two descriptors is computed as follows:

$$B = (A + B) + (B + C) - (A + B + C)$$

$$B = H(\mathbf{a}) + H(\mathbf{b}) - H(\mathbf{a},\mathbf{b}) \qquad \textbf{(6.10)}$$

$$B = 1.75 + 2.00 - 2.84 = 0.91$$

With more decimals, B = 0.90564; this value is used in the example that follows eq. 6.14. The information exclusive to each descriptor, A and C, is computed by subtraction as follows:

$$A = (A + B + C) - (B + C)$$

$$A = H(\mathbf{a}|\mathbf{b}) = H(\mathbf{a},\mathbf{b}) - H(\mathbf{b}) \qquad \textbf{(6.11)}$$

$$A = 2.84 - 2.00 = 0.84$$

and

$$C = (A + B + C) - (A + B)$$

$$C = H(\mathbf{b}|\mathbf{a}) = H(\mathbf{a},\mathbf{b}) - H(\mathbf{a}) \qquad \textbf{(6.12)}$$

$$C = 2.84 - 1.75 = 1.09$$

There is a relationship between the reciprocal information B and Wilks X^2_W statistic. It can be shown that $B = (1/n) \sum O \log_e(O/E)$ when B is computed with natural logarithms ($\log_e$), or else $B \log_e 2 = (1/n) \sum O \log_e(O/E)$ when B is in bits. Using these relationships, it is possible to calculate the probability associated with B after transforming B into a Wilks X^2_W-statistic (eq. 6.6):

$$X^2_W = 2nB \quad \text{when B is in nats} \qquad \textbf{(6.13)}$$

or $X^2_W = 2n\mathrm{B}\log_e 2 = n\mathrm{B}\log_e 4 = 1.38629\, n\mathrm{B}$ when B is in bits. **(6.14)**

For the numerical example, $X^2_W = 2n\mathrm{B}\log_e 2 = 2 \times 120 \times 0.90564 \times 0.69315 = 150.66$ before Williams' correction.

Similarity

Using the measures of information A, B and C, various reciprocal information coefficients can be computed. The *similarity* of descriptors **a** and **b** can be calculated as the amount of information that the two descriptors have in common, divided by the total information of the system:

$$S(\mathbf{a},\mathbf{b}) = \mathrm{B} / (\mathrm{A} + \mathrm{B} + \mathrm{C}) \quad \textbf{(6.15)}$$

$S(\mathbf{a},\mathbf{b}) = 0.91 / 2.84 = 0.32$, for the numerical example.

If the following steps of the analysis (clustering and ordination, Chapters 8 and 9) require that the measure of association between **a** and **b** be a metric, one may use the corresponding distance, defined as the sum of the information that the two descriptors possess independently, divided by the total information:

Rajski's metric

$$D(\mathbf{a},\mathbf{b}) = (\mathrm{A} + \mathrm{C}) / (\mathrm{A} + \mathrm{B} + \mathrm{C}) \quad \textbf{(6.16)}$$

For the numerical example, $D(\mathbf{a},\mathbf{b}) = (0.84 + 1.09) / 2.84 = 0.68$. As indicated by the structure of the formulas, $S(\mathbf{a},\mathbf{b}) + D(\mathbf{a},\mathbf{b}) = 1$.

Coherence coefficient

The distance measure in eq. 6.16 is Rajski's metric (1961). This author also proposed another measure of similarity among descriptors, the *coherence coefficient*, which is used to assess the stochastic independence of two random variables:

$$S' = \sqrt{1 - D^2} \quad \textbf{(6.17)}$$

Another version of this coefficient,

$$\mathrm{S}'' = \mathrm{B} / (\mathrm{A} + 2\mathrm{B} + \mathrm{C}) \quad \textbf{(6.18)}$$

Symmetric, asymmetric uncertainty

is available in some computer programs under the name *symmetric uncertainty coefficient*. Two *asymmetric uncertainty coefficients* have also been proposed. They are used, for example, to compare the explanatory power of a given descriptor with respect to several other descriptors: B / (A + B) controls for the total amount of information in **b**, whereas B / (B + C) controls for the total information in **a**.

The construction of an association matrix, containing any of the symmetric coefficients described above, requires calculating $p(p-1)/2$ contingency tables; this matrix is symmetric and its diagonal is $S = 1$ or $D = 0$. *Qualitative (nonordered) descriptors* can thus be used to compute *quantitative association coefficients*, which makes possible the numerical analysis of multivariate qualitative data sets. Furthermore, since quantitative or semiquantitative descriptors can be recoded into

discrete states, it is possible, using uncertainty coefficients, to compute association matrices among descriptors of mixed types.

It is only through B, which can be transformed into a X^2_W-statistic, that a probability can be associated to the uncertainty coefficients. For coefficient S above (eq. 6.15), short of computing a p-value, one can state in general terms that two descriptors are very closely related when $S(\mathbf{a},\mathbf{b}) > 0.5$; they are well associated when $0.5 > S > 0.3$; and some association exists when $S < 0.3$ without coming too close to 0 (Hawksworth *et al*., 1968).

The probability associated with a X^2-statistic, calculated on a contingency table, assesses the hypothesis that the relationship between the two descriptors is *random*. Biological associations, for example, could be defined on the basis of relationships found to be non-random between pairs of species — the relationship being defined by reference to a pre-selected probability level (e.g. $\alpha = 0.05$ or 0.01) associated with the X^2 measuring the contingency between two species (Subsection 7.5.2). The value of X^2 may itself be used as a measure of the *strength* of the relationship between species. This is also the case for the reciprocal information measures defined above. With the same purpose in mind, it is possible to use one of the following *contingency coefficients*, which are merely transformations of a X^2-statistic on a scale from 0 to 1 (Kendall & Buckland, 1960; Morice, 1968):

Contingency coefficient

$$\text{Pearson contingency coefficient}, C = \sqrt{X^2 / (n + X^2)} \qquad \textbf{(6.19)}$$

$$\text{Tschuproff contingency coefficient}, T = \sqrt{X^2 / (n\sqrt{\text{degrees of freedom}})} \qquad \textbf{(6.20)}$$

where n is the number of observations. These contingency coefficients are not frequently used in ecology, however. They can only be used for comparing contingency tables of the same sizes.

Contingency tables are the main approach available to ecologists for the numerical analysis of relationships among qualitative descriptors, or else between qualitative descriptors and ordered variables divided into classes. Contingency tables are also convenient for analysing *nonmonotonic relationships* among ordered descriptors (a relationship is monotonic when there is a constant evolution of a descriptor with respect to the other; see Fig. 5.1). Reciprocal information and X^2 coefficients are sensitive enough that they could be used even with ordered variables, when relationships among a large number of descriptors are analysed by computer. One must simply make sure that the ordered data are divided into a sufficiently large number of classes to avoid clumping together observations that one would want to keep distinct in the results. If a first analysis indicates that redefining the boundaries of the classes could improve the interpretation of the phenomenon under study (the classes used to recode quantitative variables do not need to have the same width), ecologists should not hesitate to repeat the analysis using the recoded data. This procedure is not circular; it corresponds to a progressive discovery of the structure of the information.

It is also possible to use the association coefficients described above to interpret the classifications resulting from a first analysis of the data (Chapter 8). A classification may be compared to the descriptors from which it originates, in order to determine which descriptors are mostly responsible for it; or else, it may be compared to a new series of descriptors that could potentially explain it. One can also use contingency tables to compare several classifications of the same objects, obtained through different methods. Subsection 10.2.1 describes these higher-level analyses.

Ecological application 6.2

Legendre *et al*. (1978) analysed data from a winter aerial survey of land fauna, using contingency tables. They compared the presence or absence of tracks of different bird and mammal species to a series of 11 environmental descriptors. Five of these descriptors were qualitative, i.e. bioclimatic region, plant association, nature of the dominant and sub-dominant surface materials, and category of aquatic ecosystem. The others were semiquantitative, i.e. height of the trees, drainage, topography, thickness of the surface materials, abundance of streams and wetlands. The analysis identified the descriptors that determined or limited the presence of the 10 species that had been observed with sufficient frequency to permit their analysis. This allowed the authors to describe the niches of these species.

6.3 Multiway contingency tables

When there are more than two descriptors, one might consider the possibility of analysing the data set using a series of two-way contingency tables, in which each pair of descriptors would be treated separately. Such an approach, however, would not take into account possible interactions among several descriptors and might thus miss part of the potential offered by the multidimensional structure of the data. This could lead to incorrect, or at least incomplete conclusions. Information on the analysis of multiway contingency tables can be found in Kullback (1959), Plackett (1974), Bishop *et al*. (1975), Upton (1978), Gokhale & Kullback (1978), Fienberg (1980), Sokal & Rohlf (1995), Agresti (2002), and Kroonenberg (2008).

Log-linear model

The most usual approach for analysing multiway contingency tables is to adjust to the data a *log-linear model*, where the natural logarithm ($\log_e$) of the expected frequency E for each cell of the table is estimated as a sum of main effects and interactions. For example, in the case of two-way contingency tables (Section 6.2), the expected frequencies could have been computed using the following equation:

$$\log_e E = [\theta] + [A] + [B] + [AB] \qquad \textbf{(6.21)}$$

Symbols in brackets are the *effects*. [A] and [B] are the main effects of descriptors **a** and **b**, respectively, and [AB] is the effect resulting from the interaction between **a** and **b**. [θ] is the mean of the logarithms of the expected frequencies. In a two-way table,

the hypothesis tested is that of independence between the two descriptors, i.e. H_0: [AB] = 0. The log-linear model corresponding to this hypothesis is thus:

$$\log_e E = [\theta] + [A] + [B] \quad \textbf{(6.22)}$$

since [AB] = 0. The expected frequencies E computed using eq. 6.22 are exactly the same as those computed in Section 6.2 (eq. 6.4). Hence for two-way tables, one usually computes the expected frequencies with eq. 6.4. For multiway tables, the expected frequencies are generated with an iterative proportional fitting algorithm. The advantage of log-linear models is obvious when analysing contingency tables with more than two dimensions (or *criteria*).

For a contingency table with three descriptors (**a**, **b**, and **c**), the log-linear model containing all possible effects is:

$$\log_e E = [\theta] + [A] + [B] + [C] + [AB] + [AC] + [BC] + [ABC]$$

Saturated model

Such a model is referred to as the *saturated model*. In practice, the effect resulting from the interaction among all descriptors is never included in any log-linear model, i.e. here [ABC]. This is because the expected frequencies for the saturated model are always equal to the observed frequencies ($E = O$), so that this model is useless. The general log-linear model for a three-way table is thus:

$$\log_e E = [\theta] + [A] + [B] + [C] + [AB] + [AC] + [BC] \quad \textbf{(6.23)}$$

where H_0: [ABC] = 0. In other words, the logarithm of the expected frequency for each cell of the contingency table is computed here by adding, to the mean of the logarithms of the expected frequencies, one effect due to each of the three descriptors and one effect resulting from each of their two-way interactions.

Hierarchical model

Different log-linear models may be formulated by setting some of the effects equal to zero. Normally, one only considers *hierarchical models*, i.e. models in which the presence of a higher-order effect implies that all the corresponding lower effects are also included; the order of an effect is the number of symbols in the bracket. For example, in a hierarchical model, to include [BC] implies that both [B] and [C] are also included. For a three-way contingency table, there are eight possible hierarchical models, corresponding to as many different hypotheses (Table 6.5). Models in the table all include the three main effects. Each hypothesis corresponds to different types of interaction among the three variables. In practice, one uses a program available in a computer package (for R functions, see Section 6.6), with which it is easy to estimate the expected frequencies for any hierarchical model of interest to the user.

The number of degrees of freedom (ν) depends on the interactions that are included in the model. For the general hierarchical model of eq. 6.23,

$$\nu = rst - [1 + (r-1) + (s-1) + (t-1) + (r-1)(s-1) + (r-1)(t-1) + (s-1)(t-1)] \quad \textbf{(6.24)}$$

Table 6.5 Possible log-linear models for a three-way contingency table. Hypotheses and corresponding models. All models include the three main effects [A], [B] and [C].

Hypotheses (H_0)	Log-linear models
1.[ABC] = 0	$\log_e E$ = [θ]+[A]+[B]+[C]+[AB]+[AC]+[BC]
2.[ABC] = 0, [AB] = 0	$\log_e E$ = [θ]+[A]+[B]+[C]+[AC]+[BC]
3.[ABC] = 0, [AC] = 0	$\log_e E$ = [θ]+[A]+[B]+[C]+[AB]+[BC]
4.[ABC] = 0, [BC] = 0	$\log_e E$ = [θ]+[A]+[B]+[C]+[AB]+[AC]
5.[ABC] = 0, [AB] = 0, [AC] = 0	$\log_e E$ = [θ]+[A]+[B]+[C]+[BC]
6.[ABC] = 0, [AB] = 0, [BC] = 0	$\log_e E$ = [θ]+[A]+[B]+[C]+[AC]
7.[ABC] = 0, [AC] = 0, [BC] = 0	$\log_e E$ = [θ]+[A]+[B]+[C]+[AB]
8.[ABC] = 0, [AB] = 0, [AC] = 0, [BC] = 0	$\log_e E$ = [θ]+[A]+[B]+[C]

where r, s and t are the numbers of states of descriptors **a**, **b** and **c**, respectively. If there were only two descriptors, **a** and **b**, the log-linear model would not include the interaction [AB], so that eq. 6.24 would become:

$$\nu = rs - [1 + (r - 1) + (s - 1)] = (r - 1)(s - 1)$$

which is identical to eq. 6.7. In Table 6.5, model 4, for example, does not include the interaction [BC], so that:

$$\nu = rst - [1 + (r - 1) + (s - 1) + (t - 1) + (r - 1)(s - 1) + (r - 1)(t - 1)]$$

Programs in computer packages calculate the number of degrees of freedom corresponding to each model.

It is possible to test the goodness of fit of a given model to the observed data by using one of the X^2 statistics already described for two-way tables, X^2_P or X^2_W (eqs. 6.5 and 6.6). The null hypothesis (H_0) tested is that the effects excluded from the model are null. Rejecting H_0, however, does not allow one to accept the alternative hypothesis that *all* the effects included in the model are not null. The only conclusion to be drawn from rejecting H_0 is that at least some of the effects in the model are not null. When the p-value associated with a model is larger than the significance level α, the conclusion is that the model fits the data well.

Williams' correction

As in the case of two-way contingency tables (eq. 6.8), it is recommended to divide X^2_W by a correction factor, q_{min} (Williams, 1976a), when the number of observations n is small, i.e. less than 4 or 5 times the number of cells in the table. For the general hierarchical model (eqs. 6.23 and 6.24):

$$q_{min} = 1 + (1/6\nu n)\,[r^2s^2t^2 - 1 - (r^2 - 1) - (s^2 - 1) - (t^2 - 1) - (r^2 - 1)(s^2 - 1) - (r^2 - 1)(t^2 - 1) - (s^2 - 1)(t^2 - 1)] \qquad \textbf{(6.25)}$$

In the case of two descriptors, eq. 6.25 becomes:

$$q_{min} = 1 + (1/6\nu n)\,[r^2s^2 - 1 - (r^2 - 1) - (s^2 - 1)]$$

$$q_{min} = 1 + (1/6\nu n)\,[(r^2 - 1)(s^2 - 1)$$

which is identical to eq. 6.8. For model 4 in Table 6.5, used above as example:

$$q_{min} = 1 + (1/6\nu n)\,[r^2s^2t^2 - 1 - (r^2 - 1) - (s^2 - 1) - (t^2 - 1) - (r^2 - 1)(s^2 - 1) - (r^2 - 1)(t^2 - 1)]$$

This correction cannot be applied, as such, to contingency tables containing null expected frequencies (see below). The other possible correction, which consists in adding to each cell of the table a small value, e.g. 0.5, has the same effect here as in two-way contingency tables (see Section 6.2).

Ecological application 6.3a

Legendre (1987a) analysed biological oceanographic data obtained at 157 sites in Baie des Chaleurs (Gulf of St. Lawrence, eastern Canada). The data set (observations made at 5-m depth) included measurements of temperature, salinity, nutrients (phosphate and nitrate), and chlorophyll *a* (estimated from the *in vivo* fluorescence of water pumped on board the ship). As it often happens in ecology, the numerical analysis was hampered by three practical problems. (1) The measured concentrations of nutrients were often near or below the detection limit, with the result that many of them exhibited large experimental errors (since the 1980s, the detection limits of some nutrients have been lowered by a factor 100 or 1000). (2) Relationships between variables were often nonmonotonic, i.e. they did not continuously increase or decrease but reached a maximum (or a minimum) after which they decreased (or increased). (3) Most of the variables were intercorrelated, so that no straightforward interpretation of phytoplankton (i.e. chlorophyll *a*) concentrations was possible in terms of environmental variables. Since multiway contingency table analysis can handle these three types of problems, it was decided to partition the (ordered) variables into discrete classes and analyse the transformed data using hierarchical log-linear models.

The initial model in Table 6.6 (line 1) only included the interaction among the three environmental variables, with no effect of these on chl *a*. This model did not fit the data well. Adding the interaction between chlorophyll *a* (chl *a*) and the temperature-salinity (TS) characteristics significantly improved the fit (i.e. there was a significant difference between models; line 2). The resulting model could be accepted (line 3), but adding the interaction between chl *a* and phosphate further improved the fit (significant difference, line 4) and the

Table 6.6 Multiway contingency table analysis of oceanographic data recoded into discrete classes (Legendre, 1987a). Using a hierarchy of log-linear models, the concentrations of chlorophyll *a* (symbol in this table: C, 4 classes) are analysed as a function of the temperature-salinity (TS) characteristics of the water masses (symbol in this table: T, 3 classes) and the concentrations of phosphate (P; 2 classes) and nitrate (N; 2 classes). When a higher-order effect is present, all the corresponding lower-order effects are included in the model.

Effects in the model	Interpretation	ν	X^2_W
[NTP], [C]	Chl *a* is independent of the environmental variables	30	121 *
Difference	Adding [CT] to the model significantly improves the fit	9	89 *
[NTP], [CT]	Chl *a* depends on the TS characteristics	21	**32**
Difference	Adding [CP] to the model significantly improves the fit	3	13 *
[NTP], [CT], [CP]	Chl *a* depends on the TS characteristics and on phosphate	18	**19**
Difference	Adding [CN] does not significantly improve the fit	7	5
[NTP], [CT], [CP], [CN]	The most parsimonious model does not include a dependence of chl *a* on nitrate	11	**14**

* $p \leq 0.05$; bold X^2_W values correspond to models with $p > 0.05$ of fitting the data

resulting model fitted the data well (line 5). Final addition of the interaction between chl *a* and nitrate did not improve the fit (difference not significant, line 6). The most parsimonious model (line 5) thus showed a dependence of chl *a* concentrations on the TS characteristics and phosphate. The choice of the initial model in Table 6.6 is explained in Ecological application 6.3b.

There are 8 hierarchical models associated with a three-way contingency table, 113 with a four-way table, and so forth, so that the choice of a single model, among all those possible, rapidly becomes a major problem. In fact, it often happens that several models fit the data well. Also, in many instances, the fit to the data could be improved by adding supplementary terms (i.e. effects) to the model. However, this improved fit would result in a more complex ecological interpretation because of the added interaction(s) among descriptors. It follows that the choice of a model generally involves a compromise between goodness of fit and simplicity of interpretation, as suggested by the principle of parsimony (Subsection 10.3.3). Finally, even when it is possible to test the fit of all possible models to the data, this way of proceeding involves multiple testing and the p-values require correction (Box 1.3).

To select a model, there are several methods that are both statistically acceptable and ecologically parsimonious. In practice, because no method is totally satisfactory, one could simply use, with care, those included in the available computer package.

Partitioning the X^2_W

1. A first method consists in *partitioning* the X^2_W statistics associated with a hierarchy of log-linear models. The hierarchy contains a series of models, which are made progressively simpler (or more complex) by removing (or adding) one effect at a time. It can be shown that the difference between the X^2_W statistics of two successive models in the hierarchy is itself a X^2_W-statistic, which can therefore be tested. The corresponding number of degrees of freedom is the difference between those of the two models. This is the approach used in Ecological application 6.3a (see Table 6.6). The main problem with this method is that one may find different "most parsimonious" models depending on the hierarchy chosen *a priori*. Partitioning X^2 statistics is possible only with X^2_W, not X^2_P.

Stepwise selection

2. A second family of approaches lies in the *stepwise forward selection* or *backward elimination* of terms in the model. As always with stepwise methods (see Subsection 10.3.3), (a) it may happen that forward selection lead to models quite different from those resulting from backward elimination, and (b) the tests of significance must be interpreted with caution because the computed statistics are not independent. Stepwise methods thus only provide guidance, which may be used for limiting the number of models to be considered. It often happens that models other than those identified by the stepwise approach are found to be more parsimonious and interesting, and to fit the data just as well (Fienberg, 1980: 80).

Effect screening

3. Other methods simultaneously consider all possible effects. An example of *effect screening* (Brown, 1976) is given in Dixon (1981). The approach is useful for reducing the number of models to be subsequently treated, for example, by the method of hierarchical partitioning of X^2_W statistics (see method 1 above).

When analysing multiway contingency tables, ecologists must be aware of a number of possible practical problems, which may sometimes have significant impact on the results. These potential problems concern the cells with zero expected frequencies, the limits imposed by the sampling design, the simultaneous analysis of descriptors with mixed levels of precision (i.e. qualitative, semiquantitative, and quantitative), and the use of contingency tables for the purpose of explanation or forecasting.

Cells with $E = 0$

1. Multiway contingency tables, in ecology, often include cells with expected frequencies $E = 0$. There are two types of zero expected frequencies, i.e. those resulting from sampling and those that are of structural nature.

Sampling zeros are caused by random variation, combined with small sample size relative to the number of cells in the multiway contingency table. Such zeros would normally disappear if the size of the sample was increased. The presence of cells with null observations ($O = 0$) may result, when calculating specific models, in some

expected frequencies $E = 0$. This is accompanied by a reduction in the number of degrees of freedom. For example, according to eq. 6.24, the number of degrees of freedom for the initial model in Table 6.6 (line 1) should be $\nu = 33$, since this model includes four main effects [C], [N], [P], and [T] and interactions [NP], [NT], [PT], and [NPT]; however, the presence of cells with null observations ($O = 0$) led to cells with $E = 0$, which reduced the number of degrees of freedom to $\nu = 30$. Rules to calculate the reduction in the number of degrees of freedom are given in Bishop *et al*. (1975: 116 *et seq*.) and Dixon (1981: 666). In practice, computer programs generally take into account the presence of zero expected frequencies when computing the number of degrees of freedom for multiway tables. The problem does not occur with two-way contingency tables because cells with $E = 0$ are only possible, in the two-way configuration, if all the observations in the corresponding row or column are null, in which case the corresponding state is automatically removed from the table.

Structural zeros correspond to combinations of states that cannot occur *a priori* or by design. For example, in a study where two of the descriptors are sex (female, male) and sexual maturity (immature, mature, gravid), the expected frequency of the cell "gravid male" would *a priori* be $E = 0$. Another example would be combinations of states that have not been sampled, either by design or involuntarily (e.g. lack of time, or inadequate planning). Several computer programs allow users to specify the cells that contain structural zeros, before computing the expected frequencies.

2. In principle, the methods described here for multiway contingency tables can only be applied to data resulting from *simple random sampling* or *stratified sampling* designs. Fienberg (1980: 32) gives some references in which methods are described for analysing qualitative descriptors within the context of *nested sampling* or a *combination of stratified and nested sampling* designs. Sampling designs are described in Cochran (1977), Green (1979), and Thompson (1992), for example.

Mixed precision

3. Analysing together *descriptors with mixed levels of precision* (e.g. a mixture of qualitative, semiquantitative, and quantitative descriptors) may be done using multiway contingency tables. In order to do so, continuous descriptors must first be partitioned into a small number of classes. Unfortunately, there exists no general approach to do so. When there is no specific reason for setting the class limits, it has been suggested, for example, to partition continuous descriptors into classes of equal width, or containing an equal number of observations. Alternatively, Cox (1957) describes a method that may be used for partitioning a normally distributed descriptor into a predetermined number of classes (2 to 6). For the specific case discussed in the next paragraph, where there is one response variable and several explanatory variables, Legendre & Legendre (1983b) describe a method for partitioning the ordered explanatory variables into classes in such a way as to maximize the relationships to the response variable. It is important to be aware that, when analysing the contingency table, different ways of partitioning continuous descriptors may sometimes lead to different conclusions. In practice, the number of classes of each descriptor should be as small as possible, in order to minimize the problems discussed above concerning the calculation of X^2_W (see eqs. 6.8 ad 6.25 for correction factor q_{min}) and the presence of

sampling zeros. Another point is that contingency table analysis considers the different states of any descriptor to be nonordered. When some of the descriptors are in fact ordered (i.e. originally semiquantitative or quantitative), the information pertaining to the ordering of states may be used when adjusting log-linear models (see for example Fienberg, 1980: 61 *et seq*.).

4. There is an analogy between *log-linear models* and *analysis of variance* since the two approaches use the concepts of effects and interactions. This analogy is superficial, however, since analysis of variance aims at assessing the effects of explanatory factors on a single response variable, whereas log-linear models have been developed to describe structural relationships among several descriptors corresponding to the dimensions of the table.

5. It is possible to use contingency table analysis for interpreting a *response variable* in terms of several interacting *explanatory variables*. In such a case, the following basic rules must be followed. (1) Any log-linear model fitted to the data must include by design the term for the highest-order interaction among all *explanatory variables*. In this way, all possible interactions among the explanatory variables are included in the model, because of its hierarchical nature. (2) When interpreting the model, one should not discuss the interactions among the explanatory variables. They are incorporated in the model for the reason given above, but no test of significance is performed on them. In any case, one is only interested in the interactions between the explanatory and response variables. An example follows.

Ecological application 6.3b

The example already discussed in application 6.3a (Legendre, 1987a) aimed at interpreting the horizontal distribution of phytoplankton in Baie des Chaleurs (Gulf of St. Lawrence, eastern Canada) in terms of selected environmental variables. In such a case, where a single response variable is interpreted as a function of several potentially explanatory variables, all models considered must include by design the highest-order interaction among the explanatory variables. Thus, all models in Table 6.6 included the interaction [NPT]. The simplest model in the hierarchy (line 1 in Table 6.6) only contained [NPT] and [C] as effects. In this simplest model, there was no interaction between chlorophyll and any of the three environmental variables, i.e. the model did not include [CN], [CP] or [CT]. When interpreting the model selected as best fitting the data, the author did not discuss the interaction among the explanatory variables because the presence of [NPT] prevented a proper analysis of this interaction. Table 6.6 then led to the interpretation that the horizontal distribution of phytoplankton depended on the TS characteristics of water masses and phosphate concentration.

Logistic regression

When the *qualitative response variable* is *binary*, one may use the *logistic linear* (or *logit*) *model* instead of the log-linear model. Fitting such a model to data is also called *logistic regression* (Subsection 10.3.7). In logistic regression, the explanatory descriptors do not have to be divided into classes; they may be discrete or continuous. This type of regression is available in various computer packages and in R (Section 10.7). Some programs allow the *response variable* to be *multi-state*. Efficient use of logistic regression requires that *all* the explanatory descriptors be potentially

related to the response variable. This method can replace discriminant analysis in cases discussed in Subsection 10.3.7 and Section 11.6.

Examples of successful use of multiway contingency tables in ecology include Fienberg (1970) and Schoener (1970) for the habitat of lizards, Jenkins (1975) for the selection of trees by beavers, Legendre & Legendre (1983b) for marine benthos, Fréchet (1990) for cod fishery, Schoener & Adler (1991) for spatial distributions of lizards and birds, Fedriani *et al*. (2001) for responses of coyote populations to anthropogenic food, Fingerut *et al*. (2003) for transmission of a marine parasite by swimming larvae, and Gorelick & Bertram (2010) for computation of diversity indices.

6.4 Contingency tables: correspondence

Once it has been established that two or more qualitative descriptors in a contingency table are not independent (Sections 6.2 and 6.3), it is often of interest to identify the cells of the table that account for the existing relationship between descriptors. These cells, which show how the descriptors are related, define the *correspondence* between the rows and columns of the contingency table. By comparison with parametric and nonparametric statistics (Chapters 4 and 5), the measures of contingency described in Sections 6.2 and 6.3 are, for qualitative descriptors, analogous to the *correlation* between ordered descriptors, whereas correspondence would be analogous to *regression* (Section 10.3) because it can be used to forecast the state of one descriptor using another descriptor. *Correspondence analysis* (Section 9.2) is another method that allows, among other objectives, the identification of the relationships between the rows and columns of a contingency table. This can be achieved directly through the approach described in the present section.

In a contingency table where the descriptors are not independent (i.e. the null hypothesis of independence has been rejected), the cells of interest to ecologists are those in which the observed frequencies (O_{ij}) are very different from the corresponding expected frequencies (E_{ij}). Each of these cells corresponds to a given state for each descriptor in the contingency table. The fact that $O_{ij} \neq E_{ij}$ is indicative of a stronger interaction, between the states in question, than expected under the null hypothesis which in invoked to compute E. For example, hypothesis H_0 in Table 6.4 is that of independence of descriptors **a** and **b**. This hypothesis having been rejected ($p < 0.001$), one may identify in the contingency table the observed frequencies O_{ij} that are much higher or lower than the corresponding expected frequencies E_{ij}. Values $O_{ij} > E_{ij}$ (in bold-face type in Table 6.4) give a preliminary indication of the associations between states of **a** and **b**. These values may be located anywhere in the table since contingency table analysis does not take into account the ordering of states.

When the test of the global X^2-statistic (eq. 6.5 or 6.6) supports the hypothesis of a significant relationship between the two descriptors, one can identify the cells that

Test of $O_{ij} = E_{ij}$

strongly contribute to the correspondence by testing the significance of the difference between O_{ij} and E_{ij} *in each cell* of the contingency table. Ecologists may be interested in any difference, whatever its sign, or only in the cases where O_{ij} is significantly higher than E_{ij} (preference) or significantly lower (avoidance, exclusion).

Bishop *et al.* (1975: 136 *et seq.*) describe three statistics for measuring the difference between O and E. They can be used for two-way or multiway contingency tables. The three statistics are the components of X^2_P, the components of X^2_W, and the Freeman-Tukey deviates:

$$\text{component of } X^2_P\text{: } (O - E) / \sqrt{E} \qquad \textbf{(6.26)}$$

$$\text{component of } X^2_W\text{: } 2\, O \log_e(O/E) \qquad \textbf{(6.27)}$$

$$\text{Freeman-Tukey deviate: } \sqrt{O} + \sqrt{O+1} - \sqrt{4E+1} \qquad \textbf{(6.28)}$$

These statistics are available in various computer packages. A critical value has been proposed by Bishop *et al.* (1975) for testing the significance of statistics 6.26 and 6.28:

$$\sqrt{\chi^2_{[\nu,\,\alpha]} / (\text{no.cells})}$$

E_{ij} is said to be significantly different from O_{ij} when the absolute value of the statistic, for cell (i, j), is larger than the critical value. According to Sokal & Rohlf (1995), however, the above critical value often results in a type I error much greater than the nominal α level. These authors use instead the following approximate criterion to test Freeman-Tukey deviates:

$$\sqrt{\nu\ \chi^2_{[1,\,\alpha]} / (\text{no.cells})} \qquad \textbf{(6.29)}$$

When the (absolute) value of the Freeman-Tukey deviate is larger than or equal to the criterion, one concludes that $E_{ij} \neq O_{ij}$ at significance level α for that cell. Authors often recommend to only test the cells where $5 \leq E_{ij} \leq (n - 5)$. Neu *et al.* (1974) recommended to apply a Bonferroni or Holm correction (Box 1.3) to significance level α in order to account for multiple testing. An example of this method, with Bonferroni correction for the number of tested cells, is presented in Table 6.7.

Test of standardized residuals

Alternatively, Haberman (1973) proposed a test of the components of X^2_P (eq. 6.26), which are also called *standardized residuals* and are represented by the symbol e_{ij}. The standard error of e_{ij} is the square root of the maximum likelihood estimate of its asymptotic variance:

$$var_{ij} = \left(1 - \frac{\text{row sum}_i}{n}\right)\left(1 - \frac{\text{column sum}_j}{n}\right)$$

Table 6.7 Test of Freeman-Tukey deviates (eq. 6.28) in individual cells of a contingency table. The observed and expected values are taken from Table 6.4. Only 8 of the 16 deviates are tested because the others, identified by an asterisk, had expected values smaller than 5 and could therefore not be tested. Absolute values larger than or equal to the criterion (eq. 6.29) with Bonferroni correction for 8 simultaneous tests, $[9\ \chi^2_{[1,\,0.05/8]}\ /\ 8]^{1/2} = [9 \times 7.48 / 8]^{1/2} = 2.90$, are in bold. These values identify the cells in which the number of observations (O_{ij}) significantly ($p < 0.05$) differs (higher or lower as shown by the sign) from the corresponding expected frequencies (E_{ij}). The overall null hypothesis (H_0: complete independence of descriptors **a** and **b**) had been rejected first (Table 6.4), before testing the significance of the observed values in individual cells of the table.

	b_1	b_2	b_3	b_4
a_1	**3.23**	–1.33	0.06	**–3.12**
a_2	**–4.57**	**3.49**	**–4.57**	0.91
a_3	–3.00 *	–3.00 *	–3.00 *	3.87 *
a_4	–3.00 *	–3.00 *	3.87 *	–3.00 *

* No test because $E_{ij} < 5$ (Table 6.4).

where n is the total number of observations in the contingency table. Dividing e_{ij} by $\sqrt{var_{ij}}$ produces an *adjusted residual* statistic Z_{ij}:

$$Z_{ij} = \frac{e_{ij}}{\sqrt{var_{ij}}} \qquad \textbf{(6.30)}$$

which is distributed like a standard normal deviate. That test is also described by Everitt (1977). When $|Z_{ij}|$ is larger than or equal to the critical value $z_{[1 - (\alpha\ /\ 2\ \text{no. tests})]}$ read from a table of standard normal deviates (z-table), one concludes that O_{ij} is significantly different from E_{ij} at significance level α. Division by the number of simultaneous tests is the Bonferroni correction (Box 1.3). Statistics higher than the critical value z are in bold-face type in Table 6.8. The conclusions drawn from Tables 6.7 and 6.8 may not be identical.

Comparing Table 6.4 to Tables 6.7 and 6.8 shows that considering only the cells where $O_{ij} > E_{ij}$ may lead to conclusions which, without necessarily being incorrect, are subject to some risk of error. Tables 6.7 and 6.8 show, for instance, that dominant species a_1 is significantly over-represented in environmental condition b_1 and under-represented in b_4, suggesting that b_1 is favourable whereas b_4 is adverse to the species.

Table 6.8 Test of standardized residuals using the *Z*-statistic (eq. 6.30). Only 8 of the 16 deviates are tested because the others, identified by an asterisk, had expected values smaller than 5 and could therefore not be tested. The observed and expected values are taken from Table 6.4. Absolute values of *Z* larger than or equal to the critical value $z_{[1-(0.05/2\times8)]} = z_{0.9969} = 2.73$ are in boldface; the correction is for 8 simultaneous tests. The bold values identify cells in which the number of observations (O_{ij}) significantly ($p < 0.05$) differs (higher or lower, as shown by the sign) from the corresponding expected frequency (E_{ij}).

	b_1	b_2	b_3	b_4
a_1	**6.32**	–2.11	0.00	**–4.22**
a_2	**–3.65**	**6.09**	**–3.65**	1.22
a_3	–2.39 *	–2.39 *	–2.39 *	7.17 *
a_4	–2.39 *	–2.39 *	7.17 *	–2.39 *

* No test because $E_{ij} < 5$ (Table 6.4).

Ecological application 6.4

Legendre *et al*. (1982) explored the relationship between the abundance of phytoplankton and vertical stability of the water column in a coastal embayment of Hudson Bay (Canadian Arctic). Surface waters are influenced by the plume of the nearby Great Whale River. There were intermittent phytoplankton blooms from mid-July through mid-September. In order to investigate the general relationship between phytoplankton concentrations (chlorophyll *a*) and the physical conditions, chl *a* and salinity data from 0 and 5 m depths were allocated to a contingency table (Table 6.9). The null hypothesis of independence being rejected, the correspondence between the two descriptors rests in four cells. (1) At high salinities (> 22), there is a significantly small number of high chl *a* observations and (2) a significantly high number of low chl *a* values. At intermediate salinities (18-22), (3) high chl *a* observations are significantly numerous, whereas (4) low chl *a* observations are significantly infrequent. At low salinities (< 18), the numbers observed are not significantly different from the frequencies expected under the null hypothesis of independence.

Table 6.9 shows that, on the one hand, high chl *a* concentrations were positively associated with intermediate salinities, whereas they were much reduced in waters of high salinity. On the other hand, low chl *a* concentrations were characteristically infrequent in waters of intermediate salinities and frequent at high salinities. The overall interpretation of these results, which also took into account estimates of the vertical stability of the water column (Richardson number), was as follows: (1) strong vertical mixing led to high salinities at the surface; this mixing favoured nutrient replenishment, but dispersed phytoplankton biomass over the water column; (2) low salinity conditions were not especially favourable nor adverse to phytoplankton,

Table 6.9 Contingency table: chlorophyll *a* concentrations as a function of salinity in the surface waters of Manitounuk Sound (Hudson Bay, Canadian Arctic). In each cell: observed (O_{ij}) and expected (E_{ij}, in parentheses) frequencies, and adjusted residual (Z, eq. 6.30) to test the hypothesis that $O_{ij} = E_{ij}$ ($\alpha = 0.05$) with correction for 5 simultaneous tests. Statistics in bold are larger than $z_{[1 - 0.05 / 2\times5]} = 2.58$, indicating that $O_{ij} \neq E_{ij}$. Total no. observations $n = 207$. $X^2_W = 33.47$ with Williams correction ($\nu = 2$, $p < 0.001$); hence the hypothesis of independence between chl *a* and salinity is rejected.

Chlorophyll *a* (mg m^{-3})	Salinity 6-18	18-22	22-26
	2	22	7
1.5-6.1	(3.29)	(8.09)	(19.62)
(high values)	–0.82 *	**6.17**	**–5.10**
	20	32	124
0-1.5	(18.71)	(45.91)	(111.38)
(low values)	0.82	**–6.17**	**5.10**

* Statistic not tested because $E_{ij} < 5$.

i.e. stratification was favourable, but dilution by water from the nearby river was adverse; (3) intermediate salinities were associated with intermittent conditions of stability; under such conditions, both the high nutrient concentrations and the stability of the water column were favourable to phytoplankton growth and accumulation. Intermittent summer blooms thus occurred upon stabilization of the water column, as a combined result of wind relaxation and fortnightly tides.

6.5 Species diversity

Biodiversity is a most important synthetic concept for ecology. It can be studied at all levels of organization of Life, from genes to ecosystems. Loreau (2010) gives a clear account of the importance of biodiversity science for both fundamental and applied ecology. He addresses, among other topics, the present crisis of diversity on Earth and the possibility of a sixth mass extinction, the socio-economic values of diversity within the context of ecological services, various frontiers of diversity science, the (controversial) linking of diversity science and policy, and finally, the need to build a new relationship between Humanity and Nature. The author also provides a well organised summary of different measures of diversity (see his Chapter 2). In the study of ecological communities, species diversity indices, discussed in the present section,

are synthetic biotic indices that capture multidimensional information relative to the species composition of an assemblage or a community.

Diversity is often called "biodiversity" nowadays. The addition of prefix "bio" before "diversity" has not changed the original concept or the way diversity is measured in ecology. Interested readers could look at the discussion of "diversity" versus "biodiversity" in Longhurst (2007, pp. 23-24).

The distribution of a quantitative variable is characterized by its *dispersion* around its mean, as shown in Sections 4.1 and 4.3. The parametric and nonparametric measures of dispersion are the *variance* (eq. 4.3) and the *range*, respectively. These two measures do not apply to qualitative variables, for which the *number of states* (q) may be used as a simple measure of dispersion. However, this measure does not take advantage of the frequency distribution of observations among the states, which is known in many instances. When the relative frequencies of the states are available, eq. 6.1 may be used to measure the dispersion of the qualitative variable:

$$H = -\sum_{i=1}^{q} \mathrm{p}_i \log \mathrm{p}_i$$

where p_i is the relative frequency or proportion (on a 0-1 scale) of observations in state (species) i. Species with frequency 0 disappear from the calculation because $\lim_{\mathrm{p} \to 0} (\mathrm{p} \log \mathrm{p}) = 0$. This formula can be rewritten as:

$$H = \frac{1}{n} \sum_{i=1}^{q} -(\log n_i - \log n)\, n_i$$

where n is the total number of organisms and n_i is the number of organisms belonging to species i. The latter equation is similar to the formula used to calculate the variance of n objects divided into q classes:

$$s_y^2 = \frac{1}{n-1} \sum_{i=1}^{q} (y_i - \bar{y})\, f_i$$

where f_i is the frequency of the i-th class. In ecology, H is widely used to measure the *diversity* of a species assemblage; it is generally computed for each sampling site separately (alpha diversity; see Subsection 6.5.3). In species diversity studies, the qualitative descriptor is the list of the q species present and each state of that descriptor corresponds to a species name. Both the number of species q and entropy H belong to the same family of generalized entropies (eq. 6.31, below).

In assemblages of biological species, there are generally several species represented by a single or a few individuals, and a few species that are very abundant. The few abundant species often account for many more individuals than all the rare

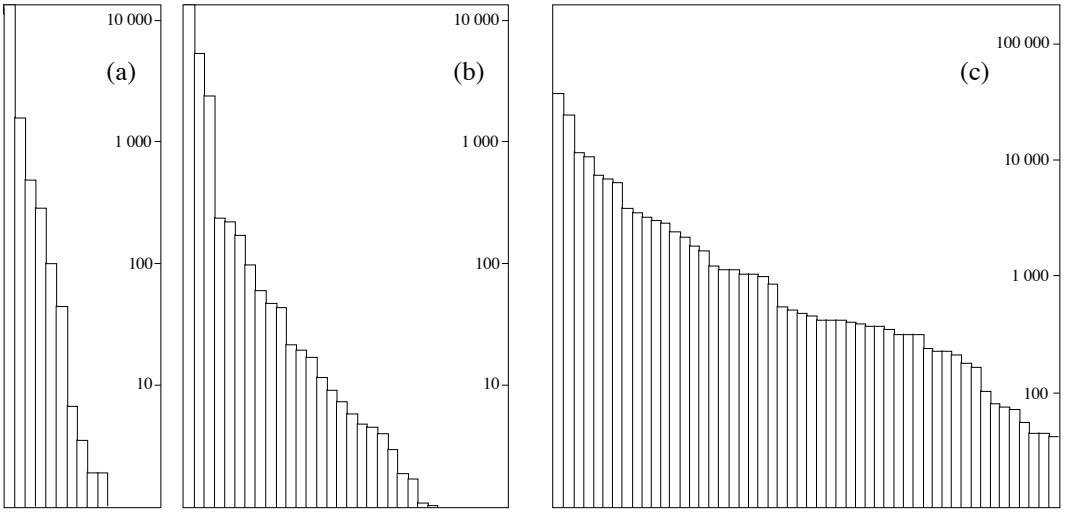


Figure 6.2 Fish catches (abundances) in (a) the Barents Sea, (b) the Indian Ocean, and (c) the Red Sea. Along the abscissa, species are arranged in order of decreasing frequencies. The histogram ordinates are logarithmic. Adapted from Margalef (1974).

species together. Figure 6.2 shows, in order of decreasing frequencies, the abundances of fish species caught in the Barents Sea, the Indian Ocean, and the Red Sea. The three water bodies clearly differ in both the number of species caught and the shape of their abundance distributions. Diversity indices must be applicable to any type of species assemblage regardless of the shape of the abundance distribution. One parameter of the distribution is clearly the *number of species*; another is the *shape of the distribution*. An alternative approach to describing a species frequency distribution with these two parameters is to combine them in a single index, e.g. the entropy measure H. Species diversity may thus be defined as a *measure of species composition, in terms of both the number of species and their relative abundances*.

Species diversity

It is generally not useful to measure species diversity of a whole community (e.g. primary, secondary, and tertiary producers and decomposers), because of the different roles played by various species in an ecosystem. It is better (Hurlbert, 1971; Pielou, 1975) to restrict the study of species diversity (and of the underlying theoretical phenomena, e.g. competition, succession) to a single *taxocene*. A taxocene is a set of species belonging to a given supraspecific taxon that make up a natural ecological community or, in other words, that represent a taxonomic segment of a community or association (Chodorowski, 1959; Hurlbert, 1971; Whittaker, 1972). The supraspecific taxon must be such that its member species are about the same size, have similar life histories, and compete over both ecological and evolutionary time for a finite amount of similar resources (Deevey, 1969). A taxocene occupies a limited

Taxocene

segment in space and in the environment. For these reasons, the following information about the reference population should accompany any measure of diversity: (1) the spatial boundaries of the region or volume within which the population is found and a description of the sampling method; (2) the temporal limits within which the observations have been made; (3) the taxocene under study (Hurlbert, 1971; Pielou, 1975).

Sampling sites may harbour species that differ much in size or role in the environment. This may occur, for example, when all plants in quadrats (ligneous and herbaceous) are counted, or when species at different developmental stages are collected (e.g. counting saplings as equivalent to adult trees). Comparisons of diversity indices with production or environmental variables may be easier in such cases if species diversity is computed, not from numbers of individuals, but instead from measures of biomass (Wilhm, 1968) or dry mass, productivity (Dickman, 1968), fecundity, or any other appropriate measure of energy transfer.

Species diversity indices may be used to compare successive observations from the same community (time series: O mode, Fig. 7.1) or sampling sites from different areas (Q mode). Coefficients in Chapter 7 compare objects by combining paired information available for each species. In contrast, diversity indices pool the multispecies information into a single value for each sampling unit, before comparing them.

Over the years, several formulae have been proposed in the ecological literature for measuring species diversity. The present section only describes the main indices that are found in the modern literature. Species diversity has been the subject of detailed discussions. Early reviews were presented in the milestone books of Pielou (1969, 1975) and Margalef (1974) and in the review paper of Peet (1974). A recent account linking species diversity to ecological theory is found in Loreau (2010).

1 — Diversity

Hill (1973a) and Pielou (1975) noted that the three diversity indices mostly used by ecologists are specific cases of the *generalized entropy* formula of Rényi (1961):

$$H_a = \frac{1}{1-a} \log \sum_{i=1}^{q} p_i^a \qquad \textbf{(6.31)}$$

where a is the order of the entropy measure ($a = 0, 1, 2, \ldots$), q is the number of species, and p_i is the relative frequency or proportion of species i. This formula gives an indeterminate result for $a = 1$. One can show, however, that the limit of this equation when a tends towards 1 from below (i.e. from 0 to 1) or from above (i.e. from 2 to 1) is the Shannon entropy formula, eqs. 6.1 and 6.34a.

Hill (1973a) prefers the corresponding diversity numbers:

Diversity number

$$N_a = \exp H_a \qquad \textbf{(6.32)}$$

The first three Rényi entropies H_a (of orders $a = 0$ to 2) and the corresponding Hill diversity numbers N_a are:

$$\text{(a) } H_0 = \log q \qquad \text{(b) } N_0 = q \qquad \textbf{(6.33)}$$

$$\text{(a) } H_1 = -\sum \mathrm{p}_i \log \mathrm{p}_i = H \qquad \text{(b) } N_1 = \exp H \qquad \textbf{(6.34)}$$

$$\text{(a) } H_2 = -\log \sum \mathrm{p}_i^2 = -\log (\lambda) \qquad \text{(b) } N_2 = \lambda^{-1} \qquad \textbf{(6.35)}$$

Hill (1973a) noted that increasing the order a diminishes the relative weights of rare species in the resulting index: when $a = 0$, the data are transformed to presence-absence form where rare and abundant species have the same importance. In a review of the topic, Peet (1974) proposed other ways of creating families of diversity indices. Let us examine the first three orders of eq. 6.31 in more detail.

Number of species

1. *Entropy of order a = 0.* — The *number of species q* (eq. 6.33b) is the index of diversity most often used in ecology. It goes back to Patrick (1949):

$$\textit{Diversity} = q \qquad \textbf{(6.36)}$$

It is more affected by the presence of rare species than higher-order indices. The number of species can also be seen as a component of other diversity indices (e.g. J, eq. 6.45 in Subsection 6.5.2).

As the size of the sampling units increases, additional rare species appear. This is a problem with all diversity indices and it is at its worst in eq. 6.36. It is incorrect to compare the diversities of sampling units that have different sizes because diversity measures are not additive (Subsection 1.4.2). This point has been empirically shown by He *et al*. (1996). The problem can be resolved by calculating the numbers of species that sampling units would contain if they all had the same size. This can be done using Sanders' (1968) *rarefaction method*, whose formula was corrected by Hurlbert (1971). In this method, a constant number of organisms is used to make the sampling units comparable, instead of the physical size of sampling unit in m^2 or litre. The formula computes the expected number of species q' in a standardized sampling unit of n'

organisms, for example 1000 organisms, from a nonstandard sampling unit containing q species, a total of n organisms, and n_i organisms belonging to each species i:

$$E(q') = \sum_{i=1}^{q} \left[1 - \frac{\binom{n - n_i}{n'}}{\binom{n}{n'}} \right] \qquad \textbf{(6.37)}$$

where $n' \leq (n - n_1)$, n_1 being the number of individuals in the most abundant species (y_1), and the terms in parentheses are combinations. For example:

$$\binom{n}{n'} = \frac{n!}{n'!\,(n-n')!}$$

Shannon's entropy

2. *Entropy of order a = 1.* — Margalef (1958) proposed to use Shannon's entropy H (eqs. 6.1 and 6.34a) as an index of species diversity:

$$H = -\sum_{i=1}^{q} p_i \log p_i$$

The properties of H as a measure of diversity are the following:

• $H = 0$ (minimum value), when the sampling unit contains a single species. H increases with the number of species.

• For a given number of species, H is maximum when the organisms are equally distributed among the q species: $H = \log q$. For a given number of species, H is lower when there is stronger dominance in the sampling unit by one or a few species (e.g. Figs. 6.1a and b). The actual value of H depends on the base of logarithms (2, e, 10, or other). This base must always be reported since it sets the scale for the H values.

• Like the variance, diversity can be partitioned into different components. In particular, the calculation of diversity can take into account not only the proportions of the different species but also those of genera, families, etc. When diversity is partitioned into a component for *genera* and a component for *species within genera*, two adaptive levels can be explored among the environmental descriptors, and diversity H can be partitioned using eqs. 6.10-6.12. The total diversity, $H = A + B + C$, which is calculated using the proportions of species without taking into account those of genera, is equal to the diversity with respect to genera, $H(G) = A + B$, plus that of species within genera, $H(S \mid G) = C$. The latter is calculated as the sum of the diversities H within genera, weighted by the proportions of individuals in the genera. The formula is:

$$H = H(G) + H(S \mid G) \qquad \textbf{(6.38)}$$

This same calculation may be extended to other systematic categories. Considering, for example, the categories family (F), genus (G), and species (S), diversity can be partitioned into the following hierarchical components:

$$H = H(F) + H(G \mid F) + H(S \mid G,F) \qquad \textbf{(6.39)}$$

Using this approach, Lloyd *et al.* (1968) measured hierarchical components of diversity for communities of reptiles and amphibians in Borneo.

Most diversity indices share the first two properties above, but only the indices derived from eq. 6.31 have the third one (Daget, 1980). The probabilistic interpretation of H refers to the *uncertainty* about the identity of an organism chosen at random in a sampling unit. The uncertainty is small when the sampling unit is dominated by a few species or when the number of species is small. These two situations correspond to low H values.

In principle, H should only be used when a sample is drawn from a theoretically infinite population, or at least a population large enough that sampling does not modify it in a noticeable way. In cases of samples drawn from small populations, or samples whose representativeness is unknown, it is theoretically better, according to Pielou (1966), to use Brillouin's formula (1956), proposed by Margalef (1958) for computing diversity H. This formula was introduced in Section 6.1 to calculate the information per symbol in a message (eq. 6.3):

$$H = (1/n) \log[n! / (n_1!\, n_2! \ldots n_i! \ldots n_q!)]$$

where n_i is the number of individuals in species i and n is the total number of individuals in the collection. Brillouin's H corresponds to sampling *without* replacement (and is thus more exact) whereas Shannon's H applies to sampling *with* replacement. In practice, H computed with either formula is the same to several decimal places, unless samples are so small that they should not be used to estimate species diversity in any case. Species diversity cannot, however, be computed on measures of biomass or energy transfer using Brillouin's formula.

3. *Entropy of order a = 2.* — Simpson (1949) proposed an index of species diversity based on the probability that two interacting individuals of a population belong to the same species. This index is frequently used in ecology. When randomly drawing, without replacement, two organisms from a sampling unit containing q species and n individuals, the probability that the first organism belong to species i is n_i/n and that the second also belong to species i is $(n_i - 1)/(n - 1)$. The combined probability of the two events is the product of their separate probabilities. Simpson's

Concentration

concentration index (λ) is the probability that two randomly chosen organisms belong to the same species, i.e. the sum of the combined probabilities for the different species:

$$\lambda = \sum_{i=1}^{q} \frac{n_i}{n} \frac{(n_i - 1)}{(n-1)} = \frac{\sum_{i=1}^{q} n_i (n_i - 1)}{n(n-1)}$$

When n is large, n_i is almost equal to $(n_i - 1)$, so that the above equation becomes:

$$\lambda = \sum_{i=1}^{q} \left(\frac{n_i}{n}\right)^2 = \sum_{i=1}^{q} p_i^2 \qquad \textbf{(6.40)}$$

which corresponds to the summation in eq. 6.35a. This index may be computed from numbers of individuals, or from measures of biomass or energy transfer. The higher is the probability that two organisms be conspecific, the smaller is the diversity of the sampling unit. For this reason, Greenberg (1956) proposed to measure species diversity as:

$$\textit{Diversity} = 1 - \lambda \qquad \textbf{(6.41)}$$

which is also the *probability of interspecific encounter* (Hurlbert, 1971). Pielou (1969) showed that this index is an unbiased estimator of the diversity of the population from which the sample has been drawn. This index is also known in ecology as the *Gini coefficient*, because it was originally proposed by economist Corrado Gini (1912) as an index of "mutability" or diversity. In the same 1912 paper, Gini also defined an index of inequality, which is widely used in economics under the name of … *Gini coefficient*. Hence the Gini coefficient of ecologists is not the same as that of economists.

Because eq. 6.41 is more sensitive than H to changes in the abundances of the few very abundant species, Hill (1973a) recommended to use instead:

$$\textit{Diversity} = \lambda^{-1} \qquad \textbf{(6.42)}$$

which is diversity number N_2 of eq. 6.35b. Hill (1973a) also showed that this index is linearly related to $\exp H$ (eq. 6.34b).

Margalef & Gutiérrez (1983) proposed the following expression, which combines eqs. 6.41 and 6.42:

$$\textit{Diversity} = \frac{1-\lambda}{\lambda} = \sum_{i \neq j} p_i p_j \Big/ \sum_{i=1}^{q} p_i^2 \qquad \textbf{(6.43)}$$

Note that each pair (i, j), for $i \neq j$, is counted twice in the expression $\sum p_i p_j$. This diversity formula is the ratio of the probability that two individuals taken at random

belong to different species, to the probability that they pertain to the same species. It is the maximum number of interspecific interactions normalized by the maximum number of intraspecific interactions.

Biodiversity indices that integrate phylogenetic information have been proposed by Helmus *et al*. (2007).

Functional diversity

Functional diversity refers to the diversity of ecological processes that maintain interactions among the components of an ecosystem. It is estimated through the diversity of species traits and functions in a study area. Several functional diversity indices have been proposed by Rao (1982), Petchey & Gaston (2002), Botta-Dukát (2005), Villéger *et al*. (2008), Laliberté & Legendre (2010), and others. In practice, these indices are computed from species functional traits (quantitative or qualitative variables) weighted by species abundances.

2 — Evenness, equitability

Several authors, for example Margalef (1974), prefer to directly interpret species diversity as a function of physical, geographical, biological, or temporal variables, whereas others consider that species diversity consists of two components, which should be interpreted separately. These two components are the *number of species* and the *evenness* of their frequency distribution. Although the concept of evenness had been introduced by Margalef (1958), it was formally proposed by Lloyd & Ghelardi (1964) for characterizing the *shape* of distributions such as those in Fig. 6.2, where the component "number of species" corresponds to the length of the abscissa. In the literature "evenness" and "equitability" are synonyms terms (Lloyd & Ghelardi, 1964; see also the review of Peet, 1974). Several indices of evenness have been proposed.

1. The simplest approach to evenness consists in comparing the measured diversity to the corresponding maximum value. When using H (eqs. 6.1 and 6.34a), diversity takes its maximum value when all species are equally represented. In such a case,

$$H_{\max} = -\sum_{i=1}^{q} \frac{1}{q} \log \frac{1}{q} = \log q \qquad \textbf{(6.44)}$$

Pielou's evenness

where q *is* the number of species. Evenness (J) is computed as (Pielou, 1966):

$$J = H / H_{\max} = \left(-\sum_{i=1}^{q} p_i \log p_i \right) / \log q \qquad \textbf{(6.45)}$$

which is a ratio, whose value is independent of the base of logarithms used for the calculation. Using the terms defined by Hill (1973a, eqs. 6.31-6.35), Daget (1980) rewrote eq. 6.45 as the ratio of entropies of orders 1 (eq. 6.34a) and 0 (eq. 6.33a):

$$J = H_1 / H_0 \qquad \textbf{(6.46)}$$

Equations 6.44 and 6.45 show that diversity H combines the *number of species* (q) and the *evenness* of their distribution (J):

$$H = JH_{max} = J\log q \quad \textbf{(6.47)}$$

Hurlbert's evenness

2. Hurlbert (1971) proposed an evenness index based on the minimum and maximum values of diversity. Diversity is minimum when one species is represented by $(n - q + 1)$ organisms and the $(q - 1)$ others by a single organism. According to Hurlbert, the following indices are independent of q:

$$J = (D - D_{min}) / (D_{max} - D_{min}) \quad \textbf{(6.48)}$$

$$1 - J = (D_{max} - D) / (D_{max} - D_{min}) \quad \textbf{(6.49)}$$

Patten's redundancy

Equation 6.48 was proposed by Patten (1962) as a measure of *redundancy* (see Section 6.1). The two indices can be computed for any diversity index D.

Broken stick model

3. Instead of dividing the observed diversity by its maximum value, Lloyd & Ghelardi (1964) proposed to use a model based on the *broken stick distribution* (Barton & David, 1956; MacArthur, 1957). To generate this distribution, a set of individuals is taken as equivalent to a stick of unit length which is broken randomly into a number of pieces (i.e. in the present case, the number of species q). The divisor in the evenness formula is the diversity computed from the lengths of the pieces of the randomly broken stick. The expected lengths (E) of the pieces of the broken stick (species) y_i are given, in decreasing order, by the successive terms of the following series (Pielou, 1975), corresponding to the successive values $i = 1, 2, \ldots, q$, for a given number of species q:

$$E(y_i) = q^{-1} \sum_{x=i}^{q} x^{-1} \quad \textbf{(6.50)}$$

For example, for $q = 3$ species, eq. 6.50 gives the following lengths for species $i = 1$ to 3: 0.6111, 0.2778, and 0.1111, respectively (R function: Section 6.6). Diversity of this series is computed using the formula for H (eq. 6.1 or 6.34a):

$$M = -\sum_{i=1}^{q} E(y_i) \log E(y_i) \quad \textbf{(6.51)}$$

The evenness index of Lloyd & Ghelardi (1964), which they called *equitability*, is similar to eq. 6.45, with M being used instead of H_{max}:

$$J = H / M \quad \textbf{(6.52)}$$

In the same paper, Lloyd & Ghelardi proposed another evenness index:

$$J = q' / q \qquad \textbf{(6.53)}$$

where q is the observed number of species and q' is the number of species for which the broken stick model predicts the observed diversity H, i.e. $H(q) = M(q')$. Values computed with eq. 6.52 or 6.53 are usually, but not always, smaller than one. Indeed, it happens that biological populations are more diversified than predicted by the broken stick model.

Functional evenness

4. Troussellier & Legendre (1981) described an *index of functional evenness*, for studying bacterial assemblages. In such assemblages, the species level is often poorly defined. The index bypasses the step of species identification, using instead as data the set of binary biochemical (and other) descriptors that characterize the microbial isolates. The authors showed that their index has the usual properties of an evenness measure. Functional evenness J of a bacterial sampling unit is defined as:

$$J = \frac{I}{I_{\max}} = \frac{1}{c \log 0.5} \sum_{i=1}^{c} [\, p_i \log p_i + (1 - p_i) \log (1 - p_i) \,] \qquad \textbf{(6.54)}$$

where I and $I_{\max}$ are measures of information, c is the number of binary descriptors used, and p_i is the proportion of positive responses to test i.

Evenness indices 6.44, 6.47, 6.51, and 6.52 all suffer from the problem that they depend on field estimation of the number of species in the population; in other words, q is not a fixed and known value but a random variable. Because the true value of q is not known and cannot be estimated from the data, there is no formula for computing a standard error (and, thus, a confidence interval) for these estimates of J. This point has been stressed by Pielou (1975) for eq. 6.45. This is not the case with eq. 6.54, where the denominator of J is a constant ($I_{\max} = c \log 0.5$ where c is the number of binary descriptors used in the calculation). Several methods may be used for computing the confidence interval of J (e.g. the jackknife, briefly described at the end of Subsection 1.2.4). Legendre *et al.* (1984b) provided examples where the computation of confidence intervals for J, measured during biodegradation experiments, showed that significant changes had taken place, at some point in time, in the structure of the bacterial assemblages involved in the biodegradation processes.

In varying environments, the ecological interpretation of the two components of diversity (eq. 6.47) could be carried out, for example, along the lines proposed by Legendre (1973). (1) The *number of species* may be a function of the stability of the environment. Indeed, a more stable environment entails a higher degree of organization and complexity of the food web (Margalef, 1958), so that such an environment contains more niches and, thus, more species. The number of species is proportional to the number of niches since, by definition, the realized niche of a species is the set of environmental conditions that this species does not share with any

other sympatric species (Hutchinson, 1957, 1965). This approach has the advantage of linking species diversity to environmental diversity. (2) The *evenness of species distribution* may be inversely related to the overall biological activity in the studied environment; the lower the evenness, the higher the biological activity (e.g. production, life cycles, energy flows among trophic levels). On a seasonal basis, another factor may contribute to lower the evenness. In an environment where interspecific competition is low (high evenness), seasonal reduction of resources or deterioration of weather conditions could induce stronger competition and thus favour some species over others, which would decrease the evenness. The same is often observed in cases of pollution.

3— Species diversity through space

A most interesting property of species diversity is its organization through space. This phenomenon, which is now well known to community ecologists, was first discussed by Whittaker in two seminal papers (1960, 1972) where he described the alpha, beta and gamma diversity levels. The development of multiscale spatial analysis of communities (Chapter 14) is grounded in Whittaker's concept of beta diversity.

Alpha diversity

Alpha (α) diversity is the diversity in species composition at individual sites i (e.g. plots, quadrats; α_i in Fig. 6.3). The indices used for alpha diversity estimate, in different ways, the variance in the species identity of individuals observed at a given site. A monoculture, for example, has the lowest possible alpha diversity because there is no variance in species identity among the individuals. Alpha diversity is measured by one of Rényi's entropy indices H_0 (eq. 6.33a), H_1 (eq. 6.34a) or H_2 (eq. 6.35a), by Hill's diversity numbers N_0 (richness, eq. 6.33b), N_1 (eq. 6.34b) or N_2 (eq. 6.35b), or by some other indices such as Fisher's α logarithmic series parameter (Fisher *et al.*, 1943). The most commonly used indices are N_0, H_1 and N_2, mentioned in Fig. 6.3.

Gamma diversity

Gamma (γ) diversity is the diversity of the whole region of interest in a study (γ in Fig. 6.3). It is usually measured by pooling the observations from a group of sampling units (which form a *sample* in the statistical sense), i.e. a large number of sites from the area of interest, except in cases where the community composition of an entire area is known, e.g. the CTFS permanent forest plots*. Gamma diversity is measured using the same indices as alpha diversity.

Beta diversity

Beta (β) diversity is of different nature: it is conceptually the variation in species composition among sites in the geographic area of interest (Legendre *et al.,* 2005, Anderson *et al.,* 2006; β in Fig. 6.3). Its value will vary with the extent of the area, the physical size of the sampling units and the sampling interval in the area under study, which form three aspects of the study scale (Section 13.0). Studies of beta diversity can actually focus on two aspects of community structure (Anderson *et al.*, 2011). The

* A map of the *Center for Tropical Forest Science* (CTFS) forest plots, and details about each plot, are available on the Web page http://www.ctfs.si.edu/.

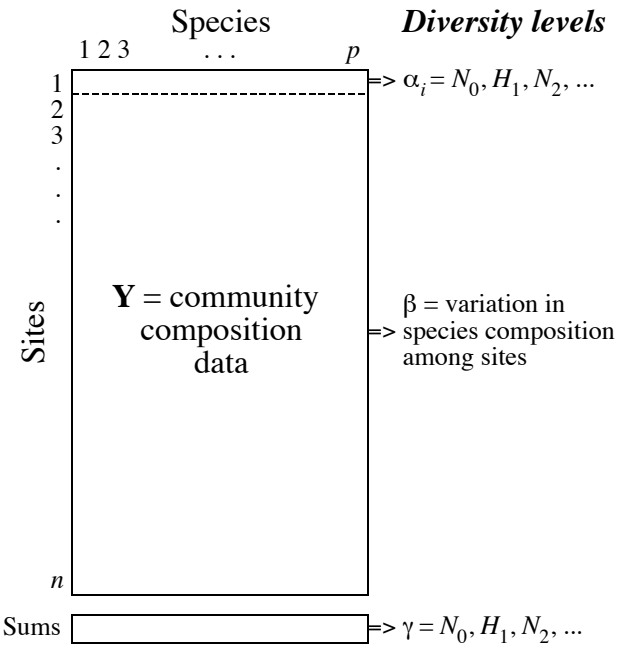


Figure 6.3 Species diversity indices are computed from the community composition data (matrix **Y**). Alpha (α) diversity indices are computed for individual sites (rows) *i*. Gamma diversity (γ) is computed from the vector of column sums of the data matrix using the same indices as for alpha diversity. Beta (β) diversity is of a different nature: it is the variation in community composition among sites. It cannot be computed with the usual entropy of diversity number indices.

first one is *turnover*, or the change in community composition between adjacent sampling units, explored by sampling along a spatial, temporal, or environmental gradient. The second is a *non-directional* approach to the study of *community variation* through space [or time]; it does not refer to any specific gradient but centres on the variation in community composition among the study units. The present section focuses on the second approach, as it links the concept of beta diversity with the analysis of the variation of community data matrices performed by the methods described in the following chapters.

If the variation in community composition is random and accompanied by biotic processes (e.g. reproduction) that generate spatial autocorrelation in the species data due to their limited dispersal (Subsection 1.1.1, model 2; Fig. 1.5, case 3), a gradient in species composition may appear (called a "false gradient" in Subsection 13.1.2) if the sampling area is small compared to the dispersal distance. Beta diversity can then be interpreted in terms of the rate of change, or *turnover*, in species composition along that gradient. Ecologists often refer to this turnover to explain beta diversity. The

community spatial structure is often more complex than a single gradient, however: if differentiation among sites is due to environmental factors, which may combine gradient-like and patchy geographic distributions, beta diversity should be analysed with respect to the hypothesized forcing environmental variables (Subsection 1.1.1, model 1; Fig. 1.5, case 4). In ecosystems, beta diversity may be caused concurrently by varying proportions of these two processes (i.e. induced spatial dependence and true autocorrelation due to biotic processes: Fig. 1.5, case 5). Chapter 14 will show how these two types of hypotheses about the processes that generate beta diversity can be disentangled.

Whittaker (1960, 1972) showed that beta diversity could be estimated using either presence-absence or quantitative species data. Ecologists use both types of measures to study beta diversity, although some researchers only refer to presence-absence data when they talk about the rate of species replacement, or turnover, along an ecological gradient. In the ordination literature, however, ecologists most often use species abundance data to study turnover rates by reference to the appearance and disappearance of species with unimodal distributions along gradients.

A first method, proposed by Whittaker (1960, 1972), for obtaining a global measure of beta diversity from species presence-absence data, is to compute the ratio of two diversity indices: $\beta = S/\bar{\alpha}$, where S is the number of species in a composite community composition vector representing the area of interest, and $\bar{\alpha}$ is the mean number of species observed at the sites that were used to compute S. This is a multiplicative approach, where S represents gamma diversity. The ratio $S/\bar{\alpha}$ indicates how many more species are present in the whole region than at an average site, and uses that value as the measure of beta diversity. Other beta diversity indices have been reviewed by Koleff *et al*. (2003), Magurran (2004), Tuomisto (2010) and Anderson *et al*. (2011).

An alternative, additive approach had been present in the literature since MacArthur *et al*. (1966), Levins (1968) and Allan (1975). It was revived by Lande (1996) and has been widely used since then (Veech *et al*., 2002). In that approach, $D_T = D_{among} - \bar{D}_{within}$ where D_T is the total (gamma) diversity. This approach can be applied to species richness N_0 (eq. 6.33b), Shannon information H_1 (eqs. 6.1 and 6.34a), or Simpson diversity $D = (1 - \lambda)$ (eq. 6.41); see Lande (1996) for details. Because diversities are variances, one recognizes an analysis of variance approach in that equation.

Whittaker (1960, 1972) suggested that beta diversity could also be estimated from distance matrices computed among sites. This approach is based on the fact that a distance between two sites, computed from community composition data, provides a measure of the variation, or beta diversity between these sites. Distance matrices computed using appropriate indices (Chapter 7) thus assess the pairwise beta diversity among all pairs of sites. To obtain an overall index of beta diversity over a group of sites, Whittaker (1972) suggested to use *the mean* (not the variance) of the distances among sites: "the mean CC [i.e. the distance coefficient that is the complement of

Jaccard's coefficient of community, $D = 1 - S_7$ in Table 7.2] for samples of a set compared with one another [...] is one expression [of] their relative dissimilarity, or beta differentiation" (Whittaker, 1972: 233). Whittaker derived his concept from the *index of biotal dispersity* suggested fifteen years before by Koch (1957). Whittaker thus acknowledged the fact that dissimilarities (i.e. distances, Chapter 7) are themselves measures of the differentiation between sites.

Total variation of **Y**

Box 6.1 shows that the total variation of a data matrix **Y**, e.g. the one shown in Fig. 6.3, can be computed either from **Y** itself or from a distance matrix **D** derived from **Y**. This equality is pertinent here as it shows the equivalence of Whittaker's overall measure of beta diversity computed from a distance matrix, **D**, and beta diversity defined as the variation in species composition among sites, which can be measured by the total variation in matrix **Y**, SS(**Y**). Indeed, SS(**Y**) can be computed from matrix **D** using eq. 6.56. For distance matrices that are not Euclidean but whose square root is Euclidean, one may use eq. 6.58. The distance functions that Whittaker (1972) was citing, i.e. $1 - S_7$ (Jaccard), $1 - S_8$ (Sørensen), D_9 (Whittaker), and D_{14} (percentage difference), pertain to that group. Box 6.1 shows that eq. 6.58 is a logical choice for the computation of SS(**Y**) for such distance functions.

To sum up, beta diversity can be estimated as the total variation in **Y** using two different equations: by computing eq. 6.55 from the raw data table **Y**, or computing eq. 6.56 on distances that have the Euclidean property, e.g. the Euclidean, chord, chi-square and Hellinger distances. Equation 6.58 is an alternative reasonable choice for distances whose square root is Euclidean, e.g. the (1 – Jaccard), (1 – Sørensen), Whittaker, and percentage difference distances.

An interesting observation is that for the chord and Hellinger distances, the maximum possible value of total variance Var(**Y**), computed by applying eq. 6.56 followed by eq. 6.57, is 1. The maximum values are obtained when all sites in table **Y** have entirely different species compositions. Similarly for community composition data transformed using the chord or Hellinger transformations (Section 7.7), the maximum possible value of Var(**Y**), computed using eq. 6.55 followed by eq. 6.57, is 1. Hence, using these transformations or distances, the estimates of beta diversity provided by Var(**Y**) are easily comparable since they fall in the range 0 to 1.

SS(Y), Var(Y) — Box 6.1

The total variation in a data matrix **Y** with n rows and p columns can be computed in two ways, which produce the same result.

• First method — Centre each column of **Y** on its mean using eq. 1.9 to obtain matrix $\mathbf{Y}_{\text{cent}} = [\, y_{\text{cent}.ij} \,]$, then compute the sum of these centred values squared:

$$\text{SS}(\mathbf{Y}) = \sum_{j=1}^{p} \sum_{i=1}^{n} (y_{ij} - \bar{y}_j)^2 = \sum_{j=1}^{p} \sum_{i=1}^{n} y_{\text{cent}.ij}{}^2 \qquad \textbf{(6.55)}$$

This is the total variation, or total sum of squares, of matrix **Y**. It is noted SS(**Y**), or e_k^2 in eq. 8.5.

• Second method — Compute a Euclidean distance matrix $\mathbf{D} = [D_{ih}]$ among the n rows of **Y** using distance function D_1 (eq. 7.32, Chapter 7). Then, calculate

$$\text{SS}(\mathbf{Y}) = \left(\sum_{i \neq h} D_{ih}^2 \right) / n \qquad \textbf{(6.56)}$$

using the $n(n-1)/2$ distances from the upper [or lower] triangular portion of **D**. SS(**Y**) computed in this way is called e_k^2 in eq. 8.6. The equivalence of these two ways of computing SS(**Y**) (Fig. 8.18) is demonstrated in Appendix 1 of Legendre & Fortin (2010).

The total variance in **Y** can be calculated from SS(**Y**) computed either way:

$$\text{Var}(\mathbf{Y}) = \text{SS}(\mathbf{Y})/(n-1) \qquad \textbf{(6.57)}$$

Besides eq. 6.56, there are three other ways of computing SS(**Y**) from **D**:

• SS(**Y**) is the trace of the Gower-centred distance matrix Δ_1 derived from **D** (eqs. 9.40 and 9.41, Chapter 9).

• SS(**Y**) is the sum of the eigenvalues of Δ_1, i.e. the eigenvalues of the principal coordinate analysis (PCoA) of **D**.

• SS(**Y**) is the total sum of squares of the principal coordinates of **D** (e.g. Table 9.9).

Box 6.1 (*continued*)

Equation 6.56 can be applied to any distance matrix **D**, Euclidean or not.

• Euclidean distances — For distances that have the Euclidean property (Tables 7.2 and 7.3), the rectangular matrix **Y**' obtained by principal coordinate analysis of **D** contains real numbers only. The distances among the rows of **Y**' computed using the Euclidean distance function D_1 (eq. 7.32) are equal to the distances in **D** (Subsection 9.3.3). Thus the total sum of squares in **Y**' computed with eq. 6.55 is equal to SS(**Y**) computed by applying eq. 6.56 to **D**.

Four of the Euclidean distance functions recommended for community composition data in Table 7.4 — the chord distance D_3 (eq. 7.35), the distance between species profiles D_{18} (eq. 7.53), the chi-square distance D_{16}, (eq. 7.55) and the Hellinger distance D_{17} (eq. 7.56) — have an additional property: eq. 6.55 computed from community composition data transformed using the chord (eq. 7.67), profile (eq. 7.68), chi-square (eq. 7.70) or Hellinger transformations (eq. 7.69) produces values of SS(**Y**) identical to those computed using eq. 6.56 with the chord, species profiles, chi-square and Hellinger distance matrices.

• Non-Euclidean distances — Examples of distance functions described in Chapter 7 that do not have the Euclidean property in their basic form are the Jaccard distance ($1 - S_7$), the Sørensen distance ($1 - S_8$), the percentage difference distance ($D_{14} = 1 - S_{17}$; D_{14} is called the Bray-Curtis distance in some computer packages), and the Whittaker distance (D_9); they may produce negative eigenvalues in principal coordinate analysis (PCoA, Section 9.3). For these distances, one can still compute eq. 6.56, but the corresponding matrix **Y**' of principal coordinates contains both real and complex (imaginary) axes (Subsection 9.3.4). Equation 6.55 can still be computed for **Y**' (McArdle & Anderson, 2001) with the result that SS(**Y**') is equal to the total sum of squares computed from **D** using eq. 6.56.

Ecologists may not be comfortable, however, in considering a matrix **Y**' that contains complex axes as a fair representation of community composition data. Luckily, there is another way: matrix $\mathbf{D}' = [D_{ih}^{0.5}]$, which contains the square roots of the distances, is Euclidean for these (and some other) distance functions, as shown in Tables 7.2 and 7.3. Hence, SS(**Y**) computed by applying eq. 6.56 to **D**' is equal to the total variation (eq. 6.55) of the rectangular data matrix **Y**" obtained by principal coordinate analysis (PCoA) of **D**', and this time **Y**" only contains real axes. So, for these non-Euclidean distance functions, because D_{ih} is equal to $\sqrt{D_{ih}^2}$, an appropriate formula for computing SS(**Y**) for the original matrix $\mathbf{D} = [D_{ih}]$ is:

$$\mathrm{SS}(\mathbf{Y}) = \left(\sum_{i \neq h} D_{ih} \right) / n \qquad \textbf{(6.58)}$$

6.6 Software

Two-way and multiway contingency table analysis are available in most commercial statistical software. The R language also offers functions implementing the methods described in this chapter.

1. In R, the standard function to conduct the Pearson chi-square test on a contingency table crossing two qualitative variables is ***chisq.test()*** of STATS; parametric and permutation tests are available in that function. Package SURVEY contains several functions to construct contingency tables and perform chi-square tests of association for survey data. Using cross-classifying factors, functions ***table()*** and ***ftable()*** of BASE construct two-way or multi-way contingency tables crossing factor levels. Function ***table.cont()*** in ADE4 plots contingency table data into a graph.

Function ***mantelhaen.test()*** of STATS performs a Cochran-Mantel-Haenszel chi-square test of interaction between two factors in three-way contingency tables. Also in STATS, function ***loglin()*** fits log-linear models to multidimensional contingency tables*.

2. R functions for studying diversity are found in packages BIODIVERSITYR, VEGAN and PICANTE. Rarefaction curves are computed by VEGAN's function ***rarefy()***. In a phylogenetic context, ***specaccum.psr()*** in PICANTE computes a rarefaction curve for phylogenetic species richness.

Function ***dbFD()*** of package FD computes seven functional diversity indices. Among these, Rao's (1982) quadratic entropy is also computed by functions ***divc()*** of ADE4 and ***raoD()*** of PICANTE. The broken-stick distribution is computed by function ***bstick()*** in VEGAN.

* A tutorial is available at the Web address http://ww2.coastal.edu/kingw/statistics/R-tutorials/loglin.html.

Chapter 7 *Ecological resemblance*

7.0 The basis for clustering and ordination

For almost a century, ecologists have collected quantitative observations to determine the resemblance between either the objects under study (sites) or the variables describing them (species or other descriptors). Objects and descriptors are defined in Section 1.4. Measuring the association (Section 2.2) between objects (Q mode) or descriptors (R mode) is the first, and sometimes the only step in the numerical analysis of ecological data. The various modes of analysis are discussed in Section 7.1. It may indeed happen that examining the association matrix suffices to elucidate the structure and thus answer the question at the origin of the investigation.

The present chapter provides a review of the main measures of association available to ecologists. Section 7.2 introduces the three types of association coefficients; the measures pertaining to each type — similarity, distance, and dependence — are described in Sections 7.3 to 7.5, respectively. In order to help ecologists choose from among this plurality of coefficients, Section 7.6 summarizes criteria for choosing a coefficient; the criteria are presented in the form of identification keys. Ecologists who do not wish to study the theory that underlies the measures of association may directly go to Section 7.6 after making themselves familiar with the terminology (Sections 7.1 and 7.2). When necessary, they may then refer to the paragraphs of Sections 7.3 to 7.5 describing the measures of interest.

Clustering

Ordination

In the next chapters, measures of resemblance between objects or descriptors will be used to cluster the objects or descriptors (Chapter 8) or to produce ordination diagrams in spaces of reduced dimensionality (Chapter 9). The clustering of objects (or descriptors) is an operation by which the set of objects (or descriptors) is partitioned in two or more subsets (clusters), using pre-established rules of agglomeration or division. Ordination in reduced space is an operation by which the objects (or descriptors) are represented in a space that contains fewer dimensions that in the original data set; the positions of the objects or descriptors with respect to one

another may also be used to cluster them. Both operations are often carried out on association matrices, which are computed as described in the following sections.

7.1 Q and R analyses

As noted by Cattell (1952), the ecological data matrix may be studied from two main viewpoints. One may wish to look at relationships among either the objects or the descriptors. The important point here is that these modes of analysis are based on different measures of association. The different types of coefficients are described in Section 7.2. Measuring the dependence between descriptors is done using coefficients like Pearson's r correlation coefficient (eq. 4.7, Section 4.2), so that studying the data matrix based on such coefficients is called *R analysis*. By opposition, studying the data matrix to uncover relationships among objects is called *Q analysis* (Fig. 2.1).

R mode
Q mode

Cattell (1966) had also observed that the *data box* (objects × descriptors × time instances; Fig. 7.1) may be looked at from other viewpoints than simply Q and R. He defined six modes of analysis:

O: among time instances, based on all observed descriptors (a single object);
P: among descriptors, based on all observed times (a single object);
Q: among objects, based on all observed descriptors (a single instance);
R: among descriptors, based on all observed objects (a single instance);
S: among objects, based on all observed times (a single descriptor);
T: among time instances, based on all observed objects (a single descriptor).

In the present chapter, the discussion of association coefficients will deal with the two basic modes only, i.e. Q measures (computed among objects) and R measures (computed among descriptors).

O-mode studies are conducted using Q measures; see, for example, Section 12.6. Similarly, P-mode studies are generally carried out with the usual R-type coefficients. When the data set forms a time series, however, P studies are based on special R-type coefficients that are discussed in Chapter 12: cross-covariance, cross-correlation, co-spectrum, coherence.

S- and T-mode studies mostly belong to autecology, i.e. studies involving a single species. S-mode comparisons among objects use the same coefficients as in P-mode analysis. Studying the relationship between "descriptor y observed at site x_1" and "the same descriptor y observed at site x_2" is analogous to the comparison of two descriptors along a time axis.

In T-mode studies, a variable is observed across several objects (sites, etc.) and different instances through time. Statistical tests of hypothesis for related samples are often applicable to these problems; see Table 5.2. In other cases, the two time instances

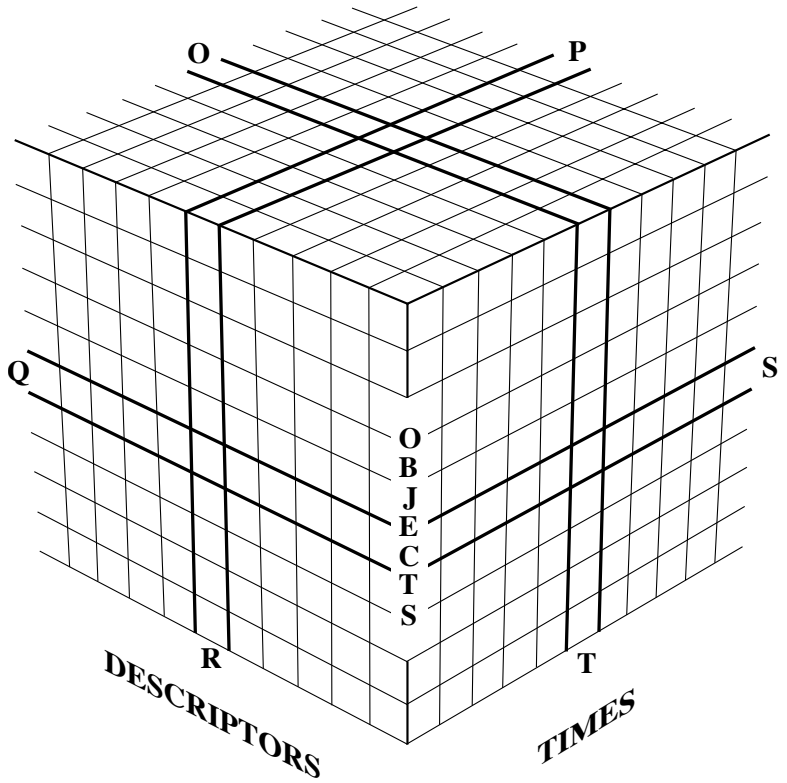


Figure 7.1 The three-dimensional data box (objects × descriptors × times). Adapted from Cattell (1966).

to be compared are considered to define two descriptors, as in the S mode, so that normal R-type measures may be used. Environmental impact studies form an important category of T-mode problems; ecologists should look at the literature on BACI designs when planning such studies (Before/After – Control/Impact: Green, 1979; Bernstein & Zalinski, 1983; Stewart-Oaten *et al*., 1986; Underwood, 1991, 1992, 1994).

Q or R?

It is not always obvious whether an analysis belongs to the Q or R mode. As a further complication, in the literature, authors define the mode based either on the association matrix or on the purpose of the analysis. Principal component analysis (Section 9.1), for instance, is based on a dispersion matrix among descriptors (R mode?), but it may be used for ordination of either the objects (Q mode?) or the descriptors (R mode?). In order to prevent confusion, in the present book, any study starting with the computation of an *association matrix among objects* is called a *Q analysis* whereas studies starting with the computation of an *association matrix among descriptors* are referred to as *R analyses*. In Chapter 9 for example, it is

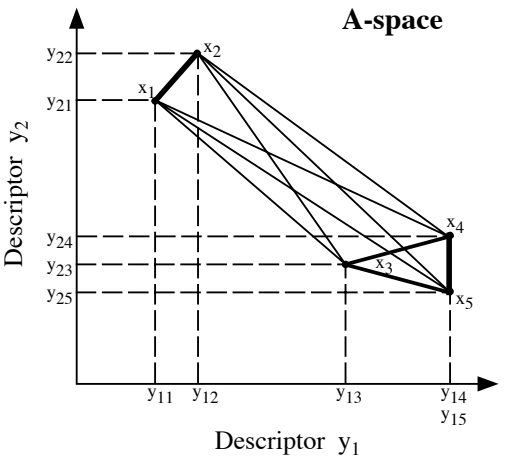


Figure 7.2 Scatter plot representation of five objects in an A-space with two descriptors. In this graph, the thickness of the lines that join the objects is proportional to their degree of resemblance with regard to the two descriptors, i.e. their proximity in the space.

possible to obtain an ordination of objects in low-dimension space using either the R method of principal component analysis (Section 9.1) or the Q method of principal coordinate analysis (Section 9.3). Interestingly, these two analyses lead to the same ordination of the objects when the principal coordinate analysis is computed from a Euclidean distance matrix (coefficient D_1, Section 7.4), although the results of Q and R analyses are not always reducible to each other.

The number of dimensions that can be represented on paper is limited to two or eventually three. Hence, one generally imagines distances among objects (Fig. 7.2) as embedded in a 2- or 3-dimensional space. Section 7.4 will show that such models can be extended to any number of dimensions. The descriptor, or *attribute space*, is called A-space. Distances and similarities computed in the present chapter will be based, in most instances, on measurements made in high-dimensional space.

A-space

Metric, Euclidean space

The A-space is called *metric* because the reference axes are quantitative, metric descriptors (Table 1.2). It is also called *Euclidean* because Euclide's geometry holds in that space. The qualifier *metric* will be used in different contexts in this book: metric variable (Table 1.2), metric space (here), metric properties of distances (beginning of Section 7.4), metric distances (Subsection 7.4.1); see also Plate 3.1, p. 142. Likewise, *Euclidean* will refer either to Euclidean space (here), to *the* Euclidean distance D_1 (Subsection 7.4.1), to the property of distances that can be embedded in Euclidean space (Tables 7.2 and 7.3), or to a property of ordination methods and plots (beginning of Section 9.3).

In addition to the methods described in the present and following chapters, there exist approaches allowing the analysis of the whole data box instead of subsets, as was the case in the six modes described above. Examples are found in Williams & Stephenson (1973), Williams *et al*. (1982), Cailliez & Pagès (1976), Marcotorchino & Michaud (1979), Kroonenberg (1983: three-way principal component analysis*), Carlier & Kroonenberg (1996: three-way correspondence analysis) and Kroonenberg (2008).

7.2 Association coefficients

The most usual approach to assess the resemblance among objects or descriptors is to first condense all (or the relevant part of) the information available in the ecological data matrix (Section 2.1) into a square matrix of association among the objects or descriptors (Section 2.2). In most instances, the association matrix is *symmetric*. Non-symmetric matrices can be decomposed into symmetric and skew-symmetric components, as described in Section 2.3; the components may then be analysed separately. Non-symmetric matrices can also be subjected to a special type of clustering called *seriation* (Section 8.10). In Chapters 8 and 9, objects or descriptors will be clustered or represented in reduced space after analysing an association matrix. It follows that *the structure resulting from a numerical analysis is that of the association matrix; the results of the analysis do not necessarily reflect all the information originally contained in the ecological data matrix*. This stresses the importance of choosing an appropriate measure of association. This choice determines the issue of the analysis. Hence, it must take into account the following considerations:

- The nature of the study (i.e. the initial question and the hypothesis) determines the kind of ecological structure to be evidenced through an association matrix, and consequently the type of measure of resemblance to be used.

- The various measures available have different mathematical constraints. The methods of analysis to which the association matrix will be subjected (clustering, ordination) may require measures of resemblance with specific mathematical properties.

- One must also consider the computational aspect, and thus preferably choose a measure available in a computer package or R function (Section 7.8), or one that can easily be programmed.

* Program 3WayPack (Kroonenberg & De Roo, 2010) for three-way principal component analysis and other three-way analyses is available from Pieter M. Kroonenberg, Leiden Institute of Education and Child Studies, Leiden University, Wassenaarseweg 52, NL-2333 AK Leiden, The Netherlands. Other three-mode software is described on the Web page: http://three-mode.leidenuniv.nl/. An overview of these methods can be found in Kroonenberg (2008).

Ecologists are, in principle, free to define and use any measure of association suitable for the ecological question under study; mathematics impose few constraints to this choice. This is why so many association coefficients are found in the literature. Some of them are of wide applicability whereas others have been developed to meet specific needs. Several coefficients have been rediscovered by successive authors and may be known under various names. Reviews of some coefficients can be found in Cole (1949, 1957), Goodman & Kruskal (1954, 1959, 1963), Dagnelie (1960), Sokal & Sneath (1963), Williams & Dale (1965), Cheetham & Hazel (1969), Sneath & Sokal (1973), Clifford & Stephenson (1975), Orlóci (1978), Daget (1976), Blanc *et al.* (1976), Prentice (1980), Gower (1985), and Gower & Legendre (1986).

1 — Similarity, distance, and dependence coefficients

In the following sections, *association* will be used as a general term to describe any measure or coefficient used to quantify the resemblance or difference between objects or descriptors, as proposed by Orlóci (1975). With *dependence* coefficients, used in the R mode, zero corresponds to no association. In Q-mode studies, *similarity* coefficients between objects will be distinguished from *distance* (or *dissimilarity*) coefficients. Similarities are *maximum* ($S = 1$) when the two objects are identical and *minimum* when the two objects are completely different; distances follow the opposite rule. Figure 7.2 shows the difference between the two types of measures: the length of the line between two objects is a measure of their distance, whereas its thickness, which decreases as the two objects get further apart, is proportional to their similarity. If needed, a similarity can be transformed into a distance, for example by computing its one-complement. For a similarity (S) measure, which takes values between 0 and 1, the corresponding distance (D) may be computed as:

Dependence
Similarity
Distance

$$D = 1 - S, \quad D = \sqrt{1 - S}, \quad \text{or} \quad D = \sqrt{1 - S^2}$$

We will see in Tables 7.2 and 7.3 and in Subsection 9.3.4 that the choice of a transformation instead of another may have consequences for the result of ordination analysis. Distances, which in several cases are not bound by a pre-determined upper value, can be normalized using eqs. 1.10 or 1.11:

$$D_{norm} = \frac{D}{D_{max}} \quad \text{or} \quad D_{norm} = \frac{D - D_{min}}{D_{max} - D_{min}}$$

where D_{norm} is the distance normalized in the interval [0, 1]; D_{max} and D_{min} are the maximum and minimum values taken by the distance coefficient, respectively. Normalized distances can be used to compute similarities by reversing the transformations given above:

$$S = 1 - D_{norm}, \quad S = 1 - D_{norm}^2, \quad \text{or} \quad S = \sqrt{1 - D_{norm}^2}$$

The following three sections describe the coefficients that are most useful with ecological data. Criteria to be used as guidelines for choosing a coefficient are

discussed in Section 7.6. Section 7.7 describes transformations for community composition data; each of these transformations is the first step in the calculation of an asymmetrical distance function described in Section 7.4. Computer programs and R functions are reviewed in Section 7.8.

2 — The double-zero problem

Niche theory (Hutchinson, 1957) states that species have ecological "preferences", meaning that they have evolved genetic adaptations to specific environmental conditions, including other species. Species are mostly found at locations where they encounter appropriate living conditions. The theory also predicts that species have *unimodal distributions* along environmental variables (Whittaker, 1967), like the Gaussian curves in Fig. 4.5: a species is found in greater abundance in some intervals along the gradients of major environmental variables or along composite axes*. The position of the mode of a species distribution along an environmental variable can be interpreted as the optimum value for the species along that variable. Along an environmental gradient, a species becomes rare and ultimately absent as one departs from its optimal conditions.

Unimodal distribution

As a consequence, community composition data sampled across a range of environmental conditions typically contain many zero values. This phenomenon is discussed in most texts of community and numerical ecology, in particular in Whittaker (1967), ter Braak (1987c) and ter Braak & Prentice (1988).

Comparison of sites is often based upon species abundance data. Species are important indicators of the apportioning of environmental resources among them. The division of resources should be reflected in the relative productivities of the species (Whittaker, 1972). The productivity of different species is not easily measured, however, and ecologists most often rely on other values of species importance such as number of individuals, biomass, coverage (for plants or corals) or basal area (for plants).

If a species is present at two sites, this is an indication of similarity between these sites since they both present conditions that are favourable or at least tolerable for the species. Likewise, the presence of a species at site 1 and not at site 2 is taken as an indication of difference in ecological conditions, notwithstanding sampling error†. However, if a species is absent from two sites, it may be because these sites have environmental conditions that are outside the niche of the species, and these conditions

* Composite environmental axes can be computed by ordination (Chapter 9), for instance as the principal components (PCA, Section 9.1) of a matrix of environmental variables.

† For a variety of reasons, species may not be observed at sites where they are present. Species may be inconspicuous, camouflaged, or hidden. With fungi for example, the carpophores of a species may not appear above ground at the time of a survey although the mycelium is present in the soil. When sampling is done "blindly", e.g. in a lake or the ocean with a plankton or fish net, many species may escape capture either randomly of by active avoidance of the sampling gear.

may be similar or very different at the two sites. Hence most ecologists do not consider that the absence of a species from two sites provides univocal or useful information. It is also understood, of course, that besides unimodal distributions and niche optimality, several reasons may explain the local absence of a species: the niche of the species may be present in one (or both) of two sites but be occupied by substitute species; absence may also be the result of the species dispersion, random local extinction, historical events, or other processes that cause stochastic variation.

The proportion of zeros in community composition data generally increases with the variability in environmental conditions among the sampling sites. If sampling has been conducted along one or several environmental axes, the species present are likely to differ at least partly from site to site. Including double zeros in the comparison between sites would result in high values of similarity for the many pairs of sites holding only a few species, these pairs presenting many double zeros; this would not provide a correct ecological assessment of the situation.

Double-zero problem

Because double zeros are not informative, their interpretation generates *the double-zero problem*: is the value of an association coefficient affected by inclusion of double zeros in its calculation? When choosing an association coefficient, ecologists must pay attention to the interpretation of double zeros: except in very limited cases (e.g. controlled experiments involving very few species and with small uncontrolled ecological variation), it is preferable to draw no ecological conclusion from the simultaneous absence of a species at two sites. In numerical terms, this means to skip double zeros when computing similarity or distance coefficients using species presence-absence or abundance data. Coefficients of this type are called *asymmetrical* because they treat double absences in a different way than double presences.

Asymmetrical coefficient

Symmetrical coefficient

In similarity coefficients (S), the handling of double zeros is clear in coefficient formulas (Section 7.3). Similarity coefficients all have a minimum value of 0 and a maximum value of 1. In *symmetrical* similarity coefficients, state zero for two objects is treated in exactly the same way as any other pair of values. This would be the correct way to handle double zeros in the case, for example, where two lakes are found to have 0 mgL^{-1} of dissolved oxygen in the hypolimnion in winter because this observation provides valuable information concerning their physical and chemical similarity and their capacity to support species. Coefficient S_{15}, for instance, would consider a double zero as an indication of resemblance between the lakes and include this information in the overall assessment of their similarity.

In distance coefficients (D, Section 7.4), however, one has to examine if the value computed for pairs of sites depends primarily on which species are present at each site (*asymmetrical coefficients*), or strictly on the numerical differences between species abundances (*symmetrical coefficients*). Symmetrical coefficients like the Euclidean distance (D_1) will be shown to lead to incorrect conclusions from an ecologist's viewpoint (see Fig. 7.8). The asymmetrical distance coefficients all have a fixed upper bound, which is either 1 or $\sqrt{2}$ in most cases.

Because ordination methods implicitly (PCA, CA, Sections 9.1 and 9.2) or explicitly (PCoA, Section 9.3) use a distance function as their metric to position objects with respect to one another in ordination space, it is important to make sure that the chosen distance coefficient is meaningful for the objects under study, especially when dealing with community composition data. By choosing an appropriate distance measure, an ecologist tries to appropriately model the relationships among the sites for the data at hand. The choice of a similarity or distance measure (Section 7.6) is an ecological, not a statistical decision.

7.3 Q mode: similarity coefficients

Similarities form a large group of coefficients in the literature. The similarity coefficients in the present section measure the association between *objects*. Similarities take values in the interval [0, 1], 1 being the similarity of two identical objects and of an object with itself. In contrast to most distance coefficients, similarity coefficients are never metric (definition at the beginning of Section 7.4) since it is always possible to find two objects, A and B, that are more similar than the sum of their similarities with another, more distant, object C. It follows that similarities cannot be used directly to position objects in a metric space (ordination; Chapter 9); they must be converted into distances using one of the transformations of Subsection 7.2.1. Which transformation to use is discussed at the beginning of Section 7.4 and in Tables 7.2 and 7.3. For clustering (Chapter 8), however, algorithms can be easily adapted to conduct the analysis on either a distance or a similarity matrix.

Similarity coefficients were first developed for binary descriptors, representing presence-absence data or answers to yes-no questions. They were later generalized to multi-state descriptors when computers made that possible. Another major dichotomy among similarity coefficients concerns how they deal with double-zeros or negative matches. This dichotomy was discussed in Subsection 7.2.2.

The remainder of this section distinguishes between binary and quantitative similarity coefficients and, for each type, those that use double-zeros or exclude them from the assessment of resemblance. Tables 7.4 and 7.5 summarize the use of the various similarity and distance coefficients in ecology.

1 — Symmetrical binary coefficients

In the simplest cases, the similarity between two sites is based on presence-absence data. Binary descriptors may describe the presence or absence of environmental

conditions (here) or species (next subsection). Observations may be summarized in a 2×2 frequency table:

Object $\mathbf{x}_1$ \ Object $\mathbf{x}_2$	1	0	
1	a	b	$a+b$
0	c	d	$c+d$
	$a+c$	$b+d$	$p=a+b+c+d$

where a is the number of descriptors for which the two objects are coded 1, d is the number of descriptors coding the two objects 0, whereas b and c are the numbers of descriptors for which the two objects are coded differently; and p is the total number of descriptors. An obvious way of computing the similarity between two objects is to count the number of descriptors that code the objects in the same way and divide this by the total number of descriptors:

$$S_1(\mathbf{x}_1, \mathbf{x}_2) = \frac{a+d}{p} \qquad \textbf{(7.1)}$$

Coefficient S_1* is called the *simple matching coefficient* (Sokal & Michener, 1958). When using this coefficient, one assumes that there is no difference between double-0 and double-1. This is the case, for instance, when any one of the two states of each descriptor could be coded 0 or 1 indifferently. A variant of this measure is the *coefficient of Rogers & Tanimoto* (1960) in which differences are given more weight than resemblances:

$$S_2(\mathbf{x}_1, \mathbf{x}_2) = \frac{a+d}{a+2b+2c+d} \qquad \textbf{(7.2)}$$

Sokal & Sneath (1963) proposed four other measures that include double-zeros. They have their counterparts in coefficients that exclude double-zeros, in the next subsection:

$$S_3(\mathbf{x}_1, \mathbf{x}_2) = \frac{2a+2d}{2a+b+c+2d} \qquad \textbf{(7.3)}$$

counts resemblances as being twice as important as differences;

* The numbers of the coefficients found in the first edition of this book (*Écologie numérique*, Masson, Paris, 1979) were not changed in subsequent editions because these numbers had rapidly been adopted by ecologists and used as coefficient identifiers in computer programs. Coefficients added since the 1983 edition have received sequential numbers.

$$S_4(\mathbf{x}_1, \mathbf{x}_2) = \frac{a+d}{b+c} \quad \textbf{(7.4)}$$

compares the resemblances to the differences, in a measure that goes from 0 to infinity;

$$S_5(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{4}\left[\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{b+d} + \frac{d}{c+d}\right] \quad \textbf{(7.5)}$$

compares the resemblances to the marginal totals;

$$S_6(\mathbf{x}_1, \mathbf{x}_2) = \frac{a}{\sqrt{(a+b)\,(a+c)}}\,\frac{d}{\sqrt{(b+d)\,(c+d)}} \quad \textbf{(7.6)}$$

is the product of the geometric means of the terms relative to a and d, respectively, in coefficient S_5.

Among the above coefficients, S_1 to S_3 are of more general interest for ecologists, but the others may occasionally prove useful to adequately handle special descriptors. Three additional measures are available in the NTSYSPC package (Section 7.8, footnote): the *Hamann coefficient*:

$$S = \frac{a+d-b-c}{p} \quad \textbf{(7.7)}$$

the *Yule coefficient*:

$$S = \frac{ad-bc}{ad+bc} \quad \textbf{(7.8)}$$

and *Pearson's* ϕ (phi):

$$\varphi = \frac{ad-bc}{\sqrt{(a+b)\,(c+d)\,(a+c)\,(b+d)}} \quad \textbf{(7.9)}$$

where the numerator is the value of the determinant of the 2×2 frequency table. φ is actually the square root of the X^2 (chi-square) statistic computed from 2×2 tables divided by n (eq. 7.61). In ecology, coefficients of this type are mostly used in R-mode analyses. These last indices are described in detail in Sokal & Sneath (1963).

2 — Asymmetrical binary coefficients

Coefficients that parallel those above can be used to compare sites based on species presence-absence data, where the comparison must exclude double-zeros. The best-known measure is Jaccard's (1900, 1901, 1908) *coefficient of community*. It is often referred to simply as *Jaccard's coefficient*:

$$S_7(\mathbf{x}_1, \mathbf{x}_2) = \frac{a}{a+b+c} \quad \textbf{(7.10)}$$

in which all terms have equal weights. The *Sørensen coefficient*[*] (1948) gives double weight to double presences:

$$S_8(\mathbf{x}_1, \mathbf{x}_2) = \frac{2a}{2a + b + c} \quad \textbf{(7.11)}$$

because (see above) one may consider that the presence of a species is more informative than its absence at one site. Absence of a species at one site may be due to various factors, as discussed in Subsection 7.2.2; it does not necessarily reflect differences in the environment. Double-presence, on the contrary, is a strong indication of resemblance. S_8 is the binary form of the Steinhaus similarity S_{17}, meaning that its value is equal to S_{17} applied to binary data. Before Sørensen, Dice (1945) had used S_8 under the name *coincidence index* in an R-mode study of species associations; this question is further discussed in Section 7.5.

The distance version of this coefficient, $D_{13} = 1 - S_8$, is a semimetric, as shown in the example that follows eq. 7.57. A consequence is that principal coordinate analysis (Section 9.3) of a S_8 or D_{13} resemblance matrix is likely to produce negative eigenvalues. Solutions to this problem are discussed in Subsection 9.3.4. The easiest solution is to base the principal coordinate analysis on square-root-transformed distances $D = \sqrt{1 - S_8}$ instead of $D = 1 - S_8$ (Table 7.2).

Another variant of S_7 gives triple weight to double presences:

$$S_9(\mathbf{x}_1, \mathbf{x}_2) = \frac{3a}{3a + b + c} \quad \textbf{(7.12)}$$

The asymmetrical counterpart to the coefficient of Rogers & Tanimoto (S_2 in the previous subsection) was suggested by Sokal & Sneath (1963). This coefficient gives double weight to differences in the denominator:

$$S_{10}(\mathbf{x}_1, \mathbf{x}_2) = \frac{a}{a + 2b + 2c} \quad \textbf{(7.13)}$$

Coefficients S_9, S_8, S_7 and S_{10} form a series S_w with weights $w = \{1/3, 1/2, 1, 2\}$ respectively; w is the weight of $(b + c)$ in the formulas, considering that a receives a weight of 1. Gower & Legendre (1986) have shown that the coefficients in this series are monotonically related, meaning that they produce the same results when used in order-invariant methods like single, complete and proportional-link linkage clustering (Section 8.5) or nonmetric multidimensional scaling (Section 9.4), which rely on the

[*] Some authors refer to this coefficient as having been first described by Czekanowski (1913). The Czekanowski (1913) paper, written in Polish, is about body part measurements and anthropology; it deals with quantitative measurement variables only. The index developed in that paper with the a, b, c symbols has nothing to do with the binary indices (Jaccard, Sørensen) described in the present section.

ordinal and not the absolute values of the similarities. Among the symmetrical binary coefficients, S_3, S_1 and S_2 form a similar series with weights $w = \{0.5, 1, 2\}$.

Russell & Rao (1940) suggested a measure that compares the number of double presences, in the numerator, to the total number of species found at all sites, including species that are absent (d) from the pair of sites considered:

$$S_{11}(\mathbf{x}_1, \mathbf{x}_2) = \frac{a}{p} \quad \textbf{(7.14)}$$

Kulczynski (1928) proposed a coefficient opposing double-presences to the sum of disagreements (presence in one site and absence in the other):

$$S_{12}(\mathbf{x}_1, \mathbf{x}_2) = \frac{a}{b+c} \quad \textbf{(7.15)}$$

Among their coefficients for presence-absence data, Sokal & Sneath (1963) mention the binary version of Kulczynski's coefficient S_{18} for quantitative data:

$$S_{13}(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{2}\left[\frac{a}{a+b} + \frac{a}{a+c}\right] \quad \textbf{(7.16)}$$

where double-presences are compared to the marginal totals $(a + b)$ and $(a + c)$.

Ochiai (1957) used, as measure of similarity, the geometric mean of the ratios of a to the number of species in each site, i.e. the marginal totals $(a + b)$ and $(a + c)$:

$$S_{14}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\frac{a}{(a+b)}\,\frac{a}{(a+c)}} = \frac{a}{\sqrt{(a+b)\,(a+c)}} \quad \textbf{(7.17)}$$

Note the relationship to the X^2 (chi-square) statistic, where the expected value of the cell containing the value a is $E(a) = (a + b)(a + c)/n$. Coefficient S_{14} is the same as S_6 with the portion that concerns double-zeros (d) excluded. S_{14} is the binary form of both the chord (D_3) and Hellinger (D_{17}) distances (Section 7.4): when applied to binary data, these two distance coefficients produce values equal to $\sqrt{2}\sqrt{1 - S_{14}}$.

Faith (1983) suggested the following coefficient for community composition data, in which disagreements b and c (presence in one site and absence in the other) are given a weight opposite to that of double presences a. The value of S_{26} is not invariant but decreases with an increasing number of double-zeros:

$$S_{26}(\mathbf{x}_1, \mathbf{x}_2) = \frac{2a+d}{2p} = \frac{a-b-c}{2p} + \frac{1}{2} \quad \textbf{(7.18)}$$

3 — Symmetrical quantitative coefficients

Ecological descriptors often have more than two states. Researchers have sometimes extended the binary coefficients described in Subsection 7.3.1 to accommodate nonordered multi-state descriptors. For example, the simple matching coefficient can be used as follows with multi-state qualitative descriptors:

$$S_1(\mathbf{x}_1, \mathbf{x}_2) = \frac{\text{agreements}}{p} \qquad \textbf{(7.19)}$$

where the numerator contains the number of descriptors for which the two objects are in the same state. For example, if a pair of objects was described by the following 10 multi-state qualitative descriptors:

	Descriptors									
Object $\mathbf{x}_1$	9	3	7	3	4	9	5	4	1	6
Object $\mathbf{x}_2$	2	3	2	1	2	9	3	2	1	6
Agreements	0	+ 1	+ 0	+ 0	+ 0	+ 1	+ 0	+ 0	+ 1	+ 1 = 4

the value of S_1 computed for these data would be:

$$S_1(\mathbf{x}_1, \mathbf{x}_2) = 4 \text{ agreements}/10 \text{ descriptors} = 0.4$$

It is possible to extend in the same way the use of all binary coefficients to multi-state descriptors. However, coefficients of this type often result in a loss of valuable information, especially in the case of ordered descriptors for which two objects can be compared on the basis of the *amount of difference* between states.

Gower (1971a) proposed a general coefficient of similarity that can combine different types of descriptors and process each one according to its own mathematical type. Although the description of this coefficient may seem a bit complex, it can be easily translated into a short computer program. The coefficient initially takes the following form (see also the final form, eq. 7.21):

$$S_{15}(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{p}\sum_{j=1}^{p} s_{12j}$$

Partial similarity

The similarity between two objects is the average, over the p descriptors, of the similarities calculated for all descriptors. For each descriptor j, the *partial similarity value* s_{12j} between objects $\mathbf{x}_1$ and $\mathbf{x}_2$ is computed as follows:

- For *binary* descriptors, $s_j = 1$ (agreement) or 0 (disagreement). Gower proposed two forms for this coefficient. The form used here is symmetrical, giving $s_j = 1$ to double-zeros. The other form, used in Gower's asymmetrical coefficient S_{19} (Subsection 7.3.4), gives $s_j = 0$ to double-zeros.

• *Qualitative* descriptors are treated following the simple matching rule stated above: $s_j = 1$ when there is agreement and $s_j = 0$ when there is disagreement. Double-zeros are treated as in the symmetrical form of the previous paragraph.

• *Semiquantitative descriptors* can be handled in various ways in the computation of S_{15}. In function ***gowdis()*** of the R package FD, three options are available for handling semiquantitative variables, which are called *ordered factors* in R.

1. With the "classic" option, eq. 7.20 (below) is used as if the values were quantitative.

2. With the "metric" option, the values are ranked across the set of objects under study. Tied values are replaced by their average ranks. Consider an example where the objects under study are a subset of a larger data base in which a semiquantitative variable has ten ordered states, but the subset only contains three of the ten states. With this option, the states are recoded as ranks 1, 2 and 3 (or by the average values of the tied ranks) before they are used in eq. 7.20. Function ***daisy()*** of package CLUSTER also recodes semiquantitative variables in that way before computing the Gower distance.

3. With the "podani" option, the states and tied ranks are recoded as in the "metric" option. Equation 7.20 (below) is modified to take tied ranks into account, as proposed by Podani (1999). The modified eq. 7.20 is shown in the documentation file of the ***gowdis()*** function.

With these options, one should make sure that distances between adjacent states are comparable in magnitude. For example, for a semiquantitative descriptor coded from 1 to 3, $|y_{1j} - y_{2j}|$ of eq. 7.20 makes sense only if the difference between states 1 and 2 can be thought of as similar to the difference between states 2 and 3. If there is too much difference, values $|y_{1j} - y_{2j}|$ are not comparable and the semiquantitative descriptor should be converted into an unordered factor.

• *Quantitative* descriptors (real numbers) are treated in an interesting way. For each descriptor, one first computes the difference between the states of the two objects $|y_{1j} - y_{2j}|$, as in the case of distance coefficients belonging to the Minkowski metric group (Section 7.4). This value is then divided by the largest difference (R_j) found for this descriptor across all sites in the study — or if one prefers, in a reference population*. Since this ratio is actually a normalized distance, it is subtracted from 1 to transform it into a similarity:

$$s_{12j} = 1 - [|y_{1j} - y_{2j}| / R_j] \qquad \textbf{(7.20)}$$

Missing values Kronecker delta

The Gower coefficient may be programmed to include an additional element of flexibility: no comparison is computed for descriptors where information is *missing* for one or the other object. This is obtained by a value w_j, called a *Kronecker's delta*, describing the presence or absence of information: $w_j = 0$ when the information about

* In most applications, the largest difference R_j is calculated for the data table under study. In epidemiological studies, for example, one may proceed to the analysis of a subset of a much larger database. To ensure consistency of the results in all the partial studies, it is recommended to calculate the largest differences (the "range" statistic of databases) observed throughout the whole database for each descriptor j and use these as values R_j when computing S_{15} or S_{19}.

y_j is missing for one or the other object, or both; $w_j = 1$ when information is present for the two objects. The final form of the *Gower coefficient* is the following:

$$S_{15}(\mathbf{x}_1, \mathbf{x}_2) = \frac{\sum_{j=1}^{p} w_{12j} s_{12j}}{\sum_{j=1}^{p} w_{12j}} \qquad \textbf{(7.21)}$$

Coefficient S_{15} produces similarity values between 0 and 1 (maximum similarity).

One last touch of complexity, which was not suggested in Gower's paper but is added here, provides weighting to the various descriptors. Instead of *0 or 1*, one can assign to w_j a value *between 0 and 1* corresponding to the weight one wishes each descriptor to have in the analysis. Descriptors with weights close to 0 contribute little to the final similarity value whereas descriptors with higher weights (closer to 1) contribute more. Giving a weight of 0 to a descriptor is equivalent to removing it from the analysis. A missing value automatically changes the weight w_j to 0.

The following numerical example illustrates the computation of coefficient S_{15}. In the example, two sites are described by eight quantitative environmental descriptors. Values R_j (the range of values among all objects, for each descriptor y_j) given in the table have been calculated for the whole database prior to computing coefficient S_{15}. Weights w_{12j} are only used in this example to eliminate descriptors with missing values (Kronecker delta function):

	Descriptors j								Sum
Object $\mathbf{x}_1$	2	2	–	2	2	4	2	6	
Object $\mathbf{x}_2$	1	3	3	1	2	2	2	5	
w_{12j}	1	1	0	1	1	1	1	1	= 7
R_j	1	4	2	4	1	3	2	5	
$\lvert y_{1j} - y_{2j} \rvert$	1	1	–	1	0	2	0	1	
$\lvert y_{1j} - y_{2j} \rvert / R_j$	1	0.25	–	0.25	0	0.67	0	0.20	
$w_{12j} s_{12j}$	0	0.75	0	0.75	1	0.33	1	0.80	= 4.63

thus $S_{15}(\mathbf{x}_1, \mathbf{x}_2) = 4.63/7 = 0.66$.

Another general coefficient of similarity was proposed by Estabrook & Rogers (1966). The similarity between two objects is, as in S_{15}, the sum of the partial similarities by descriptors, divided by the number of descriptors for which there is information for the two objects. In their original publication, the authors used state 0 to

identify missing values, but any other convention is acceptable, like *NA* in R. The general equation of this coefficient is the same as for Gower's coefficient (eq. 7.21):

$$S_{16}(\mathbf{x}_1, \mathbf{x}_2) = \frac{\sum_{j=1}^{p} w_{12j} s_{12j}}{\sum_{j=1}^{p} w_{12j}} \tag{7.22}$$

As in S_{15}, the w_j parameters may be used as weights (between 0 and 1) instead of only playing the roles of Kronecker deltas. The coefficient of Estabrook & Rogers differs from S_{15} in the computation of the partial similarities s_j.

Partial similarity

In the paper of Estabrook & Rogers (1966), the state values were positive integers and the descriptors were either ordered or unordered. The partial similarity between two objects for a given descriptor j is computed using a monotonically decreasing function of partial similarity. Among all possible functions of this type, the authors empirically chose the following function of two numbers, d and k:

$$\begin{aligned} s_{12j} &= f(d_{12j}, k_j) = \frac{2(k+1-d)}{2k+2+dk} && \text{when } d \le k \\ s_{12j} &= f(d_{12j}, k_j) = 0 && \text{when } d > k \end{aligned} \tag{7.23}$$

where d is the distance between the states of the two objects $\mathbf{x}_1$ and $\mathbf{x}_2$ for descriptor j, i.e. the same value as $|y_{1j} - y_{2j}|$ in eq. 7.20; k is a parameter, determined *a priori* by the user for each descriptor, that describes how far non-null partial similarities are permitted to go. Parameter k is equal to the largest difference d for which the partial similarity s_{12j} (for descriptor j) is allowed to be larger than 0. Values k for the various descriptors may be quite different from one another. For example, for a descriptor coded from 1 to 4, one might decide to use $k = 1$ for this descriptor; for another descriptor with code values from 1 to 50, $k = 10$ could be used. For qualitative descriptors, k is set to 0.

In order to fully understand the partial similarity function s_{12j} (eq. 7.23), readers are invited to compute s_{12j} by hand in the following numerical example. Values k, are provided for each descriptor in the table:

	Descriptors j						$S_{16}(\mathbf{x}_1, \mathbf{x}_2)$
Object $\mathbf{x}_1$	2	1	3	4	2	1	
Object $\mathbf{x}_2$	2	2	4	3	2	3	
k_j	1	0	1	2	1	1	
	↓	↓	↓	↓	↓	↓	
$s_{12j} = f(d_{12j}, k_j)$	1.0 + 0 + 0.4 + 0.5 + 1.0 + 0						= 2.9 / 6 = 0.483

Table 7.1 Values of the partial similarity function $f(d, k)$ in coefficients S_{16} and S_{20}, for the most usual values of k (adapted from Legendre & Rogers, 1972: 594).

k	d							
	0	1	2	3	4	5	6	7
0	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1	1	0.40	0.00	0.00	0.00	0.00	0.00	0.00
2	1	0.50	0.20	0.00	0.00	0.00	0.00	0.00
3	1	0.55	0.29	0.12	0.00	0.00	0.00	0.00
4	1	0.57	0.33	0 18	0.08	000	0.00	0.00
5	1	0.59	0.36	0.22	0.13	0.05	0.00	0.00

Values taken by the partial similarity function for the first values of k are shown in Table 7.1. If $k = 0$ for all descriptors, S_{16} is identical to the simple-matching coefficient for multistate descriptors (eq. 7.19).

These same values of function $f(d, k)$ are shown in Fig. 7.3a, which illustrates how the function decreases with increasing d. It is easy to see that function $f(d, k)$, which was originally defined by Estabrook & Rogers for discontinuous descriptors (coded only with integers: 0, 1, 2, 3, …), can actually be used with real-number descriptors since the function only requires that d and k be differences, i.e. natural numbers. Figure 7.3a also raises the question: could $f(d, k)$ take negative values? To accomplish that, Legendre & Chodorowski (1977) proposed to simply leave out the second line of eq. 7.23 stating that $f(d, k) = 0$ when $d > k$. This is shown in Fig. 7.3b, where the function decreases over the interval $[0, \infty)$, taking negative values when $d > (k + 1)$; differences are subtracted from resemblances in this form of the coefficient. This contributes to further separate dissimilar objects when the similarity matrix is subjected to clustering (Chapter 8).

Partial similarity matrix

The major interest of S_{16} over all other coefficients is the possibility to use predefined partial similarity matrices for environmental descriptors. Estabrook & Rogers (1966) proposed this alternative for situations where function $f(d, k)$ does not adequately describe the relationships between objects for some descriptor j. The approach consists in providing the computer program with a "do-it yourself" matrix that describes the partial similarities between all states of descriptor j. This partial similarity matrix replaces in eq. 7.22 the values $s_{12j} = f(d_{12j}, k)$ computed by eq. 7.23

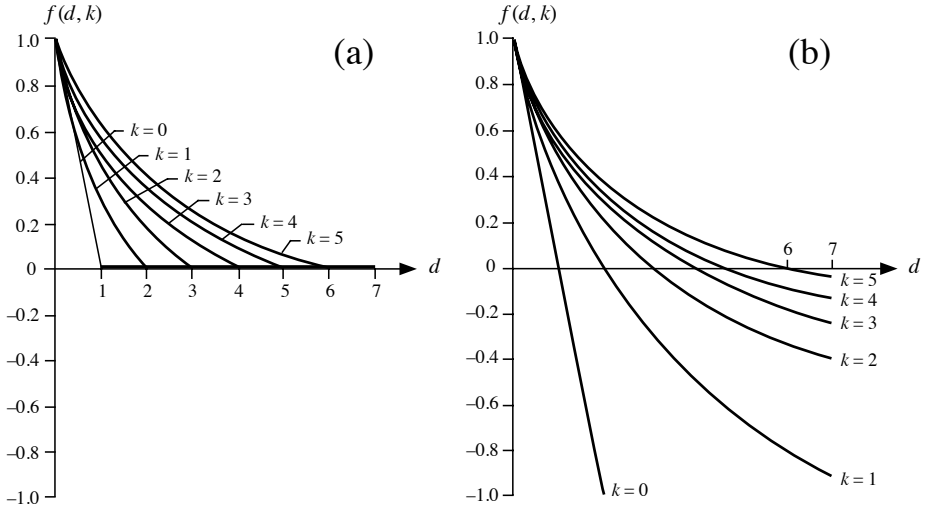


Figure 7.3 Coefficients S_{16} and S_{20}: change in $f(d, k)$ as a function of d, for six values of k, (a) under the condition $f(d, k) = 0$ when $d > k$; (b) without this condition. Adapted from Legendre & Chodorowski (1977).

for descriptor j; the partial similarity of the other descriptors is computed with eq. 7.23. Partial similarity matrices may be constructed for any or all descriptors if needed.

Ecological application 7.3a

Partial similarity matrix

In a study of terrestrial fauna, Legendre *et al*. (1978) used the following partial similarity matrix (matrix unpublished) for the descriptor "category of aquatic ecosystem", which was one of the descriptors of the ecosystems where tracks of animal had been recorded in the snow blanket:

	1	2	3	4	5	6	7	8	9
1. <5% water	1.0								
2. 5-15% water	0.4	1.0							
3. >15% water, lakes < 250 ha	0.0	0.4	1.0						
4. >15% water, lakes 250-500 ha	0.0	0.3	0.8	1.0					
5. >15% water, lakes 500-1000 ha	0.0	0.2	0.7	0.8	1.0				
6. >15% water, lakes 1000-2500 ha	0.0	0.2	0.6	0.7	0.8	1.0			
7. >15% water, lakes > 2500 ha	0.0	0.2	0.5	0.6	0.7	0.8	1.0		
8. >5% rivers	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	
9. Bordering large rivers	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	1.0

The upper triangle of the matrix is not given; it is symmetric to the lower one. The diagonal may also be left out because the partial similarity of a state with itself must be 1. It is shown here to indicate that the matrix contains similarities, not distances. The matrix means that a site from an area with less than 5% of its surface covered by water is given a partial similarity $s_j = 0.4$ with another site from an area with 5 to 15% of its surface covered by water. State 1 has partial similarity with state 2 only; lake systems only have partial similarities with other lake systems, the similarity decreasing as the difference in lake areas increases; and rivers only have partial similarities when compared to other rivers. Partial similarity matrices are especially useful with descriptors that are nonordered, or only partly ordered as is the case here.

Partial similarity matrices represent a powerful way of using unordered or partly ordered descriptors in multivariate data analyses. They are useful in the following cases:

- When, from the user's point of view, function $f(d, k)$ (eq. 7.23) does not adequately describe the partial similarity relationships.

- When the descriptor states are not fully ordered. For example, in a study on ponds, the various states of descriptor "water temperature" may be followed by state "dry pond", which is quite different from a lack of information.

- If some states are on a scale different from that of the other states. For example, 0-10, 10-20, 20-30, 30-40, and then 50-100, 100-1000, and >1000.

- With nonordered or only partly ordered descriptors (including "circular variables" such as directions of the compass card or hours of the day), if one considers that pairs of sites coded into different states are partly similar, as in Ecological application 7.3a.

4 — Asymmetrical quantitative coefficients

Subsection 7.3.3 started with an extension of coefficient S_1 to multi-state descriptors. In the same way, the binary coefficients described in Subsection 7.3.2 could be extended to accommodate multi-state species abundance data. For example, Jaccard's coefficient would become:

$$S_7(\mathbf{x}_1, \mathbf{x}_2) = \frac{\text{agreements}}{p - \text{double-zeros}}$$

where the 'agreement' quantity in the numerator is the number of species with *the exact same* abundance at the two sites. This form would obviously cause the loss of part of the information carried by species abundance data.

The classic indices of compositional similarity described in Subsection 7.3.2 are highly sensitive to sample size, especially for assemblages containing many rare species. Chao *et al*. (2005) developed new forms of the Jaccard (S_7) and Sørensen (S_8) indices, applicable to quantitative community composition data, that estimate and take into account the number of unseen shared species. A full description of these indices,

based on a probabilistic derivation, is found in the Chao *et al*. (2005) paper. Function ***vegdist()*** of the VEGAN library (with method = "chao") produces distances corresponding to the modified abundance-based Jaccard similarity index ($D = 1 - S$). Numerical simulations reported in the Chao *et al*. (2005) paper show that the new estimators of the Jaccard and Sørensen indices are considerably less biased than the corresponding classic indices (S_7, S_8) when a substantial proportion of the species are missing from the sample data.

Other measures are available for species abundance data. They are divided in two categories: the coefficients that can be used with either raw or normalized data and the measures whose application should be limited to normalized data.

• As discussed in Subsection 1.5.6 and Section 7.7, the distribution of abundances of a species across an ecological gradient may be strongly skewed. Normalization of species abundances often calls for square root, double square root, or logarithmic transformations. Another way to obtain approximately normal data is to use a scale of relative abundances with boundaries forming a geometric progression, for example a scale from 0 (absent) to 7 (very abundant). The Anderson *et al*. (2006) transformation (eq. 7.66) is an example of such a recoding method.

• Abundances thus normalized reflect the role of each species in the ecosystem better than the raw abundance data, since a species represented by 100 individuals at a site does not have a role 10 times as important in the ecological equilibrium as another species represented by 10 individuals, everything else being equal. The former is perhaps twice as important as the latter; this is the ratio obtained after applying a base-10 logarithmic transformation, and assuming that numbers 100 and 10 at the site are representative of true relative abundances in the population.

Some coefficients lessen the effect of the largest differences and may therefore be used with raw species abundances, whereas others compare the different abundance values in a more linear way and are thus better adapted to normalized data.

In the group of coefficients to be used with raw species abundances, the best-known is a coefficient attributed to the Polish mathematician H. Steinhaus by Motyka (1947) and Motyka *et al*. (1950). This measure has been rediscovered a number of times; its one-complement is known as the percentage difference, Odum, or Bray-Curtis coefficient (eq. 7.58; see note there). It is sometimes incorrectly attributed to anthropologist Czekanowski (1909 and 1913; Czekanowski's *mean character difference* coefficient is described in Subsection 7.4.1, eq. 7.45). The Steinhaus coefficient compares two sites ($\mathbf{x}_1$, $\mathbf{x}_2$) in terms of the minimum abundance of each species:

$$S_{17}(\mathbf{x}_1, \mathbf{x}_2) = \frac{W}{(A+B)/2} = \frac{2W}{(A+B)} \qquad \textbf{(7.24)}$$

where W is the sum of the minimum abundances of the various species, this minimum being defined as the abundance at the site where the species is the rarest. A and B are the sums of the abundances of all species at each of the two sites or, in other words, the

total number of specimens observed or captured at each site, respectively. Consider the following numerical example:

	Species abundances						A	B	W
Site $\mathbf{x}_1$	7	3	0	5	0	1	16		
Site $\mathbf{x}_2$	2	4	7	6	0	3		22	
Minimum	2	3	0	5	0	1			11

$$S_{17}(\mathbf{x}_1, \mathbf{x}_2) = \frac{2 \times 11}{16 + 22} = 0.579$$

This measure is closely related to the Sørensen coefficient (S_8): if presence-absence data are used instead of species counts, S_{17} becomes S_8 (eq. 7.11).

The distance version of this coefficient, $D_{14} = 1 - S_{17}$, is a semimetric, as shown in the example that follows eq. 7.58. A consequence is that principal coordinate analysis of a D_{14} resemblance matrix is likely to produce negative values. Solutions to this problem are discussed in Subsection 9.3.4. The easiest solution is to base the principal coordinate analysis on square-root-transformed distances $D = \sqrt{1 - S_{17}}$ instead of $D = 1 - S_{17}$ (Table 7.2).

The Kulczynski coefficient (1928) also belongs to the group of measures that are appropriate for raw abundance data. The sum of minima is first compared to the grand total at each site; then the two values are averaged:

$$S_{18}(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{2}\left(\frac{W}{A} + \frac{W}{B}\right) \qquad \textbf{(7.25)}$$

For presence-absence data, S_{18} becomes S_{13} (eq. 7.16). For the numerical example above, coefficient S_{18} is computed as follows:

$$S_{18}(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{2}\left(\frac{11}{16} + \frac{11}{22}\right) = 0.594$$

Coefficients S_{17} and S_{18} always produce values between 0 and 1, although Kulczynski (1928) multiplied the final value by 100 to obtain a percentage. Kulczynski's approach, which consists in computing the average of two comparisons, seems more arbitrary than Steinhaus' method, in which the sum of minima is compared to the mean of the two site sums. In practice, values of these two coefficients are almost monotonic.

The following coefficients belong to the group adapted to "normalized" abundance data, meaning here unskewed or not strongly skewed frequency distributions. These coefficients parallel S_{15} and S_{16} of the previous subsection. Concerning coefficient S_{19}, Gower (1971a) had initially proposed that his general coefficient S_{15} should exclude

double-zeros from the comparison (Subsection 7.3.3); this makes it well-suited for quantitative species abundance data. Since the differences between states are computed as $|y_{1j} - y_{2j}|$ and are thus linearly related to the measurement scale, this coefficient should be used with previously normalized data. The general form is:

$$S_{19}(\mathbf{x}_1, \mathbf{x}_2) = \frac{\sum_{j=1}^{p} w_{12j} s_{12j}}{\sum_{j=1}^{p} w_{12j}}, \text{ where} \qquad \textbf{(7.26)}$$

Partial similarity

- $s_{12j} = 1 - [|y_{1j} - y_{2j}| / R_j]$, as in S_{15},
- and $w_{12j} = 0$ when $y_{1j} = y_{2j}$ = absence of the species, i.e. $(y_{1j} + y_{2j}) = 0$;
- while $w_{12j} = 1$ in all other cases.

For binary (presence-absence) species data, S_{19} is equivalent to the Jaccard coefficient S_7. The weights w_j could be made to vary between 0 and 1, either to reflect the biomasses or biovolumes of the different species, or to compensate for selective effects of the sampling gear.

Legendre & Chodorowski (1977) proposed an asymmetrical coefficient of similarity that parallels S_{16}. This measure uses a slightly modified version of the partial similarity function $f(d, k)$ (eq. 7.23), or else an imposed matrix of partial similarities as in Ecological application 7.3a. Since S_{20} processes all differences d in the same way, irrespective of whether they correspond to high or low values in the scale of abundances, it is better to use this measure with normalized abundance data. The only difference between S_{16} and S_{20} is in the way in which double-zeros are handled. The general form of the coefficient is the sum of the partial similarity values over all species, divided by the total number of species in the combined two sites:

$$S_{20}(\mathbf{x}_1, \mathbf{x}_2) = \frac{\sum_{j=1}^{p} w_{12j} s_{12j}}{\sum_{j=1}^{p} w_{12j}}, \text{ where} \qquad \textbf{(7.27)}$$

Partial similarity

- $s_{12j} = f(d_{j12}, k_j)$ $\begin{cases} = \dfrac{2(k+1-d)}{2k+2+dk} & \text{when } d \le k \quad \text{(I)} \\ = 0 & \text{when } d > k \quad \text{(II)} \\ = 0 & \text{when } y_{j1} \text{ or } y_{j2} = 0 \text{ (i.e. } y_{j1} \times y_{j2} = 0) \quad \text{(III)} \end{cases}$

When comparing the presence of a species at a site with its absence at the other site, a similarity of 0 is imposed in point III to acknowledge a strong ecological difference.

• or else $s_{12j} = f(y_{1j}, y_{2j})$ is given by a partial similarity matrix, as in Ecological application 7.3a, in which $s_{12j} = 0$ when y_{1j} or $y_{2j} = 0$,

• and $w_{12j} = 0$ when y_{1j} or y_{2j} = absence of information, or
when $y_{1j} = y_{2j}$ = absence of the species, i.e. $(y_{1j} + y_{2j}) = 0$,

• while $w_{12j} = 1$ in all other cases. Else, w_{12j} may receive a value between 0 and 1, as explained above for S_{19}.

In summary, the properties of coefficient S_{20} are the following:

• when $d_j > k_j$, the partial similarity between sites is $s_{12j} = 0$ for species j (see $f(d, k)$, part II);

• when $d_j = 0$, then $s_{12j} = 1$ (see $f(d, k)$, part I), except when $y_{1j} = 0$ or $y_{2j} = 0$ (see $f(d, k)$, part III);

• $f(d, k)$ decreases with increasing d, for a given k;

• $f(d, k)$ increases with increasing k, for a given d;

• when $y_{1j} = 0$ or $y_{2j} = 0$, the partial similarity between sites is $s_{12j} = 0$ for species j, even if d_{12j} is not larger than k_j (see $f(d, k)$, part III);

• when $k_j = 0$ for all species j, S_{20} is the same as the Jaccard coefficient (S_7) for multi-state descriptors.

The above properties correspond to the opinion that ecologists may have on the problem of partial similarities between normalized (or at least not strongly skewed) species abundances. Depending on the scale chosen (0 to 5 or 0 to 50, for example), function $f(d, k)$ can be used to contrast to various degrees the differences between species abundances, by increasing or decreasing k_j, for each species j if necessary. An example of clustering using this measure of similarity is presented in Ecological application 8.2.

The last quantitative coefficient that excludes double-zeros is called the χ^2 *similarity*. It is the complement of the χ^2 metric (D_{15}; Section 7.4):

$$S_{21}(\mathbf{x}_1, \mathbf{x}_2) = 1 - D_{15}(\mathbf{x}_1, \mathbf{x}_2) \qquad \textbf{(7.28)}$$

The discussion of how species that are absent from two sites are excluded from the calculation of this coefficient is deferred to the presentation of D_{15}.

5 — Probabilistic coefficients

Probabilistic measures form a special category. These coefficients are based on statistical estimation of the significance of the relationship between objects.

Goodall's probabilistic coefficient (1964, 1966a) takes into account the frequency distribution of the various states of each descriptor in the whole set of objects. Indeed, it is less likely for two sites to both contain the same rare species than a more frequent species. In this sense, when estimating the similarity between sites, agreement for a rare species should be given more importance than for a frequent species. Goodall's probabilistic index, which had been originally developed for taxonomy, seems

especially meaningful for ecological classifications, because abundances of species in different sites are stochastic functions (Sneath & Sokal, 1973: 141). Orlóci (1978) suggested to use it for clustering sites (Q mode). The index has also been used in the R mode, for clustering species and identifying associations (Subsection 7.5.2).

The probabilistic coefficient of Goodall is based on the probabilities of the various states of each descriptor. The resulting measure of similarity is itself a probability, namely the complement of the probability that the resemblance between two sites is due to chance.

The probabilistic index, as formulated by Goodall (1966a), is a general taxonomic measure in which binary and quantitative descriptors can be used together. The coefficient as presented here follows the modifications of Orlóci (1978) and is limited to the clustering of sites based on species abundances. It also takes into account the remarks made at the beginning of Subsection 7.2.2 concerning double-zeros. The resulting measure is therefore a simplification of Goodall's original coefficient, oriented towards the clustering of sites. The computational steps are as follows:

(a) A partial similarity measure s_j is first calculated for all pairs of sites and for each species j. Because there are n sites, the number of partial similarities s_j to compute, for each species, is $n(n-1)/2$. If the species abundances have been normalized, one may choose either the partial similarity measure $s_{12j} = 1 - [|y_{1j} - y_{2j}| / R_j]$ from Gower's S_{19} coefficient or function s_{12j} from coefficient S_{20}, which were both described above. In all cases, double-zeros must be excluded. This is done by multiplying the partial similarities s_j by Kronecker delta w_{12j}, whose value is 0 upon occurrence of a double-zero. For raw species abundance data, Steinhaus' similarity S_{17}, computed for a single species at a time, may be used as the partial similarity measure. The chord and Hellinger distances, D_3 and D_{17}, could also be used. The outcome of this first step is a partial similarity matrix, containing as many *rows* as there are species in the ecological data matrix (p) and $n(n-1)/2$ *columns*, i.e. one column for each pair of sites; see the numerical example below.

(b) In a second table of the same size, for each species j and each of the $n(n-1)/2$ pairs of sites, one computes the proportion of partial similarity values belonging to species j that are larger than or equal to the partial similarity of the pair of sites being considered; the s_j value under consideration is itself included in the calculation of the proportion. The larger the proportion, the less similar are the two sites with regard to the given species.

(c) The above proportions or probabilities are combined into a site × site similarity matrix, using Fisher's method, i.e. by computing the product Π of the probabilities relative to the various species. Since none of the probabilities is equal to 0, there is no problem in combining these values, but one must assume that the probabilities of the different species are independent vectors. If there are correlations among species, one may use, instead of the original descriptors of species abundance (Orlóci, 1978: 62), a

matrix of component scores from a correspondence or principal coordinate analysis of the original species abundance data (Sections 9.2 and 9.3).

(d) There are two ways to define Goodall's similarity index. In the first approach, the products Π are put in increasing order. Following this, the similarity between two sites is calculated as the proportion of products that are larger than or equal to the product for the pair of sites considered:

$$S_{22}(\mathbf{x}_1, \mathbf{x}_2) = \frac{\sum_{pairs} \mathrm{d}}{n(n-1)/2} \text{ where } \begin{cases} \mathrm{d} = 1 \text{ if } \Pi \geq \Pi_{12} \\ \mathrm{d} = 0 \text{ if } \Pi < \Pi_{12} \end{cases} \quad \textbf{(7.29)}$$

(e) In the second approach, the χ^2 value corresponding to each product is computed, under the hypothesis of independence of the products:

$$\chi^2_{12} = -2 \log_e \Pi_{12}$$

This χ^2-statistic has $2p$ degrees of freedom (p is the number of species). The similarity index is the complement of the probability associated with this χ^2, i.e. the complement of the probability that a χ^2 value taken at random exceeds the observed χ^2 value:

$$S_{23}(\mathbf{x}_1, \mathbf{x}_2) = 1 - \text{prob}\ (\chi^2_{12}) \quad \textbf{(7.30)}$$

It should be clear to the reader that the value of Goodall's index for a given pair of sites may vary depending on the sites included in the computation, since it is based on the rank of the partial similarity for that pair of sites among all pairs. This makes Goodall's measure different from the other coefficients discussed so far.

The following numerical example illustrates the computation of Goodall's index. In this example, five ponds are characterized by the abundances of eight zooplankton species. Data are on a scale of relative abundances, from 0 to 5 (data from Legendre & Chodorowski, 1977).

Species	Ponds					Range R_j
	212	214	233	431	432	
1	3	3	0	0	0	3
2	0	0	2	2	0	2
3	0	2	3	0	2	3
4	0	0	4	3	3	4
5	4	4	0	0	0	4
6	0	2	0	3	3	3
7	0	0	0	1	2	2
8	3	3	0	0	0	3

(a) Gower's matrix of partial similarities has 8 rows and $n(n-1)/2 = 10$ columns which correspond to the 10 pairs of ponds:

Species	Pairs of ponds									
	212	212	212	212	214	214	214	233	233	431
	214	233	431	432	233	431	432	431	432	432
1	1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	1	0	0
3	0.33	0	0	0.33	0.67	0.33	1	0	0.67	0.33
4	0	0	0.25	0.25	0	0.25	0.25	0.75	0.75	1
5	1	0	0	0	0	0	0	0	0	0
6	0.33	0	0	0	0.33	0.67	0.67	0	0	1
7	0	0	0.50	0	0	0.50	0	0.50	0	0.50
8	1	0	0	0	0	0	0	0	0	0

(b) In the next table, one computes, for each pair of sites and each row (species), the proportion of partial similarity values in the row that are larger than or equal to the partial similarity of the pair of sites being considered. The value under consideration is itself included in the proportion. For example, for the pair of ponds (214, 233), the third species has a similarity of 0.67. In the third row, there are 3 values out of 10 that are larger than or equal to 0.67. Thus the ratio associated with the pair (214, 233) in the table is 0.3.

Species	Pairs of ponds									
	212	212	212	212	214	214	214	233	233	431
	214	233	431	432	233	431	432	431	432	432
1	0.1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	0.1	1	1
3	0.7	1	1	0.7	0.3	0.7	0.1	1	0.3	0.7
4	1	1	0.7	0.7	1	0.7	0.7	0.3	0.3	0.1
5	0.1	1	1	1	1	1	1	1	1	1
6	0.5	1	1	1	0.5	0.3	0.3	1	1	0.1
7	1	1	0.4	1	1	0.4	1	0.4	1	0.4
8	0.1	1	1	1	1	1	1	1	1	1

(c) The next table is a site × site symmetric matrix, in which are recorded the products of the terms in each column of the previous table:

Ponds	Ponds				
	212	214	233	431	432
212	–				
214	0.00035	–			
233	1.00000	0.15000	–		
431	0.28000	0.05880	0.01200	–	
432	0.49000	0.02100	0.09000	0.00280	–

(d) The first method for computing the similarity consists in entering, in a site × site matrix, the proportions of the above products that are larger than or equal to the product corresponding to each pair of sites. For example, the product corresponding to pair (212, 431) is 0.28. In the table, there are 3 values out of 10 that are larger than or equal to 0.28, hence the similarity S_{22} (212, 431) = 0.3 (eq. 7.29).

Ponds	Ponds				
	212	214	233	431	432
212	–				
214	1.0	–			
233	0.1	0.4	–		
431	0.3	0.6	0.8	–	
432	0.2	0.7	0.5	0.9	–

(e) If the chosen similarity measure is the complement of the probability associated with χ^2 (eq. 7.30), the following table is obtained. For example, to determine the similarity for pair (212, 431), the first step is to compute χ^2 (212, 431) = $-2 \log_e(0.28) = 2.5459$, where 0.28 is the product associated with that pair in the table at step (d). The value of χ^2 (212, 431) is 2.5459 and the number of degrees of freedom is $2p = 16$, so that the corresponding probability is 0.9994. The similarity is the complement of this probability: S_{23} (212, 431) = 1 – 0.99994 = 0.00006.

Ponds	Ponds				
	212	214	233	431	432
212	–				
214	0.54110	–			
233	0.00000	0.00079	–		
431	0.00006	0.00869	0.08037	–	
432	0.00000	0.04340	0.00340	0.23942	–

Even though the values in the last two tables are very different, the differences are only in term of scale; measures S_{22} computed with eq. 7.29 and S_{23} computed with eq. 7.30 are monotonic to each other.

A probabilistic similarity coefficient among sites has been proposed by palaeontologists Raup & Crick (1979) for species presence-absence data; this is the type of data usually favoured in palaeoecology. Consider the number of species in common to sites h and i; this is statistic a_{hi} of the binary coefficients of Section 7.3. The null hypothesis of the test is H_0: there is no association between sites h and i. Two variants of that null hypothesis are described in steps 2a and 2b below.

Permutation test

The association between sites, measured by a_{hi}, is tested using permutations, and the p-value is used as a distance coefficient. There are two ways of testing the significance of statistic a_{hi} depending on the precise null hypothesis one wants to use.

1. Compute the value of the number of species in common, a_{hi}, for each pair of sites h and i. This is the reference value of the statistic used in step 3. Then go to permutation method 2a or 2b.

2a. The first method, which is actually a simulation rather than a permutation method, implements the null hypothesis (H_0) that each site has received a random subset of species from the species pool, which is either the regional pool or the set of species found in a whole sediment core, while preserving the original species richness at each site. Raup & Crick (1979) called this formulation of H_0 the *random sprinkling hypothesis*. Calculate the relative frequency of each species in the whole data matrix $\mathbf{Y}$, which represents the regional or whole-core species pool. These values will be used as species weights during permutations. Consider site $\mathbf{x}_1$, with species richness s_1. To construct a vector $\mathbf{x}^*_1$ under permutation, draw s_1 species at random from the regional species pool as follows, taking the computed species weights into account.

Random sprinkling hypothesis

- Imagine a stick of length 1 that represents the sum of the weights of all species in the regional pool. For example, species 1 may be very abundant in the region and occupy the first 10% of the stick. Species 2, which is rare, may occupy the next 0.2% of the stick. And so on.

- Draw a number at random from a uniform distribution in the [0, 1] interval. Find the species whose range includes that value on the stick. This is the first species selected in vector $\mathbf{x}^*_1$.

- Remove that species from the stick and rescale the remaining species so that they fully occupy the [0, 1] interval. Draw a new random number from the uniform distribution in the [0, 1] interval. The position of that number along the stick identifies the second species in vector $\mathbf{x}^*_1$.

- Repeat the species selection process until s_1 species have been selected at random from the regional species pool. That completes the construction of vector $\mathbf{x}^*_1$.

Repeat the random species selection process for each site, creating site vectors $\mathbf{x}^*_2$, $\mathbf{x}^*_3$, …, $\mathbf{x}^*_n$ under permutation. Compute the number of species in common, a^*_{hi}, for each pair of site vectors under permutation.

2b. The second permutation method implements the null hypothesis (H_0) that each species is distributed at random among the sites. A permutation under this hypothesis is obtained by permuting at random each vector of species occurrences (i.e. each column of **Y**) independently of the other species vectors. The number of occurrences of each species in the data set is preserved during permutations, but not the original species richness at each site (which is a measure of alpha diversity, Subsection 6.5.3). Compute the number of species in common, a^*_{hi}, for each pair of sites under permutation. This variant of the permutation method for the a_{hi}-statistic was described in McCoy *et al*. (1986).

3. Repeat step 2 (a or b) a large number of times, e.g. 999 or 9999 times, to obtain the null distribution of a^*_{hi}. Add the reference value a_{hi} to the distribution, i.e. the Hope (1968) correction, which is used here to agree with the description of permutation tests in Subsection 1.2.2.

4. For each pair of sites, compare a_{hi} to the reference distribution (obtained at step 3) and calculate the probability $p(a_{hi})$ that $a^*_{hi} \geq a_{hi}$ (one-tailed test), using the procedure described in Subsection 1.2.2.

The Raup-Crick coefficient is available in distance form in function ***raupcrick()*** of VEGAN. The p-values can be computed in several ways, including method 2a above.

Numerical simulations conducted while writing this chapter to check the type I error of the two variants of the Raup-Crick test described above showed that permutation method 2a produced tests that had extremely low levels of type I error, especially when the species had unequal probabilities of occurrence in the species pool. This resulted in a great loss of power when testing the association between sites, to the point that it made the test useless for the analysis of real data because it very seldom recognized significant site associations. Permutation method 2b produced tests that still had low levels of type I error, but not as low as with method 2a, also resulting in a test that had low power to detect significant associations of pairs of sites. As a result, the Raup-Crick test does not seem useful as a test of significance of the similarity of pairs of sites. Ecologists may, however, use the similarity or distance coefficients obtained from that test as they use any other resemblance coefficient among sites, i.e. as the basis for clustering or ordination, without giving them a strict significance test interpretation.

When the test is conducted in the upper tail of the distribution of a^*_{hi} (step 3), the probability $p(a_{hi})$ is expected to be near 0 for sites h and i showing high association, i.e. with more species in common than expected under the null hypothesis. A value near 0.5 indicates that the data support the null hypothesis. One could also test in the lower tail of the distribution, looking for pairs of sites that are significantly dissimilar. The probability would then be calculated as follows: $p(a^*_{hi} \leq a_{hi})$. Significantly dissimilar sites would suggest that some process may have influenced the selection of species, so that fewer species are common to the two sites than expected under the null hypothesis. Taking this approach one step further, Chase *et al*. (2011) rescaled the p-

value in the interval [–1, 1] by subtracting 0.5 and multiplying the result by 2. This modified index indicates whether local communities are more dissimilar (approaching 1), as dissimilar (approaching 0), or less dissimilar (approaching –1) than expected by chance, providing some indication of the possible underlying mechanisms of community assembly.

The probability computed in the upper tail of the distribution of a_{hi}^{*} behaves like a distance (Section 7.4) since p-values are small for similar sites. If necessary, the p-value can be turned into a probabilistic similarity measure of association between sites $\mathbf{x}_1$ and $\mathbf{x}_2$ as follows:

$$S_{27}(\mathbf{x}_1, \mathbf{x}_2) = 1 - \mathrm{p}(a_{12}) \qquad \textbf{(7.31)}$$

Vellend (2004) and Vellend *et al*. (2007) provided a new description of the Raup & Crick (1979) permutation method (paragraph 2a above) and used the coefficient to analyse forest plant communities.

7.4 Q mode: distance coefficients

Distance coefficients are functions that take their maximum values for two objects that are entirely different, and value 0 for two objects that are identical over all descriptors. Distances, like similarities, (Section 7.3), are used to measure the association between *objects*. Distance coefficients may be subdivided in three groups. The first group consists of *metrics* which share the following four properties:

Metric properties

1. minimum 0: if $a = b$, then $D(a, b) = 0$;

2. positiveness: if $a \neq b$, then $D(a, b) > 0$;

3. symmetry: $D(a, b) = D(b, a)$;

4. triangle inequality: $D(a, b) + D(b, c) \geq D(a, c)$. The sum of two sides of a triangle drawn in Euclidean space is necessarily equal to or larger than the third side.

Some authors prefer to restrict the use of *distance* to those coefficients that satisfy the four metric properties and use *dissimilarity* as the general term for all coefficients, i.e. metric, semimetric and nonmetric; see the following paragraphs.

The second group of distances are the *semimetrics* (or *pseudometrics*). These coefficients do not obey the triangle inequality, which is a theorem in Euclidean geometry. These measures cannot directly be used to order points in a *metric* or *Euclidean space* because, for three points (a, b and c), the sum of the distances from a to b and from b to c may be smaller than the distance between a and c. Numerical examples are given in Subsection 7.4.2.

Table 7.2 Some properties of distance coefficients calculated from the similarity coefficients presented in Section 7.3. These properties (from Gower & Legendre, 1986), which will be used in Section 9.3, strictly apply when there are no missing data.

Similarity coefficient	$D = 1 - S$ metric, etc.	$D = 1 - S$ Euclidean	$D = \sqrt{1 - S}$ metric	$D = \sqrt{1 - S}$ Euclidean
$S_1 = \frac{a+d}{a+b+c+d}$ (simple matching; eq. 7.1)	metric	No	Yes	Yes
$S_2 = \frac{a+d}{a+2b+2c+d}$ (Rogers & Tanimoto; eq. 7.2)	metric	No	Yes	Yes
$S_3 = \frac{2a+2d}{2a+b+c+2d}$ (eq. 7.3)	semimetric	No	Yes	No
$S_4 = \frac{a+d}{b+c}$ (eq. 7.4)	nonmetric	No	No	No
$S_5 = \frac{1}{4}\left[\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{b+d} + \frac{d}{c+d}\right]$ (eq. 7.5)	semimetric	No	No	No
$S_6 = \frac{a}{\sqrt{(a+b)\,(a+c)}} \frac{d}{\sqrt{(b+d)\,(c+d)}}$ (eq. 7.6)	semimetric	No	Yes	Yes
$S_7 = \frac{a}{a+b+c}$ (Jaccard; eq. 7.10)	metric	No	Yes	Yes
$S_8 = \frac{2a}{2a+b+c}$ (Sørensen; eq. 7.11)	semimetric	No	Yes	Yes
$S_9 = \frac{3a}{3a+b+c}$ (eq. 7.12)	semimetric	No	No	No
$S_{10} = \frac{a}{a+2b+2c}$ (eq. 7.13)	metric	No	Yes	Yes
$S_{11} = \frac{a}{a+b+c+d}$ (Russell & Rao; eq. 7.14)	metric	No	Yes	Yes
$S_{12} = \frac{a}{b+c}$ (Kulczynski; eq. 7.15)	nonmetric	No	No	No

The third group of distances consists of *nonmetrics*. These coefficients may take negative values, thus violating the property of positiveness of metrics. Only two such coefficients are described in this book: S_4 and S_{12}.

All similarity coefficient from Section 7.3 can be transformed into distances, as mentioned in Section 7.2. The metric and Euclidean properties of distance coefficients resulting from the transformations $D = (1 - S)$ and $D = \sqrt{1 - S}$ are shown in Table 7.2. These properties determine how to use them in principal coordinate analysis (PCoA, Section 9.3). Stating that a distance coefficient is *not* metric or Euclidean actually means that the coefficient is, sometimes or often, not metric or Euclidean; it does not mean that the coefficient is never metric or Euclidean. A coefficient is *likely* to be metric or Euclidean when the binary form of the coefficient, whose code name given in the table, is known by the proof of a theorem to be metric or Euclidean, and

Table 7.2 Continued.

Similarity coefficient	$D = 1 - S$ metric, etc.	$D = 1 - S$ Euclidean	$D = \sqrt{1 - S}$ metric	$D = \sqrt{1 - S}$ Euclidean
$S_{13} = \frac{1}{2}\left[\frac{a}{a+b} + \frac{a}{a+c}\right]$ (eq. 7.16)	semimetric	No	No	No
$S_{14} = \frac{a}{\sqrt{(a+b)\,(a+c)}}$ (Ochiai; eq. 7.17)	semimetric	No	Yes	Yes
$S_{15} = \sum w_j s_j / \sum w_j$ (Gower; eq. 7.21)	metric	No	Yes	Likely* (S_1)
$S_{16} = \sum w_j s_j / \sum w_j$ (Estabrook & Rogers; eq. 7.22)	metric	No	Yes	Likely* (S_1)
$S_{17} = \frac{2W}{A+B}$ (Steinhaus; eq. 7.24)	semimetric	No	Likely* (S_8)	Likely* (S_8)
$S_{18} = \frac{1}{2}\left[\frac{W}{A} + \frac{W}{B}\right]$ (Kulczynski; eq. 7.25)	semimetric	No	No* (S_{13})	No* (S_{13})
$S_{19} = \sum w_j s_j / \sum w_j$ (Gower; eq. 7.26)	metric	No	Yes	Likely
$S_{20} = \sum w_j s_j / \sum w_j$ (Legendre & Chodorowski; 7.27)	metric	No	Yes	Likely* (S_7)
$S_{21} = 1 - \chi^2\ metric$ (eq. 7.28)	metric	Yes	Yes	Yes
$S_{22} = 2\,(\sum d)\,/\,n\,(n-1)$ (Goodall; eq. 7.29)	semimetric	No	–	–
$S_{23} = 1 - p(\chi^2)$ (Goodall; eq. 7.30)	semimetric	No	–	–
$S_{26} = (a + d/2)\,/\,p$ (Faith, 1983; eq. 7.18)	metric	No	Yes	Yes

* These results follow from the properties of the corresponding binary coefficients (coefficient numbers given), when continuous variables are replaced by binary variables.
– Property unknown for this coefficient.

Euclidean coefficient

test runs using quantitative data have never turned up cases to the contrary. A coefficient is said to be Euclidean if the distances are fully embeddable in Euclidean space; principal coordinate analysis (Section 9.3) of such a distance matrix does not produce negative eigenvalues.

For ordered descriptors, distance functions are described in Subsection 7.4.1, in addition to those derived from similarity coefficients, found in Table 7.2. The metric and Euclidean properties of these distance coefficients are shown in Table 7.3. How to use the various distance coefficients is summarized in Tables 7.4 and 7.5.

Table 7.3 Some properties of the distance coefficients described in Section 7.4.

Distance coefficient	D metric, etc.	D Euclidean	$\sqrt{D}$ metric	$\sqrt{D}$ Euclidean
D_1 (Euclidean distance; eq. 7.32)	metric	Yes	Yes	Yes
D_2 (average distance; eq. 7.34)	metric	Yes	Yes	Yes
D_3 (chord distance; eqs. 7.35, 7.36)	metric	Yes	Yes	Yes
D_4 (geodesic metric; eq. 7.37)	metric	No	Yes	Yes
D_5 (Mahalanobis generalized distance; eq. 7.38)	metric	Yes	Yes	Yes
D_6 (Minkowski metric; eq. 7.43)	metric	*	–	–
D_7 (Manhattan metric; eq. 7.44)	metric	No	Yes	Yes
D_8 (mean character difference; eq. 7.45)	metric	No	Yes	Yes
D_9 (index of association; eqs. 7.47, 7.48)	metric	No	Yes	Yes
D_{10} (Canberra metric; eq. 7.49)	metric	No	Yes	Yes
D_{11} (coefficient of divergence; eq. 7.51)	metric	Yes	Yes	Yes
D_{12} (coefficient of racial likeness; eq. 7.52)	nonmetric	No	No	No
D_{13} (nonmetric coefficient; eq. 7.57)	semimetric	No	Yes	Yes
D_{14} (percentage difference; eq. 7.58)	semimetric	No	Yes	Yes
D_{15} (χ^2 metric; eq. 7.54)	metric	Yes	Yes	Yes
D_{16} (χ^2 distance; eq. 7.55)	metric	Yes	Yes	Yes
D_{17} (Hellinger distance; eq. 7.56	metric	Yes	Yes	Yes
D_{18} (distance between species profiles; eq. 7.53)	metric	Yes	Yes	Yes
D_{19} (modified mean character difference; eq. 7.46)	semimetric	No	No	No

* The result depends on the exponent r.
– Not tested for all exponents r.

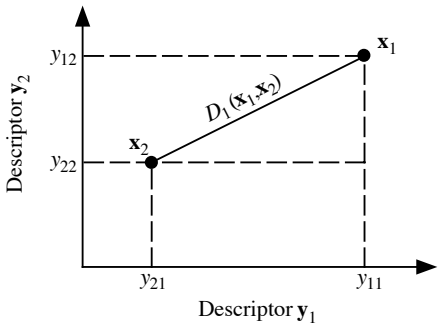


Figure 7.4 Computation of the Euclidean distance (D_1) between objects $\mathbf{x}_1$ and $\mathbf{x}_2$ in 2-dimensional space.

1 — Metric distances

Metric distances have been developed for quantitative descriptors, but they have occasionally been used with semiquantitative descriptors. Some of these measures (D_1, D_2, D_5 to D_8, D_{12}) should not be used, in general, with species abundances, as will be seen in the paradox described below, which results from the handling of double-zeros in the same way as any other value of the descriptors. Coefficients D_3, D_4, D_9 to D_{11} and D_{15} to D_{19} are, on the contrary, well adapted to species abundance data.

The most common metric measure is the *Euclidean distance*. It is computed using Pythagoras' formula from site-points positioned in a p-dimensional space called a *metric* or *Euclidean space*:

$$D_1(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{p} (y_{1j} - y_{2j})^2} \qquad \textbf{(7.32)}$$

When there are only two descriptors, this expression becomes the measure of the hypotenuse of a right-angled triangle (Fig. 7.4; Section 2.4):

$$D_1(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{(y_{11} - y_{21})^2 + (y_{12} - y_{22})^2}$$

The square of D_1 may also be used for clustering purpose. One should notice, however, that D_1^2 is a semimetric, which makes it less appropriate than D_1 for ordination:

$$D_1^2(\mathbf{x}_1, \mathbf{x}_2) = \sum_{j=1}^{p} (y_{1j} - y_{2j})^2 \qquad \textbf{(7.33)}$$

The Euclidean distance does not have an upper limit, its value increasing indefinitely with the number of descriptors. The value also depends on the scale of each descriptor: changing the scale of some or all descriptors changes the values of D_1 in a non-monotonic way. The latter problem may be avoided by using standardized variables (eq. 1.12) instead of the original data, or by restricting the use of D_1 and other distances of the same type (D_2, D_6, D_7 and D_8) to dimensionally homogeneous data matrices (Chapter 3).

Species abundance paradox

The Euclidean distance may lead to the following paradox when it is used as a measure of resemblance among sites based on species abundances: sites without any species in common may be at smaller distance than other sites sharing species. This would be incorrect from an ecologist's point of view. This paradox is illustrated by a numerical example also used in Fig. 7.8 (data modified from Orlóci (1978: 46):

Sites	Species		
	$\mathbf{y}_1$	$\mathbf{y}_2$	$\mathbf{y}_3$
$\mathbf{x}_1$	0	4	8
$\mathbf{x}_2$	0	1	1
$\mathbf{x}_3$	1	0	0

From these data, the following distances are calculated among the sites:

Sites	Sites		
	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
$\mathbf{x}_1$	0	7.6158	9.0000
$\mathbf{x}_2$	7.6158	0	1.7321
$\mathbf{x}_3$	9.0000	1.7321	0

The Euclidean distance between sites $\mathbf{x}_2$ and $\mathbf{x}_3$, which have no species in common, is smaller than the distance between $\mathbf{x}_1$ and $\mathbf{x}_2$ which share species $\mathbf{y}_2$ and $\mathbf{y}_3$. From an ecologist's point of view, this is an incorrect assessment of the relationships among sites. For environmental descriptors on the contrary, double zeros may well be a valid basis for comparing sites. D_1 should therefore not be used for comparing sites based on species abundance data. The main difficulty in ecology concerning the Euclidean distance arises from the fact that frequently used ordination methods, i.e. principal component and redundancy analyses, order objects in the multidimensional space of descriptors using D_1. The ensuing problems are discussed in Sections 7.7 and 9.1.

Various modifications of D_1 have been proposed. First, the effect of the number of descriptors may be tempered by computing an *average distance*:

$$D_2^2(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{p}\sum_{j=1}^{p}(y_{1j} - y_{2j})^2 \text{ or } D_2(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{D_2^2} \qquad \textbf{(7.34)}$$

While it is difficult to show that D_1 is sensitive to double zeros because D_1 has no upper bound, that demonstration is easy for D_2: because of the division by p, D_2 has a maximum value of 1 for presence-absence data. Consider the following example:

	Species												Sum
Object $\mathbf{x}_1$	1	1	0	1	1	1	0	1	0	0	0	0	
Object $\mathbf{x}_2$	0	0	1	1	1	0	1	1	0	0	0	0	
$(y_{1j} - y_{2j})^2$	1	1	1	0	0	1	1	0	0	0	0	0	= 5

With the first 8 columns of the data table ($p = 8$), $D_2^2 = 5/8 = 0.625$ ($D_2 = 0.79057$), whereas with all 12 columns ($p = 12$), $D_2^2 = 5/12 = 0.41667$ ($D_2 = 0.64550$). Adding double zeros has reduced the distance value; this effect would also be demonstrated with abundance data. D_2 is then a symmetrical coefficient in the sense of Subsections 7.2.2 and 7.3.1. This conclusion also applies to D_1.

Orlóci (1967b) proposed to use the *chord distance* to analyse community composition data. That distance, which is also widely used in genetics (Cavalli-Sforza & Edwards, 1967), has a maximum value of $\sqrt{2}$ for sites with no species in common and a minimum of 0 when two sites share the same species *in the same proportions* of the site vector lengths, without it being necessary for these species to be represented by the same *numbers of individuals* at the two sites. This measure is the Euclidean distance computed after scaling the site vectors to length 1 (normalization of a vector, eq. 2.7). After normalization, the Euclidean distance computed between two objects (sites) is equivalent to the length of a chord joining two points within a segment of a sphere or hypersphere of radius 1. If there are only two species involved, the normalization places the sites on the circumference of a 90° sector of a circle with radius 1 (Fig. 7.5). The chord distance may also be computed directly from non-normalized data through the following formulas:

$$D_3(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{2\left(1 - \frac{\sum_{j=1}^{p} y_{1j} y_{2j}}{\sqrt{\sum_{j=1}^{p} y_{1j}^2 \sum_{j=1}^{p} y_{2j}^2}}\right)} = \sqrt{\sum_{j=1}^{p} \left(\frac{y_{1j}}{\sqrt{\sum_{j=1}^{p} y_{1j}^2}} - \frac{y_{2j}}{\sqrt{\sum_{j=1}^{p} y_{2j}^2}}\right)^2} \quad \textbf{(7.35)}$$

The right-hand formula is a modified form of the Euclidean distance formula. The inner part of the left-hand form is the cosine of the angle (θ) between the two site vectors (eq. 2.9). So the chord distance formula can be written:

$$D_3 = \sqrt{2\,(1 - \cos\theta)} \quad \textbf{(7.36)}$$

The chord distance is maximum when the species found at two sites are completely different. In such a case, the normalized site vectors are at 90° from each other on the

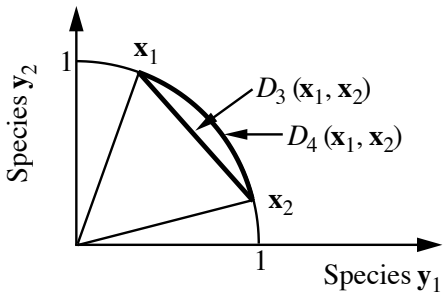


Figure 7.5 Computation of the chord distance D_3 and geodesic metric D_4 between sites $\mathbf{x}_1$ and $\mathbf{x}_2$.

circumference of a 90° sector of a circle (when there are only two species), or on the surface of a segment of a hypersphere (for p species), and the distance between the two sites is $\sqrt{2}$. This measure solves the problem caused by sites having different total abundances of species as well as the paradox explained above for D_1. Indeed, with D_3, the distances between pairs of sites for the numerical example are:

Sites	Sites		
	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
$\mathbf{x}_1$	0	0.3204	1.4142
$\mathbf{x}_2$	0.3204	0	1.4142
$\mathbf{x}_3$	1.4142	1.4142	0

The chord distance is an Euclidean metric since it is computed with the Euclidean distance formula (eq. 7.35 right). Adding any number of double zeros to a pair of sites does not change the value of D_3, which is thus an asymmetrical coefficient in the sense of Subsections 7.2.2 and 7.3.4. Since double zeros do not influence the chord distance, it can be used to compare sites described by species abundances.

A transformation of the previous measure, known as the *geodesic metric*, measures the length of the arc at the surface of the hypersphere of unit radius (Fig. 7.5):

$$D_4\,(\mathbf{x}_1, \mathbf{x}_2) \;=\; \arccos\left[1 - \frac{D_3^2\,(\mathbf{x}_1, \mathbf{x}_2)}{2}\right] \qquad \textbf{(7.37)}$$

In the numerical example, pairs of sites ($\mathbf{x}_1$, $\mathbf{x}_3$) and ($\mathbf{x}_2$, $\mathbf{x}_3$), with no species in common, are at an angle of 90°, whereas sites ($\mathbf{x}_1$, $\mathbf{x}_2$), which share two of the three species, are separated by a smaller angle (18.4°).

Mahalanobis (1936) developed a generalized distance that takes into account the covariances among descriptors; it produces identical results for variables that are standardized or not. This measure computes the distance between two points in a space whose axes are not necessarily orthogonal, in order to take into account the correlations among descriptors. The formula for the *Mahalanobis generalized distance* between two objects $\mathbf{x}_1$ and $\mathbf{x}_2$ from a data table $\mathbf{X}$ is the following:

$$D_5^2(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{d}_{12}\mathbf{V}^{-1}\mathbf{d}'_{12} \quad \textbf{(7.38)}$$

In this equation, $\mathbf{d}_{12}$ is the row vector (length = p) of the absolute value differences between two objects over the p variables and $\mathbf{V}$ is the covariance matrix over all objects in the group (matrix $\mathbf{X}$). The Mahalanobis generalized distance is the square root of $D_5^2(\mathbf{x}_1, \mathbf{x}_2)$. The principal component analysis framework (Section 9.1) will provide a geometric interpretation of Mahalanobis distances among objects (eq. 9.14). The Mahalanobis distance is also the distance preserved among group means in a canonical space of linear discriminant functions (Section 11.3).

The Mahalanobis distance is also used for comparing *groups of objects*, $\mathbf{w}_1$ and $\mathbf{w}_2$, containing n_1 and n_2 objects respectively, that are described by the same p variables. The square of the generalized distance is given by the following formula in that case:

$$D_5^2(\mathbf{w}_1, \mathbf{w}_2) = \overline{\mathbf{d}_{12}}\mathbf{V}^{-1}\overline{\mathbf{d}'_{12}} \quad \textbf{(7.39)}$$

In this equation, $\overline{\mathbf{d}_{12}}$ is the row vector (length = p) of the absolute value differences between the *means* of the p variables in the two groups of objects. $\mathbf{V}$ is the pooled within-group dispersion matrix of the two groups of objects, estimated from the matrices of sums of squares and cross products among descriptors *centred within each of the two groups*, then added up term by term and divided by $(n_1 + n_2 - 2)$, as in discriminant analysis (Table 11.8) and in multivariate analysis of variance:

$$\mathbf{V} = \frac{1}{n_1 + n_2 - 2}[(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2] \quad \textbf{(7.40)}$$

$\mathbf{S}_1$ and $\mathbf{S}_2$ are the dispersion matrices (eq. 4.6) of the two groups, so that $\mathbf{V}$ takes into account the within-group covariances among descriptors. This formula can also be used to calculate the distance between a single object and a group.

If one wishes to test D_5 for significance, the within-group dispersion matrices must be homogeneous (homoscedasticity, Box 1.4). Homoscedasticity of matrices $\mathbf{S}_1$ and $\mathbf{S}_2$ can be tested using Kullback's test (eq. 11.41) or through the multivariate generalization of Levene's test of homogeneity of variances proposed by Anderson (2006); the latter is available in function ***betadisper()*** of VEGAN. The test of significance also assumes multinormality of the within-group distributions (Sections 4.3 and 4.6) although the generalized distance tolerates some degree of deviation from this condition.

To perform the test of significance, the generalized distance is transformed into Hotelling's T^2 (1931) statistic, using the following equation:

$$T^2 = \frac{n_1 n_2}{(n_1 + n_2)} D_5^2 \tag{7.41}$$

The F-statistic is computed as follows:

$$F = \frac{n_1 + n_2 - (p + 1)}{(n_1 + n_2 - 2)\, p} T^2 \tag{7.42}$$

with p and $[n_1 + n_2 - (p + 1)]$ degrees of freedom. Statistic T^2 is a generalization of Student's t-statistic to the multidimensional case. It allows one to test the hypothesis that two groups originate from populations with similar centroids. The final generalization to several groups, called Wilks Λ (lambda), is discussed in Section 11.3 (eq. 11.42).

The Euclidean distance D_1 is the second degree ($r = 2$) of the *Minkowski metric*:

$$D_6(\mathbf{x}_1, \mathbf{x}_2) = \left[\sum_{j=1}^{p} \left| y_{1j} - y_{2j} \right|^r \right]^{1/r} \tag{7.43}$$

Forms of this metric with $r > 2$ are seldom used in ecology because powers higher than 2 give too much importance to the largest differences $\left| y_{1j} - y_{2j} \right|$. For the exact opposite reason, exponent $r = 1$ is used in many instances. The basic form,

$$D_7(\mathbf{x}_1, \mathbf{x}_2) = \sum_{j=1}^{p} \left| y_{1j} - y_{2j} \right| \tag{7.44}$$

is known as the *Manhattan metric*, *taxicab metric*, or *city-block metric*. This refers to the fact that, for two descriptors, the distance between two objects is the distance on the abscissa (descriptor y_1) plus the distance on the ordinate (descriptor y_2), like the distance travelled by a taxicab around blocks in a city with an orthogonal plan like Manhattan. This metric presents the same problem for double-zeros as in the Euclidean distance and thus leads to the same paradox.

The *mean character difference* ("durchschnittliche Differenz", in German), proposed by anthropologist Czekanowski (1909),

$$D_8(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{p} \sum_{j=1}^{p} \left| y_{1j} - y_{2j} \right| \tag{7.45}$$

has the advantage over D_7 of not increasing with the number of descriptors p. Distance D_8 is metric, since it is the Manhattan metric divided by p which is constant for a given

data matrix, but it is not Euclidean. $\sqrt{D_8}$ is, however, metric and Euclidean. When applied to presence-absence data, D_8 becomes the one-complement of the simple matching coefficient $(1 - S_1)$.

Equation 7.45 can be used with species abundances if one modifies it to exclude double-zeros from the calculations. This is done by replacing p by pp, which is the number of pairs of values in site vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ that are not double zeros:

$$D_{19}(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{pp} \sum_{j=1}^{p} |y_{1j} - y_{2j}| \qquad \textbf{(7.46)}$$

For abundance data, this *modified mean character difference* coefficient has no fixed upper limit. Neither D_{19} nor $\sqrt{D_{19}}$ are metric or Euclidean, which limits their use as the basis for principal coordinate ordination (Section 9.3). When applied to species presence-absence data, eq. 7.46 becomes the one-complement of the Jaccard coefficient $(1 - S_7)$, which is metric but not Euclidean, whereas $\sqrt{1 - S_7}$ is both metric and Euclidean (Table 7.2).

Before computing D_{19}, species abundance data must be transformed to reduce their distribution skewness. In that respect, Anderson *et al*. (2006) redescribed distance D_{19} as a modified form of the asymmetrical Gower coefficient (S_{19}). They applied it to abundance data transformed following eq. 7.66 and called this combination the *modified Gower dissimilarity.*

Whittaker's *index of association* (1952) is well adapted to species abundance data, because each species is first transformed into a fraction of the total number of individuals at the site before the subtraction. Empty data rows (e.g. sites), where no species were found, must be excluded from the calculation. The complement of this index is the following distance, which can be seen as a Manhattan-type (D_7) version of the distance between species profiles D_{18}:

$$D_9(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{2} \sum_{j=1}^{p} \left| \frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}} \right| \qquad \textbf{(7.47)}$$

where y_{1+} is the sum of values in row $\mathbf{x}_1$ and y_{2+} is the sum of values in row $\mathbf{x}_2$. The difference is zero for a species when its *relative abundances* are identical at the two sites. An equivalent formula is to compute, over all species, the sum of the smallest relative abundances at the two sites:

$$D_9(\mathbf{x}_1, \mathbf{x}_2) = \left[1 - \sum_{j=1}^{p} \min \left(\frac{y_{1j}}{y_{1+}}, \frac{y_{2j}}{y_{2+}} \right) \right] \qquad \textbf{(7.48)}$$

D_9 takes values between 0 and 1. The metric and Euclidean properties of D_9 were checked over several community composition data tables: D_9 seems to always be

metric, but it is not Euclidean. $\sqrt{D_9}$ seems, however, to always be metric and Euclidean.

The Australians Lance & Williams (1967a) developed several variants of the Manhattan metric, including their *Canberra metric* (Lance & Williams, 1966c):

$$D_{10}(\mathbf{x}_1, \mathbf{x}_2) = \sum_{j=1}^{p} \left[\frac{|y_{1j} - y_{2j}|}{(y_{1j} + y_{2j})} \right] \quad \textbf{(7.49)}$$

which must exclude double-zeros in order to avoid indetermination. This measure has no fixed upper limit. It can be shown that in D_{10}, a given difference between abundant species contributes less to the distance than the same difference between rarer species (Section 7.6). D_{10} is a non-Euclidean metric whereas $\sqrt{D_{10}}$ is both metric and Euclidean.

The Canberra metric is implemented in the ***vegdist()*** function of the VEGAN package with division of D_{10} by pp, which is the number of pairs of values that are not double zeros in the computation of a given D_{10} value. A metric coefficient taking values between 0 and 1 is thus obtained. Like D_{10}, D_{10}/pp is a non-Euclidean metric, which becomes Euclidean when taking its square root. As an ecological *similarity* measure, Stephenson *et al*. (1972) and Moreau & Legendre (1979) used the one-complement of the Canberra metric with division by pp:

$$S(\mathbf{x}_1, \mathbf{x}_2) = 1 - \frac{1}{pp} D_{10} \quad \textbf{(7.50)}$$

Clark's (1952) *coefficient of divergence* is a modification of D_{10} that uses the Euclidean distance formula:

$$D_{11}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\frac{1}{pp} \sum_{j=1}^{p} \left(\frac{y_{1j} - y_{2j}}{y_{1j} + y_{2j}} \right)^2} \quad \textbf{(7.51)}$$

D_{11} is a metric and Euclidean coefficient with a maximum value of 1. Because, in D_{11}, the difference for each descriptor is first expressed as a fraction, before squaring the values and summing them, this coefficient is appropriate for species abundance data. Double-zeros must be excluded from the computation to avoid indetermination. This coefficient was first described for multivariate taxonomic analysis, where division was by the number of characters (p) included in the calculation. For community composition analysis, division must be by pp, the number of non double-zero species included in the calculation, as in eq. 7.50.

Another coefficient related to D_{11} was developed by Pearson (1926) for anthropological studies under the name *coefficient of racial likeness*. Using this coefficient, it is possible to measure a distance between groups of objects, as with the

Mahalanobis generalized distance D_5, but without taking into account the within-group covariances among descriptors. The formula is:

$$D_{12}(\mathbf{w}_1, \mathbf{w}_2) = \sqrt{\frac{1}{p}\sum_{j=1}^{p}\left[\frac{(\bar{y}_{1j} - \bar{y}_{2j})^2}{(s_{1j}^2/n_1) + (s_{2j}^2/n_2)}\right] - \frac{2}{p}} \qquad \textbf{(7.52)}$$

for two groups of objects $\mathbf{w}_1$ and $\mathbf{w}_2$ containing respectively n_1 and n_2 objects; $\bar{y}_{ij}$ is the mean of descriptor j in group i and s_{ij}^2 is the corresponding variance. Equation 7.52 can produce negative distances when the square-root portion of the equation is smaller than $2/p$, so the coefficient is not a metric in that particular case.

Other metrics are available to calculate the distance among sites using species abundances or other types of frequency data; no negative value is allowed in the data. The first of these coefficients is the distance between species profiles. It is the Euclidean distance (D_1) computed between relative frequencies (e.g. species relative abundances computed by rows) in a frequency table. In the following example,

$$\mathbf{Y} = \begin{bmatrix} 45 & 10 & 15 & 0 & 10 \\ 25 & 8 & 10 & 0 & 3 \\ 7 & 15 & 20 & 14 & 12 \end{bmatrix} \begin{bmatrix} y_{i+} \\ 80 \\ 46 \\ 68 \end{bmatrix} \rightarrow \left[y_{ij}/y_{i+}\right] = \begin{bmatrix} 0.563 & 0.125 & 0.188 & 0.000 & 0.125 \\ 0.543 & 0.174 & 0.217 & 0.000 & 0.065 \\ 0.103 & 0.221 & 0.294 & 0.206 & 0.176 \end{bmatrix}$$

y_{i+} is the sum of the frequencies in row i. After transformation of the data into relative frequencies, the distance between pairs of rows of the right-hand matrix is computed using the Euclidean distance formula D_1 (eq. 7.32). The equation of the *distance between species profiles* is then:

$$D_{18}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{p}\left(\frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}}\right)^2} \qquad \textbf{(7.53)}$$

With this equation, the most abundant species contribute predominantly to the sum of squares. This coefficient is asymmetrical, meaning that it is insensitive to double-zeros; the upper limit of D_{18} is $\sqrt{2}$ for sites that have no species in common.

The χ^2 *metric* D_{15} is a weighted form of the distance between species profiles D_{18}. In the calculation of the sum of squares, each squared difference between relative

frequencies of a pair of rows is weighted by the inverse of the frequency of its column j, y_{+j}, computed across the whole table, as shown in the following example:

$$\mathbf{Y} = \begin{bmatrix} 45 & 10 & 15 & 0 & 10 \\ 25 & 8 & 10 & 0 & 3 \\ 7 & 15 & 20 & 14 & 12 \end{bmatrix} \begin{bmatrix} 80 \\ 46 \\ 68 \end{bmatrix} \quad \rightarrow \begin{bmatrix} y_{ij}/y_{i+} \end{bmatrix} = \begin{bmatrix} 0.563 & 0.125 & 0.188 & 0.000 & 0.125 \\ 0.543 & 0.174 & 0.217 & 0.000 & 0.065 \\ 0.103 & 0.221 & 0.294 & 0.206 & 0.176 \end{bmatrix}$$

$$\begin{bmatrix} y_{+j} \end{bmatrix} = \begin{bmatrix} 77 & 33 & 45 & 14 & 25 \end{bmatrix} \quad 194$$

The χ^2 metric is computed using the following equation:

$$D_{15}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{p} \frac{1}{y_{+j}} \left(\frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}} \right)^2} \qquad \textbf{(7.54)}$$

Thus D_{15}, as well as the chi-square distance (D_{16}, next distance), give higher weights to the rare than to the common species in the calculation of the distance. This distance is recommended when the rare species are considered to be good indicators of special ecological conditions.

For the numerical example, computation of D_{15} between the first two sites (rows) gives:

$$D_{15}(\mathbf{x}_1, \mathbf{x}_2) =$$

$$\left[\frac{(0.563 - 0.543)^2}{77} + \frac{(0.125 - 0.174)^2}{33} + \frac{(0.188 - 0.217)^2}{45} + \frac{(0 - 0)^2}{14} + \frac{(0.125 - 0.065)^2}{25} \right]^{1/2}$$

$$= 0.015$$

The fourth species, which is absent from the first two sites, cancels itself out. This is how the χ^2 metric excludes double-zeros from the calculation.

The upper limit of D_{15} is $\sqrt{2}$. This value is only obtained when there is a single species presence (with an abundance of 1) in the sites producing this value and each species has a total abundance of 1 in the data table; there may be, or not, multiple species with abundances of 1 at other sites than those producing values of $D_{15} = \sqrt{2}$. In all other situations, the distances among sites are smaller than $\sqrt{2}$. To avoid indetermination, absent species (with total abundances of 0) must be eliminated from the data table before the coefficient is computed. D_{15} is asymmetrical since it has an upper limit and its value is not affected by double-zeros.

The χ^2 metric D_{15} can be calculated either among the rows or among the columns of a frequency table. If it is computed among the rows (Q-mode analysis), the relative

frequencies y_{ij}/y_{i+} are computed across the values of each object (e.g. site). If it is computed among columns (R-mode analysis), the relative frequencies y_{ij}/y_{+j} are computed across the values of each column (e.g. species), interchanging rows for columns in eq. 7.54. For example, D_{15} was used by Roux & Reyssac (1975) to calculate distances among sites described by species abundances.

A related measure is called the χ^2 *distance* (Lebart & Fénelon, 1971). It differs from the χ^2 metric in that the terms of the sum of squares are divided by the *relative frequency* of each column in the overall table instead of its absolute frequency. In other words, it is identical to the χ^2 metric multiplied by $\sqrt{y_{++}}$ where y_{++} is the sum of all frequencies in the data table:

$$D_{16}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{p} \frac{1}{y_{+j}/y_{++}} \left(\frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}} \right)^2} = \sqrt{y_{++}} \sqrt{\sum_{j=1}^{p} \frac{1}{y_{+j}} \left(\frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}} \right)^2} \quad \textbf{(7.55)}$$

Since D_{16} is simply D_{15} multiplied by a constant, it shares the property of asymmetry of D_{15}. The maximum value of D_{16} is $\sqrt{2y_{++}}$. The χ^2 distance is the distance preserved in correspondence analysis (Section 9.2). More generally, it is used to compute the association between the rows or columns of a frequency table.

The data used above to illustrate the paradox obtained when the Euclidean distance was computed over species abundances are used again here to contrast D_{16} with D_1.

$$\begin{matrix} & & \left[y_{i+} \right] \\ \mathbf{Y} = \begin{bmatrix} 0 & 4 & 8 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \end{bmatrix} & \begin{bmatrix} 12 \\ 2 \\ 1 \end{bmatrix} & \rightarrow \left[y_{ij}/y_{i+} \right] = \begin{bmatrix} 0 & 0.333 & 0.667 \\ 0 & 0.500 & 0.500 \\ 1 & 0 & 0 \end{bmatrix} \end{matrix}$$

$$\left[y_{+j} \right] = \begin{bmatrix} 1 & 5 & 9 \end{bmatrix} 15$$

Computing D_{16} between rows (sites) 1 and 3 gives:

$$D_{16}(\mathbf{x}_1, \mathbf{x}_2) = \left[\frac{(0-1)^2}{1/15} + \frac{(0.333-0)^2}{5/15} + \frac{(0.667-0)^2}{9/15} \right]^{1/2} = 4.0092$$

The distances between all pairs of sites are:

Sites	Sites		
	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
$\mathbf{x}_1$	0	0.3600	4.0092
$\mathbf{x}_2$	0.3600	0	4.0208
$\mathbf{x}_3$	4.0092	4.0208	0

Comparison with results obtained for D_1 (after eq. 7.33) shows that the problem caused with D_1 by the presence of double-zeros does not exist here. Distance D_{16} can therefore be used directly with sites described by species abundances, contrary to D_1.

Another coefficient related to the distance between species profiles (D_{18}) is the Hellinger distance, described by Rao (1995). The formula of the Hellinger distance is:

$$D_{17}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{p} \left[\sqrt{\frac{y_{1j}}{y_{1+}}} - \sqrt{\frac{y_{2j}}{y_{2+}}} \right]^2} \qquad \textbf{(7.56)}$$

Some of its properties are discussed near the end of Section 7.6. Like D_3 and D_{18}, this distance is asymmetrical (i.e. it is insensitive to double-zeros), and its upper limit is $\sqrt{2}$. The Hellinger distance is actually the chord distance D_3 computed on square-root-transformed frequencies (e.g. species abundances). It is highly recommended for clustering or ordination of species abundance data (Prentice, 1980*; Rao, 1995). Rao (1995) recommended this measure as the basis for a new ordination method; one can obtain the same ordination by computing D_{17} among the objects and carrying out principal coordinate analysis (PCoA, Section 9.3) of the resulting distance matrix. When applied to presence-absence data, the chord (D_3) and Hellinger (D_{17}) distances are both related to the Ochiai similarity (S_{14}) as follows:

$$D_3 = D_{17} = \sqrt{2}\sqrt{1 - S_{14}}$$

2 — Semimetrics

Some distance measures do not follow the fourth property of metrics, i.e. the triangle inequality theorem described at the beginning of the present section. As a consequence, they do not allow a proper ordination of sites in a full Euclidean space. They may, however, be used for ordination by principal coordinate analysis after correction for negative eigenvalues (Subsection 9.3.4) or by nonmetric multidimensional scaling (Section 9.4). These measures are called *semimetrics* or *pseudometrics*. Some semimetrics derived from similarities are identified in Table 7.2. Other such measures are presented here.

The distance corresponding to Sørensen's coefficient S_8 was described by Watson *et al*. (1966) under the name *nonmetric coefficient*:

$$D_{13}(\mathbf{x}_1, \mathbf{x}_2) = 1 - \frac{2a}{2a + b + c} = \frac{b + c}{2a + b + c} \qquad \textbf{(7.57)}$$

* Prentice (1980) called D_{17} the "chord distance". He gave D_3 the name "cosine theta distance".

a, b and c were defined at the beginning of Subsection 7.3.1. The following numerical example shows that D_{13} does not obey the triangle inequality theorem:

Sites	Species				
	$\mathbf{y}_1$	$\mathbf{y}_2$	$\mathbf{y}_3$	$\mathbf{y}_4$	$\mathbf{y}_5$
$\mathbf{x}_1$	1	1	1	0	0
$\mathbf{x}_2$	0	0	0	1	1
$\mathbf{x}_3$	1	1	1	1	1

Distances between the three pairs of sites are:

$$D_{13}(\mathbf{x}_1, \mathbf{x}_2) = \frac{3+2}{0+3+2} = 1.00$$

$$D_{13}(\mathbf{x}_1, \mathbf{x}_3) = \frac{0+2}{(2\times 3)+0+2} = 0.25$$

$$D_{13}(\mathbf{x}_2, \mathbf{x}_3) = \frac{0+3}{(2\times 2)+0+3} = 0.43$$

Hence $0.25 + 0.43 < 1.00$, which violates the triangle inequality theorem.

Among the measures for species abundance data, the coefficients of Steinhaus S_{17} and Kulczynski S_{18} are semimetrics when transformed into distances (Table 7.2). In particular, $D_{14} = 1 - S_{17}$ was first described by Odum (1950) in distance form, who called it the *percentage difference*. It was also used by Bray & Curtis (1957) in a study of Wisconsin upland forest vegetation*. The percentage difference formula is:

$$D_{14}(\mathbf{x}_1, \mathbf{x}_2) = \frac{\sum_{j=1}^{p} |y_{1j} - y_{2j}|}{\sum_{j=1}^{p} (y_{1j} + y_{2j})} = 1 - \frac{2W}{A+B} \qquad \textbf{(7.58)}$$

* It is unclear why some computer packages refer to D_{14} as the Bray-Curtis distance. The main objective of the Bray & Curtis (1957) paper was to describe a new ordination method, which is known as the Bray-Curtis ordination. The authors never pretended that they were proposing a new S or D coefficient. They used the similarity form of the D_{14} coefficient (S_{17}, eq. 7.24) in their paper (p. 329), transforming it into a distance (p. 332) before computing their new ordination method. Bray & Curtis referred to Motyka *et al*. (1950) for the origin of this similarity coefficient, which is also described by Motyka (1947) who stated that the formula of coefficient S_{17} had first been proposed by Professor H. Steinhaus. In their study, Bray & Curtis (1957) made a very restricted application of the Steinhaus coefficient: because they analysed relative tree abundances at sampling sites, the quantities A and B were both equal to a constant and the similarity between stands was thus equal to W in their study.

The value of D_{14} does not change when double zeros are added to a pair of sites, because the values of A, B and W remain unchanged. Hence this coefficient is asymmetrical in the sense of Subsection 7.2.2. For species relative abundances $y_{ij} = y_{ij}/y_{i+}$, where y_{i+} is the row sum, $D_{14} = D_9$ where D_9 is Whittaker's index of association; D_{14} is also equal to $D_7/2$ where D_7 is the Manhattan distance. D_9 and D_7 are both metric but not Euclidean, but $\sqrt{D_9}$ and $\sqrt{D_7}$ are Euclidean. For binary data, D_{14} is equal to D_{13} or $(1 - S_8)$, which is neither metric nor Euclidean, but $\sqrt{D_{13}}$ is Euclidean.

Contrary to the Canberra metric D_{10}, differences between abundant species contribute the same to D_{14} as differences between rare species. This may be seen as a desirable property, for instance when using normalized species abundance data. Bloom (1981) compared the Canberra metric, the percentage difference and other indices to a theoretical standard. For these data, he showed that only D_{14} (or S_{17}) accurately reflected the true resemblance along its entire 0 to 1 scale, whereas D_{10}, for example, underestimated the resemblance over much of its 0 to 1 range.

The following numerical example, from Orlóci (1978: 59), shows that D_{14} does not obey the triangle inequality theorem and is thus not a metric distance:

Quadrats	Species				
	$\mathbf{y}_1$	$\mathbf{y}_2$	$\mathbf{y}_3$	$\mathbf{y}_4$	$\mathbf{y}_5$
$\mathbf{x}_1$	2	5	2	5	3
$\mathbf{x}_2$	3	5	2	4	3
$\mathbf{x}_3$	9	1	1	1	1

The distances between the three pairs of sites are:

$$D_{14}(\mathbf{x}_1, \mathbf{x}_2) = \frac{1 + 0 + 0 + 1 + 0}{17 + 17} = 0.059$$

$$D_{14}(\mathbf{x}_1, \mathbf{x}_3) = \frac{7 + 4 + 1 + 4 + 2}{17 + 13} = 0.600$$

$$D_{14}(\mathbf{x}_2, \mathbf{x}_3) = \frac{6 + 4 + 1 + 3 + 2}{17 + 13} = 0.533$$

hence $0.059 + 0.533 < 0.600$, which violates the triangle inequality theorem. Coefficient D_{14} is thus not a metric distance. Table 7.3 shows that D_{14}, which is equal to $(1 - S_{17})$, is also not Euclidean. The consequence is that ordination of sites by principal coordinate analysis (PCoA, Section 9.3) based upon D_{14} is likely to produce negative eigenvalues and complex axes. The way to obtain a distance matrix that is both metric and Euclidean before PCoA is to take the square root of D_{14} (Table 7.3).

7.5 R mode: coefficients of dependence

The main purpose of R-mode analysis is to investigate the relationships among *descriptors*; dependence (R-mode) matrices may also be used, in some cases, as the computational basis for the ordination of *objects*, e.g. in principal component or linear discriminant analyses (Sections 9.1 and 11.3). Following the classification of descriptors in Table 1.2, dependence coefficients will be described for quantitative, semiquantitative, and qualitative descriptors. This will be followed by special measures to assess the dependence between species, to be used for the identification of biological associations (Section 8.9).

Most dependence coefficients are amenable to *statistical testing*. For such coefficients, it is thus possible to associate a matrix of probabilities with the dependence matrix, if required by subsequent analyses. While it is not always legitimate to apply statistical tests of significance, it is never incorrect to compute a dependence coefficient among variables. For example, there is no objection to computing a Pearson correlation coefficient for any pair of metric variables, but these same variables must be normally distributed (Sections 4.2 and 4.3) and the sites must be independent realizations (Sections 1.1 and 1.2) to legitimately test the significance of the coefficient using the standard parametric test; a permutation test (Section 1.2) can, however, be used with non-normal data. A test of significance only allows one to reject, or not, a specific null hypothesis concerning the value of the statistic (here, the coefficient of dependence), whereas the coefficient itself measures the intensity of the relationship between descriptors. Table 7.6 summarizes the use of R-mode coefficients with ecological variables.

1 — Descriptors other than species abundances

Measures of resemblance in the present subsection, which summarizes the coefficients described in Chapters 4, 5 and 6, are used for comparing physical, chemical, geological, and other environmental variables. Measures adapted for species presence-absence and abundance data are described in the next subsection.

The resemblance between *quantitative descriptors* can be computed using parametric measures of dependence, i.e. measures based on parameters of the frequency distributions of the descriptors. These measures are the covariance and the Pearson correlation coefficient; they were described in Chapter 4. They are only adapted to descriptors whose relationships are *linear*.

The *covariance* s_{jk} between descriptors j and k is computed from centred variables $(y_{ij} - \overline{y_j})$ and $(y_{ik} - \overline{y_k})$ (eq. 4.4). The range of values of the covariance has no *a priori* upper or lower limits. The variances and covariances among a group of descriptors form their dispersion matrix **S** (eq. 4.6).

The Pearson *correlation coefficient* r_{jk} is the covariance of standardized descriptors j and k (eqs. 1.12 and 4.7). Coefficients of correlations computed among a group of descriptors form a correlation matrix **R** (eq. 4.8). Correlation coefficients range in values between –1 and +1. The significance of individual coefficients is tested using eq. 4.13, the null hypothesis being generally H_0: $r = 0$, whereas eq. 4.14 is used to test the hypothesis of complete independence among all descriptors. Pearson correlation coefficients should not be computed in Q mode (Box 7.1).

The resemblance between *semiquantitative descriptors* and, more generally between any pair of *ordered* descriptors whose relationship is *monotonic* may be determined using nonparametric measures of dependence (Chapter 5). Since *quantitative* descriptors are ordered, nonparametric coefficients may be used to measure their dependence, as long as they are monotonically related.

Two *nonparametric correlation coefficients* have been described in Section 5.3: Spearman's r and Kendall's τ (tau). In Spearman's r (eq. 5.3), quantitative values are replaced by ranks before computing Pearson's r formula. Kendall's τ (eqs. 5.5 to 5.7) measures the resemblance in a way that is quite different from Pearson's r. Values of Spearman's r and Kendall's τ range between –1 and +1. The significance of individual coefficients (the null hypothesis being generally H_0: $r = 0$) is tested using eq. 5.4 (Spearman's r) or 5.8 (Kendall's τ).

As with Pearson's r above, rank correlation coefficients should not be used in the Q mode. Indeed, even if quantitative descriptors are standardized, the same problem arises as with Pearson's r, i.e. the Q measure for a pair of objects is a function of all objects in the data set. In addition, in most biological sampling units, several species are represented by small numbers of individuals. Because these small numbers are subject to large stochastic variation, the ranks of the corresponding species are uncertain in the reference ecosystem. As a consequence, rank correlations between sites would be subject to important random variation because their values would be based on large numbers of uncertain ranks. This is equivalent to giving preponderant weight to the many poorly sampled species.

The importance of *qualitative descriptors* in ecological research is discussed in Section 6.0. The measurement of resemblance between pairs of such descriptors is based on two-way contingency tables (Section 6.2), whose analysis is generally conducted using X^2 (chi-square) statistics. Contingency table analysis is also the major approach available for measuring the dependence between *quantitative* or *semiquantitative* ordered descriptors that are not monotonically related. The minimum value of X^2 is zero, but it has no *a priori* upper limit. Its formulae (eqs. 6.5 and 6.6) and test of significance are explained in Section 6.2. If all qualitative descriptors have the same number of states, X^2 values can be transformed into contingency coefficients (eqs. 6.19 and 6.20), whose values are in the range [0, 1].

Two-way contingency tables may also be analysed using coefficients derived from information theory. In that case, the amounts of information (B) shared by two

Q-mode correlation Box 7.1

Can one compute Pearson correlation coefficients among rows, i.e. in Q mode? There are at least six objections to that.

• Because the same physical dimensions are present in the numerator and denominator of Pearson's r computed in the R mode (eq. 4.7), the resulting coefficient has no physical dimension, i.e. it is dimensionless (Chapter 3). On the contrary, correlations computed between objects (Q mode) have complex and non-interpretable physical dimensions when the descriptors are not dimensionally homogeneous. Furthermore, in Q mode, the row means $\bar{y}_i$ do not make sense for variables that have different physical dimensions so that the differences $(y_{ij} - \bar{y}_i)$ in eq. 4.7 cannot be computed.

• Physical descriptors are often expressed in arbitrary units (e.g. mm, cm, m, or km are all equally correct length measures). In R mode, the value of r remains unchanged after any arbitrary linear change of units, whereas in Q mode the same operation can dramatically change the values of correlations computed between objects, in unpredictable fashion.

• In order to avoid the two previous problems, it has been suggested to standardize the descriptors (eq. 1.12) before computing correlations in the Q mode. Consider two objects $\mathbf{x}_1$ and $\mathbf{x}_2$: their similarity should be independent of the other objects in the study; removing objects from the data set should not change the value of their similarity. Any change in object composition of the data set would, however, change the standardized variables, and so it would affect the value of the correlation computed between $\mathbf{x}_1$ and $\mathbf{x}_2$. Hence, standardization does not solve the problems because the resulting correlation between two objects would be a function of the values of all the other objects in the data set.

• In the R mode, the central limit theorem (Section 4.3) predicts that, as the number of objects increases, the means, variances, and covariances (or correlations) converge towards their values in the statistical population. Computing these same parameters in the Q mode is likely to have the opposite effect since the addition of new descriptors to the estimation of these parameters is likely to change their values in major and non-trivial ways.

• If correlation coefficients could be used as a general measure of resemblance in Q mode, they should be applicable to the simple case of the description of the proximities among sites, computed from their geographic coordinates X and Y on a map; the correlations obtained from this calculation should reflect in some way the distances among the sites. This is not the case: correlation coefficients computed among sites from their geographic coordinates are all +1 or –1. As an exercise, readers are encouraged to compute an example of their own.

• Correlation coefficients can be tested by the method of permutations, as shown in Subsection 1.2.3. In the R mode, permuting the values of a variable within a column makes physical sense: under H_0, each value could be found at any one site. In the Q mode, however, permuting values within a row of the data matrix does not make sense because, in the real world, these values could not belong to different variables. As an illustration, it would not make sense to move a salinity of 35 psu to the pH column.

Conclusion: coefficients designed for R-mode analysis should not be used in the Q mode. Sections 7.3 and 7.4 describe several Q-mode coefficients whose properties and dimensions are known or easy to determine.

descriptors j and k and exclusive to each one (A and C) are first computed. These quantities may be combined into similarity measures, such as $S(j, k) = B/(A + B + C)$ (eq. 6.15; see also eqs. 6.17 and 6.18), or into distance coefficients such as $D(j, k) = (A + C)/(A + B + C)$ (eq. 6.16). The analysis of multiway contingency tables (Section 6.3) is based on the Wilks X^2-statistic (eq. 6.6).

A *qualitative* descriptor (including a *classification*; Chapter 8) can be compared to a *quantitative* descriptor using the F-statistic of *one-way analysis of variance* (one-way ANOVA; Table 5.2 and text). The classification criterion for this ANOVA is the qualitative descriptor. As long as the assumptions underlying analysis of variance are met (i.e. normality of within-group distributions and homoscedasticity, Box 1.4), the significance of the relationship between the descriptors can be tested using the F–distribution. If the quantitative descriptor does not obey the within-group normality assumption, a permutation test of F can be used. If the comparison is between a *qualitative* and a *semiquantitative* descriptor, *nonparametric one-way analysis of variance* (Kruskal-Wallis H-test; Table 5.2) can be used.

2 — Species abundances: biological associations

The search for species associations is one of the classical problems of community ecology. How to conduct that search using clustering methods is discussed in Section 8.9. The present subsection focuses on the dependence coefficients that are appropriate for the study of species interrelationships. Measures that can be used for presence-absence data are discussed first, followed by measures for quantitative data.

1. *Species presence-absence data*. — There are several approaches in the literature for measuring the association between species based on presence-absence data. Indeed, biological associations may be defined on the basis of the *co-occurrence of species*, instead of the co-fluctuations in abundances in the quantitative approaches described below. Indeed, the definition of association may refer to the sole concept of co-occurrence, as suggested by Fager (1957) who pointed out that associations must group species that are almost always part of one another's biological environment. The reason is that quantitative data may not accurately reflect the proportions of the various species in the environment, because of problems with sampling, preservation, identification or counting, or simply because the concept of individuality is not clear (e.g. plants multiplying through rhizomes; colonial algae or animals), or because the comparison of numbers of individuals does not make ecological sense (e.g. the baobab and the surrounding herbaceous plants). The spatio-temporal aggregration of organisms may also obscure the true quantitative relationships among species, as in the case of plankton patches or reindeer herds. It follows that associations are often defined on the sole basis of the presence or absence of species.

There are many approaches in the literature for measuring the association between species on the basis of presence and absence data. These coefficients are based on the following 2 × 2 contingency table:

Species $\mathbf{y}_2$ \ Species $\mathbf{y}_1$	presence	absence	
presence	a	b	$a + b$
absence	c	d	$c + d$
	$a + c$	$b + d$	$n = a + b + c + d$

where a and d are the numbers of sites where both species are present and absent, respectively, whereas b and c are the numbers of sites where only one of the two species is present; n is the total number of sites. The measures of association between species always exclude the number of double absences, d.

Among the many binary coefficients that exclude double-zeros (Subsection 7.3.2), some have been used for assessing association between species. Jaccard's *coefficient of community* (eq. 7.10) has been used by Reyssac & Roux (1972) in the R mode:

$$S_7(\mathbf{y}_1, \mathbf{y}_2) = \frac{a}{a + b + c}$$

The corresponding distance has been used by Thorrington-Smith (1971) for the same purpose:

$$D = 1 - S_7(\mathbf{y}_1, \mathbf{y}_2) = \frac{b + c}{a + b + c} \qquad \textbf{(7.59)}$$

The Sørensen coefficient (eq. 7.11)

$$S_8(\mathbf{y}_1, \mathbf{y}_2) = \frac{2a}{2a + b + c}$$

was originally defined under the name *coincidence index* for studying species associations (Dice, 1945). The Ochiai coefficient (S_{14}) can also be used in R mode for analysing species presence-absence data.

When used in the R mode, the Sørensen (S_8) coefficient can be tested for significance using random permutations of the observations in one of the species vectors (for the test of association of two species only) or in all species vectors (for simultaneous tests of association among all species). The basic co-occurrence statistic is a, the number of sites where both species are present. For statistic a computed in R mode, a test using permutation method 2b (permutation of one or both columns), described for the Raup & Crick coefficient (eq. 7.31), is equivalent to a permutation test of the Sørensen statistic since the denominator of S_8, $(a + b) + (a + c) = 2a + b + c$,

is invariant under permutation of the values in the columns, and a permutation test of a produces the same permutational probability as a test of $2a$.

Ecological application 7.5

Clua *et al*. (2010) studied the ecology and residence patterns of a group of photo-identified adult sicklefin lemon sharks, *Negaprion acutidens*, at a shark-feeding site monitored by divers during 44 months. An objective of the study was to delineate groups of sharks that were present together and formed recognizable behavioural groups. From the observation data (presence-absence of 29 sharks during 949 dives), the authors computed co-occurrence statistics a among all pairs of sharks and tested their significance using permutation method 2b described for the Raup & Crick coefficient (S_{27}). The p-values of 0.0001, obtained after 9999 random permutations, had a corrected experimentwise p-value of 0.0406 after Holm correction for multiple testing (406 simultaneous tests; Box 1.3). The 52 edges corresponding to these p-value smaller than 0.05 were drawn on a principal coordinate ordination diagram (PCoA, Section 9.3). Five behavioural groups of sharks were recognized on the plot.

An elaborate coefficient was proposed by Fager & McGowan (1963):

$$S_{24}(\mathbf{y}_1, \mathbf{y}_2) = \frac{a}{\sqrt{(a+b)\,(a+c)}} - \frac{1}{2\sqrt{a + \max(b,c)}} \qquad \textbf{(7.60)}$$

Coefficient S_{24} replaced a probabilistic coefficient proposed earlier by Fager (1957). The first part of coefficient S_{24} is the same as the Ochiai coefficient S_{14}, i.e. the geometric mean of the proportions of co-occurrence for each of the two species; the second part is a correction for small sample size.

Krylov (1968) proposed to use the probability associated with the X^2 (chi-square) statistic of the above 2×2 contingency table to test the null hypothesis that two species are distributed independently of each other among the sites. Rejecting H_0 gives support to the alternative hypothesis of association between the two species. In the case of a 2×2 contingency table, and using Yate's correction factor for small samples, the X^2-formula is:

$$X^2 = \frac{n\,[\,|ad - bc| - (n/2)\,]^2}{(a+b)\,(c+d)\,(a+c)\,(b+d)} \qquad \textbf{(7.61)}$$

The number of degrees of freedom for the test of significance is ν = (no. rows – 1) × (no. columns – 1) = 1 (eq. 6.7). The X^2-statistic could also be tested by permutation (Section 1.2). Given that associations should be based on positive relationships between species (negative relationships reflecting competition), Krylov proposed to set $S(\mathbf{y}_1, \mathbf{y}_2) = 0$ when the expected value of co-occurrence, $E = (a + b)(a + c)/n$, is larger than or equal to the observed frequency ($E \geq a$). Following the test, two species are considered associated if the probability (p) computed for their X^2-statistic is

smaller than a pre-established significance level, for example $\alpha = 0.05$. The similarity measure between species is the complement of that probability:

$$S_{25}(\mathbf{y}_1, \mathbf{y}_2) = 1 - p(X^2), \text{ with } \nu = 1, \quad \text{when } E(a) = (a+b)(a+c)/n < a$$
$$S_{25}(\mathbf{y}_1, \mathbf{y}_2) = 0 \quad \text{when } E(a) = (a+b)(a+c)/n \geq a \quad \textbf{(7.62)}$$

The p-value itself can be used as a distance. When the number of sites n is smaller than 20 or a, b, c or d are smaller than 5, Fisher's exact probability formula should be used to compute the p-value instead of a test of X^2. The formula can be found in most textbooks of statistics.

The same formula can be derived from Pearson's ϕ (phi) (eq. 7.9), given that $X^2 = n\phi^2$. Pearson's ϕ is also called the *point correlation coefficient* because it is the general correlation coefficient (eq. 5.1) computed from presence-absence data.

2. *Species abundance data*. — For quantitative species abundance or biomass data, parametric or nonparametric correlation coefficients (Pearson r and Spearman r) can be used to assess the relationships among species (Greig-Smith, 1983; O'Connor & Aarssen, 1987; Myster & Pickett, 1992). When looking for species associations, Legendre (2005) suggested to transform the species abundances through one of the transformations described in Section 7.7 to control for differences in total abundance (or total biomass for biomass data) among sites before computing the correlations among species, in order to linearize the relationships (for Pearson r) or make them more monotonic (for Spearman r).

If the correlations must be transformed into distances before hierarchical clustering or ordination, use the transformations $D = (1 - r)$ or $D = \sqrt{1 - r}$ that were used to transform S into D in Subsection 7.2.1: two species that are identically distributed across the sites have a correlation $r = 1$, hence $D = 0$, which would be the appropriate distance measure in that case. For negative correlations, if any, these transformations produce distances larger than 1, which would cause no problem in clustering or ordination. To induce the clustering or ordination methods to put together species that are either positively or negatively correlated, use $D = (1 - r^2)^{0.5}$ to transform correlations into distances.

The chi-square metric (D_{15}) and the chi-square distance (D_{16}) are appropriate in both the Q and R modes. Both distances can be computed among species before clustering. The D_{16} matrix is obtained by transposing the data matrix so that species are now the rows; then, apply the chi-square distance transformation (eq. 7.70) and compute the Euclidean distance (D_1) on the transformed data.

Whittaker (1972) proposed a coefficient called *SC* (species correlation), constructed like his index of association (distance D_9). Despite its name, this coefficient has no relationship with the Pearson and Spearman correlation formulas. Each abundance value y_{ij} is first transformed into a relative abundance (eq. 7.68) by dividing it by the corresponding row sum y_{i+}, then the coefficient is computed using

the transformed data. As in coefficient D_9 (eqs. 7.47 and 7.48), there are two algebraic forms for the computation of *SC* between species (columns) $\mathbf{y}_1$ and $\mathbf{y}_2$:

$$SC(\mathbf{y}_1, \mathbf{y}_2) = 1 - \frac{1}{2}\sum_{i=1}^{n}\left|\frac{y_{i1}}{y_{1+}} - \frac{y_{i2}}{y_{2+}}\right| = \sum_{i=1}^{n}\min\left(\frac{y_{i1}}{y_{1+}}, \frac{y_{i2}}{y_{2+}}\right) \quad \textbf{(7.63)}$$

where y_{1+} is the sum of values in row $\mathbf{x}_{i=1}$ and y_{2+} is the sum of values in row $\mathbf{x}_{i=2}$. *SC* takes values between 0 and 1. Before clustering, *SC* can be transformed into a distance coefficient by computing

$$D_{20}(\mathbf{y}_1, \mathbf{y}_2) = (1 - SC) \quad \textbf{(7.64)}$$

Probabilistic association

Goodall's probabilistic coefficient (S_{22} or S_{23}, eqs. 7.29 and 7.30) can also be applied to species abundances in the R mode. An example is found in Legendre (1973). This probabilistic coefficient allows one to set an objective limit to species associations; indeed, one may then use a probabilistic definition of an association, such as: "all species that are related at a probability level $(1 - p) \geq 0.95$ are members of the association". Goodall's coefficient has the following meaning in R mode: given *p* species and *n* sites, the similarity of a pair of species is defined as the complement (1 – p) of the probability that any pair of species chosen at random would be as similar as, or more similar than the two species under consideration. Goodall's similarity coefficient is computed as in Subsection 7.3.5, with species interchanged with sites. In step (a), if the species data have been normalized (for example using the transformation $y' = \log(y + 1)$ in eq. 7.65, or eq. 7.66), the partial similarity of Gower's coefficient S_{19} (eq. 7.26)

$$s_{i12} = 1 - [|y_{i1} - y_{i2}| / R_i]$$

may be used to describe the similarity between species y_1 and y_2 at site i. R_i is the range of variation of the normalized species abundances at site i; R_i scales the differences between species for each site.

7.6 Choice of a coefficient

Criteria for choosing a coefficient are summarized in Tables 7.4 to 7.6. In these tables, the coefficients are identified by the names and numbers used in Sections 7.3 to 7.5. The three tables distinguish between coefficients appropriate for species (or frequency) descriptors, and those for other types of descriptors.

Levels 4 and 6 of Table 7.4 require some explanation. Coefficients differentiated in these levels are classified with respect to two criteria, i.e. (a) standardization (or not) of each object-vector prior to the comparison and (b) relative importance given by the coefficient to the abundant or rare species. This defines various types of coefficients.

Type 1 coefficients. Consider two objects, each represented by a vector of species abundances, to be compared using a Q-mode measure. With type 1 coefficients, if there is a given difference between sites for some abundant species and the same difference for a rare species, the two species contribute equally to the similarity or distance between sites. A small numerical example illustrates this property for the percentage difference (D_{14}), which is the complement of Steinhaus' similarity (S_{17}):

Species:	$\mathbf{y}_1$	$\mathbf{y}_2$	$\mathbf{y}_3$
Site $\mathbf{x}_1$	100	40	20
Site $\mathbf{x}_2$	90	30	10
$\lvert y_{1j} - y_{2j} \rvert$	10	10	10
$(y_{1j} + y_{2j})$	190	70	30

Using eq. 7.58 shows that each of the three species contributes 10/290 to the total distance between the two sites. With some coefficients (D_3, D_4, D_9, D_{17}, D_{18}), the standardization of the site-vectors, which is automatically done prior to the computation of the coefficient, may make the result unclear as to the importance given to each species. With these coefficients, the property of "equal contribution" is found only when the two site-vectors are equally important, the importance being measured in different ways depending on the coefficient (see the footnote of Table 7.4).

Type 2a coefficients. — With coefficients of this type, a difference between values for an abundant species contributes less to the distance (and, thus, more to the similarity) than the same difference for a rare species. The *Canberra metric* (D_{10}) belongs to this type. For the above numerical example, calculation of D_{10} (eq. 7.49) shows that species $\mathbf{y}_1$, which is the most abundant, contributes 10/190 to the distance, $\mathbf{y}_2$ contributes 10/70, whereas the contribution of $\mathbf{y}_3$, which is the rarest species, is the largest of the three (10/30). The total distance is $D_{10} = 0.529$. The *coefficient of divergence* (D_{11}; eq. 7.51) also belongs to this type.

Type 2b coefficients. — Coefficients of this type behave similarly to the previous ones, except that the importance of each species is calculated with respect to the whole data set instead of the two site-vectors that are compared. The χ^2 *metric* (D_{15}) is representative of this. In eq. 7.54 and accompanying example, the squared difference between conditional probabilities, for a given species, is divided by y_{+j} which is the total number of individuals belonging to this species at all sites. If this number is large, it reduces the contribution of the species to the total distance between two rows (sites) more than would happen in the case of a rarer species. *Gower's coefficient* (S_{19}; eq 7.26) has the same behaviour (unless special weights w_{12j} are used for some species), since the importance of each species is determined from its range of variation over all sites. The coefficient of Legendre & Chodorowski (S_{20}; eq 7.27) also belongs to this type when parameter k in the partial similarity function s_{12j} for each species is made proportional to its range of variation over all sites.

Legendre *et al*. (1985) suggested that it is more informative to compare dominant or well-represented species than rare taxa, because the latter are generally not well

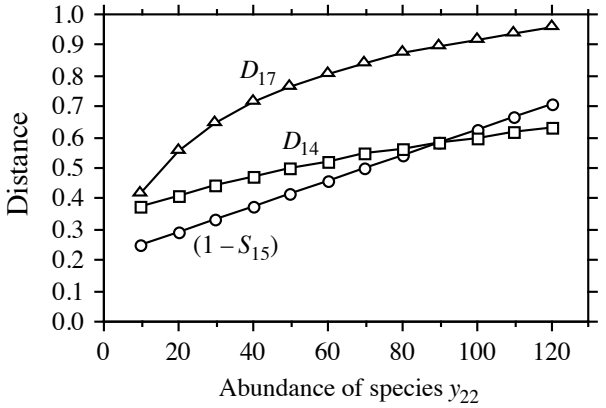


Figure 7.6 Results of an *ordered comparison case series* (OCCAS) where species y_{22} abundance varies from 10 to 120 by steps of 10. The values taken by coefficients $(1 - S_{15})$, D_{14}, and D_{17} are shown.

sampled. This provides an approach for choosing a coefficient. In immature communities, most of the species are represented by small numbers of individuals, so that only a few species are well sampled, whereas, in mature communities, several species exhibit intermediate or high abundances. When calculating similarities between species from immature communities, a reasonable approach may thus be to give more weight to the few well-sampled species (type 2 coefficients) whereas, for sites from mature communities, type 1 coefficients may be more appropriate.

Another way of choosing a resemblance coefficient is to construct an artificial data set representing contrasting situations that the similarity or distance measure should be able to differentiate. Computing several candidate coefficients for the test data will indicate which coefficient is the most appropriate for data of that type. In that spirit,

OCCAS

Hajdu (1981) constructed series of test cases, called *ordered comparison case series* (OCCAS), corresponding to linear changes in the abundances of two species along different types of simulated environmental gradients. The results are distances between sites, computed using different coefficients, for linearly changing species composition.

To illustrate the method, consider one of Hajdu's OCCAS with two species. For these species, site 1 had frequencies $y_{11} = 100$ and $y_{12} = 0$; site 2 had frequency $y_{21} = 50$ whereas y_{22} varied from 10 to 120. Figure 7.6 shows the results for three coefficients: $(1 - S_{15})$ has a completely linear behaviour across the values of y_{22}, D_{14} is not quite linear, and D_{17} is strongly curvilinear.

An ideal coefficient should change linearly when plotted against a series of test cases corresponding to a linear change in species composition, as simulated in OCCAS

Non-linearity

runs. Hajdu (1981) proposed a measure of *non-linearity*, defined as the *standard deviation of the changes* in values of distance between adjacent test cases along the series. A good distance coefficient should also change substantially along the series

and reach its maximum value when the difference in species composition is maximum.

Resolution

Resolution was defined as the *mean change* occurring in distances between adjacent test cases along the series. High linearity is desirable in ordination methods whereas high resolution is desirable in cluster analysis. The ratio of non-linearity over resolution defines a coefficient of variation that should be small for a "good" overall resemblance coefficient.

Resolutions are only comparable among coefficients that are bounded in the interval [0, 1] or [0, $\sqrt{2}$]; as a consequence, this measure should not be used to compare coefficients, such as D_1, D_2, and D_{10}, which do not have an upper bound. Non-linearity near 0 is always a good property, but, again, higher values are only comparable for coefficients that are bounded. Coefficients of variation are comparable because the scale of variation of each specific coefficient is taken into account in the calculation.

Gower & Legendre (1986) used Hajdu's OCCAS to study the behaviour of several similarity and distance coefficients and to make recommendations about their use. They studied 15 coefficients for binary data (all of which are described in the present chapter) and 10 coefficients for quantitative data (5 of them are described here). Among the binary coefficients, S_{12} (eq. 7.15) and the coefficient of Yule (eq. 7.8) were strongly non-linear and should be avoided; all the other coefficients in that study (S_1, S_2, S_3, S_5, S_6, S_7, S_8, S_{10}, S_{13}, S_{14}, as well as eqs. 7.7 and 7.9) behaved well.

The coefficients for quantitative data included in that study were S_{15}, $D_{14} = 1 - S_{17}$, D_2, D_{10} and D_{11}. Coefficients D_2 and S_{15}, which are adapted to physical descriptors (Table 7.5), behaved well. D_2 is a standardized form of the Euclidean distance D_1; they both have the same behaviour.

All coefficients adapted to species abundance data (Table 7.4) that were included in the study (D_{10}, D_{11}, D_{14}) behaved well and are recommended. Coefficients S_{15} and D_{10} had perfect linearity in all specific OCCAS runs; they are thus the best of their kinds for principal coordinate analysis (PCoA, Section 9.3), which is a metric ordination method based on distances.

A later analysis of coefficient $D = \sqrt{D_{14}} = \sqrt{1 - S_{17}}$ showed that its non-linearity was very similar to that of $D_{14} = 1 - S_{17}$; the resolution of $\sqrt{D_{14}}$ was slightly lower than that of D_{14}. Both forms are thus equally suitable for ordination whereas D_{14} may be marginally preferable for clustering purposes. The square root transformation of D_{14}, used in the latter part of Numerical example 1 (continued) in Subsection 9.3.5, offers a simple way to avoid negative eigenvalues in principal coordinate ordination.

Another comparative analysis involving the chi-square metric and related forms (D_{15}, D_{16}, D_{17} and D_{18}) showed that the best of this group for metric ordination (PCoA) is the Hellinger distance (D_{17}), which has the lowest coefficient of variation (best compromise between linearity and resolution), despite the fact that it is strongly non-linear. Other properties of resemblance coefficients have been investigated by Bloom (1981), Wolda (1981) and Hubálek (1982).

Table 7.4 Choice of an association measure among objects (Q mode), to be used with species descriptors (asymmetrical coefficients). For explanation of levels 4 and 6, see the accompanying text.

1) Descriptors: presence-absence or ordered classes on a scale of relative abundances (no partial similarities computed between classes) **see 2**

   2) Metric coefficients: *coefficient of community* (S_7) and variants (S_{10}, S_{11})

   2) Semimetric coefficients: variants of the coef. community (S_8, S_9, S_{13}, S_{14})

   2) Nonmetric coefficient: Kulczynski (S_{12}) (non-linear: not recommended)

   2) Probabilistic coefficient: S_{27}

1) Descriptors: quantitative or semiquantitative (states defined in such a way that partial similarities can be computed between them) **see 3**

   3) Coefficients for raw or normalized abundance data **see 4**

      4) No standardization by object; the same difference for either abundant or rare species, contributes equally to the similarity between sites: *coefficients of Steinhaus* (S_{17}) and *Kulczynski* (S_{18}), *percentage difference* (D_{14}), $\sqrt{D_{14}}$

      4) Standardization by object-vector; if objects are of equal importance*, same contributions for abundant or rare species to the similarity or distance between sites: *chord distance* (D_3), *geodesic metric* (D_4), *index of association* (D_9), *Hellinger dist.* (D_{17}), *dist. between profiles* (D_{18})

      4) Standardization by object-vector*; differences for abundant species (in the whole data set) contribute more than differences between rare species to the similarity (less to the distance) between sites: χ^2 *similarity* (S_{21}), χ^2 *metric* (D_{15}), χ^2 *distance* (D_{16})

   3) Limited to normalized abundances (species distributions not strongly skewed). [Normalization of species abundance data: Sections 1.5.6 and 7.7] **see 5**

      5) Coefficients without associated probability levels **see 6**

         6) Differences for abundant species (for two sites under consideration) contribute more than differences between rare species to the similarity (less to the distance) between sites: *Canberra metric* (D_{10}), *coefficient of divergence* (D_{11}). Both have low resolution: not recommended for clustering

         6) Differences for abundant species (in the whole data set) contribute more than differences between rare species to the similarity (less to the distance) between sites: *asymmetrical Gower coefficient* (S_{19}), *coefficient of Legendre & Chodorowski* (S_{20})

         6) Differences for abundant and rare species contribute the same to the distance between sites: *modified mean character difference* or *modified Gower dissimilarity* (D_{19})

      5) Probabilistic coefficient: *Goodall coefficient* (S_{23})

* D_3 and D_4: importance quantified relative to the length of the row vector $\sqrt{\sum_t y_{ij}^2}$

D_9, D_{15} to D_{18}: importance relative to the sum of individuals in the row vector $\sum_t y_{ij}$

Table 7.5 Choice of an association measure among objects (Q mode), to be used with chemical, geological physical, etc. descriptors (symmetrical coefficients, using double-zeros).

1) Association measured between individual objects **see 2**

  2) Descriptors: presence-absence or multistate (no partial similarities computed between states) **see 3**

    3) Metric coefficients: *simple matching* (S_1) and derived coefficients (S_2, S_6)

    3) Semimetric coefficients: S_3, S_5

    3) Nonmetric coefficient: S_4

  2) Descriptors: multistate (states defined in such a way that partial similarities can be computed between them) **see 4**

    4) Descriptors: quantitative and dimensionally homogeneous **see 5**

      5) Differences enhanced by squaring: *Euclidean distance* (D_1) and *average distance* (D_2)

      5) Differences mitigated: *Manhattan metric* (D_7), *mean character difference* (D_8)

    4) Descriptors: not dimensionally homogeneous; weights (equal or not, according to values w_j used) given to each descriptor in the computation of association measures **see 6**

      6) Descriptors are qualitative (no partial similarities computed between states) and quantitative (partial similarities based on the range of variation of each descriptor): *symmetrical Gower coefficient* (S_{15})

      6) Descriptors are qualitative (possibility of using matrices of partial similarities between states) and semiquantitative or quantitative (partial similarity function for each descriptor): *coefficient of Estabrook & Rogers* (S_{16})

1) Association measured between groups of objects

  7) Removing the effect of correlations among descriptors: *Mahalanobis generalized distance* (D_5)

  7) Not removing the effect of correlations among descriptors: *coefficient of racial likeness* (D_{12})

Table 7.6 Choice of a dependence measure among descriptors (R mode).

1) Descriptors: species abundances	**see 2**
2) Descriptors: presence-absence	**see 3**
3) Coefficients without associated probability levels: S_7, S_8, S_{14}, S_{24}	
3) Probabilistic coefficient: S_{25}	
2) Descriptors: multistate	
4) Data are raw abundances: χ^2 *similarity* (S_{21}), χ^2 *metric* (D_{15}), χ^2 *distance* (D_{16}), *Whittaker's SC* (D_{20})	**see 4**
4) Data are abundances in linear or monotonic relationships	**see 5**
5) Coefficients without associated probabilities: *covariance*, *Pearson r*, *Spearman r*, Pearson or Spearman correlations among chord-transformed or Hellinger-transformed data	
5) Probabilistic coefficients: *probabilities associated to Pearson r* or *Spearman r*, *Goodall coefficient* (S_{23})	
1) Descriptors: chemical, geological, physical, etc.	**see 6**
6) Coefficients without associated probability levels	**see 7**
7) Descriptors are quantitative and linearly related: *covariance*, *Pearson r*	
7) Descriptors are ordered and monotonically related: *Spearman r*, *Kendall* τ	
7) Descriptors are qualitative or ordered but not monotonically related: χ^2, *reciprocal information coefficient*, *symmetric uncertainty coefficient*	
6) Probabilistic coefficients	**see 8**
8) Descriptors are quantitative and linearly related: *probabilities associated to Pearson r*	
8) Descriptors are ordered and monotonically related: *probabilities associated to Spearman r* or *Kendall* τ	
8) Descriptors are qualitative or ordered but not monotonically related: *probabilities associated to* χ^2	

7.7 Transformations for community composition data

In communities sampled over fairly homogeneous environmental conditions, e.g. short environmental gradients, the species composition data contain few zeros, and symmetric association coefficients, including the Euclidean distance D_1, can be used for clustering or ordination. Frequency histograms of individual species may, however, display asymmetric distributions because species tend to have exponential growth when conditions are favourable. This well-known fact has been embedded in the theory of species-abundance models; see He & Legendre (1996, 2002) for a synthetic view of these models. To reduce the asymmetry of the species distributions, a species abundance variable **y** may be transformed to **y**' by taking the square root or the fourth root (equivalent to taking the square root twice), or by using a log transformation:

$$y' = y^{0.5} \quad \text{or} \quad y' = y^{0.25} \quad \text{or} \quad y' = \log(y + c) \qquad \textbf{(7.65)}$$

where y is the species abundance and c is a constant. Usually, $c = 1$ in species abundance log transformations; in this way, an abundance $y = 0$ is transformed into $y' = \log(0 + 1) = 0$ for any logarithmic base. These transformations represent the series of exponents $\gamma = 0.5, 0.25$ and 0 of the Box-Cox transformation (eq. 1.15).

Another interesting transformation that reduces the asymmetry of heavily skewed abundance data is the one proposed by Anderson *et al*. (2006). The abundance data y_{ij} are transformed as follows to a logarithmic scale that makes allowance for zeros:

$$y'_{ij} = \log_{10}(y_{ij}) + 1 \qquad \text{when } y_{ij} > 0 \qquad \textbf{(7.66)}$$

$$\text{or } y'_{ij} = 0 \qquad \text{when } y_{ij} = 0.$$

Hence, for $y_{ij} = \{0, 1, 10, 100, 1000\}$, the transformed values y'_{ij} are $\{0, 1, 2, 3, 4\}$. Note that this is *not* the $\log(y_{ij} + 1)$ transformation. This transformation is available in the ***decostand()*** function of VEGAN (method = "log") where users can choose the base of the logarithm. Changing the base of logarithms in eq. 7.65 (right) produces a linear change among the y'_{ij} values, so it does not induce any change in the relationships among the transformed values. With eq. 7.66 on the contrary, the transformations produced by different bases of logarithms are not perfectly linearly related.

Community composition data sampled over variable environmental conditions, e.g. along long environmental gradients, typically contain many zero values because species are known to generally have unimodal distributions along environmental gradients (ter Braak & Prentice, 1988) and to be absent from sites far from their optimal living conditions. The proportion of zeros is greater when the environmental conditions are more variable across the sampling sites. For association coefficients, this situation generates the double-zero problem that was discussed in Subsection 7.2.2 and leads to the selection of an asymmetrical similarity or distance coefficient for clustering or ordination.

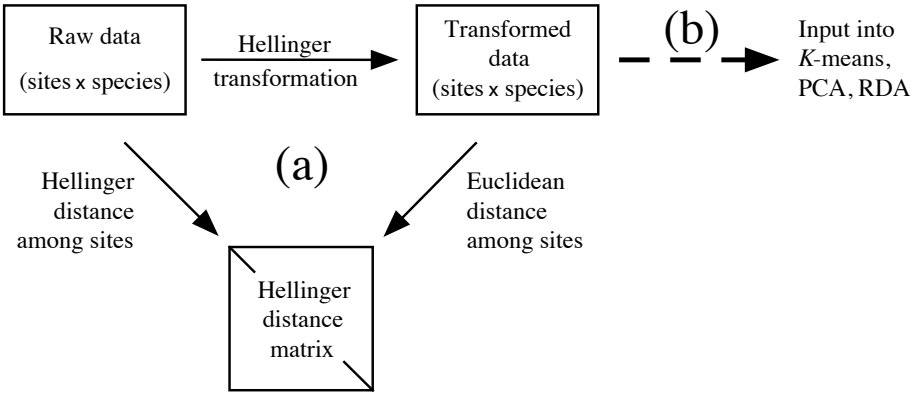


Figure 7.7 (a) Calculation of a distance matrix either directly from the raw data (left diagonal arrow) or through a two-step approach in which the raw data are transformed (horizontal arrow) before computation of the distance matrix (right diagonal arrow). The example shown here uses the Hellinger transformation to obtain the Hellinger distance matrix (D_{17}). The same approach can be used to obtain the chord (D_3), species profile (D_{18}), chi-square metric (D_{15}) and chi-square distance (D_{16}) matrices, as summarized in Fig. 7.8. (b) The transformed species data can also be used as input (dashed arrow) into linear methods of analysis, in particular PCA, RDA, and *K*-means partitioning. Modified from Legendre & Gallagher (2001).

An alternative method of computation for the asymmetrical distance coefficients D_3, D_{15}, D_{16}, D_{17} and D_{18} was proposed by Legendre & Gallagher (2001). The method consists of a transformation of the community composition data followed by the calculation of Euclidean distances (D_1) among sites. These two steps produce the distance function corresponding to the name of the transformation (Fig. 7.7). Data subjected to one of these transformations can also be used directly as input into linear methods of analysis that carry out computations in Euclidean space, such as *K*-means partitioning, PCA, and RDA (Sections 8.8, 9.1, 11.1). This approach is called transformation-based PCA (tb-PCA), transformation-based RDA (tb-RDA), and transformation-based *K*-means partitioning (tb-*K*-means).

1 — Transformation formulas

The following transformations, found in the vertical rectangle in the centre of Fig. 7.8, can be used to obtain the distance coefficients found on their left. The effect of these transformations is to remove the differences in total abundance (for abundance data) or total biomass (for biomass data) from the data, keeping the variations in relative species composition among sites. The chord and Hellinger transformations described below have been in use in community ecology and palaeoecology for a long time (e.g. Noy-Meir *et al.*, 1975; Prentice, 1980). Legendre & Gallagher (2001) showed

Species abundance paradox data ⇒
(3 sites, 3 species)

	Species 1	Species 2	Species 3
Site 1	0	4	8
Site 2	0	1	1
Site 3	1	0	0

Transformations ⇓

$$D_1(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{p} (y_{1j} - y_{2j})^2} \qquad \mathbf{D}_1 = \begin{bmatrix} 0.0000 & 7.6158 & 9.0000 \\ 7.6158 & 0.0000 & 1.7321 \\ 9.0000 & 1.7321 & 0.0000 \end{bmatrix}$$

$$D_3(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{p} \left(\frac{y_{1j}}{\sqrt{\sum_{j=1}^{p} y_{1j}^2}} - \frac{y_{2j}}{\sqrt{\sum_{j=1}^{p} y_{2j}^2}} \right)^2} \qquad y'_{ij} = \frac{y_{ij}}{\sqrt{\sum_{j=1}^{p} y_{ij}^2}} \qquad \mathbf{D}_3 = \begin{bmatrix} 0.0000 & 0.3204 & 1.4142 \\ 0.3204 & 0.0000 & 1.4142 \\ 1.4142 & 1.4142 & 0.0000 \end{bmatrix}$$

$$D_{18}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{p} \left(\frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}} \right)^2} \qquad y'_{ij} = \frac{y_{ij}}{y_{i+}} \qquad \mathbf{D}_{18} = \begin{bmatrix} 0.0000 & 0.2357 & 1.2472 \\ 0.2357 & 0.0000 & 1.2247 \\ 1.2472 & 1.2247 & 0.0000 \end{bmatrix}$$

$$D_{17}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{p} \left[\sqrt{\frac{y_{1j}}{y_{1+}}} - \sqrt{\frac{y_{2j}}{y_{2+}}} \right]^2} \qquad y'_{ij} = \sqrt{\frac{y_{ij}}{y_{i+}}} \qquad \mathbf{D}_{17} = \begin{bmatrix} 0.0000 & 0.1697 & 1.4142 \\ 0.1697 & 0.0000 & 1.4142 \\ 1.4142 & 1.4142 & 0.0000 \end{bmatrix}$$

$$D_{16}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{y_{++}} \sqrt{\sum_{j=1}^{p} \frac{1}{y_{+j}} \left(\frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}} \right)^2} \qquad y'_{ij} = \sqrt{y_{++}} \frac{y_{ij}}{y_{i+}\sqrt{y_{+j}}} \qquad \mathbf{D}_{16} = \begin{bmatrix} 0.0000 & 0.3600 & 4.0092 \\ 0.3600 & 0.0000 & 4.0208 \\ 4.0092 & 4.0208 & 0.0000 \end{bmatrix}$$

Figure 7.8 Species abundance paradox data, modified from Orlóci (1978). The paradox is that the Euclidean distance between sites 2 and 3, which have no species in common, is smaller than that between sites 1 and 2 which share species 2 and 3. This results in an incorrect assessment of the ecological relationships among sites. With the other coefficients in this figure, which are asymmetrical, the distance between sites 2 and 3 is larger than that between sites 1 and 2, and the distance between sites 1 and 3 is the same as between sites 2 and 3, or very nearly so. Distance matrix $\mathbf{D}_{15}$ (not shown) is equal to $\mathbf{D}_{16}/\sqrt{y_{++}} = \mathbf{D}_{16}/\sqrt{15}$.

that these transformations were the first step towards the calculation of one of the asymmetrical distances that are appropriate for Q-mode analysis of community data. Only five of the coefficients discussed in this chapter can be computed by the two-step procedure described in Fig. 7.7, i.e. $D_3, D_{15}, D_{16}, D_{17}$ and D_{18}.

Chord transformation

1) *Chord transformation*. — The species abundances from each object (sampling unit) are transformed into a vector of length 1 using the following equation:

$$y'_{ij} = \frac{y_{ij}}{\sqrt{\sum_{j=1}^{p} y_{ij}^2}} \qquad \textbf{(7.67)}$$

where y_{ij} is the abundance of species j in object i. This equation, called the "chord transformation" in Legendre & Gallagher (2001), is available in the program CANOCO (Centring and standardisation for "samples": *Standardise by norm)* and in the ***decostand()*** function of VEGAN (method = "normalize"). If one computes the Euclidean distance (D_1) between two rows of the transformed data table, the resulting value is identical to the chord distance (D_3, eq. 7.35) computed between the rows of the original (untransformed) species abundance data table; this is how the chord distance can be computed through the two-step calculation shown in Fig. 7.7a. As a consequence, after a chord transformation, the community composition data are suitable for PCA or RDA, as well as other methods of analysis that preserve the Euclidean distance among the objects (Fig. 7.7b).

Species profile transformation

2) *Species profile transformation*. — The data can be transformed into profiles of relative species abundances through the following equation:

$$y'_{ij} = \frac{y_{ij}}{y_{i+}} \qquad \textbf{(7.68)}$$

This is a method of data standardisation that is often used prior to analysis, especially when the sampling units are not all of the same size. Data transformed in that way are called *compositional data*. In community ecology, the species assemblage is considered to represent a response of the community to environmental, historical, or other types of forcing; the variation of any single species has no clear interpretation. Compositional data are used because ecologists feel that the vectors of relative proportions of species can lead to meaningful interpretations. Relative abundances can be transformed into percentages by multiplying the values y'_{ij} by 100. Computing Euclidean distances among rows of a data table transformed in this way produces distances among species profiles (D_{18}, eq. 7.53). The transformation to relative abundance profiles is available in the ***decostand()*** function of VEGAN (method = "total"). Statistical criteria investigated by Legendre and Gallagher (2001) show that this is not the best transformation and that the Hellinger transformation (next paragraph) is often preferable.

Abundance data transformed into profiles by eq. 7.68 have the following property: centring the data by columns to means of 0 automatically centres the rows to means of 0. Make sure that the raw abundance data contain no row that sums to 0, though.

Hellinger transformation

3) *Hellinger transformation*. — A modification of the species profile transformation produces the Hellinger transformation:

$$y'_{ij} = \sqrt{\frac{y_{ij}}{y_{i+}}} \qquad \textbf{(7.69)}$$

Computing Euclidean distances among objects of a data table transformed in this way produces a matrix of Hellinger distances among sites (D_{17}, eq. 7.56; Fig. 7.7). The Hellinger distance has good statistical properties as assessed by the criteria of R^2 and monotonicity used by Legendre and Gallagher (2001) in their comparison of transformation methods. The Hellinger transformation is available in the ***decostand()*** function of VEGAN (method = "hellinger").

Chi-square distance transformation

4) *Chi-square distance transformation*. — A more complex modification of the species profile transformation is the chi-square distance transformation:

$$y'_{ij} = \sqrt{y_{++}}\frac{y_{ij}}{y_{i+}\sqrt{y_{+j}}} \qquad \textbf{(7.70)}$$

where y_{ij} is a species presence or abundance value, y_{i+} is the sum of values over row (object) i, y_{+j} is the sum of values over column (species) j, and y_{++} is the sum of values over the whole data table. Euclidean distances computed among the rows of the transformed data table $[y'_{ij}]$ are equal to chi-square distances (D_{16}, eq. 7.55) among the rows of the original, untransformed data table. The chi-square distance transformation is available in the ***decostand()*** function of VEGAN (method = "chi.square").

The chi-square distance transformation equation reduces the value of an abundant species more than that of a rare species. Hence this transformation is interesting when one wants to give more weight to rare species; this is the case when the rare species are considered to be good indicators of special ecological conditions.

Chi-square metric transformation

5) *Chi-square metric transformation*. — The *chi-square metric* (D_{15}) only differs from the *chi-square distance* (D_{16}) by the constant $\sqrt{y_{++}}$ found in eq. 7.70. It can be obtained by the simplified transformation:

$$y'_{ij} = \frac{y_{ij}}{y_{i+}\sqrt{y_{+j}}} \qquad \textbf{(7.71)}$$

followed by calculation of the Euclidean distance. Data transformed using eq. 7.71 are smaller than the same data transformed using eq. 7.70 by a constant factor of $\sqrt{y_{++}}$.

Before applying the transformations described in the previous paragraphs, any of the standardizations investigated by Noy-Meir *et al*. (1975), Prentice (1980), and Faith *et al*. (1987) may be used if the study justifies it. These include species adjusted to equal maximum abundances or equal standard deviations, sites standardised to equal totals, or both. In particular, one may apply a square root or log transformation to the species abundances in order to reduce the asymmetry of the species distributions.

The chord and Hellinger transformations appear to be the best for general use. Legendre & Gallagher (2001) showed that the values of the corresponding distances are monotonically increasing across a simulated ecological gradient and are maximally related (R^2) to the spatial distances along the geographic gradient. Other asymmetrical distances, like D_{14}, that are useful for the analysis of community composition data cannot be obtained through the two-step process of a transformation followed by calculation of the Euclidean distance illustrated in Fig. 7.7. The chord and Hellinger transformations are closely related: chord-transformed abundance data are equal to squared abundance data that are then Hellinger-transformed.

The five transformations described above can be applied to presence-absence data. In that situation, the chord and Hellinger transformations produce identical results, and the corresponding distances, D_3 and D_{17}, are both equal to $\sqrt{2}\sqrt{1-\frac{a}{\sqrt{(a+b)\ (a+c)}}}$ where $\frac{a}{\sqrt{(a+b)\ (a+c)}}$ is the Ochiai similarity coefficient for binary data (S_{14}). Correspondence analysis, which preserves the chi-square distance, has long been used with species presence-absence data; hence the chi-square transformation can also be applied to this type of data.

2 — Numerical example

The modified Orlóci paradox data set was used in Subsection 7.4.1 to shows that the Euclidean distance function may produce misleading results when applied to assemblage composition data. Asymmetrical similarity and distance functions, which were specifically designed for the analysis of community composition data, do not have this drawback. Figure 7.8 (right-hand side) shows, for five distance functions, the distance matrices obtained for these data. From a community ecologist viewpoint, the identity of the species present at two sites is more important for assessment of the differences among these sites than their abundances. Following that conception, sites 1 and 2, which share two species, are more similar to each other than either of them is to site 3, which harbours a single species not found at sites 1 and 2.

Instead of that, Euclidean distances (D_1) show that sites 1 and 2 (D = 7.6158) are more dissimilar than sites 2 and 3 (D = 1.7321). This assessment would be considered incorrect by most community ecologists although the calculations are mathematically correct. In contrast, the four other distance matrices in Fig. 7.8 indicate that the two less dissimilar sites are 1 and 2, an answer that would be considered a correct

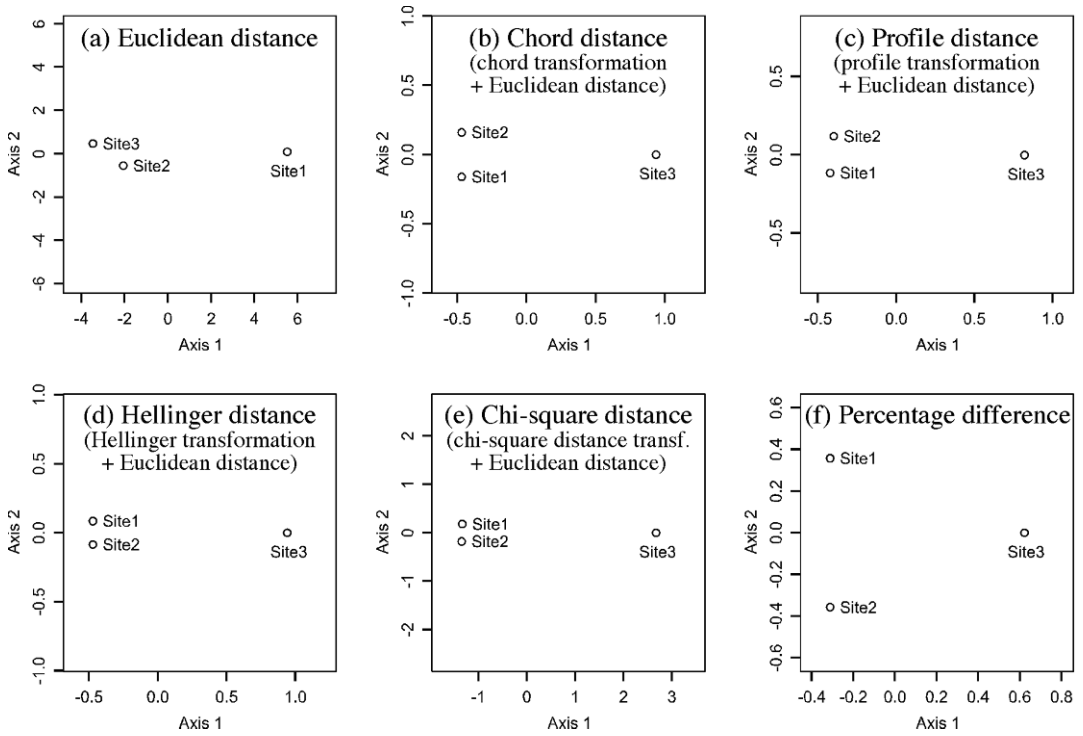


Figure 7.9 Principal coordinate ordination plots (PCoA, Section 9.3) of the distance matrices computed in Fig. 7.8: (a) D_1, (b) D_3, (c) D_{18}, (d) D_{17}, (e) D_{16}, and (f) a PCoA plot of the percentage difference (Steinhaus/Odum/Bray-Curtis) distance matrix (D_{14}) computed for the same data.

assessment of community similarity by most ecologists. Observe also that the chord and Hellinger distances produce a value of $\sqrt{2} = 1.4142$ between sites that have no species in common; this is the maximum value attainable by these distance functions, as noted in Subsection 7.4.1.

Figure 7.9 presents principal coordinate ordination plots (PCoA, Section 9.3) computed from the distance matrices in Fig. 7.8, plus a PCoA plot of the percentage difference matrix (D_{14}) computed for the same data. In the Euclidean distance ordination (Fig. 7.9a), sites 2 and 3 are the closest among the three sites, which may be seen as incorrect for the data under consideration. In all five other ordination plots (Fig. 9b-f), sites 1 and 2 are the closest. The plots also display the interesting property that the different asymmetrical distance functions deal with the differences among sites differently: sites 1 and 2 are the closest to each other in Fig. 7.9e and the farthest in Fig. 7.9f (percentage difference). The distance between sites 1 and 2 would be even larger if PCoA had been computed from square-rooted D_{14} values, which is recommended before PCoA to make percentage difference matrices Euclidean.

3 — Beals smoothing

Beals smoothing is a multivariate transformation designed for species presence/absence community data containing noise and/or many zeros. This transformation replaces the observed presence/absence values of a species by predicted probabilities of occurrence, on the basis of the co-occurrences of that species with the other species in the data set (Beals, 1984; McCune, 1994). The transformed values can be used as input in multivariate analyses. De Cáceres & Legendre (2008) studied the statistical and ecological bases underlying the Beals smoothing function and explored the factors that may affect the reliability of the transformed values using simulated data. They showed that Beals predictions are only reliable for target species that are closely related to the overall ecological structure displayed by the data set. They developed a statistical test to determine when the observed presence/absence values can be replaced with Beals smoothing predictions.

7.8 Software

Only the largest general-purpose commercial statistical packages, such as SAS, SPSS, SYSTAT, JMP, and STATISTICA, offer clustering among their methods for data analysis (Section 8.15), and functions to compute some resemblance coefficients. The smaller commercial packages offer no such facilities. Among the Q-mode coefficients found in the larger packages, one always finds the Euclidean distance. The squared Euclidean, Manhattan, Chebychev* and Minkowski distances may also be found, as well as the simple matching coefficient for multistate nominal data (eq. 7.19). For R-mode analyses, one finds Pearson's r in most packages, or related measures such as the cosine of the angle between variables, dot product, or covariance. Nonparametric correlation coefficients, as well as chi-square, uncertainty and contingency coefficients may also be found. In addition, for Q-mode analysis, SYSTAT offers several binary coefficients and some coefficients for quantitative data (Bray-Curtis, Kulczynski).

Packages written for ecological or taxonomic analysis emphasize resemblance coefficients and clustering methods. They are: NTSYSPC†, developed by F. J. Rohlf, originally for numerical taxonomy studies; CLUSTAN‡, developed by D. Wishart;

---

* In R, the Chebychev distance, $D_{\text{Chebychev}}(\mathbf{x}_1, \mathbf{x}_2) = \max_j |x_{1j} - x_{2j}|$, is computed by function ***dist()*** with method = "maximum". $D_{\text{Chebychev}}$ is a metric. This distance function does not seem to have been used in community ecology. It is described here because it is found in computer packages and in an R function, hence readers may wonder what its equation is.

† NTSYSPC is available from Exter Software Inc., 47 Route 25A, Suite 2, Setauket, New York 11733-2870, USA; http://www.exetersoftware.com.

‡ The CLUSTAN package may be ordered from CLUSTAN Limited, 16 Kingsburgh Road, Edinburgh EH12 6DZ, Scotland. See also the Web page http://www.clustan.com/.

PATN[*], developed by L. Belbin; PC-ORD[†] written under the direction of B. McCune; and SYN-TAX 2000[**] written by J. Podani. In the R language,

1. The most inclusive functions to compute distances are: ***dist()*** in STATS, ***vegdist()*** in VEGAN, ***dist.binary()*** in ADE4, and ***daisy()*** in CLUSTER. In addition, ***gowdis()*** in FD offers a complete set of options to compute the Gower distance ($1 - S_{15}$). Function ***mahalanobis()*** in STATS computes Mahalanobis distances between the objects in a data table and a vector, which can be the multivariate mean vector of the same data table. Function ***raupcrick()*** in VEGAN computes the Raup-Crick distance ($1 - S_{27}$). Function ***is.euclid()*** of ADE4 checks the Euclidean nature of distance matrices; see Tables 7.2 and 7.3.

2. In the R mode, package STATS offers functions ***var()*** and ***cov()*** to compute covariance matrices and ***cor()*** to compute correlation matrices. ***cor.test()*** is used to test the significance of correlation coefficients. The Pearson, Spearman, and Kendall correlation coefficients are available as options in both ***cor()*** and ***cor.test()***. ***chisq.test()*** of STATS provides chi-square tests of significance. ***pf()*** of STATS computes the parametric p-value associated with F-statistics.

3. Transformations for community composition data described in Section 7.7 are available in the VEGAN function ***decostand()***. Multivariate homogeneity of variances is tested by VEGAN's function ***betadisper()***. Beals smoothing is available in the VEGAN function ***beals()***. The test of significance to determine when species presence/absence values can be replaced with Beals smoothing predictions is conducted by function ***BSS.test()***[‡].

[*] PATN is available from Blatant Fabrications Pty Ltd, Carlton, Tasmania, Australia. Technical information is available on the Web page http://www.patn.com.au, or from Lee Belbin at <lee@blatantfabrications.com>.

[†] Availability: see Section 11.7 (footnote).

[‡] R code and documentation file available with the Beals smoothing function on the Web page http://sites.google.com/site/miqueldecaceres/.

Chapter 8 *Cluster analysis*

8.0 A search for discontinuities

Humans have always tried to classify the animate and inanimate objects that surround them. Classifying objects into collective categories is a prerequisite to naming them. It requires the recognition of discontinuous subsets in an environment that is sometimes discrete, but most often continuous.

To cluster is to recognize that objects are sufficiently similar to be put in the same group and to also identify distinctions or separations between groups of objects. Measures of resemblance between objects (Q mode) or descriptors (R mode) have been discussed in Chapter 7. The present chapter considers the different criteria that may be used to decide whether objects are similar enough to be allocated to the same group when several groups have been defined, and shows that different clustering strategies correspond to different definitions of a what a cluster is. The chapter also examines special clustering approaches that are used to identify species associations.

Few ecological theories predict the existence of discontinuities in nature. Evolutionary theory tells taxonomists that discontinuities exist between species, which are the basic units of evolution, as a result of reproductive barriers; taxonomists use classification methods to reveal these discontinuities. For the opposite reason, taxonomists are not surprised to find continuous differentiation at the sub-species level. In contrast, the world that ecologists try to understand is most often a continuum. In numerical ecology, methods used to identify clusters must therefore be more contrasting than in numerical taxonomy.

Given a sufficiently large group of objects, ecological clustering methods should be able to recognize clusters of similar objects while ignoring the few intermediates that often persist between clusters. Indeed, one cannot expect to find discontinuities when clustering sampling sites unless the physical environment is itself discontinuous, or unless sampling occurred at opposite ends of a gradient, instead of within the gradient (Whittaker, 1962: 88). Similarly, when looking for associations of species, small groups of densely associated species are usually found, with the other species gravitating around one or more of the association nuclei.

Typology

The result of clustering ecological objects sampled from a continuum is often called a *typology* (i.e. a system of types). In such a case, the purpose of clustering is to identify various *object types*, which may be used to describe the structure of the continuum; it is thus immaterial to wonder whether these clusters are "natural" or unique.

For readers with no practical experience in clustering, Section 8.2 provides a detailed account of single linkage clustering, which is simple to understand and is used to introduce the principles of clustering. The review of other methods includes a survey of the main dichotomies among existing methods (Section 8.4), followed by a discussion of the most widely available methods of interest to ecologists (Sections 8.5, 8.7 and 8.8). Theoretical aspects are examined in Sections 8.3 and 8.6. Section 8.9 discusses clustering algorithms useful in identifying biological associations and indicator species analysis, whereas Section 8.10 gives an overview of seriation, a method useful in particular to cluster non-symmetric resemblance matrices. Section 8.11 describes multivariate regression tree analysis (MRT), a method that involves two data sets, i.e. response and explanatory, whose output is a tree. A review of clustering statistics, methods of cluster validation, and graphical representations, completes the chapter (Sections 8.12 to 8.14). The relationships between clustering and other steps of data analysis are depicted in Fig. 10.3.

Despite the wide applicability of clustering methods, one should remember that no single family of methods can answer all questions raised in numerical ecology. Before engaging in clustering, one should be able to justify why one believes that discontinuities exist in the data or explain why one has a practical need to divide a continuous set of objects into groups.

8.1 Definitions

Clustering
Partition

Clustering is an operation of multidimensional analysis that consists in partitioning a collection of objects or descriptors. Most of the methods described in this chapter can be used to cluster descriptors instead of objects. The presentation in the chapter focuses on objects for simplicity, except in Section 8.9 where methods especially designed to cluster species into associations are described. Explanatory variables can also be clustered to identify groups of collinear variables. A *partition* is a division of a set (collection) into subsets, such that each object belongs to one and only one subset for that partition (Legendre & Rogers, 1972). The classification of objects that results may include a single partition, or several hierarchically nested partitions of the objects (or descriptors), depending on the clustering model that has been selected (Table 8.1).

The clustering methods described in this chapter belong to the class of *hard* or *crisp* clustering, where the groups are mutually exclusive and each object belongs to a single group of a partition. In *fuzzy clustering* on the contrary (Bezdek, 1987), an object may simultaneously belong, to different degrees, to two or more groups of a

Table 8.1 Example of hierarchically nested partitions of a set of objects (e.g. sampling sites). The first partition divides the objects according to the environment where they were found. The second partition, hierarchically nested within the first, describes clusters of sites in each environment.

Partition 1	Partition 2	Sampling sites
Observations in environment A	Cluster 1	7, 12
	Cluster 2	3, 5, 11
	Cluster 3	1, 2, 6
Observations in environment B	Cluster 4	4, 9
	Cluster 5	8, 10, 13, 14

partition. In the study of species associations for example (Section 8.9), this approach is interesting because a species may be partly related to two or more associations. Methods of fuzzy clustering are not described in detail in this chapter but R software to compute them is mentioned in Section 8.15.

A partition resulting from a hard clustering method has the same definition as a descriptor (Section 1.4). Each object is characterized by a state (its cluster) of the classification and it belongs to only one of the clusters. This property is useful for the interpretation of classifications (Chapter 10) since any partition may be considered as a qualitative descriptor and compared as such to any other descriptor. A clustering of objects defined in this way imposes a discontinuous structure onto the data set, even if the objects have originally been sampled from a continuum. This structure results from the grouping into subsets of objects that are recognized as sufficiently similar given the variables considered. One can then look for characteristics that differentiate the clusters from one another.

Clustering has been part of ecological tradition for a long time. It goes back to the Polish ecologist Kulczynski (1928) who needed to cluster ecological observations; he developed a method quite remote from the modern clustering algorithms. The technique, called seriation, consisted in permuting the rows and columns of an association matrix in such a way as to get the largest values near the diagonal. The method is still used in phytosociology, anthropology, the social sciences, and other fields. Analytical solutions to the seriation problem are mentioned in Section 8.10.

Most methods of clustering (this chapter) and ordination (Chapter 9) proceed from association matrices (Chapter 7). Distinguishing between clustering and ordination

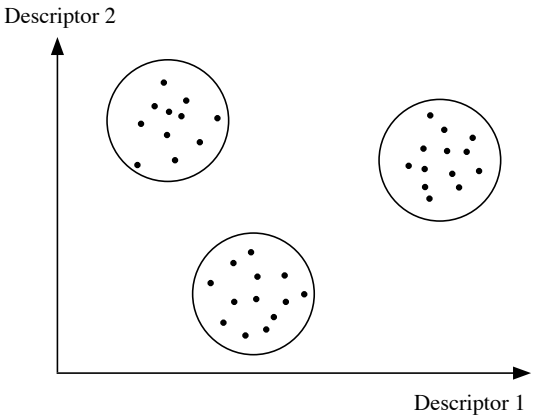


Figure 8.1 Empirically delineating clusters of objects in a scatter diagram is easy when there are no intermediate objects between the groups.

methods is somewhat recent. While ordination in reduced space goes back to Spearman (factor analysis: 1904), most modern clustering methods have only been developed since the era of second-generation computers. The first programmed method was developed by Sokal & Michener (1958) for biological purposes*. Before that, one simply plotted the objects in a scatter diagram with respect to a few variables or principal axes; clusters were then delineated manually (Fig. 8.1) following a method that, today, would be called centroid (Section 8.5) and based upon the Euclidean distances among points. This empirical clustering method still remains a valid approach when the number of variables is small and the structure to be delineated is not obscured by the presence of intermediate objects between the clusters.

Clustering is a family of methods undergoing rapid development. In their report on the literature they reviewed, Blashfield & Aldenderfer (1978) mentioned that they found 25 papers in 1964 that contained references to the basic texts on clustering; then they found 136 papers in 1970, 294 in 1973, and 501 in 1976. The number has been growing ever since. Nowadays, hundreds of mathematicians and researchers from various application fields are collaborating within national or multinational *Classification Societies* throughout the world, under the umbrella of the *International Federation of Classification Societies* founded in 1985.

* Historical note provided by Prof. F. James Rohlf: "Actually, Sokal & Michener (1958) did not use a computer for their very large study. They used an electromechanical accounting machine to compute the raw sums and sums of products. The coefficients of correlation and the cluster analysis itself were computed by hand with the use of mechanical desk calculators. Sneath did use a computer in his first study."

The commonly-used clustering methods are based on easy-to-understand mathematical constructs: arithmetic, geometric, graph-theoretic, or simple statistical models (minimizing within-group variance), leading to rather simple calculations on the dissimilarity or similarity values. It must be understood that most clustering methods are heuristic; they create groups by reference to some concept of what a group embedded in some space should be like, without reference, in most case, to the processes occurring in the application field — ecology in the present book. They have been developed by numerical taxonomists and numerical ecologists, later joined by other researchers in the physical sciences, economics and humanities. In several methods, clusters are delineated on the basis of statements such as: "$\mathbf{x}_1$ is closer to $\mathbf{x}_2$ than it is to $\mathbf{x}_3$", whereas other methods rest on probabilistic models of the type: "Chances are higher that $\mathbf{x}_1$ and $\mathbf{x}_2$ pertain to the same group than $\mathbf{x}_1$ and $\mathbf{x}_3$". In all cases, clustering models make it possible to link the points without requiring prior positioning in a graph (i.e. a metric space), which would be impractical in more than three dimensions. These models allow a graphical representation of other interesting relationships among the objects of the data set than their positions in a reference space of variables, for example the dendrogram of their hierarchical relationships. Chapter 10 will show how it is possible to combine clustering and ordination, computed with different methods, to obtain a more complete picture of the data structure.

The choice of a clustering method is as critical as the choice of an association measure. It is important to fully understand the properties of clustering methods in order to correctly interpret the ecological structure they bring out. Most of all, the methods to be used depend upon the type of clustering sought. Williams *et al*. (1971) recognized two major categories of methods. In a *descriptive clustering*, misclassifying objects is to be avoided, even at the expense of creating single-object clusters. In a *synoptic clustering*, all objects are forced into one of the main clusters; the objective is to construct a general conceptual model that encompasses a reality wider than the data under study. Both approaches have their usefulness.

Descriptive, synoptic clustering

When two or more clustering models seem appropriate to a problem, one should apply them all to the data and compare the results. Clusters that repeatedly come out of analyses that use appropriate methods are the robust solutions to the clustering problem. Differences among results must be interpreted in the light of the known properties of the clustering models, which are explained in the following sections.

8.2 The basic model: single linkage clustering

For natural scientists, a simple-to-understand clustering method (or *model*) is *single linkage* (or *nearest neighbour*) clustering (Sneath, 1957). Its logic seems natural, so that it is used to introduce readers to the principles of clustering. Its name, *single linkage*, distinguishes it from other clustering models, called complete or intermediate

linkage, detailed in Section 8.5. The algorithm for single linkage clustering is sequential, agglomerative and hierarchical, following the nomenclature of Section 8.4. Its starting point is any association matrix (distance or similarity) among the objects to be clustered. One assumes that the association measure has been carefully chosen, following the recommendations of Section 7.6. In the examples that follow, a distance matrix **D** will be used as the starting point for clustering because this is the standard in the clustering functions of the R language.

The method proceeds in two steps:

- First, the association matrix is rewritten in order of increasing distances or decreasing similarities, heading the list with the two closest objects (smallest distance) of the association matrix, followed by the second most similar pair, and proceeding until all the measures comprised in the association matrix have been listed.

- Second, the clusters are formed hierarchically, starting with the two closest objects, and then letting the objects combine into groups, and the groups aggregate to one another, as the distance increases. The following example illustrates this method.

Ecological application 8.2

Five ponds characterized by 38 zooplankton species were studied by Legendre & Chodorowski (1977). The data were counts, recorded on a relative abundance scale from 0 = absent to 5 = very abundant. These ponds have been used as example for the computation of Goodall's coefficient (S_{23}, Chapter 7; only eight zooplankton species were used in that example). These five ponds, with others (see Ecological application 10.1), were subjected to single linkage clustering after computing similarity coefficient S_{20} with parameter $k = 2$. The symmetric similarity matrix, transformed into distances using the equation $D = 1 - S$, is represented by its lower triangle. The diagonal is trivial because it contains distances of 0 by construct.

Ponds	Ponds				
	212	214	233	431	432
212	—				
214	0.400	—			
233	1.000	0.929	—		
431	1.000	0.937	0.700	—	
432	1.000	0.786	0.800	0.500	—

The first clustering step consists in rewriting the distance values in increasing order:

$D = 1 - S_{20}$	Pairs formed
0.400	212-214
0.500	431-432
0.700	233-431
0.786	214-432
0.800	233-432
0.929	214-233
0.937	214-431
1.000	212-233
1.000	212-431
1.000	212-432

Link

As the distance levels increases, pairs of objects are formed. These pairs are called "links"; they serve to link the objects or groups into a chain, as discussed below.

Connected subgraphs are one of the many possible graphical representations of cluster formation (Fig. 8.2a). As the distance increases, clusters are formed, following the list of links in the table of ordered distances above. Only the distance levels at which clusters are modified by addition of objects are represented in the figure. The first link is formed between ponds 212 and 214 at $D = 0.4$, then between 431 and 432 at $D = 0.5$. Pond 233 joins this second cluster nucleus at $D = 0.7$. Finally these two clusters merge at $D = 0.786$ due to a link formed between ponds 214 and 432. The clustering may stop at this point since, according to the single linkage rule (below), all ponds now belong to the same cluster. If the distance criterion is allowed to relax down to $D = 1$ (Fig. 8.2a), links form between members of the cluster up to a point where all ponds are linked to one another. That part of the clustering is of no interest in single linkage clustering, but these links will be of interest in the other forms of linkage clustering below.

Dendrogram

Edge

Node

A dendrogram (Fig. 8.2b) is a commonly-used representation of hierarchical clustering results. Dendrograms only display the clustering topology and object labels, not the links between objects. Dendrograms are made of branches ("edges") that meet at "nodes" which are drawn at the distance values where fusions of branches takes place. For graphical convenience, vertical lines are used in Fig. 8.2b to connect branches at the distance levels of the nodes; the lengths of these lines are of no consequence. Branches could be connected directly to nodes. The branches furcating from a node may be switched ("swivelled") without affecting the information contained in a dendrogram.

The clustering results were interpreted by Legendre & Chodorowski (1977) with respect to the conditions prevailing in the ponds. In their larger study summarized in Ecological application 10.1, all non-permanent ponds (including 212 and 214) formed a cluster while the permanent ponds (including 233, 431 and 432) formed a distinct group (Fig. 10.2).

Single linkage rule

From this example, it should be clear that the rule for assigning an object to a cluster, in single linkage clustering, requires that the object be no more distant than the considered D level from *at least one object already member of the cluster.* In complete linkage hierarchical clustering (Subsection 8.5.2), the assignment rule differs and requires the object to be no more distant than the given level from *all* the objects

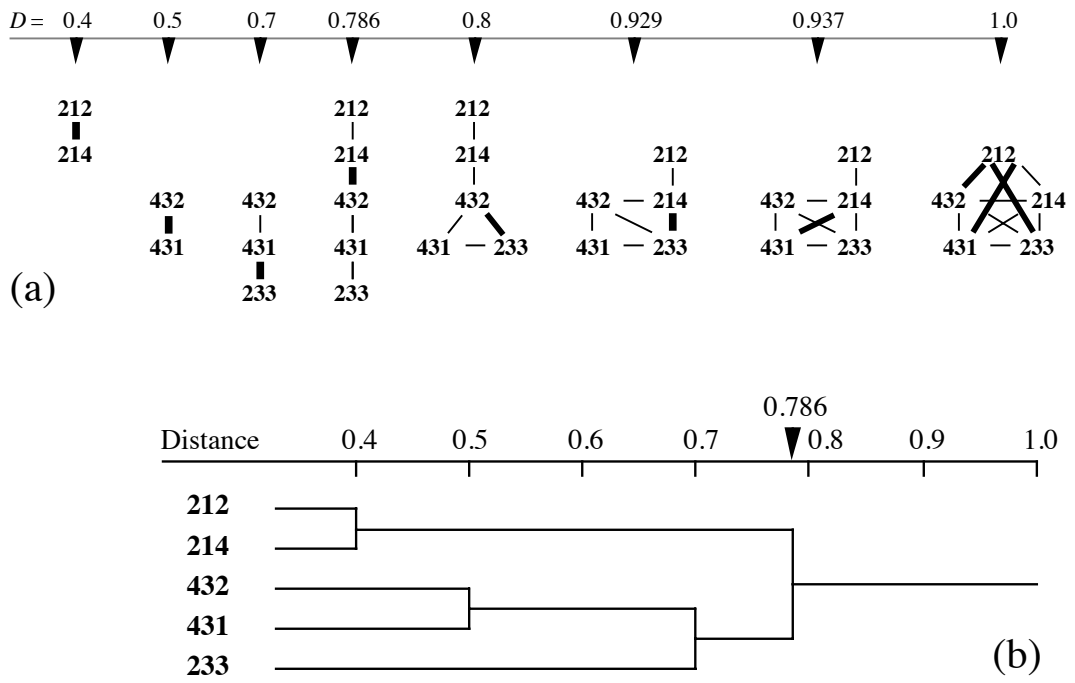


Figure 8.2 Illustrations of single linkage agglomerative clustering for the ponds in the example. (a) Connected subgraphs: groups of objects are formed as the distance level is relaxed from left to right. The levels where clusters are modified by addition of objects are represented; they are ordered along the distance scale (D). New links between ponds are represented by heavy lines; thin lines are used for links formed at previous (lower) distance levels. (b) Dendrogram representing the result of single linkage clustering.

already members of the cluster. The chaining rule used in single linkage clustering may be stated as follows: at each partition level, two objects must be allocated to the same subset if their dissimilarity (distance) is less than or equal to that of the partitioning level considered. The same rule can be formulated in terms of similarities: two objects must be allocated to the same subset if their similarity is equal to or higher than that of the partitioning level considered.

Estabrook (1966) discussed single linkage clustering using the language of graph theory. The exercise has didactic value. A cluster is defined through the following steps:

Link

a) For any pair of objects $\mathbf{x}_1$ and $\mathbf{x}_2$, a *link* is defined between them by a relation G_c:

$$\mathbf{x}_1 \, G_c \, \mathbf{x}_2 \text{ if and only if } D(\mathbf{x}_1, \mathbf{x}_2) \leq c$$

$$\text{or equally, if } S(\mathbf{x}_1, \mathbf{x}_2) \geq (1 - c)$$

assuming distances between 0 and 1. Index c in the clustering relation G_c is the distance level considered. At a distance level of 0.45, for instance, ponds 212 and 214 of the example are in relation $G_{0.45}$ since $D(212, 214) \leq 0.45$. This definition of a link has the properties of symmetry ($\mathbf{x}_1\ G_c\ \mathbf{x}_2$ if and only if $\mathbf{x}_2\ G_c\ \mathbf{x}_1$) and reflexivity ($\mathbf{x}_i\ G_c\ \mathbf{x}_i$ is always true since $D(\mathbf{x}_i, \mathbf{x}_i) = 0$). A group of links for a set of objects, such as defined by relation G_c, is called an *undirected graph*.

Undirected graph

Chain
Chaining

b) The chaining that characterizes single linkage clustering may be described by a G_c-chain. A G_c-chain is said to extend from $\mathbf{x}_1$ to $\mathbf{x}_2$ if there exist other points $\mathbf{x}_3, \mathbf{x}_4, \ldots, \mathbf{x}_i$ in the collection of objects under study, such that $\mathbf{x}_1\ G_c\ \mathbf{x}_3$ and $\mathbf{x}_3\ G_c\ \mathbf{x}_4$ and … and $\mathbf{x}_i\ G_c\ \mathbf{x}_2$. For instance, at similarity level $c = 0.786$ of the example, there exists a $G_{0.786}$-chain from pond 212 to pond 233 since there are intermediate ponds such that $212\ G_{0.786}\ 214$ and $214\ G_{0.786}\ 432$ and $432\ G_{0.786}\ 431$ and $431\ G_{0.786}\ 233$. The number of links in a G_c-chain defines the *connectedness* of a cluster (Subsection 8.12.1).

c) There only remains to delineate the clusters resulting from single linkage chaining. For that purpose, an equivalence relation R_c ("member of the same cluster") is defined as follows:

$\mathbf{x}_1\ R_c\ \mathbf{x}_2$ if and only if there exists a G_c-chain from $\mathbf{x}_1$ to $\mathbf{x}_2$ at distance level c.

In other words, $\mathbf{x}_1$ and $\mathbf{x}_2$ are assigned to the same cluster at distance level c if there exists a chain of links joining $\mathbf{x}_1$ to $\mathbf{x}_2$. Thus, at level $D = 0.786$ in the example, ponds 212 and 233 are assigned to the same cluster ($212\ R_{0.786}\ 233$) because there exists a $G_{0.786}$-chain from 212 to 233. The relationship "member of the same cluster" has the following properties: (1) it is reflexive ($\mathbf{x}_i\ R_c\ \mathbf{x}_i$) because G_c is reflexive; (2) the G_c-chains may be reversed because G_c is symmetric; as a consequence, $\mathbf{x}_1\ R_c\ \mathbf{x}_2$ implies that $\mathbf{x}_2\ R_c\ \mathbf{x}_1$; and (3) it is transitive because, by G_c-chaining, $\mathbf{x}_1\ R_c\ \mathbf{x}_2$ and $\mathbf{x}_2\ R_c\ \mathbf{x}_3$ implies that $\mathbf{x}_1\ R_c\ \mathbf{x}_3$. Each cluster thus defined is a connected subgraph, which means that the objects of a cluster are all connected in their subgraph; in the graph of all the objects, distinct clusters (subgraphs) have no links attaching them to one another.

Connected subgraph

Single linkage clustering provides an accurate picture of the relationships between pairs of objects, but its propensity to chaining is often not desirable in ecological analysis. This is because the presence of an object midway between two compact clusters, or a few intermediate objects connecting two clusters, are enough to turn them into a single cluster. Of course, clusters do not chain unless intermediates are present; so, the occurrence of chaining provides information about the data. To describe this phenomenon, Lance & Williams (1967c) wrote that this method "contracts the reference space". Picture the objects as laying in descriptor space (A-space, Fig. 7.2): the presence of a cluster increases the probability of inclusion, by chaining, of neighbouring objects into the cluster. This is as if the distances between objects were smaller in that region of the space; see also Fig. 8.24a.

Section 10.1 will show how to take advantage of the interesting properties of single linkage clustering by combining it with ordination results, while avoiding the undue influence of chaining on the clustering structure.

Minimum spanning tree

The set of edges that first connect objects to clusters or small graphs into larger graphs, in single linkage clustering, form a graph called *minimum spanning tree* (MST, Gower & Ross, 1969). For Ecological application 8.2, the first four edges represented by heavy links in the left-hand part of Fig. 8.2a, down to $D = 0.786$, form the MST.

That tree has been described a number of times in the literature and has received several names: *dendrites* (Lukaszewicz, 1951), *network* (Prim, 1957), *Prim network* (Cavalli-Sforza & Edwards, 1967), *shortest spanning tree* or *minimum-length tree* (Sneath & Sokal, 1973). MSTs are very useful when analysing clusters drawn in an ordination space (Section 10.1). If the MST is drawn on a scatter diagram of the objects, one can obtain a non-hierarchical clustering of the objects by removing the single largest or the few largest distance links. Such graphs are illustrated in Figs. 10.1 and 10.2. A MST is also used to calculate the truncation distance in the computation of spatial eigenfunctions in Chapter 14. Section 8.15 shows how to compute a MST in R.

Chain of primary connections

A related concept is the *chain of primary connections* (Legendre, 1976): this is the set of links that first connect objects to groups, or groups to one another, in any type of hierarchical clustering. For single linkage clustering, that chain is identical to the MST, but it may differ for other methods if the clustering topology they produce is different. How to compute it is described at the end of Subsection 8.5.4 for the UPGMA case.

8.3 Cophenetic matrix and ultrametric property

Any classification or partition can be fully described by a cophenetic matrix. This matrix is used for comparing different classifications of the same objects.

1 — Cophenetic matrix

Cophenetic distance

The *cophenetic distance* (or *similarity*) of two objects $\mathbf{x}_1$ and $\mathbf{x}_2$ is defined as the distance (or similarity) level at which objects $\mathbf{x}_1$ and $\mathbf{x}_2$ become members of the same cluster during the course of clustering (Jain & Dubes, 1988), as depicted by connected subgraphs or by a dendrogram (e.g. Fig. 8.2a, b). Any dendrogram can be uniquely represented by a matrix in which the distance (or similarity) for a pair of objects is their cophenetic distance (or similarity). Consider the single linkage clustering dendrogram of Fig. 8.2b. The clustering levels, read directly on the dendrogram, lead to the following distance (**D**) and similarity (**S**, where $S = 1 - D$) matrices:

D	212	214	233	431	432
212	—			(upper triangle	
214	0.400	—		symmetric to lower)	
233	0.786	0.786	—		
431	0.786	0.786	0.700	—	
432	0.786	0.786	0.700	0.500	—

S	212	214	233	431	432
212	—			(upper triangle	
214	0.600	—		symmetric to lower)	
233	0.214	0.214	—		
431	0.214	0.214	0.300	—	
432	0.214	0.214	0.300	0.500	—

Cophenetic matrix

These matrices are called *cophenetic matrices* (Sokal & Rohlf, 1962; Jain & Dubes, 1988). The ordering of objects in the cophenetic matrix is irrelevant; any order that suits the researcher is acceptable. The same applies to dendrograms; the order of the

objects may be changed at will, provided that the dendrogram is redrawn to accommodate the new ordering.

For a *partition* of the data set (as in the *K*-means method, below), the resulting groups of objects are not related through a dendrogram. A cophenetic matrix may nevertheless be computed. Consider the groups (212, 214) and (233, 431, 432) obtained by cutting the dendrogram of Fig. 8.2b at distance level $D = 0.75$, ignoring the hierarchical structure of the two clusters. The cophenetic matrices would be:

D	212	214	233	431	432
212	—		(upper triangle		
214	0	—	symmetric to lower)		
233	1	1	—		
431	1	1	0	—	
432	1	1	0	0	—

S	212	214	233	431	432
212	—		(upper triangle		
214	1	—	symmetric to lower)		
233	0	0	—		
431	0	0	1	—	
432	0	0	1	1	—

2 — Ultrametric property

If there are no *reversals* in the clustering (Fig. 8.16), a classification has the following *ultrametric property*:

$$D(\mathbf{x}_1, \mathbf{x}_2) \leq \max[D(\mathbf{x}_1, \mathbf{x}_3), D(\mathbf{x}_2, \mathbf{x}_3)] \qquad \textbf{(8.1)}$$

for every triplet of objects $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$ in the study. The cophenetic matrix is then called ultrametric. Cophenetic distances also possess the four *metric properties* of Section 7.4. The ultrametric property may be expressed in terms of similarities:

$$S(\mathbf{x}_1, \mathbf{x}_2) \geq \min[S(\mathbf{x}_1, \mathbf{x}_3), S(\mathbf{x}_2, \mathbf{x}_3)] \qquad \textbf{(8.2)}$$

As an exercise, readers can verify that the five properties apply to all doublets and triplets of distances in the cophenetic **D** matrix shown above.

8.4 The panoply of methods

Clustering algorithms have been developed using a wide range of conceptual models and for studying a variety of problems. Sneath & Sokal (1973) proposed a classification of clustering procedures. Its main dichotomies are briefly described.

1 — Sequential versus simultaneous algorithms

Most clustering algorithms are sequential in the sense that they proceed by applying a recurrent sequence of operations to the objects. The agglomerative single linkage

clustering of Section 8.2 is an example of a sequential method: the search for the equivalence relation R_c is repeated at all distance levels in the association matrix, up to the point where all objects are in the same cluster. In *simultaneous* algorithms, which are less frequent, the solution is obtained in a single step. Ordination techniques (Chapter 9), which may be used for delineating clusters, are of the latter type. This is also the case of the direct complete linkage clustering method presented in Subsection 8.9.1. The *K*-means (Section 8.8) and other non-hierarchical partitioning methods may be computed using sequential algorithms, although these methods are neither agglomerative nor divisive (next paragraph).

2 — Agglomeration versus division

Among the sequential algorithms, *agglomerative* procedures begin with the discontinuous partition of all objects, i.e. the objects are considered as being separate from one another. They are successively grouped into larger and larger clusters until a single, all-encompassing cluster is obtained. If the continuous partition of all objects is used instead as the starting point of the procedure (i.e. a single group containing all objects), *divisive* algorithms subdivide the group into sub-clusters, and so on until the discontinuous partition is reached. In either case, it is left to users to decide which of the intermediate partitions is to be retained, given the problem under study. Agglomerative algorithms are the most developed for two reasons. First, they are easier to program. Second, in clustering by division, the erroneous allocation of an object to a cluster at the beginning of the procedure cannot be corrected afterwards (Gower, 1967) unless a complex procedure is embedded in the algorithm to do so.

3 — Monothetic versus polythetic methods

Divisive clustering methods may be monothetic or polythetic. *Monothetic* models use a single descriptor at each step as the basis for partitioning, whereas *polythetic* models use several descriptors which, in most cases, are combined into an association matrix (Chapter 7) prior to clustering. Divisive monothetic methods proceed by choosing, for each partitioning level, the descriptor considered to be the best for that level; objects are then partitioned following the state to which they belong with respect to that descriptor. For example, the most appropriate descriptor at each partitioning level could be the one that best represents the information contained in all other descriptors, after measuring the reciprocal information between descriptors (Subsection 8.7.1). When a single partition of the objects is sought, monothetic methods produce the clustering in a single step.

4 — Hierarchical versus non-hierarchical methods

In *hierarchical* methods, the members of inferior-ranking clusters become members of larger, higher-ranking clusters. Most of the time, hierarchical methods produce non-overlapping clusters, but this is not a necessity according to the definition of "hierarchy" in the dictionary or the usage recognized by Sneath & Sokal (1973).

Single linkage clustering of Section 8.2 and the methods of Sections 8.5 and 8.7 are hierarchical.

Non-hierarchical methods are very useful in ecology. They produce a single partition that optimizes within-group homogeneity, instead of a hierarchical series of partitions optimizing the hierarchical attribution of objects to clusters. Lance & Williams (1967d) restrict the term "clustering" to the non-hierarchical methods and call the hierarchical methods "classification". Non-hierarchical methods include *K*-means partitioning, the ordination techniques (Chapter 9) used as clustering methods, the creation of clusters by removing edges from a graph (which may be a minimum spanning tree), the methods of matrix seriation of Section 8.10, and the algorithm described in Subsection 8.9.1 for the clustering of species into biological associations. These methods should be used in cases where the aim is to obtain a direct representation of the relationships among objects instead of a summary of their hierarchy. Hierarchical methods are easier to compute and more often available in statistical packages than non-hierarchical procedures.

Most hierarchical methods use a resemblance matrix as their starting point. This prevents their use with very large data sets because the resemblance matrix, with its $n(n-1)/2$ values, may become extremely large. Algorithms have been developed for hierarchical agglomeration of very large numbers of objects after computing only a small fraction of the distances (e.g. Jambu & Lebeaux, 1983; Rohlf, 1978, 1982a).

5 — Constrained clustering methods

In constrained clustering, external information about the sampling design is used by the clustering algorithm, in addition to the distance relationships among objects. Two forms of constrained clustering are described in this book: time-constrained (Section 12.6) and space-constrained clustering (Subsection 13.3.2).

6 — Probabilistic versus non-probabilistic methods

Probabilistic methods include a clustering model by Clifford & Goodall (1967), designed to be used in conjunction with Goodall's probabilistic index (S_{23}, Chapter 7), in which clusters are formed in such a way that the within-group association matrices have a given probability of being homogeneous. That method is described in the previous edition of this book (Legendre & Legendre, 1998, Subsection 8.9.2). This category also includes the parametric and nonparametric methods for estimating density functions in multivariate space.

Sneath & Sokal (1973) describe other dichotomies for clustering methods, which are of lesser interest to ecologists. These are: global or local criteria, direct or iterative solutions, equal or unequal weights, and adaptive or non-adaptive clustering.

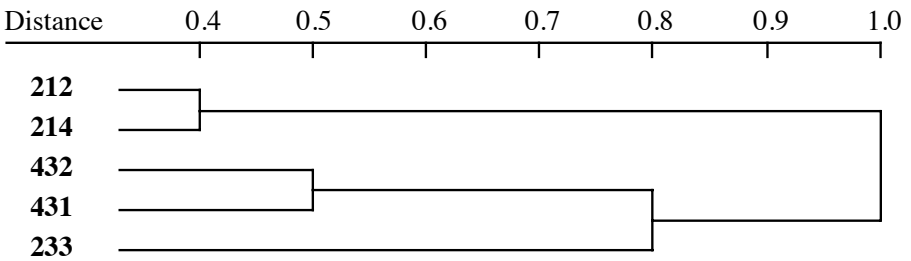


Figure 8.3 Complete linkage clustering of the ponds of Ecological application 8.2.

8.5 Hierarchical agglomerative clustering

Most methods of hierarchical agglomeration can be computed as special cases of a general model which is discussed in Subsection 8.5.9.

1 — Single linkage agglomerative clustering

In single linkage agglomeration (Section 8.2), two clusters fuse when the two objects closest to each other (one in each cluster) reach the distance level of the considered partition (Fig. 8.2). As a consequence of chaining, results of single linkage clustering are sensitive to noise in the data (Milligan, 1996), because noise changes the distance values and may thus modify the order in which objects cluster. The origin of single linkage clustering is found in a collective work by mathematicians Florek, Lukaszewicz, Perkal, Steinhaus, and Zubrzycki, published by Lukaszewicz in 1951.

2 — Complete linkage agglomerative clustering

Opposite to the single linkage approach is *complete linkage agglomeration*, also called *furthest neighbour sorting*. In this method, first proposed by Sørensen (1948), the fusion of two clusters depends on the most distant pair of objects instead of the closest. Thus, an object joins a cluster only when it is linked (relationship G_c, Section 8.2) to all the objects already members of that cluster. In the same way, two clusters can fuse only when all objects of the first are linked to all objects of the second, and vice versa.

Complete linkage rule

Coming back to the ponds of Ecological application 8.2, the steps of complete linkage clustering (Fig. 8.3) can be followed on the subgraphs shown in Fig. 8.2a. Examine the connected subgraphs and locate the D levels where completely connected groups of 2, 3, 4, and 5 objects are found. The pair (212, 214) is formed at $D = 0.4$ and the pair (431, 432) at $D = 0.5$. The next clustering step must wait until $D = 0.8$ since it is only then that pond 233 is finally linked (relationship G_c) to both ponds 431 and

432. Although there is a group of four completely linked ponds at $D = 0.937$, these ponds do not form a cluster in the agglomerative framework because pond 214 is already linked to pond 212, hence the two clusters (212, 214) and (233, 431, 432) cannot fuse at that level. It is only at $D = 1$ that ponds 212 and 214 are linked to all the ponds of cluster (233, 431, 432) and the five ponds form a single cluster.

In the complete linkage strategy, as a cluster grows, it becomes more and more difficult for new objects to join to it because the new objects should bear links with all the objects already in the cluster before being incorporated. For that reason, the growth of a cluster seems to move it away from the other objects or clusters in the analysis. According to Lance & Williams (1967c), this is equivalent to dilating the reference space in the neighbourhood of that cluster; see also Fig. 8.24c and related text. This effect is opposite to what was found in single linkage clustering, which contracted the reference space. In reference space A (Fig. 7.2), complete linkage produces maximally linked and rather spherical clusters, whereas single linkage may produce elongated clusters with loose chaining. Complete linkage clustering is often desirable in ecology, when one wishes to delineate clusters with clear discontinuities.

The intermediate (next subsection) and complete linkage clustering models have one drawback when compared to single linkage. In all cases where two incompatible candidates present themselves at the same time to be included in a cluster, algorithms use a preestablished and often arbitrary rule, called a "right-hand rule", to choose one and exclude the other. This problem does not exist in single linkage. An example is when two objects or two clusters could be included in a third cluster, while these two objects or clusters have not completed the linkage with each other. For this problem, Sørensen (1948) recommended the following: (1) choose the fusion leading to the largest cluster; (2) if equality persists, choose the fusion that most reduces the number of clusters; (3) as a last criterion, choose the fusion that minimizes the average distance within the cluster.

3 — Intermediate linkage clustering

Between the chaining of single linkage and the extreme space dilation of complete linkage, the most interesting solution in ecology may be a type of linkage clustering that approximately conserves the metric properties of reference space A; see also Fig. 8.24b. If the interest only lies in the clusters shown in the dendrogram, and not in the actual links between clusters shown by the subgraphs, the average clustering methods of Subsections 8.5.4 to 8.5.7 below could be useful since they also conserve the metric properties of the reference space.

Connectedness

In intermediate linkage clustering, the fusion criterion of an object or a cluster with another cluster is considered satisfied when a given proportion of the total possible number of links is reached. For example, if the criterion of connectedness (*Co*) is 0.5, two clusters are only required to share 50% of the possible links in order to fuse; in other words, the fusion is authorized when $£/n_1n_2 \geq Co$ where £ is the actual number of *between-group* links at sorting level *L,* while n_1 and n_2 are the numbers of objects in

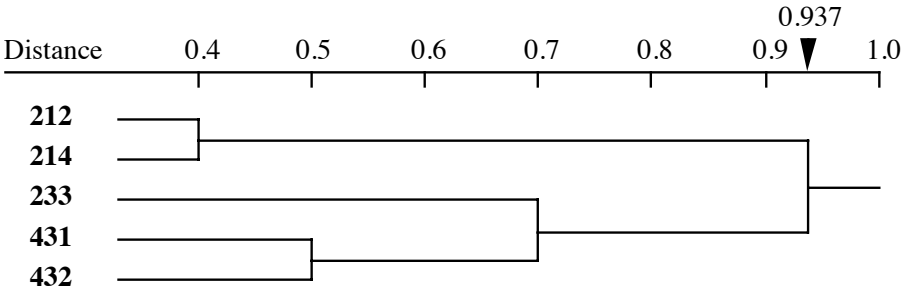


Figure 8.4 Intermediate linkage clustering, using the proportional link linkage criterion (Co = 50%), for the ponds of Ecological application 8.2.

Proportional link linkage

the two clusters, respectively. This criterion has been called *proportional link linkage* by Sneath (1966). Figure 8.4 gives the results of proportional link linkage clustering with Co = 50% for the pond example.

Sneath (1966) described three other ways of defining intermediate linkage clustering criteria: (1) by *integer link linkage*, which specifies the number of links required for the fusion of two groups (fusion when £ is larger than or equal to a fixed integer, *or else* when £ = $n_1 n_2$); (2) by their *absolute resemblance*, based on the sum of similarity links between the members of two clusters (the sum of the realized between-group similarities, $\sum S_{12}$, must reach a given threshold before fusion occurs); or (3) by their *relative resemblance*, where the sum of similarity links between the two clusters, $\sum S_{12}$, is divided by the number of between-group similarities, $n_1 n_2$ (fusion occurs at level L when the ratio $\sum S_{12}/n_1 n_2$ is greater than cL, where c is an arbitrary constant). When c equals 1, the method is called *average linkage clustering*. Similarities, not distances, must be used for criteria 2 and 3. These strategies are not combinatorial in the sense of Subsection 8.5.9.

4 — Unweighted arithmetic average clustering (UPGMA)

Average clustering

There are four methods of *average clustering* that conserve the metric properties of reference space A. These four methods were called "average linkage clustering" by Sneath & Sokal (1973), although they do not tally the links between clusters. As a consequence they are not object-linkage methods in the sense of the previous three subsections. They rely instead on the calculation of average distances among objects or the centroids of clusters. The four methods have nothing to do with Sneath's (1966) "average linkage clustering" described in the previous paragraph, so that we prefer calling them "average clustering". These methods (Table 8.2) result from the combinations of two dichotomies: (1) arithmetic average *versus* centroid clustering and (2) weighting *versus* non-weighting.

Table 8.2 Average clustering methods discussed in Subsections 8.5.4 to 8.5.7.

	Arithmetic average	Centroid clustering
Equal weights	4. Unweighted arithmetic average clustering (UPGMA)	6. Unweighted centroid clustering (UPGMC)
Unequal weights	5. Weighted arithmetic average clustering (WPGMA)	7. Weighted centroid clustering (WPGMC)

The first method in Table 8.2 is the *unweighted arithmetic average clustering* (Rohlf, 1963), also called "UPGMA" ("Unweighted Pair-Group Method using Arithmetic averages") by Sneath & Sokal (1973) or "group-average sorting" by Lance & Williams (1966a and 1967c). It is also called "average linkage" by SAS, SYSTAT and some other statistical packages, thus adding to the confusion pointed out in the previous paragraph. This is method = "average" in function ***hclust()*** of R. The lowest distance (or highest similarity) identifies the next cluster to be formed. Following this event, the method computes the arithmetic average of the distances between a candidate object and each of the cluster members or, in the case of a previously formed cluster, between all members of the two clusters. All objects receive equal weights in the computation. The distance matrix is updated and reduced in size at each clustering step. Clustering proceeds by agglomeration as the distance criterion increases, just as it does in single linkage clustering.

For the ponds of Section 8.2, UPGMA clustering proceeds as shown in Table 8.3 and Fig. 8.5. At step 1, the lowest distance value in the matrix is $D(212, 214) = 0.400$; hence the two objects fuse at level 0.400. As a consequence of this fusion, the distance values of these two objects with each of the remaining objects in the study must be averaged (values in the inner boxes in the table, step 1); this results in a reduction of the size of the distance matrix. Considering the reduced matrix (step 2), the smallest distance value is $D = 0.500$; it indicates that objects 431 and 432 fuse at level 0.500. The fused distance values are obtained by averaging the boxed values in the step 2 panel; this produces a new reduced distance matrix for the next step. In step 3, the lowest distance is 0.750; it leads to the fusion of the already-formed group (431, 432) with object 233 at level 0.750. In the example, this last fusion is the difficult point to understand. Before averaging the values, each one must be multiplied by the number of objects in the corresponding group. There is one object in group (233) and two in group (431, 432), so that the fused distance value is calculated as $[(0.9645 \times 1) + (0.93075 \times 2)]/3 = 0.942$. This is equivalent to averaging the six boxed distances in the top panel (larger box) with equal weights; the result would also be 0.942. So, this method is "unweighted" in the sense that it gives equal weights to the original

Table 8.3 Unweighted arithmetic average clustering (UPGMA) of the pond data. At each step, the lowest distance value is identified (italicized boldface value) and the two corresponding objects or groups are fused by averaging their distances as described in the text (boxes).

Objects	212	214	233	431	432	
212	—					**Step 1**
214	***0.400***	—				
233	1.000	0.929	—			
431	1.000	0.937	0.700	—		
432	1.000	0.786	0.800	0.500	—	
212-214		—				**Step 2**
233		0.9645	—			
431		0.9685	0.700	—		
432		0.8930	0.800	***0.500***	—	
212-214		—				**Step 3**
233		0.9645	—			
431-432		0.93075	***0.750***	—		
212-214		—				**Step 4**
233-431-432		***0.942***	—			

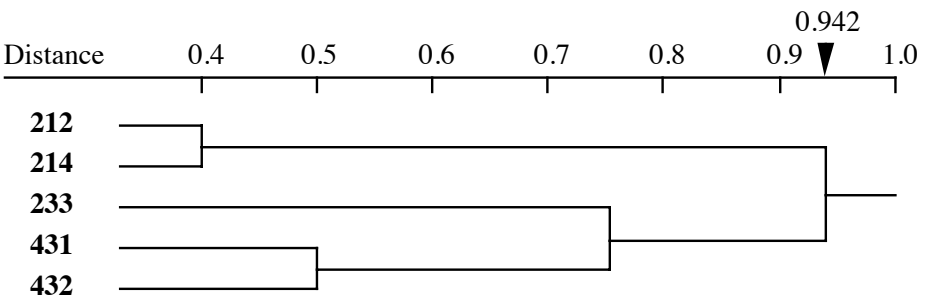


Figure 8.5 Unweighted arithmetic average clustering (UPGMA) of the ponds from Ecological application 8.2. This type of clustering only produces a dendrogram. It cannot be represented by connected subgraphs since it is not a linkage clustering as found in Figs. 8.2 to 8.4.

distances. To achieve this at step 3, one has to use weights that are equal to the number of objects in the groups. At step 4, there is a single remaining distance value; it is used to perform the last fusion at level 0.942. In the dendrogram, fusions are drawn at the identified distance levels.

Because it gives equal weights to the original distances, the UPGMA method assumes that the objects in each group form a representative sample of the corresponding larger groups of objects in the reference population under study. For that reason, UPGMA clustering should only be used in connection with simple random or systematic sampling designs if the results are to be extrapolated to a larger reference population.

Unlike the linkage clustering methods, information about the relationships between pairs of objects is lost in methods based on progressive reduction of the distance matrix, since only the relationships among groups are considered. This information can be extracted from the original distance matrix by making a list containing, for each fusion level, the lowest distance found between objects of the two groups. For the pond example, the *chain of primary connections* corresponding to the dendrogram would be made of the following links: (212, 214) for the first fusion level, (431, 432) for the second level, (233, 431) for the third level, and (214, 432) for the last level (Table 8.3). The topology obtained by UPGMA clustering may differ from that of single linkage. If this had been the case here, the chain of primary connections would have been different from that of single linkage clustering.

5 — Weighted arithmetic average clustering (WPGMA)

It often occurs in ecology that groups of objects, representing different regions of a territory, are of unequal sizes. Eliminating objects to equalize the clusters would mean discarding valuable information. However, the presence of a large group of objects, which are more similar *a priori* because of their common origin, may distort the UPGMA results when a fusion occurs with a smaller group of objects. Sokal & Michener (1958) proposed a solution to this problem, called *weighted arithmetic average clustering* ("WPGMA" in Sneath & Sokal, 1973: "Weighted Pair-Group Method using Arithmetic averages"; method = "mcquitty" in function ***hclust()*** of R). This solution consists in giving equal weights, when computing fusion distances, to the two *branches* of the dendrogram that are about to fuse. This is equivalent, when computing a fusion distance, to giving different weights to the original distances, i.e. down-weighting the distances of the largest group. Hence the name of the method.

Table 8.4 and Fig. 8.6 describe the WPGMA clustering sequence for the pond data. In this example, the only difference with UPGMA is the last fusion value. It is computed here by averaging the two distances from the previous step: (0.9645 + 0.93075)/2 = 0.947625. Weighted arithmetic average clustering increases the separation of the two main clusters, compared to UPGMA. This gives sharper contrast to the classification.

Table 8.4 Weighted arithmetic average clustering (WPGMA) of the pond data. At each step, the lowest distance value is identified (italicized boldface value) and the two corresponding objects or groups are fused by averaging their distances (boxes).

Objects	212	214	233	431	432	
212	—					**Step 1**
214	***0.400***	—				
233	1.000	0.929	—			
431	1.000	0.937	0.700	—		
432	1.000	0.786	0.800	0.500	—	
212-214		—				**Step 2**
233		0.9645	—			
431		0.9685	0.700	—		
432		0.8930	0.800	***0.500***	—	
212-214		—				**Step 3**
233		0.9645	—			
431-432		0.93075	***0.750***	—		
212-214		—				**Step 4**
233-431-432		***0.9476***	—			

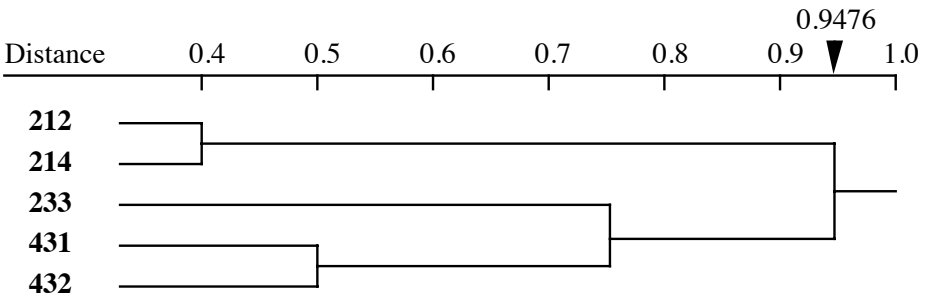


Figure 8.6 Weighted arithmetic average clustering (WPGMA) of the ponds from Ecological application 8.2. This type of clustering only produces a dendrogram. It cannot be represented by connected subgraphs since it is not a linkage clustering as found in Figs. 8.2 to 8.4.

6 — Unweighted centroid clustering (UPGMC)

Centroid

The *centroid* of a cluster of objects can be imagined as the type-object of the cluster, whether that object actually exists or is only a mathematical construct. In A-space (Fig. 7.2), the coordinates of the centroid of a cluster are computed by averaging the coordinates of the objects in the cluster.

Unweighted centroid clustering (Lance & Williams, 1967c; "UPGMC" in Sneath & Sokal, 1973: "Unweighted Pair-Group Centroid Method") is based on a simple geometric approach. This is method = "centroid" in function ***hclust()*** of R. Along a decreasing scale of distances, UPGMC proceeds to the fusion of objects or clusters presenting the lowest distance, as in the previous methods. At each step, the members of a cluster are replaced by their common centroid (i.e. "mean point"). The centroid is considered to represent a new object for the remainder of the clustering procedure; in the next step, one looks again for the pair of objects with the lowest distances, on which the fusion procedure is repeated.

Gower (1967) proposed the following formula for centroid clustering, where the distance of the centroid (**hi**) of objects or clusters **h** and **i** to a third object or cluster **g** is computed from the distances $D(\mathbf{h}, \mathbf{g})$, $D(\mathbf{i}, \mathbf{g})$, and $D(\mathbf{h}, \mathbf{i})$:

$$D(\mathbf{hi}, \mathbf{g}) = \frac{w_h}{w_h + w_i} D(\mathbf{h}, \mathbf{g}) + \frac{w_i}{w_h + w_i} D(\mathbf{i}, \mathbf{g}) - \frac{w_h w_i}{(w_h + w_i)^2} D(\mathbf{h}, \mathbf{i}) \qquad \textbf{(8.3)}$$

were the w's are weights given to the clusters. To simplify the symbolism, letters **g**, **h**, and **i** designate three objects considered in the course of clustering; they may also represent centroids of clusters obtained during previous clustering steps.

Gower's formula insures that the centroid **hi** of objects (or clusters) **h** and **i** is geometrically located on the line between **h** and **i**. In classical centroid clustering, the numbers of objects n_h and n_i in clusters **h** and **i** are taken as values for the weights w_h and w_i; these weights are 1 at the start of the clustering because there is then a single object per cluster. If initial weights are attached to individual objects, they may be used instead of 1's in eq. 8.3.

Centroid clustering may lead to reversals (Section 8.6). Some authors feel uncomfortable about reversals since they violate the ultrametric property (eq. 8.1); such violations make dendrograms more difficult to draw. A reversal is found with the pond example (Table 8.5, Fig. 8.7): the fusion distance found at step 4 is lower than that of step 3. The last fusion distance (step 4) is calculated as follows:

$$D[(233, 431\text{-}432), (212\text{-}214)] = \frac{1}{3} \times 0.8645 + \frac{2}{3} \times 0.70575 - \frac{2}{3^2} \times 0.625 = 0.61978$$

As indicated above, UPGMC clustering is geometrically interpreted as the fusion of objects into cluster centroids. Figure 8.8 presents the four clustering steps depicted

Table 8.5 Unweighted centroid clustering (UPGMC) of the pond data. At each step, the lowest distance value is identified (italicized boldface value) and the two corresponding objects or groups are fused using eq. 8.3.

Objects	212	214	233	431	432
212	—				**Step 1**
214	***0.400***	—			
233	1.000	0.929	—		
431	1.000	0.937	0.700	—	
432	1.000	0.786	0.800	0.500	—
212-214		—			**Step 2**
233		0.8645	—		
431		0.8685	0.700	—	
432		0.9730	0.800	***0.500***	—
212-214		—			**Step 3**
233		0.8645	—		
431-432		0.70575	***0.625***	—	
212-214		—			**Step 4**
233-431-432		***0.6198***	—		

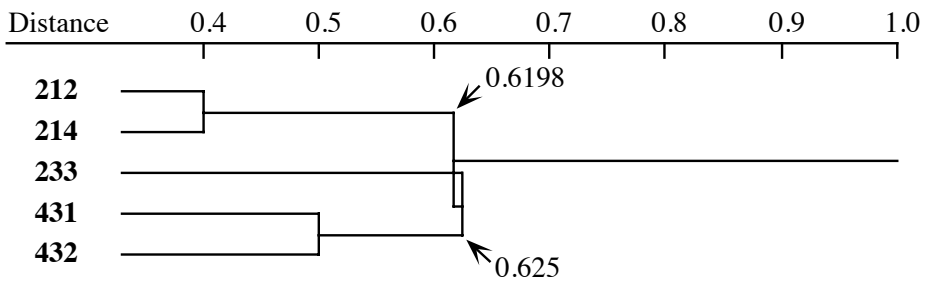


Figure 8.7 Unweighted centroid clustering (UPGMC) of the ponds from Ecological application 8.2. This type of clustering only produces a dendrogram. The reversal in the structure of the dendrogram is explained in Section 8.6.

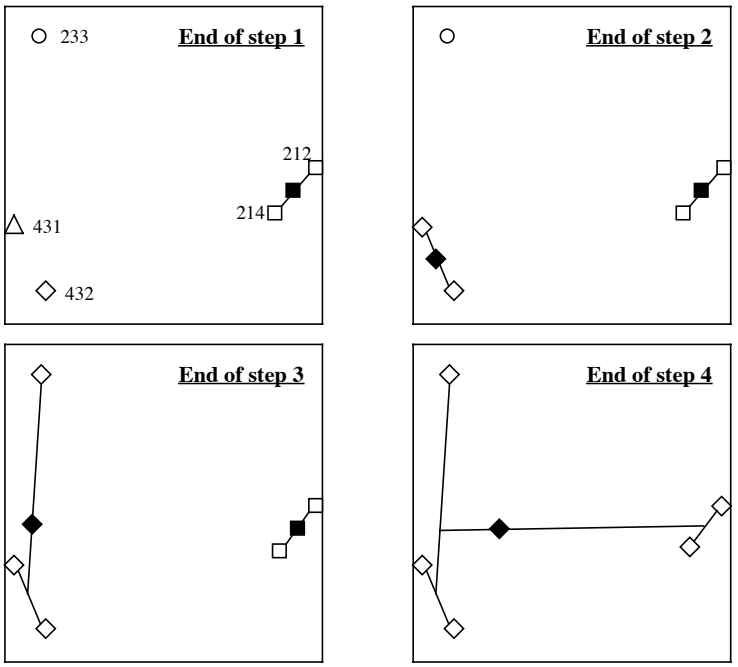


Figure 8.8 The four UPGMC clustering steps of Fig. 8.7 are drawn in A-space. Objects are represented by open symbols and centroids by dark symbols; object identifiers are shown in the first panel only. Separate clusters are represented by different symbols. The first two principal coordinates, represented here, account for 87% of the variation of the objects in the full A-space.

by the dendrogram, drawn in an A-space (Fig. 7.2) reduced to two dimensions through principal coordinate analysis (Section 9.3) to facilitate representation. At the end of each step, a new cluster is formed and its centroid is represented at the *centre of mass* of the cluster members; examine especially steps 3 and 4.

Unweighted centroid clustering may be used with any measure of distance, but Gower's formula (eq. 8.3) only retains its geometric properties for distances that are Euclidean (Table 7.2). Note also that in this clustering procedure, the links between clusters do not depend upon identifiable pairs of objects; this was also the case with clustering methods 4 and 5 above. Thus, if the chain of primary connections is needed, its links be identified by the method described at the end of Subsection 8.5.4.

The assumptions of this model with respect to representativeness of the observations are the same as in UPGMA since equal weights are given to all objects during clustering. So, UPGMC should only be used in connection with simple random or systematic sampling designs if the results are to be extrapolated to a larger reference population. When the branching pattern of the dendrogram displays asymmetry (many

more objects in one branch than in the other), this can be attributed to the structure of the reference population if the sampling design was random.

7 — Weighted centroid clustering (WPGMC)

Weighted centroid clustering was proposed by Gower (1967). This is method = "median" in function ***hclust()*** of R. It plays the same role with respect to UPGMC as WPGMA (method 5) plays with respect to UPGMA (method 4). When many observations of a given type have been included in the set to be clustered, next to other types that were not as well-sampled (sampling design other than simple random or systematic), the positions of the centroids may be biased towards the over-represented types, which in turn could distort the clustering. In *weighted centroid clustering*, which Sneath & Sokal (1973) called "WPGMC" ("Weighted Pair-Group Centroid Method"), this problem is corrected by giving equal weights to two clusters on the verge of fusing, independently of the number of objects in each cluster. To achieve this, eq. 8.3 is replaced by the following formula (Gower, 1967):

$$D(\mathbf{hi}, \mathbf{g}) = \frac{1}{2}[D(\mathbf{h}, \mathbf{g}) + D(\mathbf{i}, \mathbf{g})] - \frac{1}{4}D(\mathbf{h}, \mathbf{i}) \qquad \textbf{(8.4)}$$

The five ponds of Ecological application 8.2 are clustered as described in Table 8.6 and Fig. 8.9. The last fusion distance (step 4), for example, is calculated as follows:

$$D[(233, 431\text{-}432), (212\text{-}214)] = \frac{1}{2}[0.8645 + 0.70575] - \frac{1}{4} \times 0.625 = 0.62888$$

This value is the level at which the last fusion takes place. Note that no reversal appears in this result, although WPGMC can produce reversals like UPGMC clustering.

As indicated above, WPGMC clustering is geometrically interpreted as the fusion of objects into cluster centroids. Figure 8.10 presents the four clustering steps depicted by the dendrogram, in A-space (Fig. 7.2) reduced to two dimensions through principal coordinate analysis (Section 9.3) to facilitate representation. At the end of each step, a new cluster is formed and its centroid is represented at the *geometric centre* of the last line drawn; examine especially steps 3 and 4 and compare to Fig. 8.8.

8 — Ward's minimum variance method

Ward's (1963) minimum variance method is related to the centroid methods (Subsections 8.5.6 and 8.5.7 above) in that it also leads to a geometric representation in which cluster centroids play a key role. To form clusters, the method minimizes an *objective function* which is, in this case, the same "squared error" criterion as that used in multivariate analysis of variance.

Objective function

Table 8.6 Weighted centroid clustering (WPGMC) of the pond data. At each step, the lowest distance value is identified (italicized boldface value) and the two corresponding objects or groups are fused using eq. 8.4.

Objects	212	214	233	431	432	
212	—					**Step 1**
214	***0.400***	—				
233	1.000	0.929	—			
431	1.000	0.937	0.700	—		
432	1.000	0.786	0.800	0.500	—	
212-214		—				**Step 2**
233		0.8645	—			
431		0.8685	0.700	—		
432		0.7930	0.800	***0.500***	—	
212-214		—				**Step 3**
233		0.8645	—			
431-432		0.70575	***0.625***	—		
212-214		—				**Step 4**
233-431-432		***0.6289***	—			

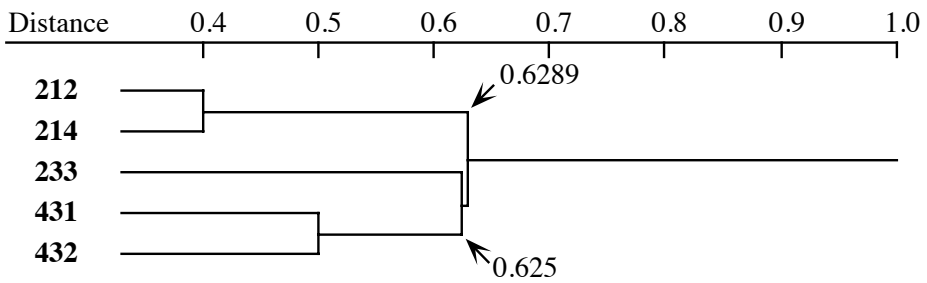


Figure 8.9 Weighted centroid clustering (WPGMC) of the ponds from Ecological application 8.2. This type of clustering only produces a dendrogram.

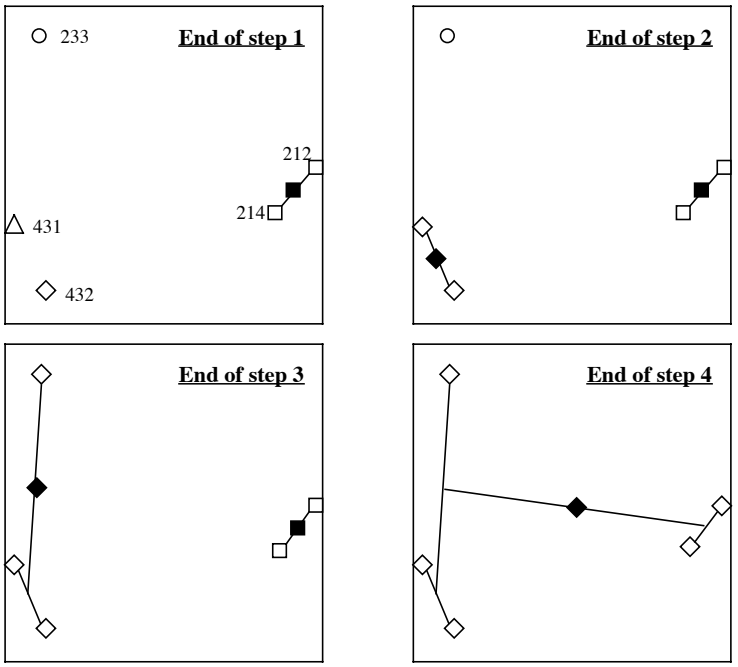


Figure 8.10 The four WPGMC clustering steps of Fig. 8.9 are drawn in A-space. Objects are represented by open symbols and centroids by dark symbols; object identifiers are shown in the first panel only. Distinct clusters are represented by different symbols. The first two principal coordinates, represented here, account for 87% of the variation of the objects in the full A-space.

At the beginning of the procedure, each objects is in a cluster of its own, so that the distance of an object to its cluster's centroid is 0; hence, the sum of all these distances is also 0. As clusters form, the centroids move away from actual object coordinates and the sums of the squared distances from the objects to the centroids increase. At each clustering step, Ward's method finds the pair of objects or clusters whose fusion increases as little as possible the sum, over all groups formed so far, of the *squared distances* between objects and cluster centroids; that sum is the total within-group sum-of-squares. The distance of object $\mathbf{x}_i$ to centroid $\mathbf{m}$ of its cluster is computed using the squared Euclidean distance formula (eq. 7.33) over the various descriptors $\mathbf{y}_j$ ($j = 1 \ldots p$):

$$\sum_{j=1}^{p} [y_{ij} - m_j]^2$$

The centroid **m** of a cluster was defined at the beginning of Subsection 8.5.6. The sum of the squared distances of all objects in cluster k to their common centroid, which is called "error" in ANOVA (hence the symbol e_k^2), is the sum of the squared Euclidean distances between the members of the cluster and its centroid:

Squared error

Error in cluster k: $$e_k^2 = \sum_{i=1}^{n_k} \sum_{j=1}^{p} [y_{ij}^{(k)} - m_j^{(k)}]^2 \quad \textbf{(8.5)}$$

where $y_{ij}^{(k)}$ is the value of descriptor $\mathbf{y}_j$ for an object i member of group (k) and $m_j^{(k)}$ is the mean value of descriptor j over all members of group k. e_k^2 is used as a measure of the tightness of a cluster. If all data points in a cluster have the same coordinates in multidimensional space, or there is a single point in a cluster, the within-cluster variation is 0. Alternatively, the within-cluster sums of squared errors e_k^2 can be computed as the mean of the squared distances among cluster members:

Error in cluster k: $$e_k^2 = \left[\sum_{h,i=1}^{n_k} D_{hi}^2 \right] / n_k \quad \textbf{(8.6)}$$

where the D_{hi}^2 are the squared distances among objects in cluster k (Table 8.7) and n_k is the number of objects in that cluster. Equations 8.5 and 8.6 have already been shown in Box 6.1 (eqs. 6.55 and 6.56); they both allow the calculation of the squared error statistic. The equivalence of these two equations is stated in a theorem whose demonstration is found in Kendall & Stuart (1963, parag. 2.22) for the univariate case and in Legendre & Fortin (2010, Appendix 1) for the multivariate case. Numerical examples illustrating the calculation of eqs. 8.5 and 8.6 are given at the end of Section 8.8 on K-means partitioning.

The sum of squared errors E_K^2, over all K clusters corresponding to a given partition, is the criterion to be minimized at each clustering step:

Sum of squared errors

Total error, K clusters: $$E_K^2 = \sum_{k=1}^{K} e_k^2 \quad \textbf{(8.7)}$$

At each clustering step, two objects or clusters **h** and **i** are merged into a new cluster **hi**, as in previous sections. Since changes occurred only in groups **h**, **i**, and **hi**, the change in the overall sum of squared errors, $\Delta E_{\mathbf{hi}}^2$, can be computed from the changes that occurred in these groups only:

Change in total error: $$\Delta E_{\mathbf{hi}}^2 = e_{\mathbf{hi}}^2 - e_{\mathbf{h}}^2 - e_{\mathbf{i}}^2 \quad \textbf{(8.8)}$$

Table 8.7 Ward's minimum variance clustering of the pond data. Step 1 of the table contains *squared distances* computed as D^2 from the distance values in the upper panels of Tables 8.3 to 8.6. At each step, the lowest squared distance is identified (italicized boldface value) and the two corresponding objects or groups are fused using eq. 8.10.

Objects	212	214	233	431	432	
212	—					**Step 1**
214	***0.16000***	—				
233	1.00000	0.86304	—			
431	1.00000	0.87797	0.49000	—		
432	1.00000	0.61780	0.64000	0.25000	—	
212-214		—				**Step 2**
233		1.18869	—			
431		1.19865	0.49000	—		
432		1.02520	0.64000	***0.25000***	—	
212-214		—				**Step 3**
233		1.18869	—			
431-432		1.54288	***0.67000***	—		
212-214		—				**Step 4**
233-431-432		***1.6795***	—			

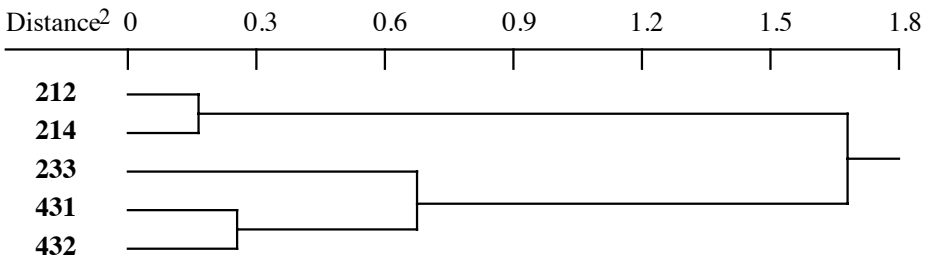


Figure 8.11 Ward's minimum variance clustering of the ponds from Ecological application 8.2. The scale of this dendrogram is here the squared distances computed in Table 8.7.

Table 8.8 Clustering steps in Ward's minimum variance clustering for the pond data. The objects are renamed 1 to 5 for shortness. K is the number of clusters, represented by underscored groups of objects. The total sum of squares (SS_{Total}) of the 5 objects is 1.37976 (eq. 8.6). SS_{Within} is also computed using eq. 8.6; $SS_{Among} = SS_{Total} - SS_{Within}$. Between clustering levels, $\Delta E^2_{\mathbf{hi}}$ is computed using eq. 8.8 or by difference between the successive values of SS_{Within} or SS_{Among}. D^2_{min} was computed in Table 8.7; D_{min} is the square root of D^2_{min}.

K	Objects	SS_{Within}	SS_{Among}	$\Delta E^2_{\mathbf{hi}}$	D^2_{min}	D_{min}
5	1 2 3 4 5	0	1.37976			
				0.08000	0.16000	0.40000
4	1 2 3 4 5	0.08000	1.29976			
				0.12500	0.25000	0.50000
3	1 2 3 4 5	0.20500	1.17476			
				0.33500	0.67000	0.81854
2	1 2 3 4 5	0.54000	0.83976			
				0.83976	1.67952	1.29596
1	1 2 3 4 5	1.37976	0			

It can be shown that this change depends only on the distance between the centroids of clusters **h** and **i** and on their numbers of objects n_h and n_i:

Change in total error: $$\Delta E^2_{\mathbf{hi}} = \frac{n_h n_i}{n_h + n_i} \sum_{j=1}^{p} \left[m_j^{(\mathbf{h})} - m_j^{(\mathbf{i})} \right]^2 \quad \textbf{(8.9)}$$

So, one way of identifying the next fusion would be to compute the $\Delta E^2_{\mathbf{hi}}$ statistic for all possible pairs and select the pair that generates the smallest value for the next fusion. An easier way is to use the following updating formula to compute the fusion distances between the new cluster **hi** and all other objects or clusters **g** in the agglomeration table (Table 8.7):

$$D^2(\mathbf{hi}, \mathbf{g}) = \frac{n_h + n_g}{n_h + n_i + n_g} D^2(\mathbf{h}, \mathbf{g}) + \frac{n_i + n_g}{n_h + n_i + n_g} D^2(\mathbf{i}, \mathbf{g}) - \frac{n_g}{n_h + n_i + n_g} D^2(\mathbf{h}, \mathbf{i}) \quad \textbf{(8.10)}$$

Wishart (1969) and Kaufman & Rousseeuw (1990) demonstrated mathematically that the smallest distance computed using this updating formula corresponds to the fusions that obeys Ward's (1963) criterion at each clustering step. Note that *squared distances* are used instead of distances in eq. 8.10 and in Table 8.7. This algorithm is called Ward.D2. Table 8.8 shows the clustering steps for the example data.

An alternative formula found in some manuals, e.g. Jain & Dubes (1988), uses distances D instead of D^2 in eq. 8.10. This formula is implemented in some programs and functions; it will be called the Ward.D algorithm*. One can show that the resulting updating formula produces cluster fusions that do not necessarily minimize the change in total error (eq. 8.8), so the clustering does not follow Ward's rule.

Ward clustering is a hierarchical agglomerative method; it proceeds sequentially by binary group fusions. Each fusion, going from (k+1) to k groups, satisfies Ward's criterion. This hierarchical method does not guarantee, however, that globally, all partitions into k = {(n–1), (n–2), ..., 4, 3, 2} groups satisfy that criterion. One should use K-means partitioning (Section 8.8) to obtain a partition into a specified number of groups (K) that minimizes the sum of residual sums-of-squares.

Dendrograms for Ward's clustering may be represented along a variety of scales although these dendrograms all represent the same clustering topology.

• In Fig. 8.11, the dendrogram is drawn using the scale of *squared distances* computed in Table 8.7.

• One can compute the *square roots of the fusion distances* of Table 8.7 and draw the dendrogram accordingly. This solution, illustrated in Fig. 8.12a, is often used in computer programs and functions, including ***agnes()*** of CLUSTER in R; it removes the distortions created by squaring the distances. It is especially suitable when one wants to compare the fusion distances of Ward's clustering to the original distances, either graphically (Shepard-like diagrams, Fig. 8.24) or numerically (cophenetic correlations, Subsection 8.12.2).

• The sum of squared errors E_K^2 (eq. 8.7) is used in some computer programs as the dendrogram scale. This statistic is also called the *total error sum of squares* (TESS) by Everitt (1980) and other authors. This solution is illustrated in Fig. 8.12b.

TESS

• The SAS package recommends two scales for Ward's clustering. The first one is the proportion of variance (R^2) accounted for by the clusters at any given partition level. It is computed as the total sum of squares (i.e. the sum of squared distances from the centroid of all objects) minus the within-cluster squared errors E_K^2 of eq. 8.7 for the given partition, divided by the total sum of squares. R^2 decreases as clusters grow. When all the objects are lumped in a single cluster, the resulting one-cluster partition does not explain any of the objects' variation so that $R^2 = 0$. The second scale recommended by SAS is called the *semipartial* R^2. It is computed as the between-

* In R, function ***hclust()*** of package STATS with method = "ward" implements the Ward.D algorithm (at least up to version 2.12.1), whereas function ***agnes()*** of package CLUSTER with method = "ward" implements the Ward.D2 algorithm. ***hclust()*** can be made to produce results corresponding to the Ward.D2 algorithm by using squared distances in the input matrix. To obtain the Ward.D2 dendrogram with correct scale, one has to modify the $height element of the output list to make it contain the square roots of the height values before calling ***plot()***.

cluster sum of squares divided by the (corrected) total sum of squares. This statistic increases as the clusters grow.

Because the Ward method minimizes the sum of within-group sums of squares (squared error criterion), the clusters tend to be hyperspherical, i.e. spherical in multidimensional A-space, and to contain roughly equal numbers of objects if the observations are evenly distributed through A-space. The same applies to the centroid methods of the previous subsections. This may be seen as either an advantage or a problem, depending on the researcher's conceptual model of a cluster.

9 — General agglomerative clustering model

Lance & Williams (1966a, 1967c) proposed a general model that encompasses all the agglomerative clustering methods presented up to now, except intermediate linkage (Subsection 8.5.3). The general model offers the advantage of being translatable into a single, simple computer program, so that it is used in most statistical packages that offer agglomerative clustering, including R. The general model allows one to select an agglomerative clustering model by choosing the values of four parameters called α_h, α_i, β, and γ that determine the clustering strategy. This model only outputs the

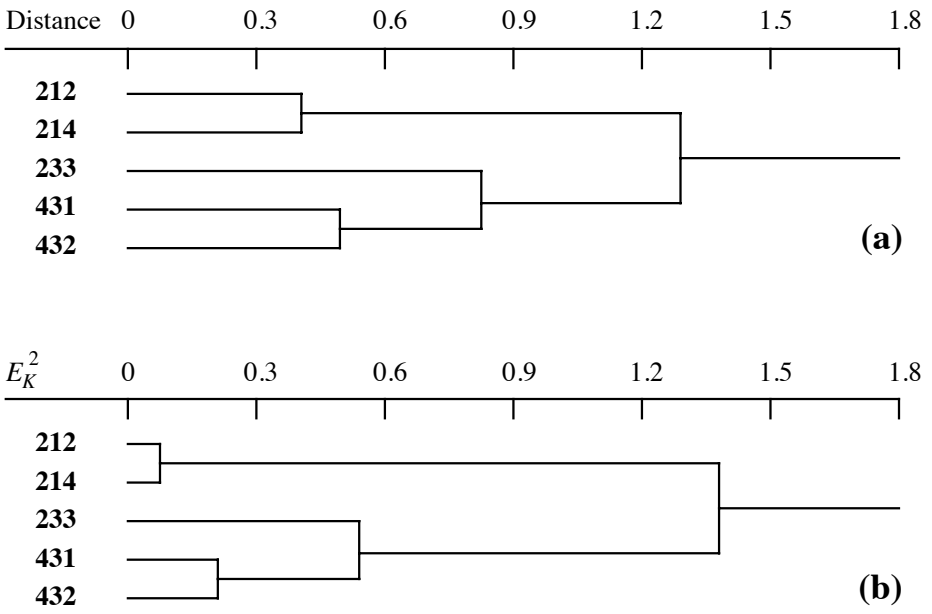


Figure 8.12 Ward's minimum variance clustering of the ponds from Ecological application 8.2. The scale of dendrogram (a) is the square root of the squared distances computed in Table 8.7; that scale can be compared to the original distances. In dendrogram (b), it is the E_K^2 (or TESS) statistic.

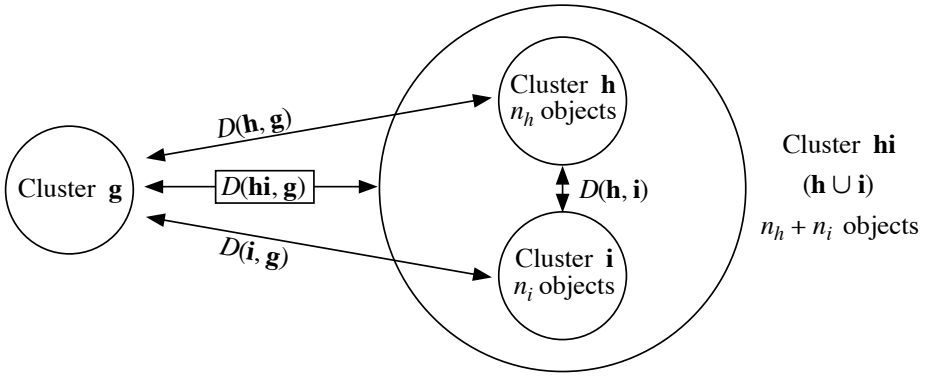


Figure 8.13 In combinatorial clustering methods, the distance between a cluster **hi**, resulting from the fusion of two previously formed clusters **h** and **i**, and an external cluster **g** is a function of the three distances between (**h** and **i**), (**h** and **g**), and (**i** and **g**), and of the number of objects in **h**, **i**, and **g**.

branching pattern of the clustering tree (the dendrogram), as it was the case for the methods described in Subsections 8.5.4 to 8.5.8. For the linkage clustering strategies (Subsections 8.5.1 to 8.5.3), the list of primary links responsible for cluster formation can be obtained afterwards by comparing the dendrogram to the distance matrix.

Combinatorial method

The model of Lance & Williams is limited to *combinatorial* clustering methods, i.e. those for which the distance $D(\mathbf{hi}, \mathbf{g})$ between an external cluster **g** and a cluster **hi**, resulting from the prior fusion of clusters **h** and **i**, is a function of the three distances $D(\mathbf{h}, \mathbf{g})$, $D(\mathbf{i}, \mathbf{g})$, and $D(\mathbf{h}, \mathbf{i})$ and also, eventually, of the numbers n_h, n_i, and n_g of objects in clusters **h**, **i**, and **g**, respectively (Fig. 8.13). Individual objects are considered to be single-member clusters. Since the distance of cluster **hi** to an external cluster **g** can be computed from the above six values, **h** and **i** can be condensed into a single row and a single column in the updated distance matrix; following that, the clustering proceeds as in the tables of the previous subsections. Since the new distances at each step can be computed by *combining* those from the previous step, it is not necessary for a computer program to retain the original distance matrix or data set. Non-combinatorial methods do not have this property. For distances, the general model for combinatorial methods is the following:

$$D(\mathbf{hi}, \mathbf{g}) = \alpha_h D(\mathbf{h}, \mathbf{g}) + \alpha_i D(\mathbf{i}, \mathbf{g}) + \beta D(\mathbf{h}, \mathbf{i}) + \gamma |D(\mathbf{h}, \mathbf{g}) - D(\mathbf{i}, \mathbf{g})| \qquad \textbf{(8.11)}$$

When using similarities, the combinatorial equation is:

$$S(\mathbf{hi}, \mathbf{g}) = (1 - \alpha_h - \alpha_i - \beta) + \alpha_h S(\mathbf{h}, \mathbf{g}) + \alpha_i S(\mathbf{i}, \mathbf{g}) + \beta S(\mathbf{h}, \mathbf{i}) - \gamma |S(\mathbf{h}, \mathbf{g}) - S(\mathbf{i}, \mathbf{g})| \qquad \textbf{(8.12)}$$

Table 8.9 Values of parameters α_h, α_i, β, and γ in Lance and Williams' general model for combinatorial agglomerative clustering. Modified from Sneath & Sokal (1973) and Jain & Dubes (1988).

Clustering method	α_h	α_i	β	γ	Effect on space A
Single linkage	1/2	1/2	0	–1/2	Contracting*
Complete linkage	1/2	1/2	0	1/2	Dilating*
UPGMA	$\frac{n_h}{n_h+n_i}$	$\frac{n_i}{n_h+n_i}$	0	0	Conserving*
WPGMA	1/2	1/2	0	0	Conserving
UPGMC	$\frac{n_h}{n_h+n_i}$	$\frac{n_i}{n_h+n_i}$	$\frac{-n_h n_i}{(n_h+n_i)^2}$	0	Conserving
WPGMC	1/2	1/2	–1/4	0	Conserving
Ward's	$\frac{n_h+n_g}{n_h+n_i+n_g}$	$\frac{n_i+n_g}{n_h+n_i+n_g}$	$\frac{-n_g}{n_h+n_i+n_g}$	0	Conserving
Flexible	$\frac{1-\beta}{2}$	$\frac{1-\beta}{2}$	$-1 \le \beta < 1$	0	Contracting if $\beta \approx 1$ Conserving if $\beta \approx -.25$ Dilating if $\beta \approx -1$

* Terms used by Sneath & Sokal (1973).

Clustering proceeds in the same way for all combinatorial agglomerative methods. As the distances increases, a new cluster is obtained by the fusion of the two closest objects or groups, after which the algorithm proceeds to the fusion of the two corresponding rows and columns in the distance (or similarity) matrix using eq. 8.11 or 8.12. The matrix is thus reduced by one row and one column at each step. Table 8.9 gives the values of the four parameters for the most commonly used combinatorial agglomerative clustering strategies. Values of the parameters for some other clustering strategies are given by Gordon (1996a).

In the case of equality between two mutually exclusive pairs, the decision may be made on an arbitrary basis (the so-called "right-hand rule" used in most computer programs) or based upon ecological criteria as, for example, Sørensen's criteria reported at the end of Subsection 8.5.2, or those described in Subsection 8.9.1.

In several strategies, $\alpha_h + \alpha_i + \beta = 1$, so that the term $(1 - \alpha_h - \alpha_i - \beta)$ becomes zero and disappears from eq. 8.12. One can show how the values chosen for the four

parameters make the general equation correspond to each specific clustering method. For single linkage clustering, for instance, the general equation becomes:

$$D(\mathbf{hi}, \mathbf{g}) = \frac{1}{2}[D(\mathbf{h}, \mathbf{g}) + D(\mathbf{i}, \mathbf{g}) - |D(\mathbf{h}, \mathbf{g}) - D(\mathbf{i}, \mathbf{g})|]$$

The last term (absolute value) corrects the largest of the two distances $D(\mathbf{h}, \mathbf{g})$ and $D(\mathbf{i}, \mathbf{g})$, making it equal to the smallest one. Hence, $D(\mathbf{hi}, \mathbf{g}) = \min[D(\mathbf{h}, \mathbf{g}), D(\mathbf{i}, \mathbf{g})]$. In other words, the distance between a newly-formed cluster **hi** and some other cluster **g** becomes equal to the smallest of the distance values previously computed between the two original clusters (**h** and **i**) and **g**.

Intermediate linkage clustering is not a combinatorial strategy. All along the clustering procedure, it is necessary to refer to the original association matrix in order to calculate the connectedness of pairs of clusters. This is why it cannot be obtained using the Lance & Williams general agglomerative clustering model.

10 — Flexible clustering

Lance & Williams (l966a, 1967c) proposed to vary parameter β (eq. 8.11 or 8.12) between –1 and +1 to obtain a series of intermediate solutions between single linkage chaining and the space dilation of complete linkage. The method is called *beta-flexible clustering* by some authors. Lance & Williams (*ibid*.) have shown that, if the other parameters are constrained a follows:

$$\alpha_h = \alpha_i = (1 - \beta)/2 \qquad \text{and} \qquad \gamma = 0$$

the resulting clustering is always ultrametric (no reversals; Section 8.6).

When β is close to 1, strong chaining is obtained. As β decreases and becomes negative, space dilation increases. The space properties are conserved for small negative values of β near –0.25. Figure 8.14 shows the effect of varying β in the clustering of 20 objects. Like weighted centroid clustering, flexible clustering is compatible with all association measures except correlation coefficients.

Ecological application 8.5

Pinel-Alloul *et al*. (1990) studied phytoplankton in 54 lakes of Québec to determine the effects of acidification, physical and chemical characteristics, and lake morphology on species assemblages. Phytoplankton was enumerated into five main taxonomic categories (microflagellates, chlorophytes, cyanophytes, chrysophytes, and pyrrophytes). The data were normalized using the generalized form of the Box-Cox method that finds the best normalizing transformation for all species (Subsection 1.5.6). A Gower (S_{19}) similarity matrix, computed among lakes, was subjected to flexible clustering with parameter $\beta = -0.25$. Six clusters were found, which were roughly distributed along a NE-SW geographic axis and corresponded to increasing concentrations of total phytoplankton, chlorophytes, cyanophytes, and microflagellates. Explanation of the phytoplankton-based lake typology was sought by comparing it to the environmental variables (Subsection 10.2.1).

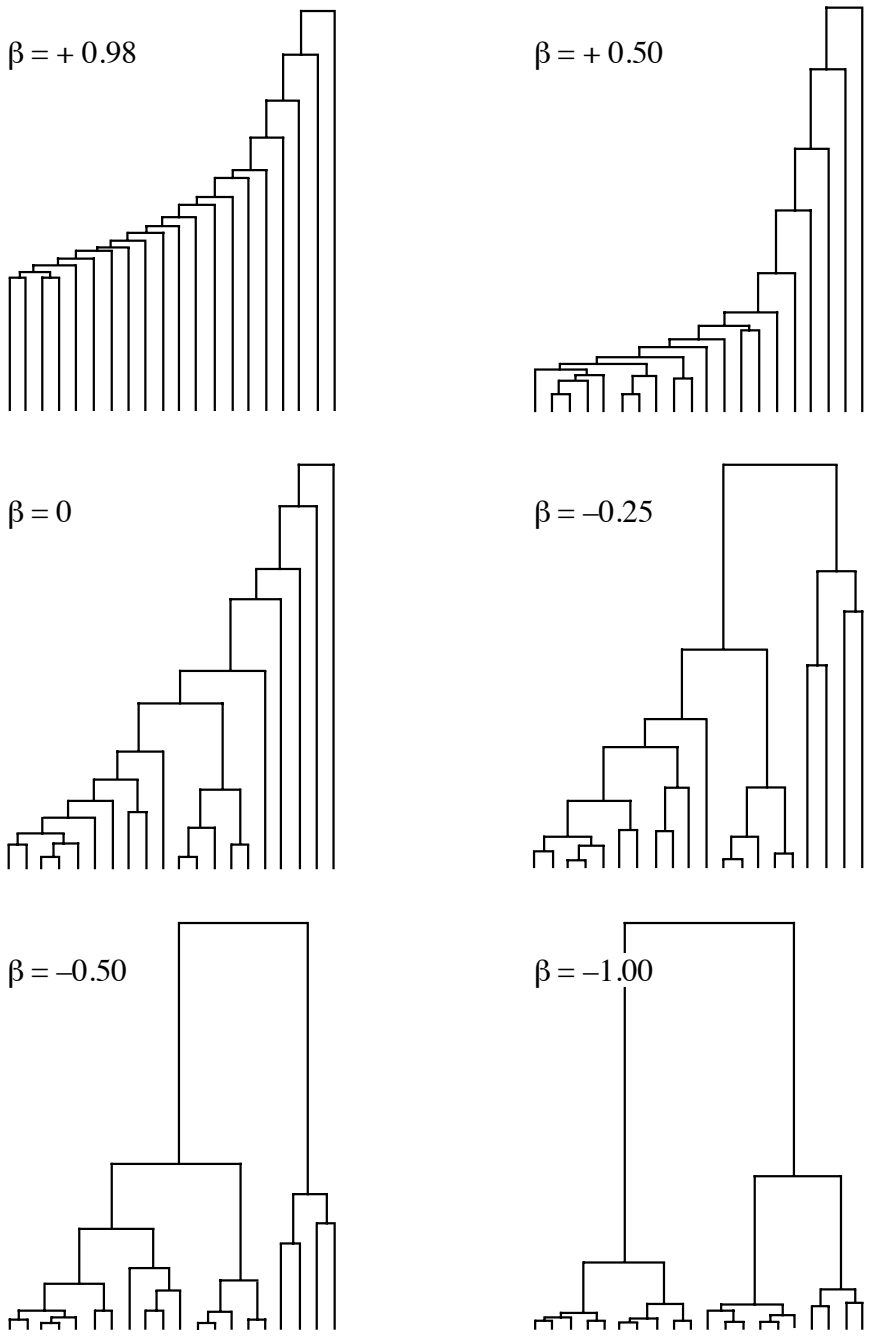


Figure 8.14 Flexible clustering of 20 objects for six values of β. The measure of association is the squared Euclidean distance D_1^2. Adapted from Lance & Williams (1967c: 376).

11 — Information analysis

Information analysis is a Q-mode clustering method developed for ecological purposes by Williams *et al*. (1966) and Lance & Williams (1966b). It does not go through the usual steps of distance calculation followed by clustering. It is a direct method of clustering based on information measures.

Entropy

Shannon's formula (eq. 6.1) can be used to measure the diversity or information in a frequency or probability distribution:

$$H = -\sum_{j=1}^{p} p_j \log p_j$$

Information analysis is a type of unweighted centroid clustering, adapted to species presence-absence data. At each step, the two objects or clusters causing the smallest gain in within-group diversity (or information) are fused. As a consequence, the clusters are as homogeneous as possible in terms of species composition.

The method could be applied to species abundance data divided into a small number of classes but, in practice, it is mostly used with presence-absence data. The information measure described below is not applicable to raw abundance data because the number of different states would then vary from one species to another, which would give them different weights in the overall measure.

To illustrate the method, the pond zooplankton counts used in Chapter 7 to illustrate the calculation of coefficient S_{23} (eq. 7.30) are transformed here into presence-absence data:

Species j	Ponds					p_j	$(1 - p_j)$
	212	214	233	431	432		
1	1	1	0	0	0	0.4	0.6
2	0	0	1	1	0	0.4	0.6
3	0	1	1	0	1	0.6	0.4
4	0	0	1	1	1	0.6	0.4
5	1	1	0	0	0	0.4	0.6
6	0	1	0	1	1	0.6	0.4
7	0	0	0	1	1	0.4	0.6
8	1	1	0	0	0	0.4	0.6

Total information in this group of ponds is computed using an information measure derived from the following reasoning (Lance & Williams, 1966b). The entropy of each species presence-absence descriptor j is calculated on the basis of the probabilities of presence p_j and absence $(1 - p_j)$ of species j, which are written in the right-hand part of

the table. The probability of presence is estimated as the number of ponds where species j is present, divided by the total number of ponds *in the cluster under consideration* (here, the group of five ponds). The probability of absence is estimated likewise, using the number of ponds where species j is absent. The entropy of species j is therefore:

$$H(j) = -[p_j \log p_j + (1 - p_j) \log(1 - p_j)] \quad \text{for } 0 < p_j < 1 \qquad \textbf{(8.13)}$$

The base of the logarithms is indifferent, as long as the same base is used throughout the calculations. Natural logarithms are used throughout the present example. For the first species, $H(1)$ would be:

$$H(1) = -[0.4 \log_e(0.4) + 0.6 \log_e(0.6)] = 0.673$$

The information of the conditional probability table can be calculated by summing the entropies per species, considering that all species have the same weight. Since the measure of *total information* in the group must also take into account the number of objects in the cluster, it is defined as follows:

$$I = -n \sum_{j=1}^{p} [p_j \log p_j + (1 - p_j) \log (1 - p_j)] \quad \text{for } 0 < p_j < 1 \qquad \textbf{(8.14)}$$

where p is the number of species represented in the group of n objects (ponds). For null probabilities, $\lim_{p \to 0} [-p \log (p)] = 0$. For the group of 5 ponds above,

$$I = -5 [8 (-0.673)] = 26.920$$

If I is to be expressed as a function of the number a_j of ponds with species j present, instead of a function of probabilities $p_j = a_j/n$, it can be shown that the following formula is equivalent to eq. 8.14:

$$I = np \log n - \sum_{j=1}^{p} [a_i \log a_j + (n - a_j) \log (n - a_j)] \qquad \textbf{(8.15)}$$

I is zero when all ponds in a group contain the exact same set of species. Like entropy H, I has no upper limit; its maximum value depends on the number of species present in the study.

At each clustering step, three series of values are computed: (a) the total information I in each group, which is 0 at the beginning of the process since each object (pond) then forms a distinct cluster; (b) the value of I for all possible combinations of groups taken two at a time; and (c) the increase of information ΔI resulting from each possible fusion. As recommended by Sneath & Sokal (1973), all these values can be written in a matrix, initially of dimension $n \times n$ which decreases as clustering proceeds. For the example data, values (a) of information in each group are

placed on the diagonal, values (b) of I in the lower triangle, and values (c) of ΔI in the upper triangle, in italics.

Ponds	Ponds				
	212	214	233	431	432
212	0	*2.773*	*8.318*	*9.704*	*9.704*
214	2.773	0	*8.318*	*9.704*	*6.931*
233	8.318	8.318	0	*4.159*	*4.159*
431	9.704	9.704	4.159	0	*2.773*
432	9.704	6.931	4.159	2.773	0

The value ΔI for two groups is found by subtracting the two corresponding values I, on the diagonal, from the value I of their combination in the lower triangle. Values on the diagonal are 0 in this first calculation matrix, so that values in the upper triangle are the same as in the lower triangle, but this will not be the case in subsequent matrices.

The first fusion is identified by the lowest ΔI value found in the upper triangle. This value is 2.773 for pairs (212, 214) and (431, 432), which therefore fuse. A new matrix of I values is computed:

Groups	Groups		
	212 214	233	431 432
212-214	2.773	*10.594*	*15.588*
233	13.367	0	*4.865*
431-432	21.134	7.638	2.773

This time, the ΔI values in the upper triangle differ from the I's in the lower triangle since there are now I values different from 0 on the diagonal. The ΔI corresponding to group (212, 214, 431, 432), for example, is computed as: 21.134 – 2.773 – 2.773 = 15.588. The lowest value of ΔI is for the group (233, 431, 432), which therefore fuses at this step at information level I = 7.638.

For the last clustering step, the only I value to calculate in the lower triangle is for the cluster containing the five ponds. This value, computed above after eq. 8.14, is 26.920. ΔI is then 26.920 – 2.773 – 7.638 = 16.509.

Groups	Groups	
	212 214	233 431-432
212-214	2.773	*16.509*
233-431-432	26.920	7.638

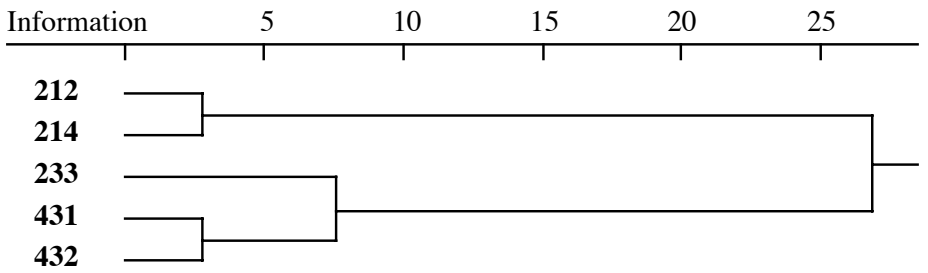


Figure 8.15 Clustering of the ponds from Ecological application 8.2, using information analysis.

The last fusion occurs at $I = 26.920$; computing ΔI is not necessary in this case. The values of I can be used as the scale for a dendrogram summarizing the clustering steps (Fig. 8.15). The same topology is obtained as in Figs. 8.3 to 8.11.

According to Williams *et al*. (1966), information analysis minimizes chaining and quickly delineates the main clusters, at least with ecological data. Field (1969) pointed out, however, that information analysis bases the similarity between objects on double absences as well as double presences. This method may therefore not be appropriate when a gradient has been sampled and the data matrix contains many zeros; see Subsections 7.2.2 and 9.2.5 for discussions of this problem.

Efficiency coefficient

The inverse of ΔI is known as the *efficiency coefficient* (Lance & Williams, 1966b). An analogue to the efficiency coefficient can be computed for dendrograms obtained using other agglomerative clustering procedures. In that case, the efficiency coefficient is still computed as $1/\Delta I$, where ΔI represents the amount by which the information in the classification is reduced due to the fusion of groups. The reduction is computed as the entropy in the classification before a fusion level minus the entropy after that fusion. In Fig. 8.2b for instance, the partition at $D = 0.60$ contains three groups of 2, 2, and 1 objects respectively; using natural logarithms, Shannon's formula (eq. 6.1) gives $H = 1.05492$. The next partition, at $D = 0.75$, contains two groups with 2 and 3 objects; Shannon's formula gives $H = 0.67301$. The difference is $\Delta = 0.38191$, hence the efficiency coefficient is $1/\Delta I = 2.61843$ for fusion level $D = 0.7$ of the dendrogram.

When $1/\Delta I$ is high, the procedure clusters objects that are mostly alike. The efficiency coefficient does not monotonically decrease as the clustering proceeds. With real data, it may decrease, reach a minimum, and increase again. If $1/\Delta I$ is plotted as a function of the successive fusion levels, the minima in the graph indicate the most informative partitions. If one wants to select a single cutting level in a dendrogram, this graph may help in deciding which partition should be selected. In Fig. 8.2b for example, one would choose the value $1/\Delta I = 1.48586$, which corresponds to the last fusion level ($D = 0.786$), as the most informative partition. The efficiency coefficient is not a rigorous decision criterion, however, since no test of significance is performed.

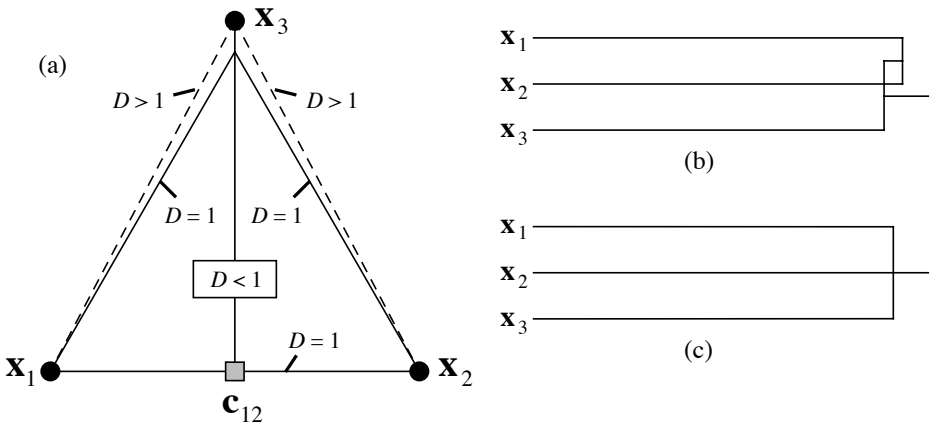


Figure 8.16 A reversal may occur in situations such as (a), where $\mathbf{x}_1$ and $\mathbf{x}_2$ cluster first because they represent the closest pair, although the distance from $\mathbf{x}_3$ to the centroid $\mathbf{c}_{12}$ is smaller than the distance from $\mathbf{x}_1$ to $\mathbf{x}_2$. (b) The result is usually depicted by a non-ultrametric dendrogram with reversal. (c) The reversal may also be interpreted as a trichotomy.

8.6 Reversals

Reversals may occasionally occur in the clustering structure when using UPGMC or WPGMC (Subsections 8.5.6 and 8.5.7), or with some unusual combinations of parameters in the general agglomerative model of Lance & Williams (Subsection 8.5.9). As an example, a reversal was produced in Fig. 8.7. Two types of situations lead to reversals:

- When $\mathbf{x}_1$ and $\mathbf{x}_2$ cluster first, because they represent the closest pair, although the distance from $\mathbf{x}_3$ to the centroid $\mathbf{c}_{12}$ is smaller than the distance from $\mathbf{x}_1$ to $\mathbf{x}_2$ (Fig. 8.16a).

- When $D(\mathbf{x}_1, \mathbf{x}_2) = D(\mathbf{x}_1, \mathbf{x}_3) = D(\mathbf{x}_2, \mathbf{x}_3)$. In such a situation, most computer programs use an arbitrary rule ("right-hand rule") and first cluster two of the three objects. A reversal appears when the third object is added to the cluster.

When this happens, the cophenetic matrix (Subsection 8.3.1) violates the ultrametric property (Subsection 8.3.2) and the dendrogram is more difficult to draw than in the no-reversal cases (Fig. 8.16b). However, departures from ultrametricity are never large in practice. For this reason, a reversal may be interpreted as nearly equivalent to a trichotomy in the hierarchical structure (Fig. 8.16c). They may also indicate true trichotomies, as discussed above; this can be checked by examination of the distance matrix.

A clustering method is said to be *monotonic* (i.e. without reversals) if

$$D(\mathbf{x}_1 \cup \mathbf{x}_2, \mathbf{x}_3) \geq D(\mathbf{x}_1, \mathbf{x}_2)$$

or

$$S(\mathbf{x}_1 \cup \mathbf{x}_2, \mathbf{x}_3) \leq S(\mathbf{x}_1, \mathbf{x}_2)$$

Assuming that $\alpha_h > 0$ and $\alpha_i > 0$ (Table 8.9), necessary and sufficient conditions for a clustering method to be monotonic in all situations are the following:

$$\alpha_h + \alpha_i + \beta \geq 1$$

and

$$\gamma \geq -\min(\alpha_h, \alpha_i)$$

(Milligan, 1979; Jain & Dubes, 1988). Some authors use the term *classification* only for hierarchies *without reversals* or for non-overlapping partitions of the objects (Table 8.1, Section 8.8).

8.7 Hierarchical divisive clustering

Contrary to the agglomerative methods of Section 8.5, hierarchical divisive techniques use the whole set of objects as the starting point. They divide it into two or several subgroups, after which they consider each subgroup and divide it again, until the criterion chosen to end the divisive procedure is met (Lance & Williams, 1967b).

In practice, hierarchical divisive clustering can only be achieved in the monothetic case or when working in an ordination space. In monothetic divisive methods, the objects are divided, at each step of the procedure, according to the states of a single descriptor. This descriptor is chosen because it best represents the whole set of descriptors (next subsection). Polythetic algorithms have been developed, but it will be seen that they are not satisfactory.

An alternative is to use a partitioning method (Section 8.8) for several numbers of groups from $K = 2$ and up and assemble the results into a graph. There is no guarantee, however, that the groups will be nested and form a hierarchy, unless the biological or ecological processes that have generated the data are themselves hierarchical.

1 — Monothetic methods

Association analysis

The clustering methods that use only one descriptor at a time are less than ideal, even when the descriptor is chosen after considering all the others. In ecology, the best-known monothetic method is Williams & Lambert's (1959) *association analysis*, originally described for species presence-absence data. Association analysis may actually be applied to any binary data table, not only species. The problem is to identify, at each step of the procedure, which descriptor is the most strongly associated

with all the others. First, X^2 (chi-square) statistics are computed for 2 × 2 contingency tables comparing all pairs of descriptors in turn. X^2 is computed using the usual formula:

$$X^2 = n\,(ad - bc)^2/[(a + b)\,(c + d)\,(a + c)\,(b + d)]$$

The formula may include Yates' correction for small sample sizes, as in similarity coefficient S_{25}. The X^2 values relative to each descriptor k are summed up:

$$\sum_{j=1}^{p} X_{jk}^2 \quad \text{for} \quad j \neq k \qquad \textbf{(8.16)}$$

The largest sum identifies the descriptor that is the most closely related to all the others. The first partition is made along the states of that descriptor; a first cluster is made of the objects coded 0 for the descriptor and a second cluster for the objects coded 1. The descriptor is eliminated from the study and the procedure is repeated, separately for each cluster. Division stops when the desired number of clusters is attained or when the sum of X^2 values no longer reaches a previously set threshold.

This method has been adapted by Lance & Williams (1968) to the information statistic I of Subsection 8.5.11. Lance & Williams (1965) also suggested using the point correlation coefficient $\varphi = \sqrt{X^2/n}$ (eq. 7.9) instead of X^2. This may prevent aberrant or unique objects in the study from determining the first partitions. This is analogous to the problem encountered with the higher powers of Minkowski's metric (D_6), which could give too much weight to the largest differences; this problem was less severe when using power 1, which is the Manhattan metric (D_7). One then looks for the descriptor that maximizes the sum $\Sigma\varphi_{jk}$ ($j \neq k$; see eq. 8.16). Gower (1967) suggested to use, for division, the species that has the largest R^2 with all the other species (eq. 10.20), instead of the one with the largest sum of simple correlations. He also suggested to use the largest variance inflation factor (*VIF*) as criterion, instead of the largest R^2, because *VIF* is monotonically related to R^2 (eq. 10.17). *VIF* can be computed by a single matrix operation for all species (sentence that follows eq. 10.17).

The principles of association analysis may be applied to descriptors with multiple states (semiquantitative or qualitative), by computing X^2 statistics between descriptors using the usual X^2 formulas (eqs. 6.5 and 6.6). Raw species abundance data should not be analysed in this way, however, because the large number of different abundance values makes the contingency tables meaningless.

Legendre & Rogers (1972) proposed a monothetic divisive method similar to association analysis, in which the choice of the descriptor best representing all the others is made with the help of an information statistic computed on contingency tables. For each descriptor k, two quantities developed by Christanson (*in* Brill *et al.*, 1972) are computed: SUMRAT (k) and SAMRAT (k) ("sum of ratios"). SUMRAT (k) is the sum of the fractions representing the amount of information that k has in common with

each descriptor j ($j \neq k$), divided by the amount of information in j. In SAMRAT (k), the divisor is the amount of information in k instead of j. Using the symbolism of Section 6.2:

$$\text{SUMRAT}\ (k) = \sum_{j=1}^{p} \frac{H(k) - H(k|j)}{H(j)} \quad \text{for} \quad j \neq k \qquad \textbf{(8.17)}$$

$$\text{SAMRAT}\ (k) = \sum_{j=1}^{p} \frac{H(k) - H(k|j)}{H(k)} \quad \text{for} \quad j \neq k \qquad \textbf{(8.18)}$$

which can be recognized as sums of asymmetric uncertainty coefficients, $\sum$ B/(B + C) and $\sum$ B/(A + B), respectively (Section 6.2). SUMRAT (k) and SAMRAT (k) both have the property of being high when k has much information in common with the other descriptors in the study. The descriptor that best represents the divisive power of all descriptors is expected to have the highest SUMRAT and SAMRAT values. However, SUMRAT (k) and SAMRAT (k) are also influenced by the number of states in k, which may unduly inflate $H(k)$, thus causing SUMRAT (k) to increase and SAMRAT (k) to decrease. This factor must be taken into account if there is conflict between the indications provided by SUMRAT and SAMRAT as to the descriptor that best represents the whole set. This peculiarity of the method requires the user's intervention at each division step, in the present state of development of the equations.

Since the information measures on which SUMRAT and SAMRAT are based are at the same exponent level as X^2 (Section 6.2), one could compute instead:

$$\text{SUMRAT}\ (k) = \sum_{j=1}^{p} \sqrt{\frac{H(k) - H(k|j)}{H(j)}} \quad \text{for} \quad j \neq k \qquad \textbf{(8.19)}$$

$$\text{SAMRAT}\ (k) = \sum_{j=1}^{p} \sqrt{\frac{H(k) - H(k|j)}{H(k)}} \quad \text{for} \quad j \neq k \qquad \textbf{(8.20)}$$

thus minimizing the effect of single objects on the first partitions, as indicated above.

Williams & Lambert (1961) have suggested using association analysis in the R mode for identifying species associations. This approach does not seem, however, to be based on an acceptable operational concept of association (see Section 8.9).

2 — Polythetic methods

There is no satisfactory algorithm for the hierarchical division of objects based on the entire set of descriptors.

The method of Edwards & Cavalli-Sforza (1965) tries all possible divisions of the set of objects into two clusters, looking for the division that maximizes the distance

between the centroids. Using sums of squared distances to centroids, one first computes *SS*, which is the sum of squares of the Euclidean distances of all objects to the centroid of the whole set of objects, divided by the number of objects *n*; this value is the total sum of squares of a single classification analysis of variance (eqs. 6.56 and 8.6). Then, for each possible partition of the objects into two groups **h** and **i**, the sums of squares of the distances to the centroids are computed within each cluster, using eq. 8.6, to obtain $SS(\mathbf{h})$ and $SS(\mathbf{i})$, respectively. The distance between the two clusters is therefore $SS - SS(\mathbf{h}) - SS(\mathbf{i})$. This is the quantity to be maximized for the first partition. Then each cluster is considered in turn and the operation is repeated to obtain subsequent divisions. Like *K*-means partitioning of Section 8.8, this method can only be applied to quantitative data because it is based on Euclidean distances.

This method may seem attractive but, apart from the theoretical objections that he raised about it, Gower (1967) noted that investigating all possible partitions to find the best one is a NP–hard computational problem (footnote in Section 8.8). He calculated that, before obtaining the first partition of a cluster of 41 objects, 54000 years of computing time would be required using a computer with an access time of 5 microseconds, to try all $(2^{40} - 1)$ possible partitions of 41 objects into two groups. 5 microseconds was the typical access time of computers in 1967. The problem remains with modern computers, even though they have much smaller access times (in the realm of nanoseconds at the beginning of the years 2010). The heuristic algorithms used to solve the *K*-means problem (Section 8.8) could, however, be applied here instead of the complete search through all possible solutions.

Dissimilarity analysis

The *dissimilarity analysis* of Macnaughton-Smith *et al*. (1964) first looks for the object that is the most different from all the others and removes it from the initial cluster. One by one, the most different objects are removed. Two groups are defined: the objects removed and the remaining ones, between which a distance is calculated. Objects are removed up to the point where the distance between clusters can no longer be increased. Each of the two clusters thus formed is subdivided again, using the same procedure. The first partition of a cluster of *n* objects requires at most $3n^2/4$ operations instead of the $(2^{n-1} - 1)$ operations required by the previous method. Other authors have developed special measures of distance to be used in dissimilarity analysis, such as Hall's (1965) *singularity index* and Goodall's (1966b) *deviant index*. Although attractive, dissimilarity analysis may produce strange results when many small clusters are present in the data, in addition to major clusters of objects.

A major disadvantage of all hierarchical divisive methods is that a division of the objects in two major clusters may also split the members of some minor cluster, which cannot be fused again unless special procedures are included in the algorithm for that purpose (Williams & Dale, 1965).

3 — Division in ordination space

Computer-efficient polythetic hierarchical divisive clustering can be obtained by partitioning the objects according to the axes of an ordination space. Using principal

component analysis (PCA, Section 9.1), the set of objects may be partitioned in two groups: those that have positive values along the first PCA axis and those that have negative values. The PCA analysis is repeated for each of the groups so obtained and a new partition of each group is performed. The process is repeated until the desired level of resolution is obtained (Williams, 1976b).

Following a similar suggestion by Piazza & Cavalli-Sforza (1975), Lefkovitch (1976) developed a hierarchical classification method for very large numbers of objects, based on principal coordinate analysis (PCoA, Section 9.3). The dendrogram is constructed from the successive principal coordinate axes, the signs of the objects on the coordinate axes indicating their membership in one of the two groups formed at each branching step. The objects are partitioned in two groups according to their signs along the first PCoA axis; each group is then divided according to the positions of the objects along the second axis; and so on. This differs from the method used with PCA above, where the analysis is repeated for each group before a new division takes place. To calculate the principal coordinates of a large number of objects, Lefkovitch proposed to first measure the similarity among objects by an equation which, like the covariance or correlation, is equivalent to the product of a matrix with its transpose. He described such a measure, applicable if necessary to combinations of binary, semiquantitative, and quantitative descriptors. The association matrix among objects is obtained by the matrix product $\mathbf{YY}'$ (order $n \times n$). In situations where there are many more objects than descriptors, computation of the eigenvalues and eigenvectors of the association matrix among *descriptors,* $\mathbf{Y}'\mathbf{Y}$, represents an important saving of computer time because $\mathbf{Y}'\mathbf{Y}$ (order $p \times p$) is much smaller than $\mathbf{YY}'$ (order $n \times n$). After Rao (1964) and Gower (1966), Lefkovitch showed that the principal coordinates $\mathbf{V}$ of the association matrix among *objects* can then be found, using the relation $\mathbf{V} = \mathbf{YU}$ where $\mathbf{U}$ is the matrix of the principal coordinates among *descriptors*. The principal coordinates thus calculated allow one to position the objects, numerous as they may be, in the reduced space. Principal coordinates can be used for the binary hierarchical divisive classification procedure that was Lefkovitch's goal.

A divisive algorithm of the same type is used in TWINSPAN (next subsection). It is based upon an ordination obtained by correspondence analysis instead of PCA or PCoA.

4 — TWINSPAN

Two Way INdicator SPecies ANalysis (TWINSPAN*) (Hill, 1979a) is fundamentally a method for hierarchical divisive classification of communities, based on progressive refinement of a single ordination axis obtained by correspondence analysis (CA) or

* Available as part of the package PC-ORD (distribution: footnote in Section 11.7). TWINSPAN is also available from Micro-computer Power: <http://www.microcomputerpower.com>. The TWINSPAN source code in FORTRAN and an executable version for Windows are available on Jari Oksanen's page <http://cc.oulu.fi/~jarioksa/softhelp/ceprog.html>. An executable program for Windows, WINTWINS, is available on the page http://www.canodraw.com/wintwins.htm.

detrended correspondence analysis (DCA) (Section 9.2) of a (sites × species) data matrix. Hill (1979a) also called the method a *dichotomized ordination analysis*.

An attractive feature of the output is a two-way table where the sites (columns) are sorted according to the splits of the hierarchical classification. The species (rows) are also sorted so as to form blocks corresponding to the groups of sites of the classification. A dendrogram representing the classification of the sites can easily be drawn, if required, from the TWINSPAN output table. In addition, the method computes an indicator values index (I) for the species for every split of the hierarchical classification of the sites.

Pseudo-species

To model the concept of *differential species* (i.e. species with clear ecological preferences), which is qualitative, TWINSPAN creates *pseudospecies*. Each species is recoded into a set of dummy variables (pseudospecies) corresponding to relative abundance levels; these classes are cumulative. If, for example, the pseudospecies cutting levels are 1%, 11%, 26%, 51% and 76%, a relative abundance of 18% at a site will fill the first and second dummy pseudospecies vectors with "1" (= presence). Cutting levels are arbitrarily decided by users. A (sites × pseudospecies) data table is thus created.

The TWINSPAN procedure is rather complex. A detailed description is given by Kent & Coker (1992). It may be summarized as follows.

1. After ordination by CA or DCA of the original (sites × species) data table, the objects are divided in two groups according to their signs along the first ordination axis. This is called the primary ordination.

2. TWINSPAN then computes an indicator values index (I) for the species, for every split of the hierarchical classification of the sites. According to Kent & Coker (1992), the index is computed as follows using the pseudospecies data:

Indicator value index

$$I_j = \frac{n_j^+}{n^+} - \frac{n_j^-}{n^-}$$

where n^+ and n^- are respectively the number of sites on the arbitrarily chosen positive and negative sides of the split, whereas n_j^+ and n_j^- are the number of sites on the positive and negative sides, respectively, that contain pseudospecies j. A pseudospecies present in every site on the positive side and in none of the sites on the negative side obtains $I_j = 1$, and -1 if it is found in every site on the negative side and in none on the positive side. A pseudospecies that occurs in all sites on both sides of the split obtains $I_j = 0$. In TWINSPAN, the indicator value describes the preference of a pseudospecies for one or the other side of the partition. The pseudospecies with the highest indicator absolute value is counted as the best indicator for that species. Then, only one pseudospecies of a single species is declared an indicator of a split, and that is the pseudospecies that has the highest absolute value of I. n_j^+/n^+ is the measure of *fidelity* to a group used in the *INDVAL* method described in Subsection 8.9.3.

Fidelity

3. Further steps lead to a refined ordination of the objects. After taking care of misclassifications, borderline cases, and other problems, a final division of the sites is obtained. Then, each subset is divided into smaller subsets by repeating the procedure. This goes on until groups become very small. Typically, groups of 4 objects or less are not partitioned further.

Problems with TWINSPAN are the following: (1) To identify species groups or compute indicator values, one cannot introduce some other classification of the sites in the program; only the classification produced by TWINSPAN, which is based on correspondence analysis (CA, Section 9.2) or detrended correspondence analysis (DCA, Subsection 9.2.5), can be used to delineate species groups. (2) The pseudospecies concept is based on species relative abundances. The relative abundance of a species depends on the absolute abundances of the other species present at a site. Such relative frequencies may be highly biased, in particular, when sampling mobile organisms: all species are not sampled with the same efficiency because of differences in behaviour. So, the coding of species abundances into pseudospecies may be highly unstable.

TWINSPAN has also been criticized by Belbin and McDonald (1993) on two grounds: (1) The method assumes the existence of a strong gradient dominating the data structure, so that it may fail to identify secondary gradients or other types of structures in data sets. (2) The cutting points along the dominant axis for the whole group, and then for subgroups, are always chosen to be the centroid of the group to be split instead of a point where a large gap occurs in the data. This problem has been alleviated by a modification to the method proposed by Rolecek *et al*. (2009).

An alternative method to obtain a reordered species-by-sites table is seriation (Section 8.10). In R (Section 8.15), a plot ("heat map") can be produced using functions ***hmap()*** of package SERIATION or ***heatmap()*** of STATS.

8.8 Partitioning by *K*-means

Partitioning consists in finding a single partition of a set of objects (Table 8.1). Jain & Dubes (1988) stated the problem in the following terms: given n objects in a p–dimensional space, determine a partition of the objects into K groups, or clusters, such that the objects within each cluster are more similar to one another than to objects in the other clusters. The number of groups, K, is determined by the user. This problem was first stated in statistical terms by Fisher (1958) who proposed solutions for a single variable (with or without contiguity constraint; see Sections 12.6 and 13.3). *K*-means partitioning is available in several R functions; see Section 8.15.

The difficulty is to define what 'more similar' means. Several criteria have been suggested; they can be divided into global and local criteria. A *global criterion* would be, for instance, to represent each cluster by a type-object (on *a priori* grounds, or

using the centroids obtained by agglomerative clustering, Subsections 8.5.6 and 8.5.7) without consideration for local densities of objects and assign each object to the nearest type-object. A type object representing a cluster is called a *medoid* (Kaufman & Rousseeuw, 1990). A *local criterion* uses the local structure of the data to delineate clusters; groups are formed by identifying high-density regions in the data represented in A-space (Fig. 7.2). The *K*-means method, described in the next paragraphs, is the most commonly used of the latter type. *K*-means belongs to a larger class of methods called *K*-centroid cluster analysis, which is briefly described in Section 8.15.

Medoid

Objective function

In *K*-means, the objective function that the partition should minimize is the same as in Ward's agglomerative clustering method (Subsection 8.5.8): the total error sum of squares (E_K^2, or TESS). The major problem encountered by the algorithms is that the solution on which the computation eventually converges depends to some extent on the initial positions of the centroids. This problem does not exist in Ward's method, which proceeds iteratively by hierarchical agglomeration. However, even though Ward's algorithm guarantees that the *increase* in sum of squared errors ($\Delta E_{\mathbf{hi}}^2$, eq. 8.8) is minimized at each step of the agglomeration (so that any order of entry of the objects should lead to the same solution, except in cases of equal distances where a "right-hand" programming rule may prevail), there is no guarantee that any given Ward's partition is optimal in terms of the E_K^2 criterion — surprising at this may seem. This same problem occurs with all stepwise statistical methods.

Local minimum

The problem of the final solution depending on the initial positions of the centroids is known as the "local minimum" problem in algorithms. The concept is illustrated in Fig. 8.17, by reference to a *solution space*. It may be explained as follows. Solutions to the *K*-means problem are the different ways to partition n objects into, say, $K = 4$ groups. If a single object is moved from one group to another, the corresponding two solutions will have slightly different values for the criterion to be minimized (E_K^2). Imagine that all possible solutions form a "space of solutions". The different solutions can be plotted as a graph with the E_K^2 criterion as the ordinate. It is not essential to accurately describe the abscissa to understand the concept; it would actually be a multidimensional space. A *K*-means algorithm starts at some position in that space, the initial position being assigned by the user (see below). It then tries to navigate the space to find the solution that minimizes the objective criterion (E_K^2). The space of solutions is not smooth, however. It may contain *local minima* from which the algorithm may be unable to escape. When this happens, the algorithm has not found the overall minimum and the partition is not optimal in terms of the objective criterion.

Overall minimum

Several approaches may be used to help a *K*-means algorithm converge towards the overall minimum of the objective criterion E_K^2. They involve either selecting specific objects as "group seeds" at the beginning of the run, or attributing the objects to the K groups in some special way. Here are some commonly-used approaches:

- Provide an initial configuration corresponding to an (ecological) hypothesis. The idea is to start the algorithm in a position in the solution space that is, hopefully, close to the final solution sought. This ideal situation is seldom encountered in real studies.

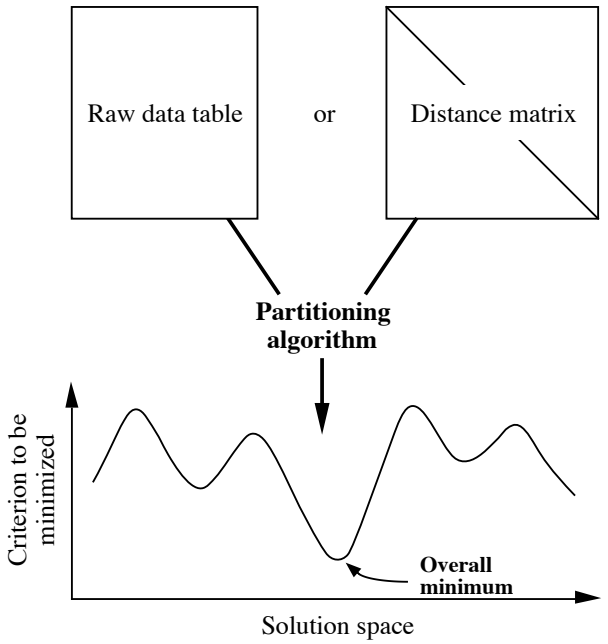


Figure 8.17 *K*-means algorithms search the space of solutions, trying to find the overall minimum (arrow) of the objective criterion to be minimized, while avoiding local minima (troughs).

• Provide an initial configuration corresponding to the result of a hierarchical clustering, obtained from a space-conserving method (Table 8.9). One simply chooses the partition into K groups found on the dendrogram and lists the objects pertaining to each group. The K-means algorithm will then be asked to rearrange the group membership and look for a better overall solution (lower E_K^2 statistic).

• If the program allows it, select as "group seed", for each of the K groups to be delineated, some object located near the centroid of that group. For very large problems, Lance & Williams (1967d) suggested to use as starting point the result of a hierarchical clustering of a *random subset of the objects*, using as "group seeds" either the centroids of K clusters, or objects located near these centroids.

• Attribute the objects at random to the various groups. All K-means computer programs offer this option. Find a solution and note the $E_{K_2}^2$ value. It is possible that the solution found corresponds to a local minimum of E_K^2. So, repeat the whole procedure a number of times (for example, 100 times), starting every time from a different random configuration. Retain the solution that minimizes the E_K^2 statistic. One is more confident that this solution corresponds to the overall minimum when the corresponding value of E_K^2 is found several times across the runs.

NP-hard

Several algorithms have been proposed to solve the *K*–means problem, which is but one of a family of problems known in computer sciences as the *NP–complete* or *NP–hard problems*[*]. In all these problems, the only way to be sure that the optimal solution has been found would be to try all possible solutions in turn. This is impossible, of course, for real-size problems, even with modern-day computers, as explained in Subsection 8.7.2. Classical references to *K*-means algorithms are Anderberg (1973), Hartigan (1975), Späth (1975, 1980), Everitt (1980), Jain & Dubes (1988) and Kaufman & Rousseeuw (1990). Milligan & Cooper (1987) reviewed the most commonly used algorithms and compared them for structure recovery, using artificial data sets. One of the best algorithms available is the following; it frequently converges to the solution representing the overall minimum for the E_K^2 statistic. It is a very simple alternating least-squares algorithm, which iterates between two steps:

- Compute cluster centroids and use them as new cluster seeds.
- Assign each object to the nearest seed.

At the start of the program, K observations are selected as "group seeds". Each iteration reduces the sum of squared errors E_K^2, if possible. Since only a finite number of partitions are possible, the algorithm eventually reaches a partition from which no improvement is possible; iterations stop when E_K^2 can no longer be improved. The FASTCLUS procedure of the SAS package, mentioned here because it can handle very large numbers of objects, uses this algorithm. Options of the program can help deal with outliers if this is a concern. The SAS manual (SAS Institute, 2011) provides more information on the algorithm and the available options.

K-means partitioning was originally proposed in a pioneering paper by MacQueen (1967) who gave the method its name: *K*-means. Lance & Williams made it popular by recommending it in their review paper (1967d). In the MacQueen paper, group centroids are recomputed after each addition of an object; this is also an option in SAS. MacQueen's algorithm contains procedures for the fusion of clusters, if centroids become very close, and for creating new clusters if an object is very distant from existing centroids.

K-means partitioning may be computed from either a table of raw data or a distance matrix, because the total error sum of squares E_K^2 (eq. 8.7) is equal to the sum of squares of the distances from the points to their respective centroids (eq. 8.5; Fig. 8.18a) and to the sum (over groups) of the mean squared within-group distances[†] (eq. 8.6; Fig. 8.18b). It is especially advantageous to compute it on raw data when the number of objects is large because, in such a situation, the distance matrix may

[*] *NP* stands for *Non-deterministic Polynomial*. In theory, these problems can be solved in polynomial time (i.e. some polynomial function of the number of objects) on a (theoretical) non-deterministic computer. NP-hard problems are probably not solvable by efficient algorithms.

[†] As shown in eq. 8.6, the mean squared distance within group k is computed as the sum of the squared within-group distances divided by the number of objects n_k in the group.

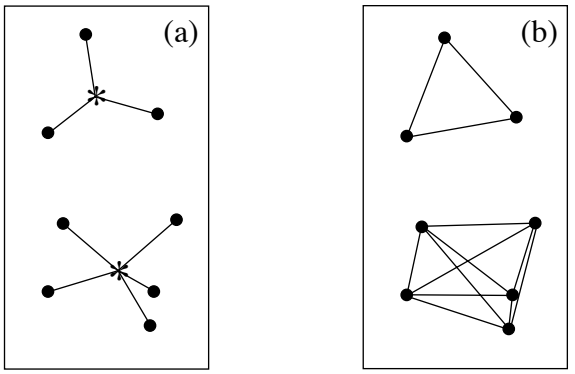


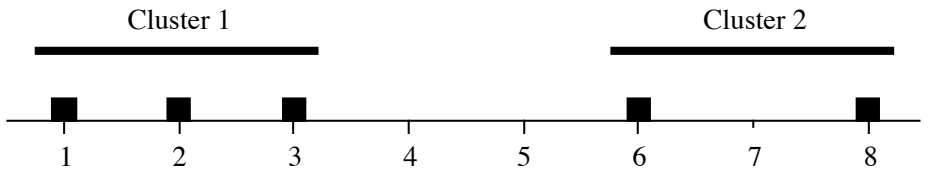
Figure 8.18 The total error sum of squares (TESS, E_K^2, eq. 8.7)) is equal (a) to the sum of squares of the distances from the points to their respective centroids (eq. 8.5). (b) It is also equal to the sum (over groups) of the mean squared within-group distances (eq. 8.6).

become very cumbersome or even impossible to store and search. In contrast, when using a table of original data, one only needs to compute the distance of each object to each group centroid, rather than to all other objects.

The disadvantage of using a table of raw data is that the only distance function among points, available during *K*-means partitioning, is the Euclidean distance (D_1, Chapter 7) in A-space. This is not suitable for species counts and other types of frequency data (Fig. 7.8). Two solutions are possible when the Euclidean distance is unsuitable: (1) one may transform the species data using one of the transformations described in Section 7.7 and use the transformed data in the *K*-means analysis; or (2) one may first compute a suitable distance matrix among objects (see Tables 7.4 and 7.5), decompose the distance matrix into eigenvectors by principal coordinate analysis (PCoA, Section 9.3), and run *K*-means partitioning using the table of eigenvectors (principal coordinates).

Following are two numerical examples that illustrate the behaviour of the E_K^2 criterion computed using eqs. 8.5 and 8.6.

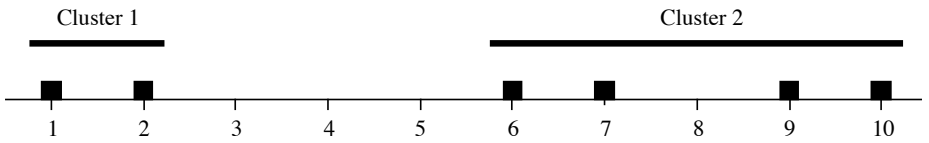
Numerical example 1. For simplicity, consider a single variable. The best partition of the following five objects (dark squares) in two clusters (boldface horizontal lines) is obviously to put objects with values 1, 2 and 3 in one group, and objects with values 6 and 8 in the other:



This example is meant to illustrate that the E_K^2 criterion can be computed from either raw data (eq. 8.5) or distances among objects (eq. 8.6). Using raw data (left-hand column, below), the group centroids are at positions 2 and 7 respectively; deviations from the centroids are calculated for each object, squared, and added within each cluster. Distances among objects (right-hand column, below) are easy to calculate from the object positions along the axis; the numbers of objects (n_k), used in the denominators, are 3 for cluster 1 and 2 for cluster 2.

$$e_1^2 = (1^2 + 0^2 + (-1)^2) = 2 \qquad e_1^2 = (2^2 + 1^2 + 1^2)/3 = 2$$
$$e_2^2 = (1^2 + (-1)^2) = 2 \qquad e_2^2 = 2^2/2 = 2$$
$$E_K^2 = 4 \qquad E_K^2 = 4$$

Numerical example 2. Considering a single variable again, this example examines the effect on the E_K^2 statistic of changing the cluster membership. There are six objects and they are to be partitioned into $K = 2$ clusters. The optimal solution is that represented by the boldface horizontal lines:



Calculations are as above. Using raw data (left-hand column, below), the group centroids are at positions 1.5 and 8 respectively; deviations from the centroids are calculated for each object, squared, and added within each cluster. Distances among objects (right-hand column, below) are easy to calculate from the object positions along the axis; the numbers of objects (n_k), used in the denominators, are 2 for cluster 1 and 4 for cluster 2.

$$e_1^2 = (0.5^2 + (-0.5)^2) = 0.5 \qquad e_1^2 = 1^2/2 = 0.5$$
$$e_2^2 = (2^2 + 1^2 + (-1)^2 + (-2)^2) = 10.0 \qquad e_2^2 = (1^2 + 3^2 + 4^2 + 2^2 + 3^2 + 1^2)/4 = 10.0$$
$$E_K^2 = 10.5 \qquad E_K^2 = 10.5$$

Consider now a sub-optimal solution where the clusters would contain the objects located at positions (1, 2, 6, 7) and (9, 10), respectively. The centroids are now at positions 4 and 9.5 respectively. Results are the following:

$$e_1^2 = (3^2 + 2^2 + (-2)^2 + (-3)^2) = 26.0 \qquad e_1^2 = (1^2 + 5^2 + 6^2 + 4^2 + 5^2 + 1^1)/4 = 26.0$$
$$e_2^2 = (0.5^2 + (-0.5)^2) = 0.5 \qquad e_2^2 = 1^2/2 = 0.5$$
$$E_K^2 = 26.5 \qquad E_K^2 = 26.5$$

This example shows that the E_K^2 criterion quickly increases when the cluster membership departs from the optimum.

In some studies, the number of clusters K to be delineated is determined by the ecological problem, but this it is not often the case. The problem of determining the most appropriate number of clusters has been extensively discussed in the literature. Over 30 different indices, called "stopping rules", have been proposed to do so. Milligan & Cooper (1985) compared them through an extensive series of simulations using artificial data sets with known numbers of clusters; the results of that study are also reported in Milligan (1996). Some of these rules recover the correct number of clusters in most instances, but others are appallingly inefficient. The best of the criteria investigated in that paper is the Calinski-Harabasz index (C-H, Calinski & Harabasz, 1974), which is the multivariate F-statistic (eq. 11.7) of a RDA in which $m = (K - 1)$ dummy variables are used to represent a partition into K groups. When the groups identified by a clustering method are well separated in A-space, the F-statistic becomes large. This statistic cannot be tested for significance, however, because the groups are derived from the same data that would be used for testing.

C-H index

SAS has implemented two among the best rules studied by Milligan: the Calinski-Harabasz F-statistic (called pseudo-F in SAS manuals) and the cubic clustering criterion. Fourteen stopping indices are available in function ***clustIndex()*** of package CCLUST in R. Cross-validation, used in multivariate regression tree analysis (MRT, Section 8.11) to decide about the size of trees, can also be used to determine the optimal number of clusters found in a series of K-means analyses involving different numbers of groups. However, none of these indices correctly identifies the correct solution when a single cluster is present in the data.

8.9 Species clustering: biological associations

Most of the methods discussed in the previous sections may be applied to clustering descriptors as well as objects. When searching for species associations, however, it is important to cluster species using methods that model as precisely as possible a clearly formulated concept of association. The present section (1) attempts to define an operational concept of association and (2) shows how to identify species associations in that framework.

Several concepts of species association have been developed since the nineteenth century; Whittaker (1962) wrote a remarkable review about them. These concepts are not always operational, however. In other words, they cannot always be translated into a series of well-defined analytical steps that would lead to the same result if they were applied by two independent researchers, using the same data. In general, the concept of association refers to a group of species that are "significantly" found together, without this implying necessarily any positive interaction among them. An association, in the statistical sense, is a recurrent group of co-occurring (presence-absence data) or correlated (abundance data) species (Legendre & Legendre, 1978). Associations of taxa belonging to categories other than species may also be defined.

Species association

Several procedures have been proposed for the identification of species associations. Quantitative algorithms have progressively replaced the empirical methods, as they have in other areas of science. All these methods, whether simple or elaborate, have two goals: first, identify the species that occur together and, second, minimize the likelihood that the co-occurrences so identified be fortuitous. The search for valid associations obviously implies that the sampling be random and planned in accordance with the source of variability under study (i.e. geographical, temporal, experimental), which defines the framework within which the groups of species, found repeatedly along the sampling axes, are called associations; one then speaks of association of species over geographic space, or in time, etc. The criterion is the recurrence of a group of species along the study axes.

Species distributions may be correlated through space or time because they have common (or opposite) environmental requirements, or as the result of biotic interactions. These two families of processes produce identifiable spatial or temporal patterns, as described in Subsection 1.1.1. For the first type, positive associations occur when species have the same ecological requirements, and negative associations when their requirements differ. The second type refers to biotic interactions among species, which include predator-prey relationships, competition, and mutualism; it can also lead to positive or negative associations among species. These processes provide grounding theory and hypotheses for the search of species associations, which is one of the classical problems of community ecology (Roxburgh & Chesson, 1998).

Correlation analysis in one form or another, for presence-absence or abundance data, has proven useful to identify species associations (Greig-Smith, 1983; O'Connor & Aarssen, 1987; Myster & Pickett, 1992; Roxburgh & Chesson, 1998). Interspecific associations are recognized when two or more species co-occur (for presence-absence data) either more or less frequently than expected by chance, or when their quantitative variation is correlated. One cannot, however, distinguish between the hypotheses of environmental control and biotic interactions from the results of an association analysis alone (Rejmánek & Leps, 1996). Finer analyses using multiscale correlation methods, e.g. multiscale codependence analysis (Subsection 14.5.2), may help decide between competing hypotheses about to the causes of species associations.

Under the hypothesis of environmental control, when associations have been found, one can concentrate on finding the ecological requirements common to most or all species of an association instead of having to describe the biology and habitat of each species individually. In an inverse approach, species associations may be used to predict environmental characteristics or as indicators of environmental quality (Legendre, 2005). Associations may be better predictors of environmental quality than individual species because they are less subject to sampling error. In certain cases, trophic groups or size classes may also be used for the same purpose.

As mentioned at the beginning of this section, a simple and operational statistical definition is that a species association is *a recurrent group of co-occurring or correlated species*. Using this definition, one can select clustering methods that are

appropriate to delineate species associations. Appropriate measures of resemblance in R mode were described in Chapter 7 (Table 7.6). A great variety of clustering methods have been used for the identification of associations, although the choice of a given method often appears to have been based on the availability of a program on the local computer instead of a good understanding of the properties and limitations of the various techniques. An alternative to standard clustering techniques was proposed by Lambshead & Paterson (1986) who used numerical cladistic methods to delineate species associations. Among the ordination methods, principal component and correspondence analyses may not produce clearly identifiable clusters of species except in the most simple cases (e.g. Fig. 8.20), even though these analyses may be very useful to investigate other multivariate ecological problems (Chapter 9).

After selecting the most appropriate coefficient of dependence for the data at hand (Table 7.6), one must next make a choice among the usual hierarchical clustering methods discussed in the previous sections of this chapter, including TWINSPAN (Subsection 8.7.4). Partitioning by *K*-means (Section 8.8) should also be considered after transformation of the species data (Section 7.7). In addition, there are two specialized methods to delineate species associations described in Subsections 8.9.1 and 8.9.2 below. When the analysis aims at identifying hierarchically-related associations, hierarchical clustering methods are appropriate. When one simply looks for species associations without implying that they should form a hierarchy, partitioning methods are in order. Hierarchical clustering may also be used in that case but one must decide, using a dendrogram or another graphical representation, which level of partition in the hierarchy best corresponds to the associations to be identified. One must take into account the level of detail required and the limits of significance or interpretability of the species clusters found. In any case, space-conserving or space-dilating methods should be preferred to single linkage, especially when one is trying to delimit groups of species from data sampled along an ecological continuum.

The search for species associations is based on the often untested assumption that species have non-random patterns of association, these associations being due to environmental control or biotic interactions. Jackson *et al*. (1992) discussed several null models that may be used to test the non-randomness of species co-occurrence across sites.

Ecological application 8.9a

Thorrington-Smith (1971) identified 237 species of phytoplankton in water samples from the West Indian Ocean. 136 of the species were clustered into associations by single linkage hierarchical clustering of a Jaccard (S_7) association matrix among species. The largest of the 11 associations contained 50 species; its distribution mostly corresponded to the equatorial subsurface water. This association was dominant at all sites and was considered typical of the endemic flora of the West Indian Ocean. Other phytoplankton associations represented seasonal or regional differences, or characterized currents or nutrient-rich regions. Since phytoplankton associations did not lose their identities even when they were mixed, the study of associations in zones of water mixing seemed a good way of tracing back the origins of water masses.

1 — Non-hierarchical complete linkage clustering

Fager's (1957) *non-hierarchical complete linkage clustering* is a specialized partitioning method designed for discovering species associations. It is well-adapted to probabilistic measures of dependence among species and other measures of dependence for which a critical or significance level can be set. This method differs from hierarchical complete linkage clustering in that one looks for clusters formed at a stated threshold of similarity without taking into account the hierarchical cluster structure that may exist at other similarity levels. For probabilistic similarity coefficients, e.g. S_{25} (eq. 7.62), the threshold is usually the significance level $\alpha = 0.05$ or $\alpha = 0.01$, which corresponds to $S \geq 0.95$ or $S \geq 0.99$. With the non-probabilistic similarity coefficient S_{24} (eq. 7.60), Fager & McGowan (1963) used $S \geq 0.5$ as the clustering threshold.

Computer programs that make the method operational have been written, but it is also possible to implement it without a special program. If a similarity coefficient was used to compute the resemblance among the species, select a threshold similarity level and draw a graph (as in Fig. 8.2a) of the species with link edges corresponding to all values of $S \geq$ (threshold). Then, delineate the species associations on the graph as the groups meeting the complete-linkage criterion, i.e. the groups in which all objects are linked to all others at the stated similarity level (Subsection 8.5.2). In case of conflicts, use the following decision rules.

Fager and Krylov rules

1. Complete-linkage clusters of species, obtained by this method, must be independent of one another, i.e. they must have no species in common. Between two possible species partitions, *form first the clusters containing as many species as possible*. For instance, if a cluster of 8 species has two species in common with another cluster of 5 species, create clusters of 8 and 3 species instead of clusters of 6 and 5 species. Krylov (1968) adds that no association should be recognized that contains fewer than three species.

If non-independent clusters remain (i.e. clusters with objects in common), consider rules 2 and 3, in that order.

2. Between several non-independent clusters *containing the same number of species*, choose the partition that maximizes the size of the resulting independent clusters. For example, if there are three clusters of 5 species each where clusters 1 and 2 have one species in common and clusters 2 and 3 also have one species in common, select clusters 1 and 3 with five species each, leaving 3 species into cluster 2. One thus creates three clusters with membership 5, 3, and 5, instead of three clusters with membership 4, 5, and 4.

3a. If the above two criteria do not solve the problem, between two or more non-independent clusters having about the same number of species, select the one *found at the largest number of sites* (Fager, 1957). One has to go back to the original data matrix in order to use this criterion.

3b. Krylov (1968) suggested replacing this last criterion with the following one: among alternative species, the species to include in a cluster is the one *that has the least affinity* with all the other species that are not members of that cluster, i.e. the species that belongs to the cluster more exclusively. This criterion may be decided from the graph of link edges among the species.

This form of non-hierarchical complete linkage clustering led Fager (1957), Fager & McGowan (1963), and Krylov (1968) to identify meaningful and reproducible plankton associations. Venrick (1971) explains an interesting additional step of Fager's computer program; this step answers an important problem of species association studies. After having recognized independent clusters of completely linked species, the program associates the remaining species, by single linkage clustering, to one or several of the main clusters. These *satellite species* do not have to be associated with all members of an association. They may also be satellites of several associations. This reflects adequately the organizational complexity of biological communities.

This last point shows that *overlapping* clustering methods could be applied to the problem of delineating species associations. The mathematical bases of these methods have been established by Jardine & Sibson (1968, 1971) and Day (1977).

Ecological application 8.9b

Fager's non-hierarchical complete linkage clustering was used by Legendre & Beauvais (1978) to identify fish associations in 378 catches from 299 lakes of northwestern Québec. Their computer program provided the list of all possible complete linkage clusters formed at a user-selected similarity level. Species associations were determined using the criteria listed above. The similarity between species was established by means of the probabilistic measure S_{25} (Subsection 7.5.2), based on presence-absence data.

At similarity level $S_{25} \geq 0.989$, the program identified 25 non-independent species clusters, involving 26 of the 29 species in the study. Each subgroup of at least three species could eventually become an association since the clustering method was complete linkage. Many of these clusters overlapped. The application of Fager's decision rules (with rule 3b of Krylov) led to the identification of five fish associations, each one completely formed at the similarity level indicated to the right. Stars indicate the internal strength of the associations (*** all links ≥ 0.999, ** all links ≥ 0.99, * all links ≥ 0.95).

1)	Lake whitefish	*Coregonus clupeaformis*	$S_{25} \geq 0.999$ ***
	Longnose sucker	*Catostomus catostomus*	
	Lake trout	*Salvelinus namaycush*	
	Round whitefish	*Prosopium cylindraceum*	
	Lake chub	*Couesius plumbeus*	
2)	Northern pike	*Esox lucius*	$S_{25} \geq 0.995$ **
	White sucker	*Catostomus commersoni*	
	Walleye	*Stizostedion vitreum*	
	Shallowwater cisco	*Coregonus artedii*	
	Yellow perch	*Perca fluviatilis*	

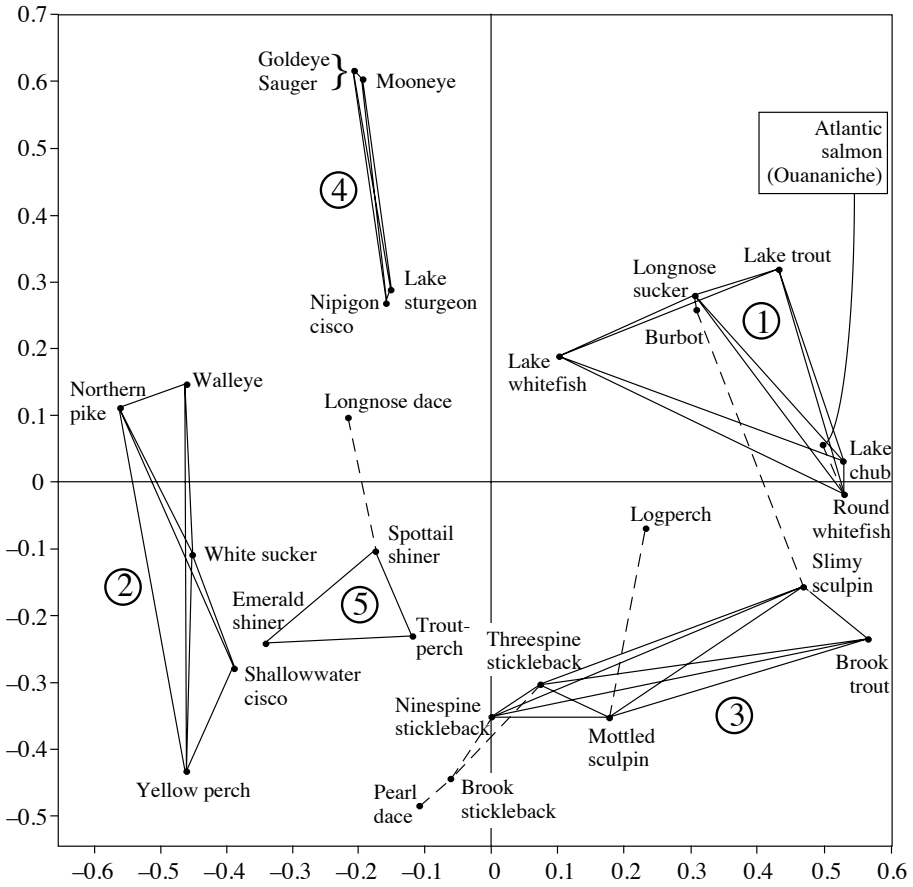


Figure 8.19 Fish associations drawn on a two-dimensional principal coordinate ordination of the species. Axes I (abscissa) and II (ordinate) explain together 55% of the variability among species. Full lines link species that are members of associations identified by non-hierarchical complete linkage clustering at $S \geq 0.989$. Dashed lines attach satellite species to the most closely related species that is a member of an association. The five associations are identified by circled numbers. Redrawn from Legendre & Beauvais (1978).

3)	Brook trout	*Salvelinus fontinalis*	$S_{25} \geq 0.991$ **
	Ninespine stickleback	*Pungitius pungitius*	
	Mottled sculpin	*Cottus bairdi*	
	Threespine stickleback	*Gasterosteus aculeatus*	
	Slimy sculpin	*Cottus cognatus*	

4)	Nipigon cisco	*Coregonus nipigon*	$S_{25} \geq 0.991$ **
	Lake sturgeon	*Acipenser fulvescens*	
	Goldeye	*Hiodon alosoides*	
	Mooneye	*Hiodon tergisus*	
	Sauger	*Stizostedion canadense*	
5)	Trout-perch	*Percopsis omiscomaycus*	$S_{25} \geq 0.989$ *
	Spottail shiner	*Notropis hudsonius*	
	Emerald shiner	*Notropis atherinoides*	

The six remaining species were attached as satellites, by single linkage chaining, to the association containing the closest species. Figure 8.19 shows the species associations drawn on a two-dimensional principal coordinate ordination (Section 9.3) of the species. Three of these associations can be interpreted ecologically. Association 1 was characteristic of the cold, clear, low-conductivity lakes of the Laurentide Shield. Association 2 characterized lakes with warmer and more turbid waters, found in the lowlands. Association 4 contained species that were all at the northern limit of their distributions; they were found in the southern part of the study area.

2 — Concordance analysis

Concordance analysis, which is based upon Kendall's coefficient of concordance (Section 5.4), is useful to delineate groups of species that form statistically significant associations. Described by Legendre (2005), the method proceeds in three steps.

1. Perform a correlation analysis to identify groups of positively correlated species. The most widely used method is to compute Ward's agglomerative clustering (Subsection 8.5.8) of a matrix of correlations among the species. In detail:

1.1. Transform the species abundances using one of the transformations described in Section 7.7. Several transformations may be tried in turn and the results compared.

1.2. Compute a Pearson or Spearman correlation matrix $\mathbf{R} = [r_{hi}]$ among the species. This is done to make the clustering results compatible with concordance analysis, which is based on correlations. Turn matrix **R** into a distance matrix by computing $\mathbf{D} = [1 - r_{hi}]$.

1.3. Carry out Ward's hierarchical clustering of that matrix.

1.4. Cut the dendrogram in two groups and retrieve the vector of species membership.

1.5. After steps 2 and 3 below, one may have to come back and try divisions of the species into 3, 4, 5, ... groups.

In simple cases, a principal component analysis (PCA, Section 9.1) of the standardized transformed species abundance data may be sufficient to delineate species groups on which steps 2 and 3 can be carried out. Because the transformed species data are standardized by columns, the PCA will be computed on the correlation matrix, making the results compatible with concordance analysis, which is based on correlations.

2. Compute global tests of significance of the concordance within the two (or more) groups (Subsection 5.4.2) using the matrix of transformed species abundances (e.g. after Hellinger transformation). Groups that are not globally significant must be refined (step 1.5) or abandoned.

3. Compute *a posteriori* tests of the contribution of individual species to the concordance of their group (Subsection 5.4.3). If the mean of the Spearman correlations of a species with all the other species of its group is negative, this indicates that this species clearly does not belong to the group, hence that group is too inclusive. Go back to step 1.5 and cut the dendrogram more finely. Groups can be refined (i.e. cut into smaller groups) separately from other groups, independently of the levels along the dendrogram.

Use corrections for multiple testing (Box 1.3) at all steps of this analysis. R functions to carry out these tests are described in Section 8.15.

This method should only be applied to species abundance data because Kendall's concordance analysis is meaningless for presence-absence data. Other methods should be developed to deal with presence-absence data. The Kendall concordance approach is useful in environmental studies where researchers are interested in identifying groups of concordant species that are indicators of some property of the environment. In some applications, significantly concordant species can be combined into environmental quality indices (Siegel, 1956), in particular in situations of pollution or contamination, and used to produce indicator maps.

Ecological application 8.9c

Legendre (2005) used the Kendall coefficient of concordance (W) to identify species associations in a multi-species community of oribatid mites (35 species, 70 soil cores[*]; Borcard & Legendre, 1994). The mite data were subjected to the Hellinger transformation (eq. 7.69) at the beginning of the analysis. Ward's agglomerative clustering (Subsection 8.5.8) and K-means partitioning (Section 8.8) both suggested the presence of two groups of mites, one including 24 species and the other 11 (Fig. 8.20). Kendall coefficients of concordance computed over each group separately indicated that both groups had significant concordance. *A posteriori* tests showed that 20 species of the first group and 8 of the second group significantly contributed to the concordance of their group, at the 5% significance level and after Holm correction for multiple testing (Box 1.3). The abundances of the species that were significant members of a group were summed over each group and the sums were plotted on maps of the study area. These indices were related to environmental variables by multiple regression (Section 10.3), which produced highly significant environmental models.

The PCA ordination diagram (Section 9.1) shown in Fig. 8.20 was computed after standardizing the Hellinger-transformed species vectors (eq. 1.12); the two mite associations identified by clustering followed by concordance analysis are represented by symbols on the plot. An alternative graphical presentation of the clustering results would be a heat map (Section 8.10) with dendrograms added to the sides; R functions to produce heat maps are listed in Section 8.15. When environmental descriptors are available, computing an RDA (Section 11.1) instead of a PCA will produce a plot providing an interpretation of the differences among the groups of species. Another example (fish associations) is presented in Section 4.10.2 of Borcard *et al*. (2011).

[*] The mite data are available on the Web page of the Borcard *et al*. (2011) book, http://numericalecology.com/NEwR.

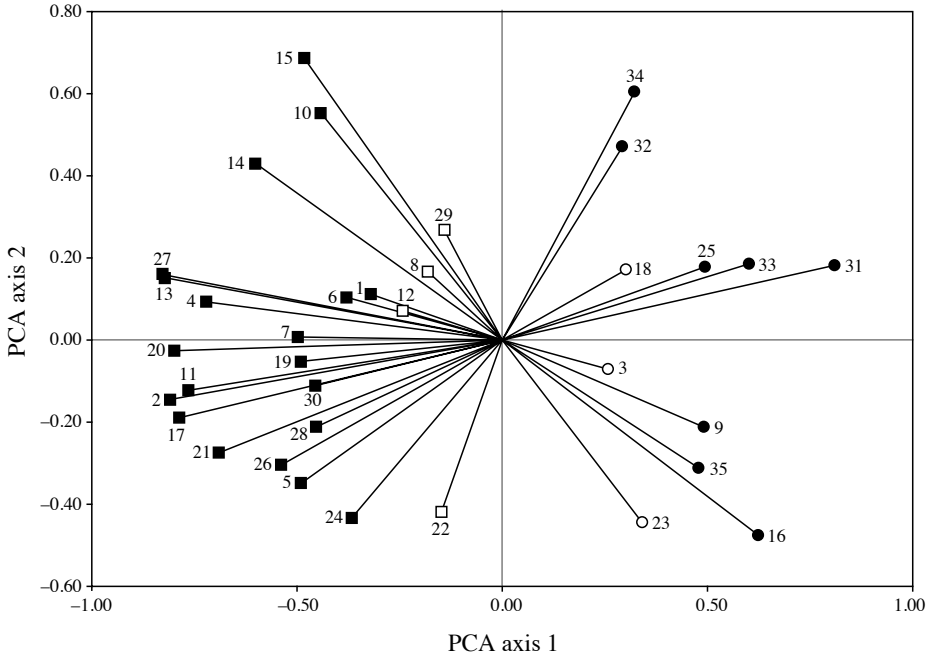


Figure 8.20 Principal component ordination diagram showing the species vectors projected in the space formed by PCA axes 1 (28.8% of the variation) and 2 (9.2%). Mite group 1 resulting from the preliminary Ward clustering is represented by squares and group 2 by circles. Solid symbols: species that are significant members of their respective associations. Modified from Legendre (2005, Fig. 4).

3 — Indicator species

The identification of indicator (or characteristic) species is a traditional question in ecology and biogeography. Field studies describing sites or habitats usually mention one or several species that characterize each habitat. For many years, the most widely used statistical method for identifying indicator species was TWINSPAN (Hill, 1979a; Subsection 8.7.4). There is clearly a need for the identification of characteristic or indicator species in the fields of monitoring, conservation, and management, as discussed below. Because indicator species add ecological meaning to groups of sites discovered by clustering, they provide criteria to compare typologies derived from data analysis, to identify where to stop dividing clusters into subsets, and to point out the main levels in a hierarchical classification of sites. *Indicator species* differ from *species associations* in that they are indicative of particular groups of sites. Good indicator species should be found mostly in a single group of a typology and be present

at most of the sites belonging to that group. This duality is of ecological interest; yet it is seldom exploited in indicator species studies.

Dufrêne & Legendre (1997) proposed an alternative to TWINSPAN in the search for indicator species and species assemblages characterizing groups of sites. Like TWINSPAN, their method is *asymmetric*, meaning that species are analysed on the basis of a *prior* partition of the sites. The first original characteristic of the method is that it derives indicator species from any hierarchical or non-hierarchical classification of the objects (sampling sites), contrary to TWINSPAN where indicator species can only be derived for classifications obtained by splitting sites along correspondence analysis (CA) or detrended correspondence analysis (DCA) axes (Subsection 8.7.4). The second original characteristic lies in the way the indicator value of a species is measured for a group of sites. The *indicator value index* (*INDVAL*) is based only on within-species abundance and occurrence comparisons; its value is not affected by the abundances of other species. The significance of the indicator value of each species is assessed by a randomization procedure (Section 1.2).*

The *indicator value* (*INDVAL*) index is defined as follows. For each species j in each cluster of sites k, one computes the product of two values, A_{kj} and B_{kj}. A_{kj} is a measure of *specificity* based on abundance values whereas B_{kj} is a measure of *fidelity* computed from presence data:

Specificity $$A_{kj} = Nindividuals_{kj} / Nindividuals_{+k}$$

Fidelity $$B_{kj} = Nsites_{kj} / Nsites_{k+}$$

$$INDVAL_{kj} = A_{kj} B_{kj} \qquad \textbf{(8.21)}$$

• In the formula for specificity (A_{kj}), $Nindividuals_{kj}$ is the mean abundance of species j across the sites pertaining to cluster k and $Nindividuals_{+k}$ is the sum of the mean abundances of species j within the various clusters. The *mean* number of individuals in each cluster is used, instead of summing the individuals across all sites of a cluster, because this removes any effect of variations in the number of sites belonging to the various clusters. Differences in abundance among sites of a cluster are not taken into account in the calculation. A_{kj} is maximum when species j is present in cluster k only.

• In the formula for fidelity (B_{kj}), $Nsites_{kj}$ is the number of sites in cluster k where species j is present and $Nsites_{k+}$ is the total number of sites in that cluster. B_{kj} is maximum when species j is present at all sites of cluster k.

• Quantities A and B must be combined by multiplication because they represent independent information (i.e. specificity and fidelity) about the distribution of species j.

* How to compute *INDVAL* in R is described in Section 8.15. The *INDVAL* index is also available in package PC-ORD; distribution: see footnote in Section 11.7.

In De Cáceres & Legendre (2009), *specificity* is called *positive predictive value* and *fidelity* is called *sensitivity*.

The indicator value of species j for a partition of sites is the largest value of $INDVAL_{kj}$ observed over all clusters k of that partition:

Indicator value

$$INDVAL_j = \max [INDVAL_{kj}] \quad \textbf{(8.22)}$$

The index is maximum (its value is 1) when the individuals of species j are observed at all sites belonging to a single cluster. A random permutation procedure of the sites among the site groups is used to test the significance of $INDVAL_j$ (Section 1.2). A correction for multiple testing (Box 1.3) is in order before reporting the results since multiple tests (i.e. for p species) are conducted. The index can be computed for any given partition of sites, or for all levels of a hierarchical classification of sites.

Numerical example. Table 8.10 describes the example given by Dufrêne & Legendre (1997) to illustrate the computation of the *INDVAL* index, slightly modified. The data represent three species observed at 25 sites, which are divided into 5 groups. To facilitate comparisons, the sums of the mean group abundances are 20 for all three species. For species 1, $INDVAL_{k1}$ has the highest value (0.30) for group $k = 3$, so $INDVAL_1 = 0.30$. Following similar reasoning, $INDVAL_2 = 0.40$ and $INDVAL_3 = 0.90$. The permutational p-values computed by functions ***indval()*** of LABDSV or ***multipatt()*** of INDICSPECIES in R are significant in all three cases.

De Cáceres & Legendre (2009) described several other statistics that can be used to measure the indicator value of species. They are divided into *correlation indices*, which are used for determining the ecological preferences of species among a set of alternative site groups or site group combinations, and *indicator value indices*, including *INDVAL*, which are used for assessing the predictive values of species as indicators of the conditions prevailing in site groups, e.g. for field determination of community types or ecological monitoring. Each of these categories of indices comes in different types: there are indices for *presence-absence* and for *quantitative* species data; there are also *non-equalized indices* that give equal weights to individual sites and *group-equalized indices* that give equal weights to all groups whatever the number of sites they contain. For studies involving several groups of sites, De Cáceres *et al*. (2010) showed how to improve the interpretation of indicator value analysis by computing the statistics for all possible combinations of site groups. An application of that method is found in Moretti *et al*. (2010).

De Cáceres *et al*. (2010) present a detailed discussion of the limitations of indicator value analysis. In particular, they point out that more indicator species will be found than expected by chance when the classification of sites has been obtained from the same species composition data that are used for *INDVAL* analysis. In that case, p-values must be interpreted with caution: they are not the result of a genuine test of significance, where the classification of sites has to be independent of the species data used in the test.

Table 8.10 Numerical example: abundance of three species at 25 sites divided into 5 groups. Modified from Dufrêne & Legendre (1997). Top panel: data. Bottom panel: calculation of the specificity (A_{kj}), fidelity (B_{kj}) and $INDVAL_{kj}$ index for each species (j) in each group of sites (k). The maximum value of $INDVAL_{kj}$ for each species is in bold.

Groups	Group 1					Group 2					Group 3					Group 4					Group 5				
Sites	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Species 1	4	4	4	4	4	5	5	5	5	5	6	6	6	6	6	3	3	3	3	3	2	2	2	2	2
Species 2	8	8	8	8	8	4	4	4	4	4	6	6	6	6	6	4	4	2	0	0	0	0	0	0	0
Species 3	18	18	18	18	18	2	2	2	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

	Group 1	Group 2	Group 3	Group 4	Group 5
Species 1					
A_{k1}	4/20 = 0.20	5/20 = 0.25	6/20 = 0.30	3/20 = 0.15	2/20 = 0.10
B_{k1}	5/5 = 1	5/5 = 1	5/5 = 1	5/5 = 1	5/5 = 1
$INDVAL_{k1}$	0.20	0.25	**0.30**	0.15	0.10
Species 2					
A_{k2}	8/20 = 0.40	4/20 = 0.20	6/20 = 0.30	2/20 = 0.10	0/20 = 0.00
B_{k2}	5/5 = 1	5/5 = 1	5/5 = 1	3/5 = 0.6	0/5 = 0
$INDVAL_{k2}$	**0.40**	0.20	0.30	0.06	0.00
Species 3					
A_{k3}	18/20 = 0.90	2/20 = 0.10	0/20 = 0.00	0/20 = 0.00	0/20 = 0.00
B_{k3}	5/5 = 1	5/5 = 1	0/5 = 0	0/5 = 0	0/5 = 0
$INDVAL_{k3}$	**0.90**	0.10	0.00	0.00	0.00

Podani & Csányi (2010) proposed variants of the *INDVAL* index. Instead of using specificity and fidelity alone, they proposed to define the indicator value of a species as the product of two among three quantities: specificity A_{kj} (that they renamed concentration), specificity (new equation, with allowance for positive or negative species preferences), and fidelity B_{kj}. They provided formulas based on either presence-absence or abundance data for each of these three quantities.

McGeoch & Chown (1998) found the indicator value method important to conservation biology because it is conceptually straightforward and allows researchers to identify bioindicators for any combination of habitat types or areas of interest, e.g. existing conservation areas, or groups of sites based on the outcome of a classification procedure. In addition, it may be used to identify bioindicators for groups of sites classified using the target taxa, as in Ecological application 8.9d, or using non-target taxa, e.g. insect bioindicators of plant community classifications.

Because each *INDVAL* index is calculated independently of other species in the assemblage, comparisons of indicator values can be made between taxonomically unrelated taxa, taxa in different functional groups, or those in different communities. Comparisons across taxa are robust to differences in abundance that may or may not be due to differences in detectability or visibility, or to sampling methods. The method is also robust to differences in the numbers of sites between site groups, to differences in abundance among sites within a particular group, and to differences in the absolute abundances of very different taxa that may exhibit similar trends.

When a group of sites for which indicator species are sought corresponds to a delimited geographic area, superposition of the distribution maps for the indicator species of that group should help delineate the core conservation areas for these species, even when little other biological information is available. McGeoch & Chown (1998) also consider the *indicator measure of a species absence* to be of value. The species absence *IndVal* provides a method for improving the objectivity with which species transient to an assemblage can be identified. Species with high values for this absence index may also be of ecological interest as indicators of peculiar ecological conditions where the species is seldom or never present.

Taxa proposed as bioindicators in the literature are often merely the favourite taxa of their proponents; ornithologists prefer birds, lepidopterists butterflies, and coleopterists beetles. According to McGeoch & Chown (1998), *IndVal* provides an objective method for addressing this problem by enabling the assessment of the relative merits of different taxa as bioindicators for a given study area. The species that do emerge from this procedure as the most useful indicators for a group of sites should prove useful in practical conservation for monitoring site change and disturbance. Two groups of species collected at the same sites can be compared by co-inertia analysis (CoIA, Section 11.5, see Ecological application 11.5) and several groups by multiple factor analysis (MFA, Section 11.5).

Ecological application 8.9d

In order to illustrate the indicator value method, Dufrêne & Legendre (1997) used a large data set of Carabid beetle distributions in open habitats of Belgium (189 species collected in pitfall traps, for a total of 39984 specimens). The data represented 123 year-catch cycles at 69 locations; a year-catch cycle cumulates catches at a site during a full year; 54 sites were studied during two years and 15 sites were sampled during a single year. The typology of sites was computed by distance-based *K*-means partitioning computed as follows: first, a distance matrix

(percentage difference D_{14}, eq. 7.58) was computed from the log-transformed species abundance data; this distance matrix was subjected to principal coordinate analysis (PCoA, Section 9.3); all principal coordinates were then used as input data into *K*-means partitioning. Although the clusters produced by *K*-means had not been forced to be hierarchically nested, they showed a strong hierarchical structure for $K = 2$ to 10 groups. This allowed the authors to represent the relationships among partitions as a dendrogram. The $K = 10$ level corresponded to the main types of habitat, recognized *a priori*, where sampling had been conducted.

Indicator values were computed for each species and partitioning level. Some species were found to be stenotopic (narrow niches) while others were eurytopic (species with wide niches, present in a variety of habitats). Others characterized intermediate levels of the hierarchy. The best indicator species ($INDVAL > 0.25$) were assembled into a two-way indicator table; this tabular representation displayed the hierarchical relationships among species.

Results of the indicator value method were compared to TWINSPAN. Note that the partitions of sites used in the two methods were not the same; the TWINSPAN typology was obtained by partitioning correspondence analysis ordination axes (Subsection 8.7.4). TWINSPAN identified, as indicators, pseudospecies pertaining to very low cut-off levels. These species were not particularly useful for prediction because they were simply known to be present at all sites of a group. Several species identified by TWINSPAN as indicators also received a high indicator value from the *INDVAL* procedure, for the same or a closely related habitat class. The *INDVAL* method identified several other indicator species, with rather high indicator values, that also contributed to the specificity of the groups of sites but had been missed by TWINSPAN. So, the *INDVAL* method appeared to be more sensitive than TWINSPAN to the fidelity and specificity of species.

Here are some more examples of the many applications of indicator species analysis found in the literature. Borcard (1996) and Borcard & Vaucher-von Ballmoos (1997) present applications of the indicator value method to the identification of the Oribatid mite species that characterize well-defined zones in a peat bog of the Swiss Jura. The indicator values of beetle species characterizing different types of forests have been studied by Barbalat & Borcard (1997). Tuomisto *et al*. (2003) used constrained clustering (Subsection 12.6.4) to group 86 sampling units, each 500 m long, forming a 43-km long transect in the Amazonian rain forest into spatial clusters, on the basis of satellite image pixel values. They also surveyed in the field the ferns and *Melastomaceae* observed in the 86 sampling units. Then they used the *INDVAL* method to determine the species that were good indicators of the spatial clusters.

Legendre *et al*. (2009) used multivariate regression tree analysis (MRT, Section 8.11) to identify habitat types that were similar in topographic conditions and in species composition in a Chinese permanent forest plot divided in 20×20 m quadrats; then they used the *IndVal* method to identify the nine, among 159 tree species, that were statistically significant indicators of the five main habitat types. De Cáceres *et al*. (2010) carried out indicator species analysis of the vegetation of the Barro Colorado Island (BCI) permanent forest plot in Panama, also divided in 20×20 m quadrats, grouped into seven habitat types identified in a previous paper. Among 307 tree species, they identified 44 indicator species of individual habitats and 64 for habitat combinations. In the first of these papers, the species used for *IndVal* analysis had been used to obtain the classification of the sites, so that the p-values had to be

interpreted with caution. In the second paper, the classification of the sites was independent of the species analysed for indicator value.

8.10 Seriation

Before clustering methods were developed, the structure of an ecological resemblance matrix was often studied by *matrix rearrangement* (Orlóci, 1978). In this approach, the order of the objects is modified in such a way as to concentrate the lowest distances (or the highest similarities) near the main diagonal of the resemblance matrix. This is a special case of an approach called *seriation* in archaeology, where the rows and columns of a *rectangular* matrix of (artefacts × descriptors) are rearranged in such a way as to bring the highest values near the main diagonal, in order to evidence the temporal seriation of the artefacts; see Kendall (1988) for a review. This technique was developed by anthropologists Petrie (1899) and Czekanowski (1909) and was first applied to ecological data by Kulczynski (1928). An interesting aspect of seriation for ecologists, nowadays, lies in the fact that the technique can be applied to the special case of similarity or distance matrices that are not symmetric, as explained below.

The statistical theory of seriation is now well developed. Papers and syntheses are found mostly in the archaeological literature, e.g. Renfrew & Bahn (2008). Hahsler *et al*. (2008) describe different seriation methods that can be used to visualize (objects × descriptors) or distance (**D**) matrices, reordered or not according to clustering results, and cite the relevant literature. These methods can only be used with small data sets. That paper is also an introduction to the R package SERIATION (Section 8.15).

Trellis diagram
Heat map

A rearranged resemblance matrix can be represented by a *trellis diagram*, called a *heat map* in recent software, which is a shaded or colour-coded matrix. Figure 8.21a gives an example where half of the matrix is represented by shades of gray corresponding to distance values, and Fig. 8.21c shows a heat map of the same distance matrix computed by an R function. Heat maps provide an interesting representation of a raw data or distance matrix, before or after clustering, when the number of objects is small, e.g. 30 or less. In R (Section 8.15), heat maps without or with dendrograms can be produced by functions ***heatmap()*** of package STATS and ***hmap()*** of SERIATION. Function ***coldiss()*** (Section 8.15) plots side by side an original and reordered **D** matrix without dendrogram. Examples are given in Borcard *et al*. (2011, Subsections 3.3.2 and 4.7.3.7).

Seriation works best when there is a single gradient in the data. For symmetric matrices, the order of the objects in any agglomerative clustering dendrogram can be used as the seriation order. A minimum spanning tree (Section 8.2) computed for the **D** matrix (Fig. 8.21b) provides details about the structure and shows if the single-gradient assumption holds for at least part of the objects ordered in a trellis diagram. At the end of the seriation procedure (Fig. 8.21a), the lowest distances, which are now found close to the diagonal, indicate the first important clusters of objects. The first

axis of an ordination diagram (Chapter 9) provides another optimal order of objects, which can be used in a trellis diagram.

Non-symmetric matrix

Seriation is an interesting approach for the analysis of non-symmetric distance matrices. Non-symmetric matrices, in which $D(\mathbf{x}_1, \mathbf{x}_2) \neq D(\mathbf{x}_2, \mathbf{x}_1)$, are rare in ecology. They may, however, be encountered in cases where the resemblance is a direct measure of the influence of an organism on another, or in behavioural studies where the attraction of an organism for another can be used as a similarity measure. They are also common in taxonomic and phylogenetic analysis (e.g. serological data, DNA pairing data).

An analytical solution to seriation that can be applied to non-symmetric as well as symmetric matrices was proposed by Beum & Brundage (1950). The algorithm starts with a similarity matrix (**S**) among objects, provided in any order. The diagonal values are excluded from the calculation, so that the "similarities" in the matrix can be any quantitative indications of preference, not necessarily with a maximum value of 1. In each column j, the products of the elements s_{ij} by the *inverse order numbers of the rows* are summed and divided by the sum of the elements in column j. These average weights are used to determine the new order of the rows and columns, from which the procedure starts over again until convergence is reached. The algorithm may at times end up alternating between two equally optimal final solutions. An R function is available to carry out the Beum-Brundage seriation procedure; see Section 8.15.

Besides seriation, non-symmetric matrices can be decomposed into symmetric and skew-symmetric components, as described in Subsection 2.3, before analysis by clustering and/or ordination methods.

Ecological application 8.10a

Kulczynski (1928) studied the phytosociology of a region in the Carpathian Mountains, southeastern Poland. He recognized 37 plant associations, listed the species found in each, and computed a similarity matrix among them. Part of that similarity matrix, turned into a **D** matrix, is reproduced in Fig. 8.21a. The order of the associations shown in the figure is the one that Kulczynski found when he performed seriation by hand. He interpreted that order as representing a series from association 22 (*Varietum pinetosum czorsztynense*) to association 13 (*Seslerietum variae normale*). The blocs of higher (darker) values near the diagonal allow one to recognize two main groups: associations (22, 21) and (11, 15, 14, 17, 18). Association 13 and 16 seem less related with the others. A minimum spanning tree computed for the same **D** matrix (Fig. 8.21b) provides more detail about the structure of the data. The corresponding heat map is shown in Fig. 8.21c. The dendrogram shown along both axes was obtained by complete linkage clustering of matrix **D**. *Note*: seriation produces a clear one-dimensional ordination when there is a single gradient in the data, which is not the case here.

Ecological application 8.10b

Wieser (1960) studied the meiofauna (small benthic metazoans) at three sites (6 or 7 cores per site) in Buzzards Bay, Massachusetts, USA. After representing the resemblance among cores as

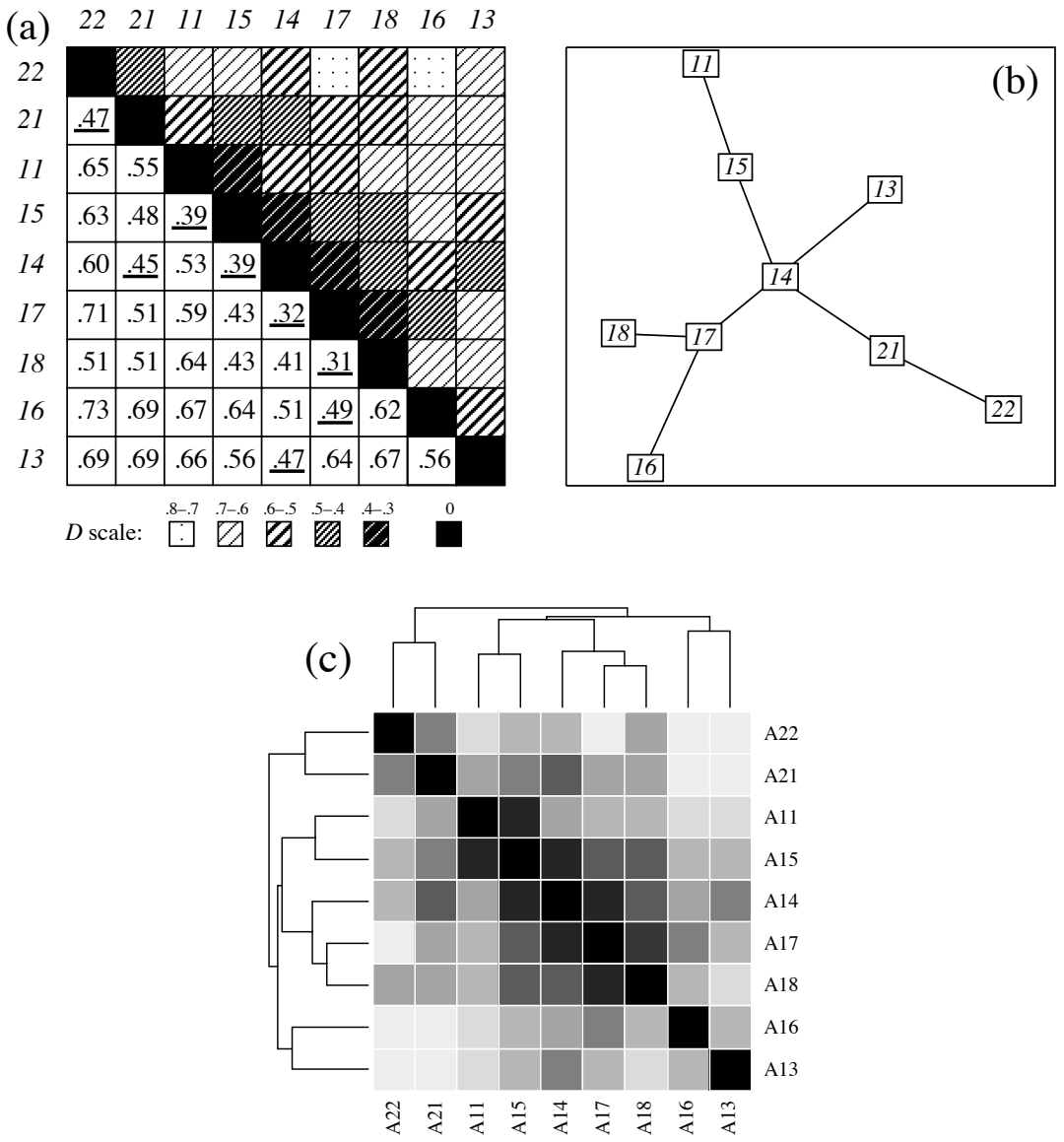


Figure 8.21 (a) Distance matrix (lower half) and trellis diagram (upper half) for part of Kulczynski's (1928) plant associations of the Carpathian Mountains. Numbers in italics, in the margins, identify the associations. In the trellis diagram, the distances are represented by shadings, as indicated underneath the matrix. (b) Minimum spanning tree computed for the same **D** matrix with function ***spantree()*** of the VEGAN R package. The edges of the tree correspond to the underscored distance values in (a). (c) Heat map of the distance matrix with dendrograms shown along both axes. The picture produced by function ***heatmap()*** was in colour.

a similarity matrix (using Whittaker's index of association, $1 - D_9$) and a trellis diagram, he found that although the three sites differed in species composition, the two sandy sites were more similar to each other than they resembled the third site where the sediment contained high concentrations of fine deposits.

The classical study reported in Ecological application 8.10b encouraged other applications of trellis diagrams in benthic ecology. Among these is Sanders' (1960) representation of an ecological time series, also from Buzzards Bay, using a trellis diagram. Inspired by these applications to benthic ecology, Guille (1970) and Soyer (1970) used the method of trellis diagrams to delineate benthic communities (macrofauna and harpacticoid copepods, respectively) along the French Catalonian coast of the Mediterranean Sea near Banyuls-sur-Mer.

Wieser's (1960) study offers an opportunity to come back to the warning of Section 8.0, that not all problems of data analysis belong to the clustering approach. Nowadays, one would not have to seriate or cluster the sites before comparing the species to the sediment data. One would directly compare the species abundance to the sediment data, or to a factor representing the three study sites, using canonical analysis (Chapter 11).

8.11 Multivariate regression trees (MRT)

Univariate *classification tree* analysis (CT) refers to situations where a qualitative response variable is to be predicted by a decision tree (defined below), whereas in *regression tree* analysis (RT) the response variable is quantitative. *Classification and regression tree* analysis (CART, Breiman *et al.,* 1984) combines these two procedures. A decision tree is a forecasting or predictive tree-like diagram resulting from recursive partitioning of the response data, with indication of the influence of the explanatory variables at each split; examples are given below. These univariate forms of analysis are not discussed further in the present chapter.

Proposed by De'ath in 2002 and Larsen & Speckman in 2004, *multivariate regression tree* analysis (MRT) is an extension of CART to multivariate response data. The method could have been presented in Chapter 11 devoted to canonical analysis since, like RDA and CCA, it involves a response and an explanatory data set. It is presented here instead because its output is a tree.

Figure 8.22a shows a simple example with a multivariate response data set **Y** on the left and a matrix of explanatory variables on the right. There are three explanatory variables in **X**; $\mathbf{x}_1$ and $\mathbf{x}_2$ are quantitative in this example and $\mathbf{x}_3$ is qualitative (three levels or states: A, B and C). For the first split, the analysis will search for the best partition of **Y** in two groups, constrained by each variable **x** in turn.

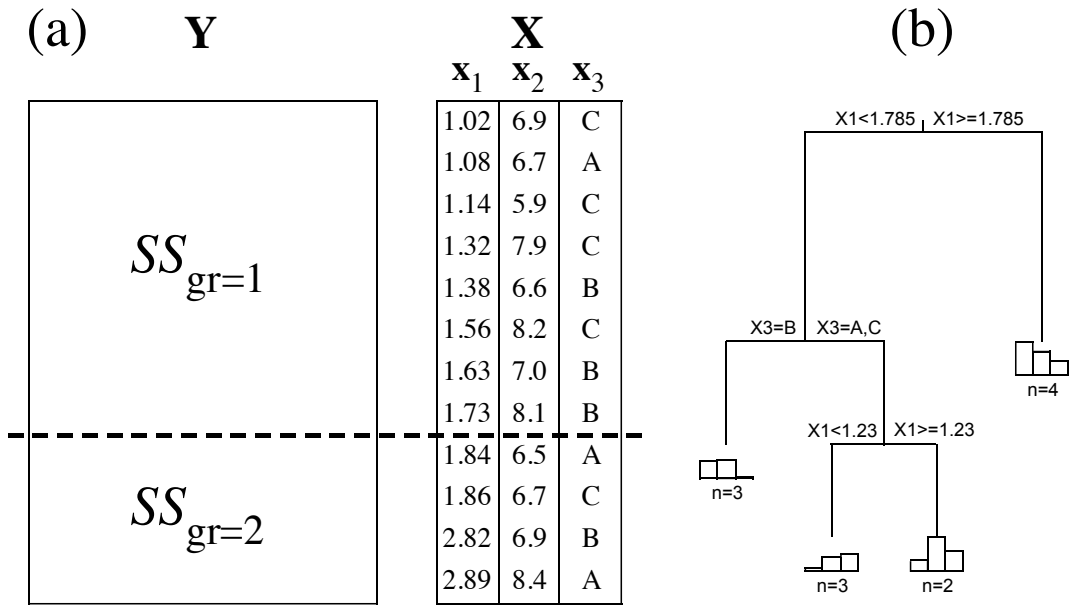


Figure 8.22 Schematic description of MRT analysis. (a) Data: **Y** is the response data set. There are three explanatory variables in **X**: $\mathbf{x}_1$ and $\mathbf{x}_2$ are quantitative in this example and $\mathbf{x}_3$ is qualitative (three factor levels or qualitative states). The dashed horizontal line indicates a cut-point along the values of $\mathbf{x}_1$. The line is extended across **Y**, which is thus divided into two groups. (b) Multivariate regression tree computed by function ***mvpart()*** of the MVPART package; there were 3 species in **Y** (not shown) in this analysis. Variable $\mathbf{x}_1$ controls the first split (the split occurs at the position of the dashed line in panel a); variable $\mathbf{x}_3$ controls the second split (objects with level 1 on the left, those with levels 2 and 3 on the right); variable $\mathbf{x}_1$ is used again for the third split. The number of objects in each group is shown underneath each leaf (terminal group) of the tree, together with a histogram showing the relative abundances of the three species in **Y**.

• For variable $\mathbf{x}_1$, imagine that the rows of the two data sets, **Y** and **X**, are ordered by increasing $\mathbf{x}_1$ values, as in Fig. 8.22a (the actual programming may differ from the description that follows). The program tries in turn all possible cut-points of variable $\mathbf{x}_1$. For each cut-point between successive but different values of $\mathbf{x}_1$, imagine a line drawn across **Y** (dashed line in Fig. 8.22a); it divides **Y** in two groups. $SS_{gr=1}$ is the sum of within-group sums-of squares (also called squared error) for the top group (gr=1), computed using eq. 8.5, and $SS_{gr=2}$ is the sum of within-group sums-of squares for the bottom group (gr=2). So the total within-group sum-of-squares, or total error, for that split of the objects is $E^2 = SS_{gr=1} + SS_{gr=2}$ (eq. 8.7). Because of the equivalence of eqs. 8.5 and 8.6 for the computation of squared error, one can compute MRT from a raw data file **Y** or from a distance matrix **D** computed from **Y**.

• The function tries in turn all possible cut-points along $\mathbf{x}_1$, making no cut between identical (tied) values, and computes $E^2_{\mathbf{x}1}$. It notes the position of the cut where E^2 is minimum for variable $\mathbf{x}_1$ and the value of $E^2_{\mathbf{x}1}$ at that point.

• The process is repeated for variable $\mathbf{x}_2$: the rows of the two data matrices are reordered in such a way that the values of $\mathbf{x}_2$ are in increasing order, all possible cut-points between non-identical values are tried in turn, and the cut that produces the smallest value of $E^2_{\mathbf{x}2}$ is noted.

• The third variable in Fig. 8.22a is a qualitative variable or ANOVA factor. All possible combinations of factor levels are tried in turn. In this example, only three solutions need to be studied: the group defined by state A *versus* the other objects, the group defined by state B, and finally the group defined by state C. The combination that produces the smallest value of $E^2_{\mathbf{x}3}$ is noted. (In the example, the second split separated the rows with level B from those with levels A and C.)

• All values of $E^2_{\mathbf{x}}$ (three in this illustration) are compared: min($E^2_{\mathbf{x}1}$), min($E^2_{\mathbf{x}2}$), and min($E^2_{\mathbf{x}3}$). The smallest of these values is used to draw the first split of the regression tree (Fig. 8.22b, top), which is the first split of data set **Y**.

• Each branch of the tree is then analysed separately (a branch is a group formed by a split). The search for a meaningful split is first carried out for the left branch of the tree. All explanatory variables in **X** are tried in turn and the variable that produces the split with the smallest value of $E^2_{\mathbf{x}}$ is used for the next split of the left-hand side of the tree. Similarly, the search is carried out for the objects in the right branch of the tree and the variable of **X** that produces the split with the smallest value of $E^2_{\mathbf{x}}$ is used for the next split of the right-hand side of the tree. Any variable may be used for several splits. Figure 8.22b shows a tree that was produced for a data set **Y** with 3 species (data not shown).

The process could go on until the tree is fully resolved and individual objects form the terminal groups (leaves of the tree). Users, however, are usually not interested in the fully resolved tree, but instead in a tree that has informative partitions. That shorter tree is found by pruning the tree, an operation that consists in removing the smallest branches. The optimal size of the tree is decided by a resampling analysis called cross-validation. In that analysis, the data are randomly divided into a number of approximately equal-sized *test groups*, e.g. 10% of the objects. Each test group is left aside in turn while the tree is reconstructed using the remaining objects, e.g. 90%. Then, distances are computed from each object of the test group to the multivariate centroids of the groups forming the *leaves* (terminal groups of objects) of the tree. The objects of the test group are attributed to the closest leaf of the reconstructed tree.

Leaf

The objects in the test group are attributed to the closest leaf of a tree with 2 groups, 3 groups, etc., considering the distances of the objects to the centroids of the groups. An overall relative error statistic (cross-validation relative error, *CVRE*) is

computed as follows for each partition size using all n objects (that is done by using the predictions made for the members of all test groups in the formula):

Cross-validation error

$$CVRE = \frac{\sum_{i=1}^{n}\sum_{j=1}^{p}\left(y_{ij(k)} - \hat{y}_{j(k)}\right)^2}{\sum_{i=1}^{n}\sum_{j=1}^{p}\left(y_{ij} - \bar{y}_j\right)^2} \quad \textbf{(8.23)}$$

where $y_{ij(k)}$ is the value of variable j for object i belonging to test group k, $\hat{y}_{j(k)}$ is the value of that same variable at the centroid of the leaf that is closest to object i, whereas the denominator is the overall sum of squares of the **Y** data. *CVRE* is then the ratio of the variation unexplained by the tree to the total variation in **Y**. *CVRE* varies from zero for a perfect set of predictors chosen for the splits of a tree, to close to one for poor predictors; its value can actually exceed 1.

Cross-validation is repeated a number of times, e.g. 100 times, for successive and independent divisions of the objects into random test groups. Then, for each partition size (number of groups), the mean and standard error of all *CVRE* estimates is computed. The cross-validation procedure is described in more details by Borcard *et al*. (2011, Section 4.11) and Ouellette *et al*. (2012).

Should one retain a tree with a single split (2 groups), 2 splits (3 groups), or more splits? *CVRE* is used to indicate the optimal size of the tree. One can select the tree that has the smallest *CVRE* value; alternatively, and following Breiman *et al*. (1984), one may prefer a more parsimonious solution (i.e. a tree with fewer splits) whose *CVRE* value is within one standard error of the smallest *CVRE* value. In any case, *CVRE* is simply a criterion that helps researchers select the optimal tree; in the end, one can opt for a tree with fewer or more leaves (groups) than proposed by the *CVRE* criterion.

MRT belongs to the family of Euclidean methods because it is based on sums of squared deviations from means, just like ANOVA and *K*-means partitioning. The appropriateness of MRT analysis for the analysis of species data tables containing many zeros may be enhanced by transforming them following Section 7.7; this could improve the interpretability and usefulness of the trees as explanatory models of community response data.

Cascade multivariate regression tree analysis (CascadeMRT) is an extension of MRT developed by Ouellette *et al*. (2012). Users can assess their explanatory hypotheses in a hierarchical (nested) manner, carrying out MRT analyses using explanatory data matrices in the order corresponding to the hierarchy of their hypotheses. The nested hypotheses may, for example, correspond to processes operating at different spatial or temporal scales. An R package implementing cascade MRT is described in Section 8.15.

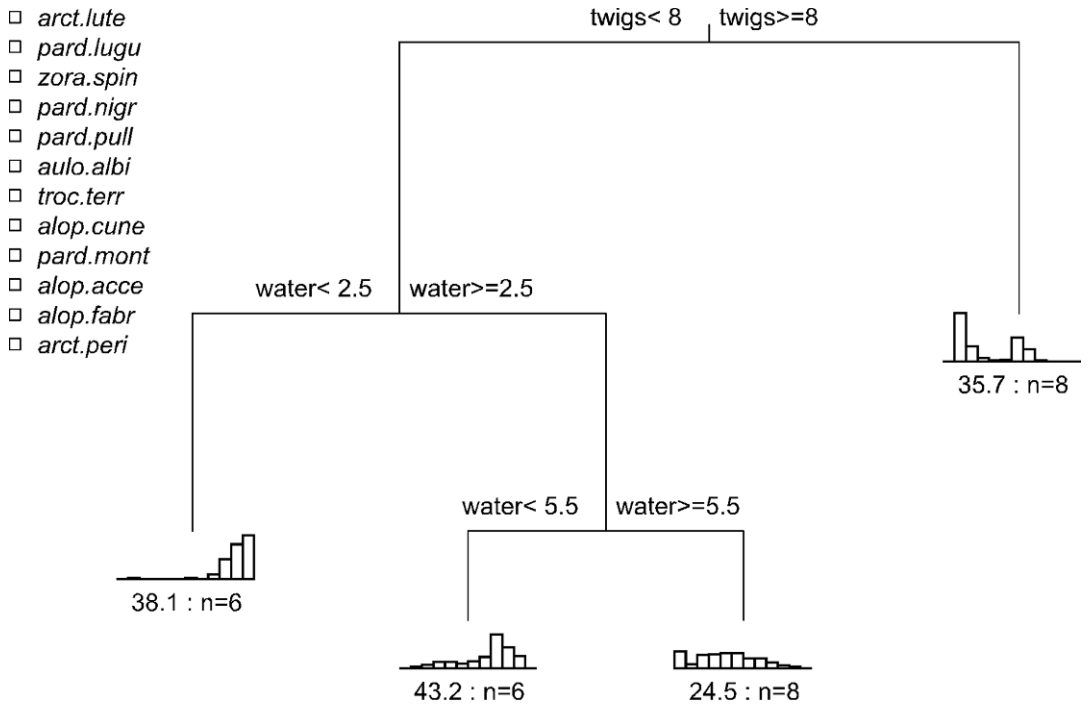


Figure 8.23 Multivariate regression tree for the hunting spider data analysed by De'ath (2002). The relative abundances of the 12 species are shown in histograms positioned at the tips of the branches, with the species in the same order as in the **Y** input file; the species names are shown in the upper-left portion of the plot as they appear in the **Y** data file. The squares in the species list and bars in the histograms have colours in the R-produced ***mvpart()*** plot. Under each histogram, *n* is the number of sites in the leaf (group); the value before *n* is the sum of squared errors for the group (eq. 8.5).

Ecological application 8.11

De'ath (2002) reanalysed the hunting spider data of Aart & Smeenk-Enserink (1975), using the spider and environmental data transformed and recoded by ter Braak (1986, Table 3); ter Braak had used these data to illustrate canonical correspondence analysis in his seminal paper. The recoded data are available in a data file of package MVPART (De'ath, 2011): 28 sites, 12 species and 6 environmental variables (water, sand, moss, light reflection, twigs, and herbs, transformed into classes from 0 to 9). Following De'ath (2002), the species data were transformed by dividing each abundance value by its column mean, then by the row mean recomputed on the resulting file. The size of the tree was selected by cross-validation: the minimum value of the cross-validation error (*CVRE* = 0.483) was used to decide on the size of the tree (4 groups, Fig. 8.23). The R^2 of that tree (1 – relative error) was 0.788. The first split separated a group of 8 sites that had more twigs (≥ 8) than the other sites; that group had higher abundances of species 2 and 7 than the other sites. The second split isolated a group of 6 sites found on dryer ground

(water < 2.5); it had higher abundances of the last two species. The last split separated two groups (n = 6 and 8 respectively) according to soil humidity (water < 5.5 *versus* ≥ 5.5); the left-hand group is dominated by species 9, while the right hand group is the only one to show substantial abundances of species 1, 4, 5 and 6. De'ath (2002) confirmed the predictive values of these spider species to the groups by indicator value analysis (Subsection 8.9.3). An identical partition of the sites into four groups was obtained by applying MRT to the chi-square transformed spider data (eq. 7.70).

8.12 Clustering statistics

This section is devoted to clustering statistics. These include connectedness and isolation, and the correlation between a cophenetic matrix and the original distance matrix.

1 — Connectedness and isolation

The connectedness within clusters and their degree of isolation can be quantified using clustering statistics. Some of these measures are described here.

Connectedness

The basic statistic of a cluster k is its *number of objects*, n_k. In linkage clustering, a measure of link density of a cluster in A-space is obtained by comparing the number of objects to the number of links among them. Link density increases with the *degree of connectedness* of a cluster. Connectedness can be measured as follows (Estabrook, 1966):

$$Co = \frac{\text{number of links in a cluster}}{\text{maximum possible number of links}} \qquad \textbf{(8.24)}$$

Cohesion index

where the maximum possible number of links is $n_k(n_k - 1)/2$, with n_k being the number of objects in cluster k. This measure of connectedness varies between 0 and 1. Day (1977) proposed other related measures. One of them is the *cohesion index*, which considers only the links that exceed the minimum number of links necessary for the cluster to be connected. If this minimum number is called m, the cohesion index can be written as follows:

$$\frac{\text{No. links} - m}{[n_k (n_k - 1)/2] - m} \qquad \textbf{(8.25)}$$

For single linkage clustering, the minimum number of links necessary for n_k objects to be connected is $n_k - 1$, so that the cohesion index becomes:

$$\frac{2\,(\text{No. links} - n + 1)}{(n-1)\,(n-2)} \qquad \textbf{(8.26)}$$

which is Estabrook's (1966) normalized connectedness index. Other possible measures of cluster density are the maximum distance or minimum similarity within a cluster, and the mean distance or similarity (Estabrook, 1966).

Isolation

The degree of isolation of clusters in metric A-space (Fig. 7.2) can be measured as the distance between the two closest objects in different clusters. It may also be measured as the mean distance between all objects in one cluster and all objects in another, or else the ratio of the distance between the two closest objects to the distance between the centroids of the two clusters. These measures are ways of quantifying the distances between clusters; a clustering or ordination *of clusters* can be computed using these distances. In the context of linkage clustering without reference to a metric A-space, Wirth *et al*. (1966) used as measure of isolation the difference between the similarity at which a cluster is formed and the similarity at which it fuses with another cluster.

2 — Cophenetic correlation and related measures

Pearson's correlation coefficient, computed between the values in a cophenetic matrix (Subsection 8.3.1) and those in the original resemblance matrix (excluding the values on the diagonal), is called the *cophenetic correlation* (Sokal & Rohlf, 1962), *matrix correlation* (Sneath & Sokal, 1973) or *standardized Mantel* (1967) *statistic* (Subsection 10.5.1). It measures the extent to which the clustering result corresponds to the original resemblance matrix. When the clustering perfectly corresponds to the coefficients in the original matrix, the cophenetic correlation is 1. In R, the cophenetic distance matrix corresponding to a hierarchical clustering is computed by function ***cophenetic()*** of the STATS package. Following that, the cophenetic correlation between the original and cophenetic distance matrices can be computed using ***cor()***.

Besides the cophenetic correlation, which compares the original distances [or similarities] to those in a cophenetic matrix, matrix correlations are useful in the following situations:

- To compare any pair of resemblance matrices, such as the original distance matrix **D** of Ecological application 8.2, and a matrix of distances among the objects in a space of reduced dimension obtained from **D** by principal coordinate analysis (Section 9.3).

- To compare two distance [or similarity] matrices obtained by computing different resemblance measures on the same data.

- To compare the results of two clustering methods applied to a resemblance matrix.

- To compare different clustering levels in a dendrogram. The ultrametric matrix representing a given clustering level only contains zeros and ones in that case, as shown in Subsection 8.3.1.

Correlations take values between –1 and +1. The cophenetic correlation is expected to be positive if the original distances are compared to cophenetic distances (or similarities to similarities) and negative if distances are compared to similarities. The higher the absolute value of the cophenetic correlation, the better the correspondence between the two matrices that are compared. Ecologists might prefer to use a non-parametric correlation coefficient (Kendall's τ or Spearman's r) instead of Pearson's r, if the interest lies more in the geometric structure of the dendrogram than the actual lengths of its branches.

A cophenetic correlation cannot be tested for significance because the cophenetic matrix is not independent of the original distance or similarity matrix; one comes from the other through the clustering algorithm. In order to test the significance of a cophenetic correlation, one would have to pretend that, under H_0, the two matrices may be independent of each other, i.e. that the clustering algorithm is likely to have a null efficiency. On the contrary, the relationship between two hierarchical classifications of *different data sets* about the same objects, e.g. community composition and environmental, measured by matrix correlation or some other measure of consensus (Rohlf, 1974, 1982b), can be tested for significance (Section 10.2, Fig. 10.4).

Other coefficients have been proposed to measure the goodness-of-fit between matrices. For instance, Gower's (1983) distance is the sum of the squared differences between values in the original distance matrix and the cophenetic distance matrix:

Gower distance

$$D_{\text{Gower}} = \sum_{i,j} (original\ D_{ij} - cophenetic\ D_{ij})^2 \qquad \textbf{(8.27)}$$

This measure, also called *stress 1* (Kendall, 1938), takes values in the interval $[0, \infty)$; it is used in standardized form as a measure of goodness-of-fit in nonmetric multidimensional scaling (eq. 9.49). Small values indicate high fit. Like the cophenetic correlation, this measure only has relative value when comparing clustering results obtained from the same original distance matrix. Several other such functions are listed by Rohlf (1974).

Other measures have been proposed for comparing different *partitions* of the same objects. Consider in turn all pairs of objects and determine, for each one, whether the two objects are placed in the same group, or not, by the partition. One can construct a 2×2 contingency table, similar to the one shown at the beginning of Subsection 7.3.1, comparing the pair assignments made by two partitions. The simple matching coefficient (eq. 7.1), computed on this contingency table, is called the Rand index (1971). Hubert & Arabie (1985) suggested a modified form that corrects the Rand index as follows: if the relationship between two partitions is comparable to that of partitions picked at random, the corrected Rand index returns a value near 0. The *modified Rand index* is widely used for comparing partitions.

Modified Rand index

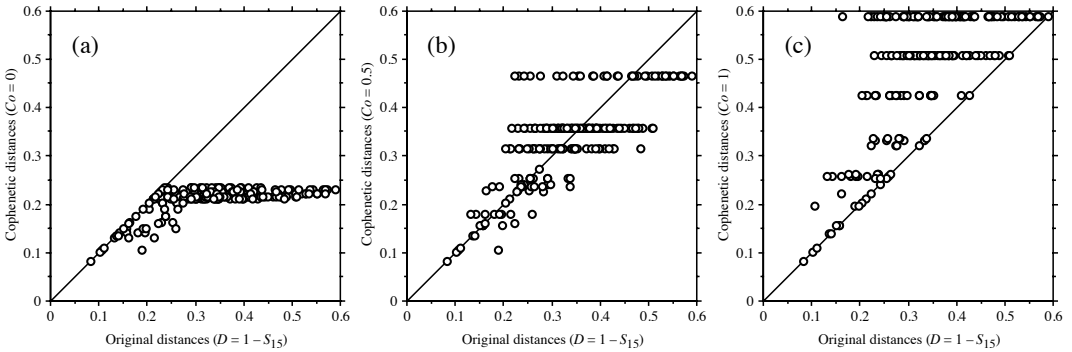


Figure 8.24 Shepard-like diagrams comparing cophenetic distances to the original distances for 21 objects analysed using three clustering methods: (a) single linkage ($Co = 0$, cophenetic $r = 0.64$, $\tau = 0.45$), (b) proportional link linkage ($Co = 0.5$, cophenetic $r = 0.75$, $\tau = 0.58$), and (c) complete linkage ($Co = 1$, cophenetic $r = 0.68$, $\tau = 0.51$). Co is the connectedness of the linkage clustering method (Subsection 8.5.3). There are 210 points (i.e. 210 distance pairs) in each graph. The diagonal lines are visual references.

Shepard-like diagram

A Shepard diagram is a scatter plot comparing distances in a space of reduced dimension, obtained by ordination methods, to distances in the original association matrix (Fig. 9.1). This type of diagram has been proposed by Shepard (1962) in the paper where he first described nonmetric multidimensional scaling (Section 9.4). Shepard-like diagrams can be constructed to compare the distances (or similarities) of the cophenetic matrix (Section 8.3) to the distances (or similarities) of the original resemblance matrix (Fig. 8.24). Such a plot may help choose between parametric and nonparametric cophenetic correlation coefficients: if the relationship between the original and cophenetic distances is curvilinear in the Shepard-like diagram, as it is the case in Figs. 24a and c, a nonparametric correlation coefficient should be used.

Figure 8.24 also helps in understanding the space-contraction effect of single linkage clustering, where the cophenetic distances are always smaller than or equal to the original distances; the space-conservation effect of intermediate linkage clustering with connectedness values around $Co = 0.5$; and the space-dilation effect of complete linkage clustering, in which cophenetic distances can never be smaller than the original distances. There are $(n - 1)$ clustering levels in a dendrogram. This limits to $(n - 1)$ the number of different values that can be found in a cophenetic matrix and, hence, along the ordinate of a Shepard-like diagram. This is why points form horizontal bands in Fig. 8.24.

Following are three measures of goodness-of-fit between the single linkage clustering results and the original distance matrix, for the pond example of Ecological application 8.2:

Pearson r cophenetic correlation = 0.9409
Kendall τ_b cophenetic correlation = 0.7736
Gower distance (D_{Gower}) = 0.1906

8.13 Cluster validation

Users of clustering methods may wonder whether the result of a clustering program run is valid or not, i.e. whether the clusters are "real", or simply artefacts of the clustering algorithm. Indeed, clustering algorithms may produce misleading results. except in simple situations where the clusters are well separated. On the one hand, most hierarchical clustering (or partitioning) algorithms will give rise to a hierarchy (or a partition), whether the objects are, or not, hierarchically interrelated (or pertaining to distinct clusters). On the other hand, different clustering algorithms may produce markedly different results because clustering methods impose different models onto the data, as shown in the present chapter: compare the dendrograms of Figs. 8.2, 8.7, 8.9, 8.11 and 8.15. Finally, different clustering methods are variously sensitive to noise (error) in the data. A simulation study comparing several clustering and partitioning methods under different levels of noise can be found in Milligan (1980); see also the review paper of Milligan (1996).

It is important to validate the results of cluster analyses. One has to show that a clustering structure departs from what may be expected from unstructured data. Unfortunately, most of the validation methods summarized below are not presently available in standard clustering packages or in R functions. Readers are referred to Chapter 4 of Jain & Dubes (1988) for details, and to the review papers of Perruchet (1983a, b), Bock (1989, 1996), Gordon (1994, 1996a, 1996b) and Milligan (1996). Lapointe (1998) reviewed the validation methods used in phylogenetic studies.

Validation may be carried out in nonstatistical or statistical ways. Statistical ways involve tests of hypotheses, whereas nonstatistical assessment accepts weaker evidence for the presence of clusters. Commonly-used nonstatistical methods are:

- Plot the clusters onto an ordination diagram and look for separation of the clusters (Section 10.1). This method is often used to assess the degree of refinement of hierarchical clustering results that one should consider for interpretation.

- Compare the results of several clustering algorithms, either informally (using visual examination, identify the partition levels that are found in most or all trees being compared), or formally (calculate consensus indices or construct a compromise "consensus" tree: below).

Different issues can be considered in cluster validation:

- The most general hypothesis is that of complete absence of classification structure in the data. In principle, such tests should be carried out before cluster analysis is attempted. Several methods have been proposed to assess the positions of the objects distributed in multidimensional space (random position hypothesis) and test for either uniform or unimodal distributions (i.e. greater density of objects near the centre of the distribution); see Gordon (1996a, 1996b). There are also tests that are carried out on graphs linking the objects, and others that involve only the object labels.

- Other methods are available to test (1) for the presence of a hierarchical structure in the data, (2) for partitions (are there distinct clusters in the data? how many clusters?), or (3) for the validity of individual clusters.

For any one of these hypotheses, validation may be carried out at different conceptual levels.

1. *Internal validation using* **Y**. — *Internal validation* methods allow the assessment of the *consistency* of a clustering topology. Internal validation consists in using the original data (i.e. matrix **Y** containing the data originally used for clustering) to assess the clustering results. One approach is to resample the original data set. One repeatedly draws subsets of objects at random, using sampling with or without replacement, to verify that the original clusters of objects are found by the clustering method for the different subsets. Nemec & Brinkhurst (1988) present an ecological application of this method to species abundance data. Another approach is to randomize the original data set, or generate random simulated data with similar distribution parameters, and compute the classification a large number of times to obtain a null distribution for some clustering statistic of interest, which can be tested using the null distribution; one may use one of the statistics discussed in Subsection 8.12.2, or the *U* statistic of Gordon (1994) described at the end of Subsection 10.5.3. The test of cluster fusion in chronological clustering (Subsection 12.6.5) is an example of an internal validation criterion. Using simulations, Milligan (1981) compared 30 internal validation criteria that may be used in this type of study. One *must not*, however, use a standard hypothesis testing procedure such as ANOVA or MANOVA on the variables used to determine the clusters. This approach would be incorrect because the *alternative hypothesis* of the test would be constructed to fit the group structure since it would be computed from the same data that would now be used for testing the null hypothesis. As a consequence, such a test would almost necessarily (subject to type II error) result in significant differences among the groups. To illustrate this point, one can generate multivariate data at random using the uniform distribution and carry out clustering: a MANOVA comparing the clusters to the original data would produce a significant result in most cases even though the data are random and thus have no structure.

2. *External validation comparing* **Y** *to* **X**. — *External validation* methods involve the comparison of two different data tables. The clustering results derived from data matrix **Y**, e.g. species, are compared to a matrix of explanatory variables,

e.g. environmental, which is called **X** in the contexts of regression (Chapter 10) and canonical analysis (Chapter 11). Comparisons can be made at different levels. One may compare a partition of the objects based on **Y** to matrix **X** using linear discriminant analysis (Table 10.1; Section 11.3). Else, the whole hierarchical tree structure may be coded using binary variables (Baum, 1992; Ragan, 1992), in the same way as nested factors in ANOVA; this matrix is then compared to the explanatory matrix **X** using RDA or CCA (Sections 11.1 and 11.2). A third way is to compare the cophenetic matrix (Section 8.3) that represents the hierarchical tree structure to a distance or similarity matrix computed from matrix **X**, using a Mantel test (Subsection 10.5.1; Hubert & Baker, 1977). Contrary to the cophenetic correlations considered in Subsection 8.12.2, testing is legitimate here because matrix **X** is independent of the data matrix **Y** used to construct the classification, but note that the Mantel test has low power compared to the other methods mentioned above.

3. *External validation comparing two or several matrices* **Y**, *same variables*. — Confirmation of the presence of a clustering structure in the data can be obtained by repeating the cluster analysis using different sets of objects (data matrices $\mathbf{Y}_1$, $\mathbf{Y}_2$, etc., all with the same descriptors) and comparing the results. Consider the situation where replicate data are available. If, for example, lakes can be selected at random from different geographic regions, one can conduct independent cluster analyses of the regions using one lake per region, different lakes being used in the separate runs, followed by a comparison of the resulting partitions or dendrograms representing the classifications of regions. Methods are available for comparing independently-obtained dendrograms representing the same objects (Fig. 10.4 and references in Section 10.2). A second approach is to take the classification of regions obtained from the first set of lakes (matrix $\mathbf{Y}_1$) as a model to be validated, using discriminant analysis, by comparing it to a second, independent set of lakes (matrix $\mathbf{Y}_2$) representing the same regions.

A third approach is *replication analysis*, where external validation is carried out for data that are not replicate observations of the same objects. One finds a classification using matrix $\mathbf{Y}_1$, determines group centroids, and assigns the data points in $\mathbf{Y}_2$ to the nearest centroid (McIntyre & Blashfield, 1980). Then, the data in $\mathbf{Y}_2$ are clustered without considering the result from $\mathbf{Y}_1$. The independently obtained classification of $\mathbf{Y}_2$ is compared to the first one using some appropriate measure of consensus (point 4 below).

In studies where data are costly to obtain, this approach is, in most cases, not appealing to researchers who are more interested in using all the available information in a single cluster analysis, instead of dividing the data set into two or several analyses. This approach is only feasible when the objects are numerous.

4. *External validation comparing two or several matrices* **Y**, *same objects*. — Several groups of descriptors may be available about the same objects, and one may wish to conduct separate cluster analyses on them. An example would be sites where data are available about several groups of arthropods (e.g. matrices $\mathbf{Y}_1$ = acarians, $\mathbf{Y}_2$ = insects,

and $\mathbf{Y}_3$ = spiders), besides physical or other variables of the environment which would form a matrix **X** of explanatory variables. Classifications may be obtained independently for each matrix **Y**. Measures of resemblance between trees, called *consensus indices* (Rohlf, 1982b), may be calculated. The cophenetic correlation coefficient of the previous subsection can be used as a consensus index; other indices are available, that only take the classification topologies into account. Alternatively, one may compute a compromise tree, called a *consensus tree*, which represents the areas of agreement among trees. Several criteria have been proposed for constructing consensus trees: majority rule, strict consensus, average consensus, etc. (Leclerc & Cucumel, 1987). Tests of significance are available for comparing independently-obtained dendrograms that describe relationships among the same objects (Fig. 10.4 and references in Section 10.2).

Cluster validation has progressed in important ways during the last decade, with new methods and packages being made available. Summarizing these developments, Rendón *et al*. (2011) described and compared a large number of internal and external cluster validation indexes. Because cluster validity indices (CVI) are numerous and no single CVI always outperforms the others, Kryszczuk & Hurley (2010) proposed composite validation indices combining different approaches. Brock *et al*. (2008) wrote the R package CLVALID for cluster validation; see Section 8.15.

8.14 Cluster representation and choice of a method

This section summarizes the most usual graphical representations of clustering results. More complete reviews of the subject are found in Sneath & Sokal (1973) and Chambers & Kleiner (1982).

Hierarchical clustering results are represented, in most cases, as dendrograms, e.g. Fig. 8.2b. They can also be represented as plots of connected subgraphs, e.g. Fig. 8.2a; the construction of these informative graphs, which would be difficult to draw by computer, was explained in Section 8.2. The branches of dendrograms may point upwards, downwards or sideways; the horizontal representation is an easier way of plotting a dendrogram that contains a large number of objects and fitting it into a page. Dendrograms are graduated in distances or similarities; the branching pattern indicates the distance or similarity of bifurcating branches. Usually, the names of the objects (or descriptors when descriptors are clustered), or their code numbers, are written at the tips of the branches. The ordinate (on horizontal dendrograms) has no specified ordering, except in TWINSPAN. Bifurcating branches are not fixed; they may be swivelled as required by the presentation of results without altering the nature of the ultrametric information in the dendrogram.

Dendrogram

Dendrograms clearly illustrate the clusters formed at each partition level, but in linkage clustering they do not allow the identification of the exact links among objects that generate cluster fusions. With some clustering methods, this information is not

directly available and must be found *a posteriori* when needed; how to compute the chain of primary connections was described at the end of Subsection 8.5.4 for the UPGMA clustering case. In synoptic clustering, which only aims at recognizing major clusters of objects, connecting links are not required.

Connected subgraphs

Series of connected subgraphs, as drawn in Fig. 8.2a, may be used to represent all the information of the distance or similarity matrix. Complex information may be represented by different types of lines; colours may also be used. When they become numerous, objects can be placed at the rim of a circle; distance links are drawn as lines between them. In each subgraph, the relative positions of the objects are of little importance. They are merely arranged in such a way as to simplify the paths of the links connecting them. The objects may have been positioned beforehand in a two-dimensional ordination space, which may be obtained by principal coordinate analysis or nonmetric scaling of the association matrix (Sections 9.3 and 9.4). Figures of connected subgraphs, informative as they may be, are quite time consuming to draw and difficult to publish.

Skyline plot
Tree
Icicle plot

Some programs still use "skyline plots" (Ward, 1963; Wirth *et al*., 1966), which may also be called "trees" or "icicle plots". These plots may be imagined as negative pictures of dendrograms. They contain the same information as dendrograms, but they are rather odd to read and interpret. In Fig. 8.25a for instance (UPGMA clustering of the pond data, see Fig. 8.5), the object names are sitting on the lines *between* columns of X's; the ordinate of the plot is a scale of distances or similarities. Since the value $D = 0$ is at the bottom of the graph, this is where the hierarchical agglomeration begins. The first clustering step is materialized by the first horizontal row of X's, at distance $D = 0.4$, which joins objects 212 and 214. It is drawn like the lintel of a door. The surface above the lintel of X's is filled with X's; these are without meaning. The next clustering step is at distance $D = 0.5$; it consists in a row of X's joining ponds 431 and 432. The third clustering step is more interesting. Note how a new lintel of X's, at $D = 0.75$, goes from pond 233, and right across the column of X's already joining ponds 431 and 432. The final clustering step is at $D = 0.942$. This new lintel crosses the last remaining gap, uniting the two columns of X's corresponding to the two already-formed clusters.

A skyline plot can be directly transformed into a dendrogram (Fig. 8.25b, c). Working from the bottom to the top, proceed as follows:

- Identify lintels and draw lines across the column of X's. The lintel lines should not extend beyond the row of X's.

- When all the horizontal lintel lines have been drawn, draw vertical lines from individual objects up to the first lintel encountered, and from the *centre* of a lower lintel up to the one above. Erase the overhanging part of the upper lintel. Repeat the operation for the next lintel up.

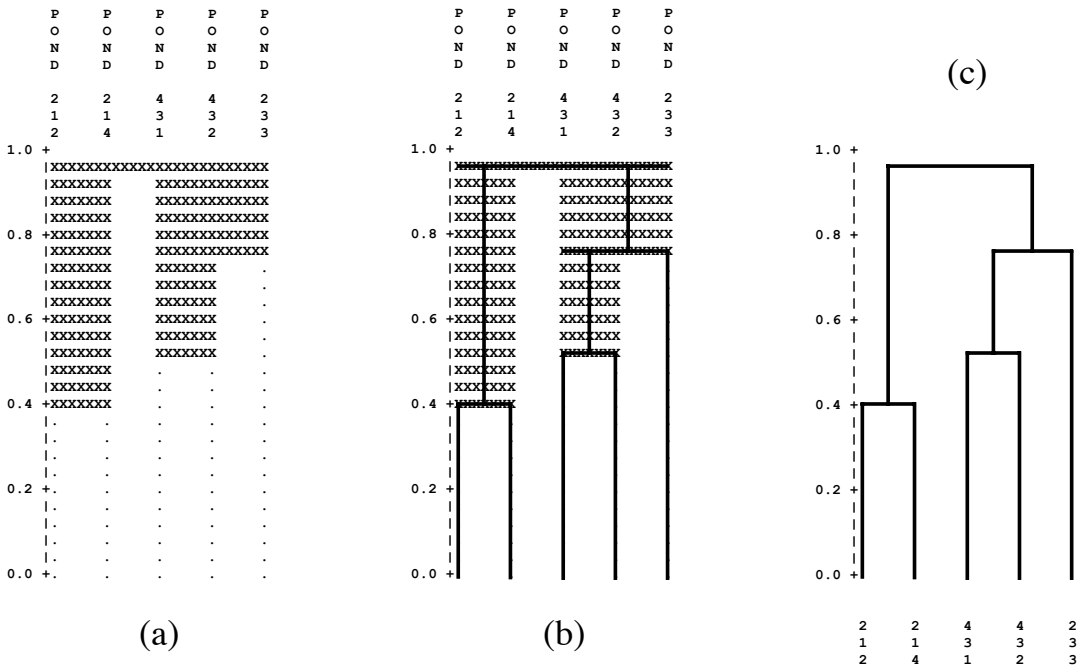


Figure 8.25 A skyline plot (a) can be transformed into a dendrogram (c) by going through the drawing steps described in (b). Vertical scale: distances. The skyline plot was computed using SAS.

• Erase the X's. The result is a standard dendrogram (Fig. 8.25c). It is identical to the dendrogram representing the same clustering results in Fig. 8.5, but it is drawn here vertically instead of horizontally, and pond 233 is swivelled to the right instead of being between ponds 214 and 431.

Heat maps (see Section 8.15) can be used to represent **D** matrices graphically, before or after clustering, e.g. Fig. 8.21. Dendrograms can be represented along the axes of heat maps as in Fig. 8.21c.

Section 10.1 shows how to superimpose clustering results onto an ordination of the same objects. This often helps evidence the structure when ecological objects form a continuum. When it comes to representing the results of a partition, the objects are represented in an ordination space and symbols can be used to represent the groups; else, envelopes can be drawn around points corresponding to the groups.

Table 8.11 summarizes, in a comparative way, the various clustering methods discussed in the present chapter. Some advantages and disadvantages of each method are pointed out.

Table 8.11 Synoptic summary of the clustering methods presented in Chapter 8.

Method	Pros & cons	Use in ecology
Hierarchical agglomeration: linkage clustering	Pairwise relationships among the objects are known.	
Single linkage	Computation simple; contraction of space (chaining); combinatorial method.	Good complement to ordination.
Complete linkage (see also: species associations)	Dense nuclei of objects; space expansion; many objects cluster at high distance; arbitrary rules to resolve conflicts; combinatorial method.	To increase the contrast among clusters.
Intermediate linkage	Preservation of reference space A; non-combinatorial: not included in Lance & Williams' general model.	Preferable to the above two methods if only one clustering method is to be used.
Hierarchical agglomeration: average clustering	Preservation of reference space A; pairwise relationships between objects are lost; combinatorial method.	
Unweighted arithmetic average (UPGMA)	Fusion of clusters when the distance reaches the mean inter-cluster distance value.	For a collection of objects obtained by simple random or systematic sampling.
Weighted arithmetic average (WPGMA)	As UPGMA, with adjustment for group sizes.	Preferable to the previous method in all other sampling situations.
Unweighted centroid (UPGMC)	Fusion of clusters with closest centroids; may produce reversals.	For simple random or systematic samples of objects.
Weighted centroid (WPGMC)	As UPGMC, with adjustment for group sizes; may produce reversals.	Preferable to the previous method in all other sampling situations.
Ward's method	Minimizes the within-group sum of squares.	When looking for hyperspherical clusters in space A.
Hierarchical agglomeration: flexible clustering	Allows contraction, conservation, or dilation of space A; pairwise relationships between objects are lost; combinatorial method.	All combinatorial methods, including this one, are implemented using the simple Lance & Williams algorithm.
Hierarchical agglomeration: information analysis	Minimal chaining; only for Q-mode clustering based upon presence-absence of species.	Ecological use is unclear: distances reflect double absences as well as double presences.

Table 8.13 Continued.

Method	Pros & cons	Use in ecology
Hierarchical division	Danger of incorrect separation of members of minor clusters near the beginning of clustering.	
Monothetic	Division of the objects following the states of the "best" descriptor at each step of the procedure.	Useful to split data into large clusters, inside which clustering depends on different phenomena.
Polythetic	For small number of objects only.	Computation impossible for sizable data sets.
Division in ordination space	Binary division along each axis of ordination space; no search is done for high concentrations of objects in space A.	Efficient algorithms for large data sets, when a coarse division of the objects is sought.
TWINSPAN	Dichotomized ordination analysis; ecological justification of several steps unclear.	Produces an ordered two-way table classifying sites and species.
***K*-means partitioning**	Minimizes within-group sum of squares; different rules may suggest different optimal numbers of clusters.	Produces a partition of the objects into K groups, K being determined by the user.
Species associations	Non-hierarchical methods; clustering at a pre-selected level of similarity or probability.	Concept of association based on co-occurring or correlated species.
Non-hierarchical complete linkage	Species associated by complete linkage (no overlap); satellite species joined by single linkage (possible overlap).	Straightforward concept; no easily available software.
Concordance analysis	Find groups of species that form statistically significant associations.	Clear, easy to apply method; R functions available.
Multivariate regression tree	A multivariate response table is constrained by a table of explan. variables, producing a tree.	Two-matrix method related to canonical analysis.
Seriation	One-dimensional ordination along the main diagonal of a distance matrix.	Useful to analyse non-symmetric association matrices.
Indicator species		
TWINSPAN	Only for classifications of sites obtained by splitting CA axes; justification of some steps unclear.	Gives indicator values for the pseudospecies.
Indicator value index	For any hierarchical or non-hierarchical classification of sites; *IndVal* for a species is not affected by the other species in the study.	Gives indicator values for the species under study; the *IndVal* index is tested by permutation.

8.15 Software

Several, but not all statistical packages offer clustering capabilities: SAS, SPSS, SYSTAT, JMP, STATISTICA, and NTSYSPC offer clustering among their methods for data analysis. All packages with clustering procedures offer at least a Lance & Williams algorithm capable of carrying out the clustering methods listed in Table 8.9. Many also have a *K*-means partitioning algorithm. Few offer proportional-link linkage or additional forms of clustering. Some methods are available in specialized packages only: clustering with constraints of temporal (Section 12.6) or spatial contiguity (Section 13.3); fuzzy clustering (algorithms described e.g. in Bezdek, 1987); or clustering by neural network algorithms (algorithms described e.g. in Fausett, 1994).

Functions in the R language are available to carry out all analyses described in this chapter.

1. Several R functions are devoted to clustering. Hierarchical clustering is computed using ***hclust()*** in STATS and ***agnes()*** in CLUSTER using the Lance & Williams general agglomerative algorithm. Functions for constrained hierarchical clustering are listed in Sections 12.8 and 13.6. A cophenetic distance matrix corresponding to a hierarchical clustering is computed by function ***cophenetic()*** of STATS.

Minimum spanning trees can be computed by several functions including ***mstree()*** in ADE4, ***mst()*** in APE, ***mstree()*** in SPDEP and ***spantree()*** in VEGAN. Function ***cophenetic()*** in STATS computes the cophenetic matrix corresponding to a hierarchical clustering. Function ***clustIndex()*** of CCLUST computes stopping indices for clustering.

2. *K*-means partitioning is available in functions ***kmeans()*** of STATS, ***cclust()*** of CCLUST, ***kkmeans()*** of KERNLAB, ***KMeans()*** of RCMDR and ***cascadeKM()*** of VEGAN; the latter function automatically repeats *K*-means partitioning using a range of values of *K*.

Heat map

3. Seriation is obtained by function ***seriate()*** of package SERIATION, which offers several calculation methods. *Heat maps* in colour can be obtained using function ***heatmap()*** of STATS, or by function ***hmap()*** of SERIATION, which calls ***heatmap()*** to produce the plot. Heat maps are also produced by function ***coldiss()*** available on the Web page of the Borcard *et al*. (2011) book, http://numericalecology.com/NEwR. With ***coldiss()***, a **D** matrix is represented by an unordered and a reordered colour heat maps, the new ordering being the result of single linkage chaining. Function ***seriation()***[*] carries out the Beum-Brundage seriation procedure for non-symmetric or symmetric matrices.

4. Multivariate regression tree analysis is available in ***mvpart()*** of package MVPART. Package MVPARTWRAP contains additional functions for multivariate regression tree

[*] Available on the Web page http://numericalecology.com/rcode.

analysis, including ***CascadeMRT()*** that carries out two MRT analyses in sequence, using explanatory matrices in the order specified by the researcher.

5. Other clustering methods have been described in the statistical literature. For instance, *K*-means partitioning is a member of a larger class of methods called *K*-centroids, where the Euclidean distance is replaced by other distances; for example, using the Manhattan distance instead of the Euclidean produces *K*-medians clustering. Package FLEXCLUST offers different types of clustering, including function ***kcca()*** that computes various types of *K*-centroid cluster analysis (*K*-means, *K*-medians and others). For distances other than the Euclidean, the *K*-centroid approach is also called *partitioning around medoids* (Kaufman & Rousseeuw, 1990); it is implemented in function ***pam()*** of the CLUSTER package. An example of partitioning around medoids is presented in Subsection 4.8.2 of the Borcard *et al*. (2011) book.

6. Fuzzy partitioning is available in functions ***fanny()*** of package CLUSTER and ***cmeans()*** of package E1071. An example of analysis in Q mode is presented in Subsection 4.12.1 of Borcard *et al*. (2011). Function ***vegclust()*** of package VEGCLUST offers three forms of fuzzy partitioning (fuzzy *c*-means, probabilistic *c*-means, and noise clustering) in addition to hard *K*-means.

7. Concordance analysis to search for species associations is available in functions ***kendall.global()*** and ***kendall.post()*** of the VEGAN package.

8. Indicator value indices (*INDVAL*, EQ. 8.21) can be computed by functions ***strassoc()*** and ***multipatt()*** of INDICSPECIES and function ***indval()*** of LABDSV. The functions in INDICSPECIES offer a choice of several different indicator statistics described in De Cáceres & Legendre (2009).

9. Function ***clValid()*** of the CLVALID package computes validation measures for clustering results, including internal validation and stability measures, plus biological measures for genetic data. The package is described in Brock *et al*. (2008). Function ***randIndex()*** computes the Rand and modified Rand indices quantifying the agreement of two partitions.

Chapter

9 *Ordination in reduced space*

9.0 Projecting data sets in a few dimensions

Ordination

Ordination (from the Latin *ordinatio*, the action of setting in order) is the arrangement of units in some order. Gower (1984) points out that the term *ordination*, widely used in multivariate statistics, actually comes from ecology where it refers to the representation of objects (sites, stations, relevés, etc.) as points along one or several reference axes. In 1954, vegetation ecologist David Goodall was the first to apply factor analysis in community ecology. Goodall proposed the term "ordination" to designate this type of analysis, a term now widely used in community ecology textbooks and publications. Ordination consists in plotting object-points along an axis representing an ordered relationship, or forming a scatter diagram with two or more axes. The ordered relationships are usually quantitative, but it would suffice for them to be of the type "larger than", "equal to" or "smaller than" (semiquantitative relations) to serve as the basis for ordinations, as it is the case in nMDS (Section 9.4).

In ecology, several descriptors are usually observed for each object under study. In most instances, ecologists are interested in characterizing the main trends of variation of the objects with respect to all descriptors, not only a few of them. Looking at scatter plots of the objects with respect to all possible pairs of descriptors is a tedious approach, which generally does not shed much light on the problem at hand. In contrast, the multivariate approach consists in representing the scatter of objects in a multidimensional diagram, with as many axes as there are descriptors in the study. It is not possible to draw such a diagram on paper with more than two or eventually three dimensions, however, even though it is a perfectly valid mathematical construct. For the purpose of analysis, ecologists therefore project the multidimensional scatter diagram onto bivariate graphs whose axes are known to be of particular interest. The axes of these graphs are chosen to represent a large fraction of the variability of the multidimensional data matrix, in a space with reduced (i.e. lower) dimensionality relative to the original data set. Methods for *ordination in reduced space* also allow

Table 9.1 Domains of application of the ordination methods presented in this chapter.

Method	Distance preserved	Variables
Principal component analysis (PCA)	Euclidean distance	Quantitative data, linear relationships (beware of double-zeros)
Correspondence analysis (CA)	χ^2 distance	Non-negative, dimensionally homogeneous quantitative or binary data; species frequencies or presence/absence data
Principal coordinate analysis (PCoA), metric (multidimensional) scaling, classical scaling	Any distance measure	Quantitative, semiquantitative, qualitative, or mixed
Nonmetric multidimensional scaling (nMDS)	Any distance measure	Quantitative, semiquantitative, qualitative, or mixed

one to derive quantitative information on the quality of the projections, and study the relationships among descriptors as well as among objects.

Ordination in reduced space is often referred to as *factor* (or *inertia*) *analysis* since it is based on the extraction of the eigenvectors or *factors* of the association matrix. Factor analysis *sensu stricto* is mainly used in the social sciences; it aims at representing the covariance structure of the descriptors in terms of a hypothetical causal model. It is not discussed further in this book.

The domains of application of the techniques discussed in the present chapter are summarized in Table 9.1. Section 9.1 is devoted to principal component analysis (PCA), a powerful technique for ordination in reduced space which is, however, limited to quantitative descriptors. Because it preserves Euclidean distances, PCA results are sensitive to the presence of double-zeros. Section 9.2 discusses correspondence analysis (CA), an ordination method useful to analyse species presence/absence or abundance data. Sections 9.3 and 9.4 are devoted to principal coordinate analysis (metric scaling, PCoA) and nonmetric multidimensional scaling (nMDS), respectively. Both methods project, in reduced space, distance matrices among objects computed prior to ordination, based on user-chosen distance measures (Chapter 7); in some of these measures, the descriptors may be of any mathematical type. PCA, CA and PCoA are eigenvector-based methods, not nMDS. The presentation of various forms of canonical analysis, which are also eigenvector-based, is deferred to Chapter 11.

It often happens that the structure of the objects under study is not continuous. In such a case, an ordination in reduced space, or a scatter diagram produced using two important variables, may be sufficient to evidence the group structure of the objects. Ordination methods may thus sometimes be used to delineate clusters of objects (Fig. 8.1, Subsection 8.7.3). Ordinations can also be used as complements to cluster analyses. The reason is that clustering investigates pairwise distances among objects, looking for fine relationships, whereas ordination in reduced space considers the variability of the whole association matrix and thus brings out general gradients. Different methods for superimposing the results of clustering onto ordinations of the same objects are described in Section 10.1.

Reduced space

Ecologists generally use ordination methods to study the relative positions of objects in reduced space. An important aspect to consider is the representativeness of the representation in reduced space, which usually has $d = 2$ or 3 dimensions. To what extent does the reduced space preserve the distance relationships among objects? To answer this, one can compute the distances between all pairs of objects, both in the multidimensional space of the original p descriptors and in the reduced d-dimensional space. The resulting values are plotted in a scatter diagram such as Fig. 9.1. When the projection in reduced space accounts for a high fraction of the variance, the distances between projections of the objects in reduced space are quite similar to the original distances in multidimensional space (case a). When the projection is less efficient, the distances between objects in reduced space are much smaller than in the original space. Two situations may then occur. When the objects are at *proportionally* similar distances in the two spaces (case b), the projection is still useful even if it accounts for a small fraction of the variance. When, however, the relative object positions are not the same in the two spaces (case c), the projection is useless. Ecologists often disregard the interpretation of ordinations when the reduced space does not account for a high fraction of the variance. This is not entirely justified, since a projection in reduced space may be informative even if that space only accounts for a small fraction of the variance (case b).

Shepard diagram

The scatter diagram of Fig. 9.1, which is often referred to as a Shepard diagram (Shepard, 1962; diagrams in Shepard's paper had their axes transposed relative to Fig. 9.1), may be used to estimate the representativeness of ordinations obtained using any reduced-space ordination method. In principal component analysis (Section 9.1), the distances among objects, in both the multidimensional space of original descriptors and the reduced space, are calculated using Euclidean distances (D_1, eq. 7.32). The **F** matrix of principal components (eq. 9.4 below) gives the coordinates of the objects in the reduced space. In principal coordinate analysis (Section 9.3) and nonmetric multidimensional scaling (Section 9.4), Euclidean distances among the objects in reduced space are compared to distances D_{hi} found in matrix **D** used as the basis for computing the ordination. In correspondence analysis (Section 9.2), it is the χ^2 distance (D_{16}, eq. 7.55) among objects that is used on the abscissa of the Shepard diagram. Shepard-like diagrams can also be constructed for cluster analysis (Fig. 8.24).

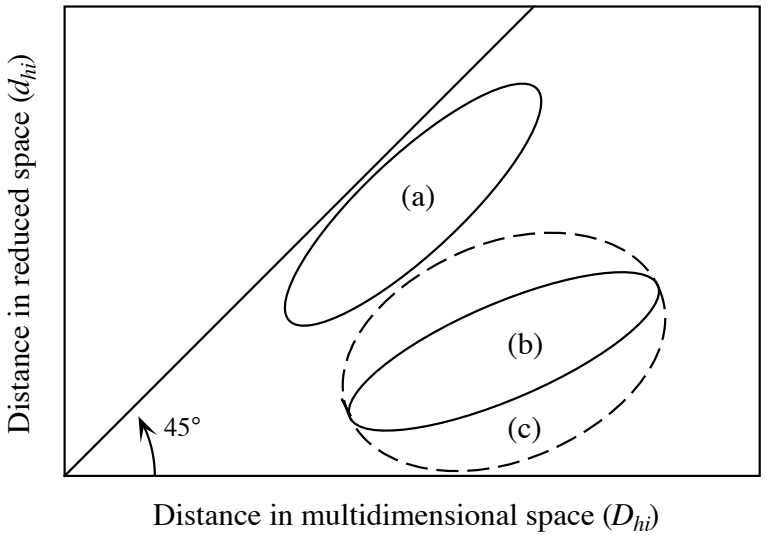


Figure 9.1 Shepard diagram. Three situations encountered when comparing distances among objects, in the *p*-dimensional space of the *p* original descriptors (abscissa) *versus* the *d*-dimensional reduced space (ordinate). The figure only shows the contours of the scatters of points. (a) The projection in reduced space accounts for a high fraction of the variance; the relative positions of objects in the *d*-dimensional reduced space are similar to those in the *p*-dimensional space. (b) The projection accounts for a small fraction of the variance, but the relative positions of the objects are similar in the two spaces. (c) Same as (b), but the relative positions of the objects differ in the two spaces. Adapted from Rohlf (1972). Compare to Fig. 8.24.

The following sections discuss the ordination methods most useful to ecologists. The sections are written to be easily understood by ecologists, so that they may not entirely fulfil the expectations of statisticians. Many programs are available to carry out ordination analysis; several of them are described by Michael Palmer*. R functions are listed in Section 9.5. For detailed discussions on the theory or computing methods, one can refer to ter Braak (1987c) and Legendre & Birks (2012). Important references about correspondence analysis are Benzécri and coll. (1973), Hill (1974), Greenacre (1983), and ter Braak (1987c). Gower (1984, 1987) reviewed the ordination methods described in this chapter, plus a number of other techniques developed by psychometricians. Some of these are progressively finding their way into numerical ecology. They include methods of metric scaling other than principal coordinate analysis, multidimensional unfolding, orthogonal Procrustes analysis (Subsection 11.5.2) and its generalized form, scaling methods for several distance matrices, and a method for ordination of non-symmetric matrices.

* Web page: http://ordination.okstate.edu/.

Ordination vocabulary* Box 9.1

Major axis. Axis in the direction of maximum variance of a scatter of points.

First principal axis (of the concentration ellipsoid in a multinormal distribution; Fig. 4.9). Line passing through the greatest dimension of the ellipsoid; major axis of the ellipsoid.

Principal components. New variates (*variates* = random variables) specified by the axes of a rigid rotation of the original system of coordinates, and corresponding to the successive directions of maximum variance of the scatter of points. The principal components give the positions of the objects in the new system of coordinates.

Principal-component axes (also called *principal axes* or *component axes*). System of axes resulting from the rotation described above.

*Adapted from Morrison (1990, pp. 87 and 323-325).

9.1 Principal component analysis (PCA)

In this book, principal component analysis* is defined as the eigenanalysis of the dispersion matrix $\mathbf{S} = (n-1)^{-1}\mathbf{Y}_c'\mathbf{Y}_c$, where $\mathbf{Y}_c$ is matrix $\mathbf{Y}$ column-centred. In other books, it may be defined as the eigenanalysis of $\mathbf{Y}_c'\mathbf{Y}_c$ without division by $(n-1)$. How to compute the principal axes (Box 9.1) of $\mathbf{S}$ was explained in Section 4.4. In a nutshell, in a multinormal distribution, the first principal axis is the line that goes through the greatest dimension of the concentration ellipsoid describing the distribution. The following principal axes (orthogonal to one another and successively shorter) go through the following greatest dimensions of the p-dimensional ellipsoid. A maximum of p principal axes can be derived from a data matrix containing p variables (Fig. 4.9). The principal axes of a dispersion matrix $\mathbf{S}$ are found by solving

$$(\mathbf{S} - \lambda_k\mathbf{I})\,\mathbf{u}_k = \mathbf{0} \qquad \textbf{(9.1)}$$

(eq. 4.23) whose characteristic equation

$$\left|\mathbf{S} - \lambda_k\mathbf{I}\right| = 0 \qquad \textbf{(9.2)}$$

* Because the pronunciation of "principal" and "principle" is similar in English, the erroneous name "*Principle* component analysis" is sometimes found in the literature, even recent.

Eigenvalue
Eigenvector

is used to compute the *eigenvalues* λ_k. The eigenvectors $\mathbf{u}_k$ associated with the eigenvalues λ_k are found by putting the different values λ_k in turn into eq. 9.1. These eigenvectors are the *principal axes* of dispersion matrix **S** (Section 4.4). The eigenvectors are normalized (i.e. scaled to unit length, Section 2.4) before computing the *principal components,* which give the coordinates of the objects on the successive principal axes. Principal component analysis (PCA) was originally described by Pearson (1901) although it is more often attributed to Hotelling (1933) who proposed it independently. The method and several of its implications for data analysis are presented in the seminal paper of Rao (1964). PCA possesses the following properties, which make it a powerful tool for the analysis of ecological data:

Principal components

1) Since any dispersion matrix **S** is symmetric, its principal axes $\mathbf{u}_k$ are *orthogonal* to one another. In other words, they correspond to *linearly independent directions* in the concentration ellipsoid of the distribution of objects (Section 2.9).

2) The eigenvalues λ_k of a dispersion matrix **S** are all positive or null because **S** is positive semidefinite (Section 4.1, Table 2.2). PCA does not produce negative eigenvalues. The eigenvalues represent the amounts of *variance* of the data along the successive principal axes (Section 4.4).

3) Because of the first two properties, principal component analysis can often *summarize, in a few dimensions, most of the variability* of a dispersion matrix of a large number of descriptors. It also provides a measure of the amount of variance explained by these few independent principal axes.

The present section shows how to compute the relationships among objects and among descriptors, as well as the relationships between the principal axes and the original descriptors. A simple numerical example is developed, involving five objects and two quantitative descriptors:

$$\mathbf{Y} = \begin{bmatrix} 2 & 1 \\ 3 & 4 \\ 5 & 0 \\ 7 & 6 \\ 9 & 2 \end{bmatrix} \quad \text{After centring on the column means, } \mathbf{Y}_c = [y - \bar{y}] = \begin{bmatrix} -3.2 & -1.6 \\ -2.2 & 1.4 \\ -0.2 & -2.6 \\ 1.8 & 3.4 \\ 3.8 & -0.6 \end{bmatrix}$$

where $\mathbf{Y}_c$ is the matrix of column-centred data. In practice, principal component analysis is never used for two descriptors only; in such a case, the objects can simply be represented in a two-dimensional scatter diagram (Fig. 9.2a). A two-dimensional example is used here for simplicity, in order to show that the main result of principal component analysis is to rotate the axes, using the centroid of the objects as pivot.

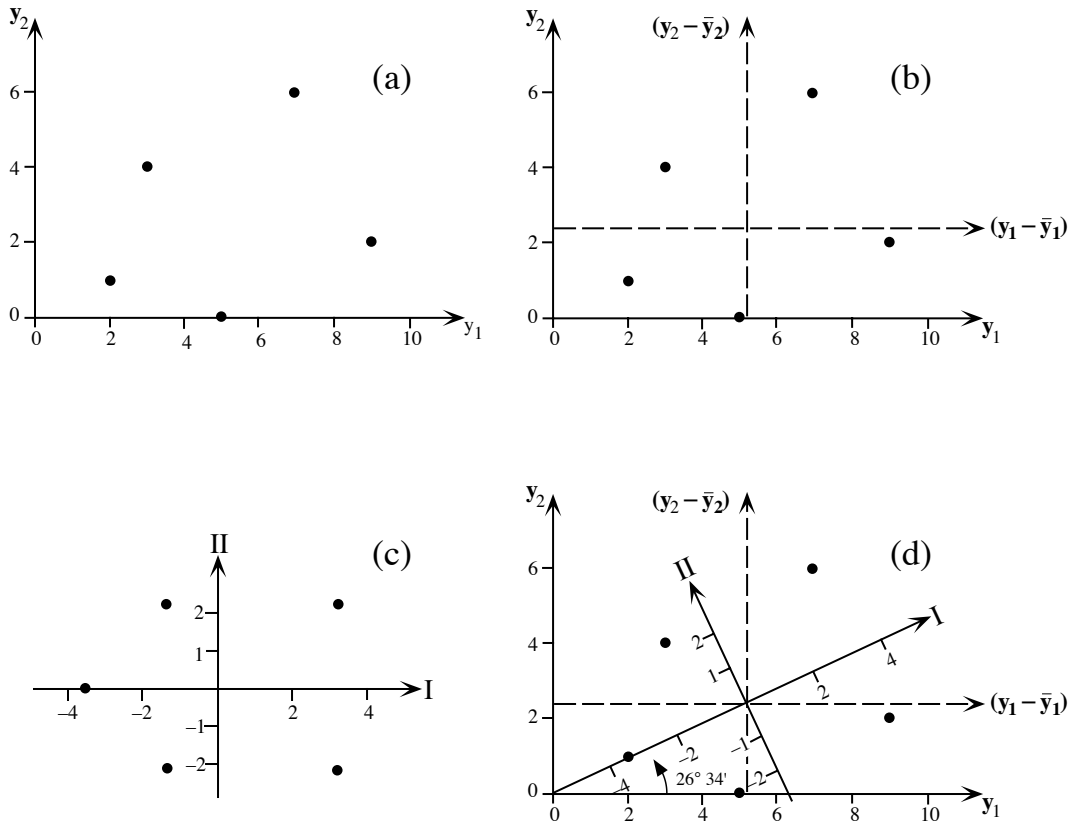


Figure 9.2 Numerical example of principal component analysis. (a) Five objects are plotted with respect to descriptors $\mathbf{y}_1$ and $\mathbf{y}_2$. (b) After centring the data, the objects are now plotted with respect to $(\mathbf{y}_1 - \bar{\mathbf{y}}_1)$ and $(\mathbf{y}_2 - \bar{\mathbf{y}}_2)$, represented by dashed axes. (c) The objects are plotted with reference to principal axes I and II, which are centred with respect to the scatter of points. (d) The two systems of axes (b and c) can be superimposed after a rotation of 26°34'.

1 — Computing the eigenvectors of a dispersion matrix

The dispersion matrix (eq. 4.6) of the descriptors in the above example is:

$$\mathbf{S} = \frac{1}{n-1}\,\mathbf{Y}_c'\,\mathbf{Y}_c = \begin{bmatrix} 8.2 & 1.6 \\ 1.6 & 5.8 \end{bmatrix}$$

The corresponding characteristic equation (eq. 2.23) is:

$$|\mathbf{S} - \lambda_k \mathbf{I}| = \left| \begin{bmatrix} 8.2 & 1.6 \\ 1.6 & 5.8 \end{bmatrix} - \begin{bmatrix} \lambda_k & 0 \\ 0 & \lambda_k \end{bmatrix} \right| = 0$$

It has two eigenvalues, $\lambda_1 = 9$ and $\lambda_2 = 5$. The total variance (sum of diagonal values) in the matrix of eigenvalues is the same as in **S**, but it is partitioned in a different way: the sum of the variances in **S**, (8.2 + 5.8 = 14), is equal to the sum of the eigenvalues, (9 + 5 = 14). $\lambda_1 = 9$ accounts for 64.3% of the variance and λ_2 makes up for the difference (35.7%). There are as many eigenvalues as there are descriptors. The successive eigenvalues account for progressively smaller fractions of the variance. Introducing, in turn, the λ_k's in matrix equation 9.1:

$$(\mathbf{S} - \lambda_k \mathbf{I})\, \mathbf{u}_k = \mathbf{0}$$

provides the eigenvectors associated with the eigenvalues. Once these vectors have been normalized (i.e. scaled to unit length, $\mathbf{u}'\mathbf{u} = 1$) they become the *columns* of matrix **U**:

$$\mathbf{U} = \begin{bmatrix} 0.8944 & -0.4472 \\ 0.4472 & 0.8944 \end{bmatrix}$$

If a different sign had been arbitrarily assigned to one of the terms of matrix **U** during calculation of the eigenvectors, a mirror image would have been produced for Figs. 9.2c. That image would have been as good at representing the data as Fig. 9.2c.

Orthogonality

It is easy to check the orthogonality of the two eigenvectors: their cross-product $\mathbf{u}'_1\mathbf{u}_2 = (0.8944 \times (-0.4472)) + (0.4472 \times 0.8944) = 0$. Moreover, Section 4.4 has shown that the elements of **U** are direction cosines of the angles between the original descriptors and the principal axes. Using this property, one finds that the system of principal axes specifies a rotation of (arc cos 0.8944) = 26°34' of the system of reference defined by the original descriptors. Hence, Figure 9.2 shows that principal component analysis has performed a rotation of the system of axes (descriptors) without changing the positions of the objects with respect to one another.

2 — Computing and representing the principal components

Loading

Principal component

The elements of the eigenvectors are also weights, or *loadings* of the original descriptors, in the linear combination of descriptors from which the principal components are computed. The *principal components* give the positions of the objects with respect to the new system of principal axes. Thus the position of an object $\mathbf{x}_i$ on the first principal axis is given by the following function, or linear combination:

$$f_{i1} = (y_{i1} - \bar{y}_1)\, u_{11} + \ldots + (y_{ip} - \bar{y}_p)\, u_{p1} = [y - \bar{y}]_i \mathbf{u}_1 \qquad \textbf{(9.3)}$$

The values $(y_{ij} - \bar{y}_j)$ are the coordinates of object $\mathbf{x}_i$ on the various centred descriptors j and the values u_{j1} are the loadings of the descriptors on the first eigenvector. The positions of all objects with respect to the system of principal axes is given by matrix **F** of the transformed variables. It is also called the *matrix of principal components*:

Matrix of principal components

$$\mathbf{F} = \mathbf{Y}_c \mathbf{U} \qquad \textbf{(9.4)}$$

where **U** is the matrix of eigenvectors and $\mathbf{Y}_c$ is the matrix of centred observations. The system of principal axes is centred with respect to the scatter of point-objects. This would not be the case if **U** had been multiplied by **Y** instead of the centred matrix $\mathbf{Y}_c$, as in some special forms of principal component analysis (*non-centred PCA*). For the numerical example, the principal components are computed as follows:

$$\mathbf{F} = \begin{bmatrix} -3.2 & -1.6 \\ -2.2 & 1.4 \\ -0.2 & -2.6 \\ 1.8 & 3.4 \\ 3.8 & -0.6 \end{bmatrix} \begin{bmatrix} 0.8944 & -0.4472 \\ 0.4472 & 0.8944 \end{bmatrix} = \begin{bmatrix} -3.578 & 0 \\ -1.342 & 2.236 \\ -1.342 & -2.236 \\ 3.130 & 2.236 \\ 3.130 & -2.236 \end{bmatrix}$$

The variance of the two columns of **F** are $\lambda_1 = 9$ and $\lambda_2 = 5$ respectively. Since the two columns of the matrix of component scores are the coordinates of the five objects with respect to the principal axes, they can be used to plot the objects with respect to principal axes I and II (Fig. 9.2c). It is easy to verify (Fig. 9.2d) that, in this two-descriptor example, the objects are positioned by the principal components in the same way as in the original system of descriptor-axes. Principal component analysis has simply rotated the axes by 26° 34' in such a way that the new axes correspond to the two main components of variability. When there are more than two descriptors, as it is usually the case in ecology, principal component analysis still only performs a rotation of the system of descriptor-axes, but now in multidimensional space. In that case, principal components I and II define the plane allowing the representation of the largest amount of variance. The objects are projected on that plane in such a way as to preserve, as much as possible, the relative Euclidean distances they have in the multidimensional space of the original descriptors.

Euclidean distance

The relative positions of the objects in the rotated p-dimensional space of principal components are the same as in the p-dimensional space of the original descriptors (Fig. 9.2d). This means that *the Euclidean distances among objects* (D_1, eq. 7.32) *have been preserved through the rotation of axes*. This important property of principal component analysis is noted in Table 9.1. The quality of the representation in a reduced Euclidean space with m dimensions only ($m \leq p$) may be assessed using the following ratio:

R^2-like ratio

$$\left(\sum_{k=1}^{m} \lambda_k \right) \Big/ \left(\sum_{k=1}^{p} \lambda_k \right) \qquad \textbf{(9.5)}$$

This ratio is the equivalent of a coefficient of determination (R^2, eq. 10.20) in regression analysis. The denominator of eq. 9.5 is actually equal to the trace of matrix **S** (sum of the diagonal elements). Thus, with the current numerical example, a representation of the objects, along the first principal component only, would account for a proportion 9/(9+5) = 0.643 of the total variance in the data matrix. This value is identical to that given in Subsection 9.1.1 for the fraction of the variance of **Y** that is accounted for by λ_1.

When the observations have been made along a temporal or spatial axis, or on a geographic surface (i.e. a map giving the coordinates of the sampling sites), one may plot the principal component values along the sampling axis, or on the geographic map. Figure 9.15 shows an example of such a map for the first ordination axis of a detrended correspondence analysis. The same approach can be used with the results of a principal component analysis, or any other ordination method.

3 — Contributions of the descriptors

Principal component analysis provides the information needed to understand the role of the original descriptors in the formation of the principal components. It may also be used to show the relationships among the original descriptors. The role of the descriptors in principal component analysis is now examined under various aspects.

1. *The matrix of eigenvectors **U**.* — In Subsection 9.1.1, the relationships among the *normalized eigenvectors*, which are the columns of matrix **U**, were studied using an expression of the form **U**'**U**. For the numerical example:

$$\mathbf{U'U} = \begin{bmatrix} 0.8944 & 0.4472 \\ -0.4472 & 0.8944 \end{bmatrix} \begin{bmatrix} 0.8944 & -0.4472 \\ 0.4472 & 0.8944 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \mathbf{I}$$

The diagonal terms of **U**'**U** result from the scalar product of the eigenvectors with themselves. These values are the (length)^2 of the eigenvectors, here equal to unity because the eigenvectors were scaled to 1. The nondiagonal terms, resulting from the multiplication of two different eigenvectors, are equal to zero because the eigenvectors are orthogonal. This result would be the same for any matrix **U** of normalized eigenvectors computed from a symmetric matrix. Matrix **U** is a square *orthonormal matrix* (Section 4.4); several properties of such matrices are described in Section 2.8.

Orthonormal matrix

In the same way, the relationships among *descriptors*, which correspond to the rows of matrix **U**, can be studied through the product **U**'. The diagonal and nondiagonal terms of **U**' have the same meaning as in **U**'**U**, except that they now concern the relationships among descriptors. This is PCA *scaling 1*, which is explained in more details in Subsection 9.1.4. The relationships among the rows of a square orthonormal matrix are the same as among the columns (Section 2.8, property 7), so that:

Scaling 1

$$\mathbf{U'} = \mathbf{I} \qquad \textbf{(9.6)}$$

The descriptors are therefore of unit lengths in the multidimensional space and they lie at 90° of one another (orthogonality).

Principal component analysis is simply a rotation, in the multidimensional space, of the original system of axes (Figs. 9.2 and 9.3a, for a two-dimensional space). It therefore follows that, after the analysis (rotation), the original descriptor-axes are still at 90° of one another. Furthermore, normalizing the eigenvectors simultaneously normalizes the descriptor-axes (the lengths of the row and column vectors are given outside the matrix):

$$\mathbf{U} = \begin{bmatrix} u_{11} & \cdots & u_{1p} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ u_{p1} & \cdots & u_{pp} \end{bmatrix} \begin{matrix} \sqrt{\Sigma u_{1k}^2} = 1 \\ \cdot \\ \cdot \\ \cdot \\ \sqrt{\Sigma u_{pk}^2} = 1 \end{matrix} \qquad \text{(9.7)}$$

$$\sqrt{\Sigma u_{j1}^2} = 1 \quad \cdots \quad \sqrt{\Sigma u_{jp}^2} = 1$$

There is a second approach to the study of the relationships among descriptors. It consists in scaling the eigenvectors in such a way that the cosines of the angles between descriptor-axes be proportional to their *covariances*. In this approach, the angles between descriptor-axes are between 0° (maximum positive covariance) and 180° (maximum negative covariance); an angle of 90° indicates a null covariance (orthogonality). This result is achieved by scaling each eigenvector k to a length equal to its standard deviation $\sqrt{\lambda_k}$ *. This is PCA *scaling 2*, explained in Subsection 9.1.4. With this scaling for the eigenvectors, the Euclidean distances among objects are not preserved.

Scaling 2

Using the diagonal matrix $\mathbf{\Lambda}$ of eigenvalues (eq. 2.20), the new matrix of eigenvectors, called $\mathbf{U}_{sc2}$ (i.e. $\mathbf{U}$ for scaling 2), can be directly computed by means of the expression $\mathbf{U\Lambda}^{1/2}$. For the numerical example:

$$\mathbf{U}_{sc2} = \mathbf{U\Lambda}^{1/2} = \begin{bmatrix} 0.8944 & -0.4472 \\ 0.4472 & 0.8944 \end{bmatrix} \begin{bmatrix} \sqrt{9} & 0 \\ 0 & \sqrt{5} \end{bmatrix} = \begin{bmatrix} 2.6833 & -1.0000 \\ 1.3416 & 2.0000 \end{bmatrix} \qquad \text{(9.8)}$$

In scaling 2, the relationships among descriptors are the same as in the dispersion matrix $\mathbf{S}$ (on which the analysis is based), since

$$(\mathbf{U\Lambda}^{1/2})\,(\mathbf{U\Lambda}^{1/2})' = \mathbf{U\Lambda U'} = \mathbf{U\Lambda U}^{-1} = \mathbf{S} \qquad \text{(9.9)}$$

* In some computer packages, PCA only scales the eigenvectors to length $\sqrt{\lambda}$ and only provides a plot of the descriptor-axes; no plot of the objects in reduced space is available.

Equation $\mathbf{U\Lambda U}^{-1} = \mathbf{S}$ is derived directly from the general equation of eigenvectors $\mathbf{SU} = \mathbf{U\Lambda}$ (eq. 2.27). In other words, the new matrix $\mathbf{U\Lambda}^{1/2}$ is of the following form (the lengths of the row and column vectors are given outside the matrix):

$$\mathbf{U}_{sc2} = \mathbf{U\Lambda}^{1/2} = \begin{bmatrix} u_{11}\sqrt{\lambda_1} & \cdots & u_{1p}\sqrt{\lambda_p} & \sqrt{\Sigma\,(u_{1k}\sqrt{\lambda_k})^2} = s_1 \\ \vdots & & \vdots & \vdots \\ u_{p1}\sqrt{\lambda_1} & \cdots & u_{pp}\sqrt{\lambda_p} & \sqrt{\Sigma\,(u_{pk}\sqrt{\lambda_k})^2} = s_p \end{bmatrix} \qquad \textbf{(9.10)}$$

$$\sqrt{\Sigma\,(u_{j1}\sqrt{\lambda_1})^2} = \sqrt{\lambda_1} \quad \cdots \quad \sqrt{\Sigma\,(u_{jp}\sqrt{\lambda_p})^2} = \sqrt{\lambda_p}$$

This equation shows that, when the eigenvectors are scaled to the lengths of their respective standard deviations $\sqrt{\lambda_k}$, the lengths of the descriptor-axes are $\sqrt{s_j^2} = s_j$ (i.e. their standard deviations) in multidimensional space. The product of two descriptor-axes, which corresponds to their angle in the multidimensional space, is therefore equal to their *covariance* s_{jl}.

2. *Projection of descriptors in reduced space, scaling 1: matrix* **U**. — When matrix **U** is used to project the descriptor-axes in a PCA plot, the descriptor-axes are of unit lengths and at right angles in multidimensional space (Fig. 9.3a). The angles between descriptor-axes and principal axes are projections of the *rotation angles* corresponding to the elements of matrix **U** (Fig. 4.10). For the numerical example, the angles between descriptors and principal axes are computed as in Section 4.4 using matrix **U**:

$$\mathbf{U} = \begin{bmatrix} 0.8944 & -0.4472 \\ 0.4472 & 0.8944 \end{bmatrix} \xrightarrow{\text{arc cos}} \begin{bmatrix} 26°34' & 116°34' \\ 63°26' & 26°34' \end{bmatrix}$$

The values of angles in the inset of Fig. 9.3a are thus: $\beta = 26°34'$, $\gamma = 63°26'$, $\delta = 26°34'$. The correlations between descriptors *j* and principal axes *k* are the same as in scaling 2 (below) because the two scalings only differ by the stretching of the axes. In scaling 1, the correlations among descriptors are equal to 0 because descriptors are orthogonal (i.e. at right angles) in this representation.

Projection u_{jk} of a descriptor-axis *j* on a principal axis *k* is *proportional* to *the covariance* of that descriptor with the principal axis. The proportionality factor is different for each principal axis, so that it is not possible to compare the projection of a descriptor on one axis to its projection on another axis. It is correct, however, to compare the projections of different descriptor-axes on the same principal axis. It can be shown that an isogonal projection (with respectively equal angles) of *p* orthogonal axes of unit lengths gives a length $\sqrt{d/p}$ to each axis in *d*-dimensional space. In Fig. 9.4, the equilibrium projection of each of the three orthogonal unit axes, in two-

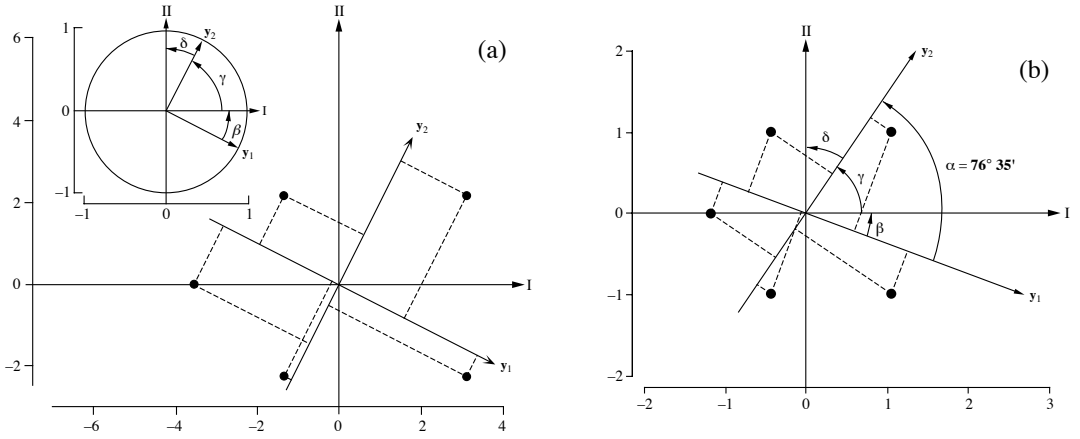


Figure 9.3 Numerical example from Fig. 9.2. Distance and correlation biplots are discussed in Subsection 9.1.4. (a) Distance biplot (scaling 1). The eigenvectors are scaled to lengths 1. Inset: descriptors (matrix **U**). Main graph: descriptors (matrix **U**; arrows) and objects (matrix **F**; dots). The interpretation of the object-descriptor relationships is not based on their proximity, but on orthogonal projections (dashed lines) of the objects on the descriptor-axes or their extensions. The lengths of the arrows were multiplied by 4 for clarity of the diagram. (b) Correlation biplot (scaling 2). Descriptors (matrix $\mathbf{U}\boldsymbol{\Lambda}^{1/2}$; arrows) with a covariance angle of 76°35'. Objects (matrix **G**; dots). Projecting the objects orthogonally on a descriptor (dashed lines) reconstructs the values of the objects along that descriptors, to within a multiplicative constant.

dimensional space, has a length of $\sqrt{2/3}$. This is due to the fact that an isogonal projection results in an equal association of all descriptor-axes with the principal axes.

Equilibrium circle

An *equilibrium circle of descriptors,* with radius $\sqrt{d/p}$, may be drawn as reference to assess the contribution of each descriptor to the formation of the reduced space (Fig. 9.4),. The circle is also drawn in the inset of Fig. 9.3a; its radius is $\sqrt{2/2} = 1$ because, in the numerical example, both the reduced space and the total space are two-dimensional. If one was only interested in the equilibrium contribution of descriptors to the first principal axis, the one-dimensional "circle" would then have a "radius" of $\sqrt{1/2} = 0.7071$. For the example, the projection of the first descriptor on the first principal axis is equal to 0.8944 (examine matrix **U** and Fig. 9.3a), so that this descriptor contributes in an important way to the formation of axis I. This is not the case for the second descriptor, whose projection on the first axis is only 0.4472.

3. *Projection of descriptors in reduced space, scaling 2: matrix* $\mathbf{U}_{sc2}$. — Ecologists using principal component analysis are not interested in the whole multidimensional space but only in a simplified *projection* of the objects in a *reduced space* (generally a two-dimensional plane). The elements $u_{jk}\sqrt{\lambda_k}$ of the eigenvectors scaled to $\sqrt{\lambda_k}$ are the coordinates of the projections of descriptors j on the different principal axes k.

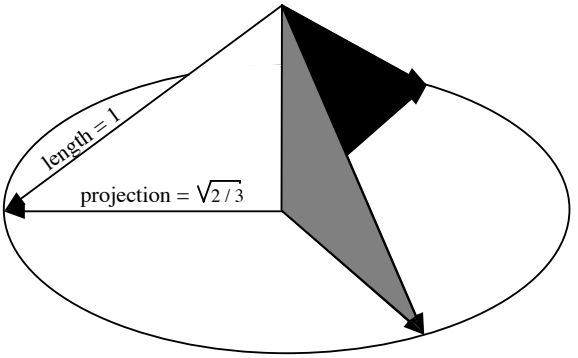


Figure 9.4 Equilibrium projection, in a plane, of three orthogonal vectors with unit lengths, and equilibrium circle of the three descriptors.

They are scaled in such a way that the projections of descriptor-axes can be drawn in the reduced space formed by the principal axes (Fig. 9.3b). As in scaling 1, the descriptors are represented by arrows since they are *axes*. In a reduced-dimension plane, projections of descriptor-axes are shorter than or equal to their lengths in the multidimensional space. In the case of Fig. 9.3b, the lengths are the same in the projection plane as in the original space because the latter only has two dimensions.

In the reduced-space plane, the angles between descriptors are *projections* of their true covariance angles. It is thus important to consider only the descriptors that are well represented in the projection plane. To do so, one must recognize, in the multidimensional space, the descriptors that form small angles with the reduced plane; they are the descriptors whose projections approach their real lengths s in the multidimensional space. Since the length of the *projection* of a descriptor-axis j is equal to or shorter than s_j, one must choose a criterion to assess the value of the representations in the projection plane.

If a descriptor j was equally associated with each of the p principal axes, all elements of row j (which is of length s_j) of matrix $\mathbf{U}_{sc2}$ would be equal, their values being $s_j\sqrt{1/p}$. The length of the descriptor-axis would be $[p\,(s_j\sqrt{1/p})^2]^{1/2} = s_j$ in multidimensional space. The length of the projection of this descriptor-axis in a reduced space with d dimensions would therefore be $s_j\sqrt{d/p}$. The latter expression defines, in the d-dimensional space, a measure of the *equilibrium contribution of a descriptor* to the various axes of the whole multidimensional space. When applied to scaling 1, where the length of the projection of a descriptor in p-dimensional space is 1 (eq. 9.7) instead of s_j (eq. 9.10), the formula for the equilibrium projection of p orthogonal axes in d-dimensional space becomes $\sqrt{d/p}$, as shown above.

Equilibrium contribution

The actual length of a descriptor in reduced space can be compared to that measure, to help judge whether the contribution of the descriptor to the reduced space is larger or smaller than it would be under the hypothesis of an equal contribution to all principal axes. For the numerical example, the lengths of the rows of matrix $\mathbf{U}_{sc2}$ (eq. 9.8), in two-dimensional space, are:

$$\text{length of the first descriptor (row)} = \sqrt{2.6833^2 + (-1.0000)^2} = 2.8636 = s_1$$

$$\text{length of the second descriptor (row)} = \sqrt{1.3416^2 + 2.0000^2} = 2.4083 = s_2$$

Because this simple numerical example has two dimensions only, these lengths are equal to their equilibrium contributions in the two-dimensional space. This is easily verified, using the variances of the descriptors, which are known (Subsection 9.1.1):

$$\text{equilibrium projection of first descriptor} = s_1\sqrt{2/2} = \sqrt{8.2}\sqrt{2/2} = 2.8636$$

$$\text{equilibrium projection of second descriptor} = s_2\sqrt{2/2} = \sqrt{5.8}\sqrt{2/2} = 2.4083$$

In real studies, where ecological data sets are multidimensional, the lengths of descriptors in the reduced space are not equal to their equilibrium contributions.

In scaling 1 above, the angular interpretation of the product of two descriptor-axes was simple: the descriptor-axes were at right angles in multidimensional space. In scaling 2, the angle between two descriptors can be found by applying eq. 2.9 to the rows of the matrix of eigenvectors: the scalar product of two rows of matrix $\mathbf{U}_{sc2}$ (eq. 9.10) divided by the product of the lengths of the rows (which are the standard deviations s_j), gives the cosine of that angle.

The scalar product of two rows of matrix $\mathbf{U}_{sc2}$ is related to the correlation coefficient of the corresponding descriptors. The angles among all descriptors are obtained by reducing to unity (i.e. = 1) the lengths of the row vectors of matrix $\mathbf{U}_{sc2} = \mathbf{U}\boldsymbol{\Lambda}^{1/2}$, then computing the matrix of scalar products among the rows:

$$[\mathbf{D}(s)^{-1}\mathbf{U}\boldsymbol{\Lambda}^{1/2}]\,[\mathbf{D}(s)^{-1}\mathbf{U}\boldsymbol{\Lambda}^{1/2}]' = \mathbf{D}(s)^{-1}\underbrace{\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}'}_{\mathbf{S}}\mathbf{D}(s)^{-1} = \mathbf{D}(s)^{-1}\,\mathbf{S}\,\mathbf{D}(s)^{-1} = \mathbf{R} \qquad \textbf{(9.11)}$$

The result of this equation is the correlation matrix among the descriptors. In the last step of the equation, the correlation matrix $\mathbf{R}$ is connected to the dispersion matrix $\mathbf{S}$ by the diagonal matrix of standard deviations $\mathbf{D}(s)$, following eq. 4.10.

The cosine of the angle α_{jl} between two descriptors $\mathbf{y}_j$ and $\mathbf{y}_l$, in multidimensional space, is therefore related to their *correlation* (r_{jl}); it can actually be shown that $\cos(\alpha_{jl}) = r_{jl}$. This angle is the same as that of the *covariance* because standardization of the rows to unit lengths has only changed the lengths of the descriptor-axes and not their positions in multidimensional space. For the numerical example, the correlation between the two descriptors is equal to $1.6/\sqrt{8.2 \times 5.8} = 0.232$ (eq. 4.7). The angle

corresponding to this correlation is (arc cos 0.232) = 76°35', which is the same as the angle of the covariance in Fig. 9.3b.

In the same way, the angle between a descriptor j and a principal axis k, in multidimensional space, is the arc cosine of the correlation between descriptor j and principal component k. The correlation r_{jk} is element jk of the matrix of eigenvectors (eq. 9.10) normalized by row (the length of a row vector in eq. 9.10 is s_j):

$$r_{jk} = u_{jk}\sqrt{\lambda_k}/s_j \qquad \textbf{(9.12)}$$

In other words, the correlation is calculated by weighting the element of the eigenvector by the ratio of the standard deviation of the principal component to that of the descriptor. For the numerical example, these correlations and corresponding angles are computed using matrix $\mathbf{U\Lambda}^{1/2}$ (calculated above) and the standard deviations of the two descriptors ($s_1 = 2.8636$, $s_2 = 2.4083$):

$$[r_{jk}] = [u_{jk}\sqrt{\lambda_k}/s_j] = \begin{bmatrix} 0.9370 & -0.3492 \\ 0.5571 & 0.8305 \end{bmatrix} \xrightarrow{\text{arc cos}} \begin{bmatrix} 20°26' & 110°26' \\ 56°09' & 33°51' \end{bmatrix}$$

The values of angles in Fig. 9.3b are thus: $\beta = 20°26'$, $\gamma = 56°09'$, $\delta = 33°51'$. These correlations may be used to study the contributions of the descriptors to the various components, the scale factors of the descriptors being removed. The highest correlations (absolute values), in the correlation matrix between descriptors and components, identify the descriptors that contribute most to each eigenvector. The significance of the correlations between descriptors and components cannot be tested using a standard test for Pearson correlation coefficients, however, because the principal components are linear combinations of the descriptors themselves.

When the descriptor-axes of matrix $\mathbf{U\Lambda}^{1/2}$ are scaled to unit lengths, which is done by computing $[\mathbf{D}(s)^{-1}\mathbf{U\Lambda}^{1/2}]$ as in eq. 9.11, drawing their projections in the principal axes space is not recommended. This is because the rescaled eigenvectors are not necessarily orthogonal and may be of any lengths:

$$[\mathbf{D}\,(s)^{-1}\,\mathbf{U\Lambda}^{1/2}]\,[\mathbf{D}\,(s)^{-1}\,\mathbf{U\Lambda}^{1/2}] \neq \mathbf{I} \qquad \textbf{(9.13)}$$

The principal axes are therefore not necessarily at right angles.

The projections of the descriptor-axes of matrix $\mathbf{U\Lambda}^{1/2}$ may be examined, in particular, with respect to the following points:

- The coordinates of the projection of a descriptor-axis specify the position of the apex of this descriptor-axis in the reduced space. It is recommended to use arrows to represent projections of descriptor-axes. Some authors call them point-descriptors or point-variables and represent them by *points* in the reduced space. This representation is ambiguous and misleading. It is acceptable only if the nature of the point-descriptors

is respected; they actually are *apices of descriptor-axes,* so that the relationships among them are defined in terms of angles representing their correlations, not in terms of proximities (Fig. 9.5).

• The projection $u_{jk}\sqrt{\lambda_k}$ of a descriptor-axis j on a principal axis k shows its covariance with the principal axis and, consequently, its positive or negative contribution to the position of the objects along the axis. It follows that a principal axis may often be qualified by the names of the descriptors that are mostly contributing, and in a preferential way, to its formation. Thus, in Fig. 9.5, principal axis I is formed mainly by descriptors 6 to 10 and axis II by descriptors 1 to 4.

• The descriptors that contribute most to the formation of the reduced space are those whose projected lengths reach or exceed the values of their equilibrium contributions. Descriptor-axes that are clearly shorter than these values contribute little to the formation of the reduced space under study and, therefore, contribute little to the structure that may be found in the projection of the *objects* in that reduced space.

• The correlations among descriptors are expressed by *the angles between descriptor-axes, not by the proximities between their apices*. In the reduced space, one can often identify groups of descriptor-axes that form small angles with one another, or have angles close to 180° ($\cos 180° = -1$, which would reflect a perfect negative correlation). One must remember, however, that projections of correlation angles in a reduced space do not render the complete correlations among variables. Thus, it may be informative to cluster descriptors by cluster analysis (Chapter 8) of a distance matrix computed as the one-complement of the correlations ($\mathbf{D} = 1 - \text{cor}(\mathbf{Y})$) or the one-complement of the absolute values of the correlations.

• Objects in scaling 2 (or correlation) biplots can be projected at right angles onto the descriptor-axes to approximate their values along the descriptors (Fig. 9.3b). The distances among objects in a scaling 2 biplot *are not* approximations of their Euclidean distances; they approximate their Mahalanobis distances (Subsection 9.1.4).

The main properties of a principal component analysis of centred descriptors are summarized in Table 9.2.

4. *Cumulative fit tables*. — How well is the variance of each descriptor explained, or *fitted*, by 1, 2, or more axes of the PCA solution? To obtain that information, R^2 coefficients (eq. 10.20) can be computed between the descriptors (which have possibly been standardized or transformed in some other way prior to PCA) and principal components 1, 2, …, k found in matrices **F** or **G** (Subsection 9.1.4). Then, for each descriptor, a table of *Cumulative fit per descriptor* is created. The R^2 coefficient for axis 1 alone is written in column 1; the cumulated sum of the R^2 coefficients over axes 1 and 2 is written in column 2; the cumulated sum of the R^2 coefficients over axes 1, 2 and 3 is written in column 3; and so on. For the numerical example data of

Cumulative fit per descriptor

Table 9.2 Principal component analysis. Main properties for centred descriptors j.

	Scaling 1 (distance biplot)	**Scaling 2** (correlation biplot)
Length of the scaled eigenvectors	1	$\sqrt{\lambda_k}$
Length of descriptor j in $\mathbf{U}$ or $\mathbf{U}_{sc2}$	1	s_j
Angles in reduced space	90°, i.e. rigid rotation of the system of axes	projections of covariances (correlations)
Length of equilibrium contribution	circle with radius $\sqrt{d/p}$	$s_j\sqrt{d/p}$
Projection on principal axis k	u_{jk} i.e. proportional to the covariance with k	$u_{jk}\sqrt{\lambda_k}$ i.e. covariance with component k
Correlation with principal component k	$u_{jk}\sqrt{\lambda_k}/s_j$	$u_{jk}\sqrt{\lambda_k}/s_j$

the present section, which produce a PCA solution in two dimensions, the table has two columns only:

	Cumul. axis 1	**Cumul. axis 2**
Descriptor $\mathbf{y}_1$	0.8780	1.0000
Descriptor $\mathbf{y}_2$	0.3103	1.0000

The values of R^2 in the last column are always 1 in PCA. Identical results can be computed directly from matrix $\mathbf{U}_{sc2}$: the coefficients found in that matrix (eq. 9.8) are squared and summed cumulatively from left to right; then the cumulated sum for row j is divided by the total variance of descriptor j. That table proves very useful for interpretation of analyses involving many variables, in particular in the case of species-rich assemblages in community studies: it allows one to decide which species are well fitted and should be represented, for example, in a two-dimensional PCA biplot (Subsection 9.1.4). This output table is available in program CANOCO where it is called "Cumulative fit per species as fraction of variance of species".

Objects are vectors in multivariate A-space (Fig. 7.2) and vectors have lengths (Section 2.4). The squared length of each object, computed as the sum of the squared values in matrix $\mathbf{Y}_c$ subjected to PCA, is the reference value. An identical total squared length can be computed using matrix $\mathbf{F}$ instead of $\mathbf{Y}_c$. Use matrix $\mathbf{F}$ to compute the

squared length of each object in 1, 2, 3 … PCA dimensions. For example, the total squared length of object 2 of the numerical example is $(-2.2^2 + 1.4^2) = 6.8$ and the squared length of the projection of object 2 on PCA axis I, read from matrix **F** (Subsection 9.1.2), is $(f_{21})^2 = (-1.342)^2 = 1.8$. Hence the proportion of fit for that object along PCA axis I is 1.8/6.8 = 0.2647. The squared residual length of object 2 after fitting it along axis I is 6.8 – 1.8 = 5. For the numerical example data, the table of *cumulative percent fit of the objects* is:

Cumulative fit of objects

	Cumul. axis 1	**Cumul. axis 2**
Object 1	1.0000	1.0000
Object 2	0.2647	1.0000
Object 3	0.2647	1.0000
Object 4	0.6622	1.0000
Object 5	0.6622	1.0000

This type of output table is useful to decide which objects are well represented in a PCA plot: the distances between well-represented objects can be trusted and interpreted. A related table called "Squared residual length per sample" is available in the output of program CANOCO; it gives squared residual values instead of relative values or percent fit.

4 — PCA biplots

The previous two subsections have shown that, in principal component analysis, both the descriptor-axes and object-vectors can be plotted in the reduced space. This led Jolicoeur & Mosimann (1960) to plot these projections together in the same diagram. Gabriel (1971) proposed the name *biplot* for these diagrams and developed the theory of biplots in Gabriel (1971, 1982). Other important contributors to the theory of biplots are ter Braak (1983 and other papers) and Gower (1990 and other papers). Mathematical details about the theory of PCA biplots are found in Greenacre (2010) and Gower *et al*. (2011); these books offer R functions to produce various types of biplots.

Two types of biplots may be used to represent PCA results (Gabriel, 1982; ter Braak, 1994). *Distance biplots* graph together matrices **U** (eigenvectors scaled to lengths 1) and **F** (eq. 9.4); in **F**, the variance of principal component (column) k is λ_k. *Correlation biplots* use matrix $\mathbf{U}_{sc2}$ for descriptors, where eigenvector k is scaled to length $\sqrt{\lambda_k}$, and matrix **G** for objects, where

$$\mathbf{G} = \mathbf{F}\mathbf{\Lambda}^{-1/2} \qquad \textbf{(9.14)}$$

The columns of matrix **G** have unit variances. The Euclidean distances among the objects in matrix **G** are equal to the Mahalanobis distances (D_5, eq. 7.38) among the objects in the original data matrix **Y**, so that the distances in a correlation biplot are

Mahalanobis distance

projections of these Mahalanobis distances, not of the original Euclidean distances. Since Mahalanobis distances are independent of the scaling of descriptors, it follows that the Euclidean distances among objects in matrix **G** are the same for a PCA conducted on unstandardized or standardized descriptors when considering all axes.

Matrices **F** and **U**, or **G** and $\mathbf{U}_{sc2}$, can be used together in biplots because the products of the eigenvectors with the object score matrices reconstruct the original (centred) matrix **Y** perfectly:

$$\mathbf{F}\mathbf{U}' = \mathbf{Y} \quad \text{and} \quad \mathbf{G}(\mathbf{U}\mathbf{\Lambda}^{1/2})' = \mathbf{Y}.$$

Actually, the eigenvectors and object score vectors may be multiplied by any constant without changing the interpretation of a PCA biplot.

Distance biplot (scaling 1)

• *Distance biplot, scaling 1 (Fig. 9.3a).* — The main features of a distance biplot are the following: (1) Distances among objects in the biplot are approximations of their Euclidean distances in multidimensional space. (2) Projecting an object at right angle on a descriptor approximates the position of the object along that descriptor. (3) Since descriptors have lengths of 1 in the full-dimensional space (eq. 9.7), the length of the projection of a descriptor in reduced space indicates how much it contributes to the formation of that space. (4) The angles among descriptor-axes are meaningless.

Correlation biplot (scaling 2)

• *Correlation biplot, scaling 2 (Fig. 9.3b).* — The main features of a correlation biplot are the following: (1) Distances among objects in the biplot are approximations of their Mahalanobis distances in multidimensional space; they *are not* approximations of their Euclidean distances. (2) Projecting an object at right angle on a descriptor approximates the position of the object along that descriptor. (3) Since descriptors have lengths s_j in full-dimensional space (eq. 9.10), the length of the projection of a descriptor in reduced space is an approximation of its standard deviation. (4) The angles between descriptors in the biplot reflect their correlations. (5) When the distance relationships among objects are important for interpretation, this type of biplot is inadequate; a distance biplot should be used.

For the numerical example, matrix **G** is computed from eq. 9.14 as follows:

$$\mathbf{G} = \mathbf{F}\mathbf{\Lambda}^{-1/2} = \begin{bmatrix} -3.578 & 0 \\ -1.342 & 2.236 \\ -1.342 & -2.236 \\ 3.130 & 2.236 \\ 3.130 & -2.236 \end{bmatrix} \begin{bmatrix} 0.3333 & 0 \\ 0 & 0.4472 \end{bmatrix} = \begin{bmatrix} -1.193 & 0.000 \\ -0.447 & 1.000 \\ -0.447 & -1.000 \\ 1.044 & 1.000 \\ 1.044 & -1.000 \end{bmatrix}$$

One can check that the columns of **G** have unit variances. In this particular example, the relationships between objects and descriptors are fully represented in a two-dimensional space. Readers are invited to repeat the PCA using standardized descriptors and verify the fact that the Euclidean distances among objects in matrix **G**

are the same for unstandardized and standardized descriptors: in both cases, they are the Mahalanobis distances among the objects in matrix **Y**.

The descriptor coordinates must often be multiplied by a constant to produce a clear visual display. In Fig. 9.3a for instance, the lengths of the descriptor arrows would be too short for visual appraisal if they were plotted in the same system of coordinates as the objects. In computer software, rescaling of the descriptors is done either by the PCA function or by the plotting function. Some researchers have been tempted to interpret the relationships between objects and descriptors in terms of their proximity in the reduced space, whereas a correct interpretation requires the projection of the objects on the descriptor-axes (centred with respect to the scatter of points) or on their extensions (Fig. 9.3a). In Fig. 9.2a for example, it would not come to mind to interpret the relationship between the objects and descriptor $\mathbf{y}_1$ in terms of the distance between the object-points and the apex (head of the arrow) of axis $\mathbf{y}_1$. In Fig. 9.3a, the position of the apex of $\mathbf{y}_1$ is arbitrary and depends on the multiplicative constant used. Projections of objects onto an axis specify the coordinates of the objects with respect to that descriptor-axis, taking the multiplicative constant into account.

5 — Principal components of a correlation matrix

Principal component analysis performs a partitioning of the total variance of matrix **Y**. If the variables are not in the same physical dimensions, their variances, which are expressed in the squared units of the variables, cannot be added (Section 3.2). As a consequence, before adding the variances of the p variables of **Y**, one must make sure that the variables are dimensionally homogeneous, i.e. expressed in the same physical dimensions. If they are not, they must be standardized (eq. 1.12). The analysis is then described as a PCA carried out on a correlation matrix **R** since correlations are covariances of standardized descriptors (Section 4.2).

In an **R** matrix, all diagonal elements are equal to 1. It follows that the sum of eigenvalues, which corresponds to the total variance of the dispersion matrix, is equal to the order of **R**, given by the number of descriptors p. Before computing the principal components, it may be a sound practice to check that $\mathbf{R} \neq \mathbf{I}$ (eq. 4.14).

Principal components extracted from correlation matrices are not the same as those computed from dispersion matrices. [Beware: some computer packages only allow the computation of principal components from correlation matrices; this is inappropriate for many studies.] Consider the basic equation for the eigenvalues and eigenvectors, $(\mathbf{S} - \lambda_k \mathbf{I})\, u_k = 0$. The sum of the eigenvalues of **S** is equal to the sum of variances s^2, whereas the sum of eigenvalues of **R** is equal to p, so that the eigenvalues of the two matrices, and therefore also their eigenvectors, are necessarily different. This is due to the fact that distances between objects are not the same in the two analyses.

In the case of correlations, the descriptors are standardized. It follows that the distances among objects are independent of the measurement units, whereas those in

the space of the original descriptors vary according to measurement scales. When the descriptors are all of the same type and order of magnitude, and have the same units, it is clear that matrix **S** must be used to compute PCA. In that case, the eigenvectors, on the one hand, and the correlation coefficients between descriptors and components, on the other hand, provide complementary information. The former give the loadings of descriptors and the latter quantify their relative importance. When the descriptors are of different types or orders of magnitude, or have different units, one must conduct PCA on matrix **R** instead of matrix **S**.

S or **R** matrix?

Ecologists who wish to study the relationships among objects in a reduced space of principal components may base their decision of conducting the analysis on **S** or **R** on the answer to the following question:

- If one wanted to cluster the objects in the reduced space, should the clustering be done with respect to the original descriptors (or any transformation of these descriptors; Section 1.5), thus preserving their differences in magnitude? Or, should all descriptors contribute equally to the clustering of objects, independently of the variance exhibited by each one? In the second instance, one should proceed from the correlation matrix. An alternative in this case is to transform the descriptors by ranging, using eq. 1.10 for relative-scale descriptors and eq. 1.11 for interval-scale descriptors, and carry out the analysis on matrix **S** of the transformed descriptors.

Another way to look at the same problem was suggested by Gower (1966):

- The Euclidean distance (eq. 7.32) is the distance preserved among objects through principal component analysis. Is it with the raw data (covariances) or with the standardized data (correlations) that the spatial configuration of the objects, in terms of Euclidean distances, is the most interesting for interpretation? In the first case, conduct PCA on matrix **S**; in the second case, use matrix **R**.

The principal components of a correlation matrix are computed from matrix **U** of the eigenvectors of **R** and the matrix of standardized observations:

$$\mathbf{F} = \left[\frac{y - \bar{y}}{s_y}\right]\mathbf{U} \qquad \textbf{(9.15)}$$

Principal component analysis is still only a rotation of the system of axes (Subsection 9.1.2). However, since the descriptors are now *standardized*, the objects are not positioned in the same way as if the descriptors had simply been *centred* (i.e. principal components computed from matrix **S** in the previous subsections).

As far as the representation of descriptors in the reduced space computed from matrix **R** is concerned, the conclusions of Subsection 9.1.3, which concerned matrix **S**, can be used here, after replacing *covariance* by *correlation*, s_{jl} by r_{jl}, and *dispersion matrix* **S** by *correlation matrix* **R**.

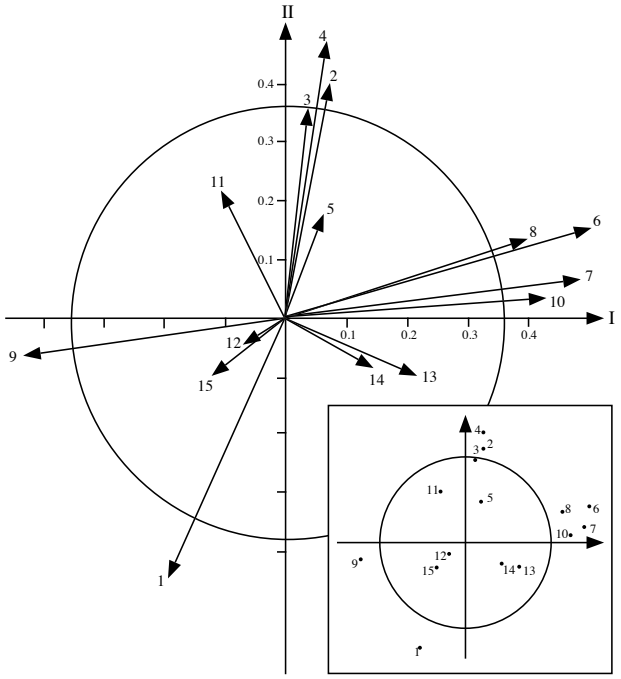


Figure 9.5 Fifteen descriptors plotted in the plane determined by the first two principal axes. The coordinates of each descriptor are the first two elements of the corresponding row of matrix $\mathbf{U}\boldsymbol{\Lambda}^{1/2}$ (i.e. the eigenvectors of $\mathbf{R}$ scaled to $\sqrt{\lambda}$). The circle of equilibrium descriptor contribution is drawn at $\sqrt{2/15} = 0.365$. The inset figure shows the same descriptor-axes using only the apices of the vectors. This representation, which is sometimes encountered in the ecological literature, must be avoided because of possible confusion with point-objects.

The variances, and therefore also the standard deviations, of the *standardized* descriptors are equal to unity (i.e. = 1), which leads to some special properties for the $\mathbf{U}\boldsymbol{\Lambda}^{1/2}$ matrix. First, $\mathbf{D}(s) = \mathbf{I}$, so that $\mathbf{U}\boldsymbol{\Lambda}^{1/2} = \mathbf{D}(s)^{-1}\mathbf{U}\boldsymbol{\Lambda}^{1/2}$, i.e. the coefficients $u_{jk}\sqrt{\lambda_k}$ are the correlation coefficients between descriptors j and components k. In addition, the equilibrium contribution corresponding to each descriptor, in the reduced space of $\mathbf{U}\boldsymbol{\Lambda}^{1/2}$, is $s_j\sqrt{d/p} = \sqrt{d/p}$ (since $s_i = 1$). It is therefore possible to judge whether the contribution of each descriptor to the reduced space is greater or smaller than expected under the hypothesis of an equal contribution to all principal axes, by comparing the lengths of their projections to an equilibrium circle with radius $\sqrt{d/p}$ (Fig. 9.5).

The main properties for standardized descriptors are summarized in Table 9.3, which parallels Table 9.2 for centred descriptors.

Table 9.3 Principal component analysis. Main properties for standardized descriptors j.

	Scaling 1 (distance biplot)	**Scaling 2** (correlation biplot)
Length of the scaled eigenvectors	1	$\sqrt{\lambda_k}$
Length of descriptor j in $\mathbf{U}$ or $\mathbf{U}_{sc2}$	1	1
Angles in reduced space	90°, i.e. rigid rotation of the system of axes	projections of correlations
Radius of equilibrium contribution circle	$\sqrt{d/p}$	$\sqrt{d/p}$
Projection on principal axis k	u_{jk} i.e. proportional to the correlation with k	$u_{jk}\sqrt{\lambda_k}$ i.e. correlation with component k
Correlation with principal component k	$u_{jk}\sqrt{\lambda_k}$	$u_{jk}\sqrt{\lambda_k}$

6 — *The meaningful components*

The successive principal components correspond to progressively smaller fractions of the total variance. One problem is therefore to determine how many components are meaningful in ecological terms or, in other words, what should be the number of dimensions of the reduced space. The best approach may be to visually check the representativeness of the projections in reduced space for two, three, or more dimensions, using Shepard diagrams (Fig. 9.1). However, principal component analysis being a form of variance partitioning, researchers may wish to test the significance of the variance associated with the successive principal axes.

There are a number of classical statistical approaches to this question, such as Bartlett's (1950) test of sphericity. These approaches have been reviewed by Burt (1952) and Jackson (1993). The problem is that these formal tests require normality of all descriptors, a condition that is rarely met by ecological data.

Kaiser-Guttman criterion

There is an empirical rule suggesting that one should only interpret a principal component if the corresponding eigenvalue λ is larger than the mean of the λ's. In the particular case of standardized data, where $\mathbf{S}$ is a correlation matrix, the mean of the λ's is 1 so that, according to the rule, only the components whose λ's are larger than 1 should be interpreted. This is the so-called Kaiser-Guttman criterion. Ibanez (1973) has provided a theoretical framework for this empirical rule. He showed that, if a variable made of randomly selected numbers is introduced among the descriptors, it is

not possible to interpret the eigenvectors that follow the one on which this random-number variable has the highest loading. One can show that this random-number variable, which has covariances near zero with all the other descriptors, introduces in the analysis an eigenvalue of 1 if the descriptors have been standardized. For non-standardized descriptors, this eigenvalue is the mean of the λ's if the variance of the random-number variable is made equal to the mean variance of the other descriptors.

Broken stick

Frontier (1976) proposed to compare the list of decreasing eigenvalues to the decreasing values of the broken stick model (Subsection 6.5.2). This comparison is based on the following idea. Consider the variance shared among the principal axes to be a resource embedded in a stick of unit length. If principal component analysis had divided the variance at random among the principal axes, the fractions of total variation explained by the various axes would be about the same as the relative lengths of the pieces obtained by breaking the unit stick at random into as many pieces as there are axes. If a unit stick is broken at random into $p = 2, 3, \ldots$ pieces, the expected values (E) of the relative lengths of the successively smaller pieces (j) are given by eq. 6.50:

$$E(\text{piece}_j) = \frac{1}{p}\sum_{x=j}^{p}\frac{1}{x} \qquad \textbf{(9.16)}$$

The expected values are equal to the lengths that would be obtained by breaking the stick at random a large number of times and calculating the mean length of the longest pieces, the second longest pieces, etc. A stick of unit length may be broken at random into p pieces by placing on the stick $(p - 1)$ random break points selected using a uniform [0, 1] random number generator. An R function is available to compute the expected values of the broken stick distribution for any number of pieces (Section 9.5).

Coming back to the eigenvalues, it would be meaningless to interpret the principal axes that explain a fraction of the variance as small as or smaller than that predicted by the broken stick null model. The test may be carried out in two ways. One may compare individual eigenvalues to individual predictions of the broken stick model and select for interpretation only the eigenvalues that are larger than the values predicted by the model. Or, to decide whether eigenvalue λ_k should be interpreted, one may compare the *sum of eigenvalues*, from 1 to k, to the sum of the values from 1 to k predicted by the model. This test usually recognizes the first two or three principal components as meaningful, which corresponds to the experience of ecologists.

After an empirical study using a variety of matrix types, using simulated and real ecological data, Jackson (1993) concluded that two methods consistently pointed to the correct number of ecologically meaningful components in data sets: the broken-stick model and a bootstrapped eigenvalue-eigenvector method proposed in his paper.

Chapter 10 will discuss how to use explanatory variables to ecologically interpret the first few principal components that are considered to be meaningful according to one of the criteria mentioned in the present subsection.

7 — Misuses of principal component analysis

Given the power of principal component analysis, some applications have used it in ways that exceed the limits of the model. Some of these limits may be transgressed without much consequences, while others are more critical. The most common errors are: the use of descriptors for which a measure of covariance is not appropriate, and the interpretation of relationships between descriptors, in reduced space, based on the relative positions of the apices of axes instead of the angles between them.

Principal component analysis was originally defined for data with *multinormal distributions* (Section 4.4), so that its optimal use (Cassie & Michael, 1968) calls for normalization of the data (Subsection 1.5.6). Deviations from normality do not necessarily bias the analysis, however (Ibanez, 1971). It is only important to make sure that the descriptors' distributions are reasonably unskewed. Typically, in analyses conducted with strongly skewed distributions, the first few principal components only separate a few objects with extreme values from the remaining objects, instead of displaying the main axes of variation of all objects in the study.

A full-rank dispersion matrix **S** cannot be estimated using a number of observations n smaller than or equal to the number of descriptors p. When $n \leq p$, since there are $n - 1$ degrees of freedom in total, the rank of the resulting dispersion matrix of order p is $(n - 1)$. In such a case, the eigen-decomposition of **S** produces $(n - 1)$ real and $[p - (n - 1)]$ null eigenvalues. Indeed, positioning n objects while respecting their distances requires $(n - 1)$ dimensions only. The PCA of a data matrix where $n \leq p$ produces $(n - 1)$ eigenvalues larger than 0 and the $(n - 1)$ corresponding eigenvectors and principal components. To obtain a full-rank dispersion matrix **S** and p principal components, the number of objects n must be larger than p.

Principal components are computed from the eigenvectors of a dispersion matrix. This means that the method is to be used on a matrix of covariances (or possibly correlations) with the following properties: matrix **S** (or **R**) has been computed among descriptors that are quantitative, and for which valid estimates of the covariances (or correlations) may be obtained. These conditions are violated in the following cases:

1) Some authors have transposed the data matrix and computed correlations among the *objects* (i.e. Q mode) instead of among the descriptors (R mode). Their aim was to position the descriptors in the reduced space of the objects. There are several reasons why this operation is incorrect, the least being that it is useless considering that principal component analysis provides information about the relationships among both objects and descriptors. The reasons why correlations should not be computed in Q-mode are explained in Box 7.1, where points 1, 2 and 4 also apply to covariances.

In the literature, the expression "components in Q mode" may sometimes designate a rightful analysis conducted on an R matrix. This expression comes from the fact that one can use principal component analysis primarily as a method for positioning objects

in reduced space. The meanings of "Q mode" and "R mode" are variable in the scientific literature; their meanings in numerical ecology are defined in Section 7.1.

Rao (1964), Gower (1966), and Orlóci (1967a) have shown that, *as a computational technique*, principal components can be obtained by computing the eigenvalues and eigenvectors of a Q-mode matrix. The steps are the following:

- Starting with matrix **Y** centred by columns, $\mathbf{Y}_c$, compute matrix $\mathbf{C}_{np} = \mathbf{Y}_c / \sqrt{n-1}$. This matrix is such that $\mathbf{C}'\mathbf{C} = \mathbf{S}_{pp}$, which is the usual variance-covariance matrix of **Y**.
- Compute the cross-product matrix $\mathbf{Q}_{nn} = \mathbf{C}\mathbf{C}'$ instead of $\mathbf{S}_{pp} = \mathbf{C}'\mathbf{C}$.
- Determine the non-zero eigenvalues of **Q** and their associated eigenvectors.
- Scale each eigenvector k to length $\sqrt{\lambda_k}$, then multiply each value by $\sqrt{n-1}$.
- The eigenvalues of matrix **Q** are the same as those of matrix **S**, and the scaled eigenvectors are matrix **F** of the principal components of **Y**. This perfectly valid computational technique is different from the approach criticised in the previous paragraph.

2) Covariances and correlations are defined for quantitative descriptors only (Section 7.5). This implies, in particular, that one must not use multistate qualitative descriptors in analyses based upon covariance matrices, because means and variances computed from non-ordered states are meaningless.

Precision of data

Principal component analysis is very robust, however, to variations in the *precision of data*. Variables may be recoded into a few classes without noticeable change to the results (Frontier & Ibanez, 1974; Dévaux & Millerioux, 1976a).

Spearman correlation

Pearson correlation coefficients calculated using semiquantitative data are equivalent to Spearman's rank correlation coefficients (eq. 5.3). In a discussion of principal component analysis computed using semiquantitative data, Lebart *et al.* (1979) provide, for various numbers of objects and descriptors, values above which the λ's of the first two principal components may be considered significant. Gower (1966) has also shown that, with binary descriptors, principal component analysis positions the objects, in multidimensional space, at distances that are proportional to the square roots of the complements of simple matching coefficients, i.e. $D = \sqrt{1 - S_1}$ (S_1: eq. 7.1).

3) When calculated over data sets with many double-zeros, coefficients such as the covariance and correlation lead to PCA ordinations with inadequate estimates of the distances among sampling sites. The problem arises from the fact that the principal-component rotation preserves the Euclidean distance among objects (Table 9.1, Fig. 9.2d). The double-zero problem has been discussed in Subsection 7.2.2 and the paradox associated with the Euclidean distance has been presented after eq. 7.32. With untransformed species abundance data, principal component analysis should only be used when the sampling sites cover short gradients (see Subsection 9.1.10). For longer ecological gradients, the species data must be transformed using one of the

transformations of Section 7.7. Else, ordinations can be obtained using correspondence analysis (CA, Section 9.2) when the chi-square distance is appropriate, or by principal coordinate analysis (PCoA, Section 9.3) or nonmetric multidimensional scaling (nMDS, Section 9.4) using other adequate distances.

This last remark explains why, in the ecological literature, principal component analysis has at times not provided interesting results, for example in studies of species associations (e.g. Margalef & Gonzalez Bernaldez, 1969; Ibanez, 1972; Reyssac & Roux, 1972). This problem had also been noted by Whittaker & Gauch (1973). The search for species association is discussed in Section 8.9.

Attempts to interpret the *proximities between the apices* of species-axes in the reduced space, instead of considering the angles separating these descriptor-axes (e.g. Fig. 9.5, inset), may also led to incorrect conclusions and useless results.

Table 9.4 summarizes, with reference to the text, the various questions that may be addressed in the course of a principal component analysis.

8 — Ecological applications

Ecological application 9.1a

From 1953 to 1960, pitfall traps were set at 100 sites in four valleys in the Meijendel dune area north of the Hague, in The Netherlands. They were visited weekly during 365 weeks. In the 36500 relevés, approximately 425 animal species were identified, about 90% of them being arthropods. Aart (1973) studied the wolf spiders (*Lycosidea* and *Pisauridae*: 45030 specimens) to assess how lycosid species shared the multidimensional space of resources (see Section 1.0 for the concept of niche). The Aart (1973) paper reports a PCA based on a data table of 100 sites × 12 species obtained by adding the values from the different week catches for each trap; two of the 14 species were eliminated because they had been found only twice and once, respectively. PCA was applied to the standardized species data, which contained about 30% zero values. Previous editions of the present book reproduced the separate PCA plots of species and sites found in the Aart (1973) paper.

Another set of pitfall traps were set at 100 sites for 60 weeks, in 1969-1970, in Bierlap, one of the dune valleys of the previous survey. Eleven of the 12 spider species were the same as in the Aart (1973) paper. Environmental descriptors were obtained for 28 of these sites. The spider data (28 sites × 12 species, data cumulated over the weeks) were analysed by Aart & Smeenk-Enserink (1975) and related to the environmental variables using canonical correlation analysis (CCorA, Section 11.4); these data are reanalyzed in Ecological application 11.1b using redundancy analysis (RDA, Section 11.1) instead of CCorA.

The spider data from the 28 sites* are analysed here by PCA to illustrate the interest of data transformations. Figure 9.6a shows the results of the analysis of the raw species abundance data and Fig. 9.6b the biplot resulting from the analysis of the same data after a $\log(y + 1)$ transformation, the same transformation that had been used by Aart & Smeenk-Enserink (1975).

* The species data file is available electronically. See footnote in Ecological application 11.1b.

Table 9.4 Questions that can be addressed in the course of a principal component analysis and the answers found in Section 9.1.

Before starting a principal component analysis	*Pages*
1) Are the descriptors appropriate?	
⇒ Quantitative descriptors; multinormality; not too many zeros.	450-452
2) Are the descriptors dimensionally homogeneous?	
⇒ If YES, conduct the analysis on the dispersion matrix	442, 446
⇒ If NO, conduct the analysis on the correlation matrix	445-448
3) Purpose of the ordination in reduced space:	
⇒ To preserve and display the relative positions of the objects: scale the eigenvectors to unit lengths to obtain matrix $\mathbf{U}$	432
Draw a distance biplot (descriptors: $\mathbf{U}$; objects: $\mathbf{F} = \mathbf{YU}$)	444
⇒ To display the correlations among descriptors: scale the eigenvectors to $\sqrt{\lambda}$ to obtain matrix $\mathbf{U}_{sc2}$	436
Draw a correlation biplot (descriptors: $\mathbf{U}_{sc2}$; objects: $\mathbf{G} = \mathbf{F}\boldsymbol{\Lambda}^{-1/2}$) (beware: Euclidean distances among objects are not preserved)	444
While examining the results of a principal component analysis	
1) How informative is a representation of the objects in an m-dimensional reduced space?	
⇒ Compute eq. 9.5	433
2) Are the distances among objects well preserved in the reduced space?	
⇒ Compare Euclidean distances using a Shepard diagram	427-428
3) Which eigenvalues are important?	
⇒ Is λ_k larger than the mean of the λ's?	448
⇒ Is the percentage of the variance corresponding to λ_k larger than the corresponding value in the broken stick model?	449
4) What are the descriptors that contribute the most to the formation of the reduced space?	
⇒ Compute the equilibrium contribution of descriptors and, when appropriate, draw the circle	437, 439, 442, 448
⇒ Compute correlations between descriptors and principal axes	440, 442, 448
⇒ Compute the table of "Cumulative fit per descriptor"	441-442
5) How to represent the objects in the reduced space?	
⇒ Scaling 1: $\mathbf{F} = \mathbf{Y}_c\mathbf{U}$; scaling 2: $\mathbf{G} = \mathbf{F}\boldsymbol{\Lambda}^{-1/2}$	433-435, 443-444
⇒ Compute the table of "Cumulative percent fit of the objects"	443

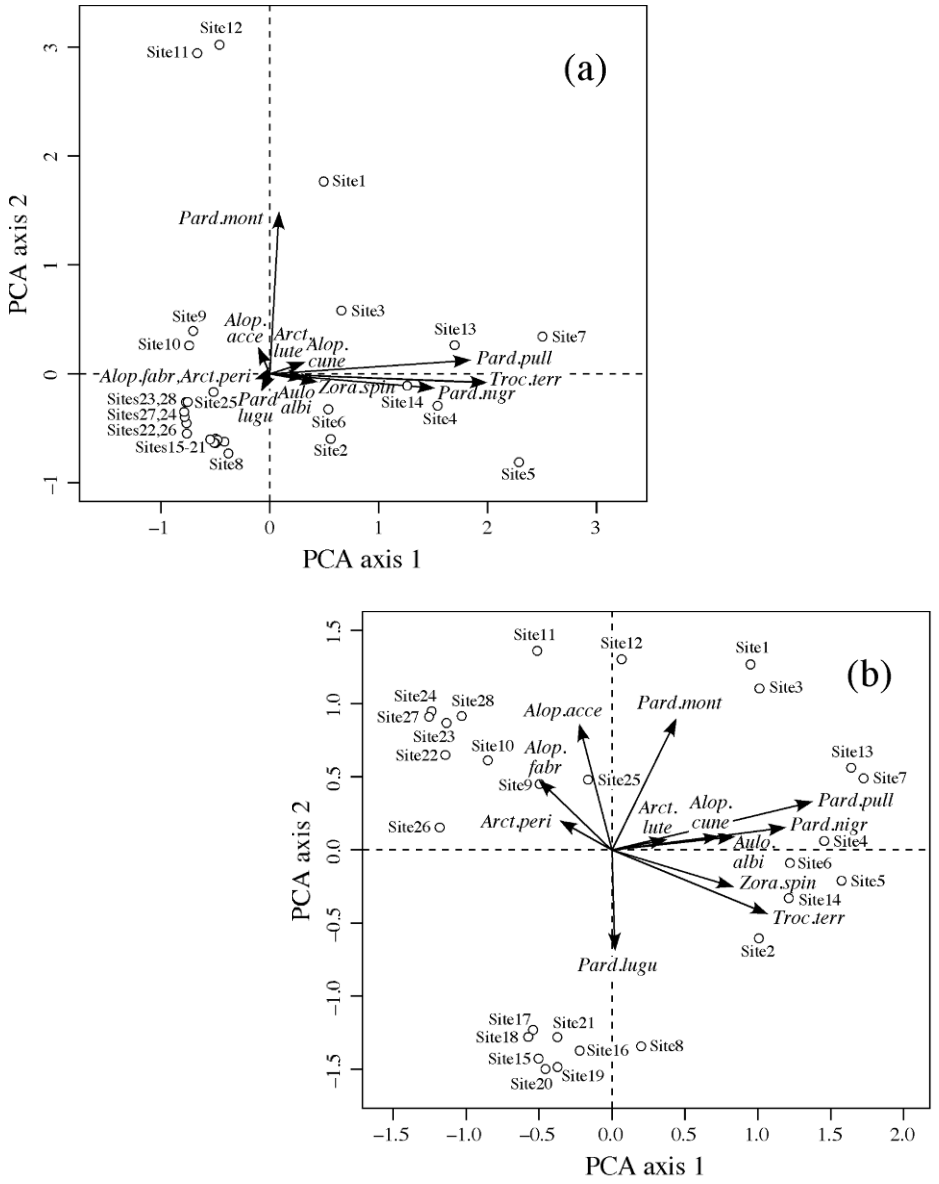


Figure 9.6 PCA correlation biplots of (a) the raw (untransformed) and (b) the log-transformed spider data (28 sites, 12 species). Scaling type 2 was used in both biplots to emphasize the covariances among species. The species are: *Alopecosa accentuata* (abbreviation: *Alop.acce*), *Alopecosa cuneata* (*Alop.cune*), *Alopecosa fabrilis* (*Alop.fabr*), *Arctosa lutetiana* (*Arct.lute*), *Arctosa perita* (*Arct.peri*), *Aulonia albimana* (*Aulo.albi*), *Pardosa lugubris* (*Pard.lugu*), *Pardosa monticola* (*Pard.mont*), *Pardosa nigriceps* (*Pard.nigr*), *Pardosa pullata* (*Pard.pull*), *Trochosa terricola* (*Troc.terr*) and *Zora spinimana* (*Zora.spin*).

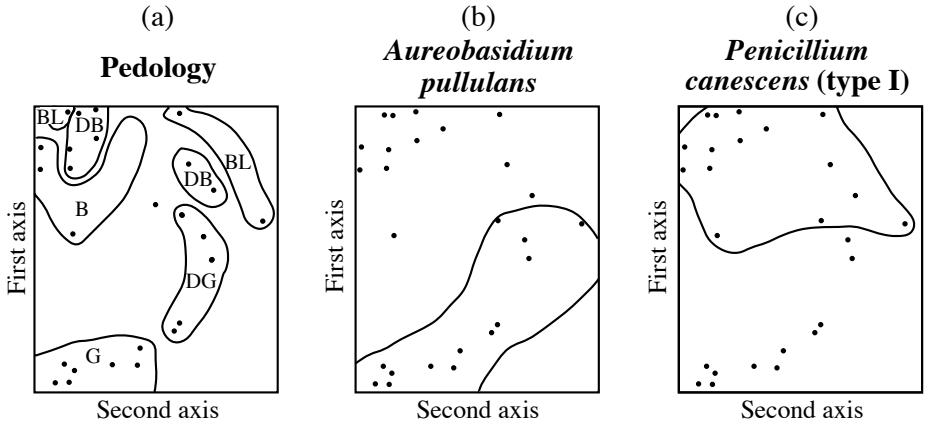


Figure 9.7 Principal component analysis computed from presence-absence of 51 soil microfungi. (a) Pedological information drawn on the ordination of the 26 sampling sites, plotted in the reduced space of the first two principal components. From north to south, soil types are: G = grey, DG = dark grey, BL = black, DB = dark brown, B = brown. (b) and (c) Distributions of the sites (envelopes) where two microflora species were present. Modified from Morrall (1974).

It is clear that Fig. 9.6b is easier to interpret than Fig. 9.6a. The reason is found by examining the "Cumulative fit per descriptor" table described in Subsection 9.1.3: in the raw data analysis, principal components I and II explained more than 60% of the variation for only four of the 12 species, whereas in the analysis of the log-transformed data, the first two principal components explained more than 60% of the variation for 10 of the 12 species.

Readers are invited to compare Fig. 9.6b to the results on the canonical analysis (RDA) in Fig. 11.7; the latter figure provides an interpretation of the site and species clusters using the environmental variables. Compare also the species clusters (species with small angles) in Fig. 9.6b to the species associations described in Ecological application 11.1b.

Ecological application 9.1b

A study of soil microfungi living in association with the aspen *Populus tremuloides* Michx. provides another type of utilization of principal component analysis. This study by Morrall (1974) covered 26 stations with 6 sites each, scattered throughout the Province of Saskatchewan (Canada). It evidenced relationships between the distributions of some species and soil types.

Among the 205 species or taxonomic entities that were identified, 51 were included in the ordination study. The others were not, considering criteria aimed at eliminating rare taxa which could have been either ephemeral constituents of the soil microflora or even contaminants of the laboratory cultures. Observations were transformed into presence-absence data.

Following principal component analysis, the 26 sampling sites were plotted in the reduced space of the first two principal components, onto which information about the nature of the soils was superimposed (Fig. 9.7a). Soils of Saskatchewan may be classified into 5 types, i.e. (G) the

grey wooded soils of the northern boreal forest, followed southward by the dark grey (DG) transition soils and the black soils (BL). Further south are dark brown soils (DB), which give way to the brown soils (B) of the grasslands. Since the principal components were computed from presence-absence data, the distribution of the sites in the reduced space is expected to reflect that of the fungus species. The author tested this for the most abundant species in the study, by plotting, in the reduced space, distributions of the sites where some fungus species were present; two examples are given in Figs. 9.7b and c. The author could then compare these distributions to that of the soil types.

9 — Algorithms

Three different methods are available for computing the eigenvalues and eigenvectors of a real, symmetric matrix, such as a covariance matrix **S**.

Householder

1. The most widely used method of eigen-decomposition is Householder reduction. This is the method implemented in function ***eigen()*** of R. It is very efficient for cases in which *all* eigenvalues and eigenvectors must be computed.

2. Clint & Jennings (1970) published a pioneering paper describing how to compute a *subset* only of the eigenvalues and corresponding eigenvectors of a real symmetric matrix, using an iterative method. Hill (1973b) used this idea to develop a "reciprocal averaging" algorithm for correspondence analysis; Hill's work will be further discussed in Section 9.2 on correspondence analysis. Building on these bases, ter Braak (1987c) proposed a *two-way weighted summation algorithm* (TWWS) for principal component analysis. This algorithm is described in detail here for three reasons: (1) it is closely linked to the basic equations of the PCA method, so that it may help readers understand them; (2) it is easy to program; (3) using it, one can compute the first few components only, when these are the ones of interest. The algorithm is summarized in Table 9.5.

TWWS algorithm

The numerical example worked out in Table 9.6 should help understand how the algorithm computes the principal components, the eigenvectors, and the eigenvalues. The data are those of the numerical example presented at the beginning of Section 9.1 and used in Subsections 9.1.1 to 9.1.4. The procedure starts with the matrix of centred data, $\mathbf{Y}_c$, shown in a box in the upper left-hand corner of Table 9.6.

To estimate principal component I, arbitrary scores are first assigned to the rows of the centred data matrix (Table 9.6, column R0); values $[f_{i1}] = [1\ 2\ 3\ 4\ 5]'$ are used here. Any other initial choice would lead to the same estimate for the first principal component $[f_{i1}]$ although the number of iterations necessary to reach it may differ. The only choice to avoid is to make all initial f_{i1} values equal. From these, column scores are found by multiplying the transpose of the data matrix by the row scores (Table 9.6, row C1):

$$[\text{column scores}_{1j}] = [y - \bar{y}]' [f_{i1}] \qquad \textbf{(9.17)}$$

Table 9.5 Two-way weighted summation (TWWS) algorithm for PCA. Modified from ter Braak (1987c).

a) Iterative estimation procedure

Step 1: Consider a table of n objects (rows) × p variables (columns). Centre each variable (column) on its mean.

Decide how many eigenvectors are needed and, for each one. **DO** the following:

Step 2: Take the row order as the arbitrary initial object scores (1, 2, …). Set the initial eigenvalue estimate to 0.

Iterative procedure begins

Step 3: Compute new variable loadings: $colscore(j) = \sum y(i,j) \times rowscore(i)$

Step 4: Compute new object scores: $rowscore(i) = \sum y(i,j) \times colscore(j)$

Step 5: For the second and higher-order axes, make the object scores uncorrelated with all previous axes (Gram-Schmidt orthogonalization procedure: see *b* below).

Step 6: Scale the vector of object scores to length 1 (normalization procedure *c*, below); obtain S.

Step 7: Upon convergence, the eigenvalue is $S/(n-1)$ where n is the number of objects. So, at the end of each iteration, $S/(n-1)$ provides an estimate of the eigenvalue. If this estimate does not differ from that of the previous iteration by more than a small quantity ("tolerance", set by the user), go to step 8. If the difference is larger than the tolerance value, go to step 3.

End of iterative procedure

Step 8: Normalize the eigenvector (variable loadings), i.e. scale it to length 1 (procedure *c*, below). Rescale the principal component (object scores) to variance = eigenvalue.

Step 9: If more eigenvectors are to be computed, go to step 2. If not, continue with step 10.

Step 10: Return the eigenvalue, % variance, cumulative % variance, eigenvector (variable loadings), and principal component (object scores).

b) Gram-Schmidt orthogonalization procedure

DO the following, in turn, for all previously computed principal components k:

Step 5.1: Compute the scalar product $SP = \sum [rowscore(i) \times v(i,k)]$ of the current object score vector estimate with previous component k, where vector $v(i,k)$ contains the object scores of component k, scaled to length 1. This product varies between 0 (if the vectors are orthogonal) and 1.

Step 5.2: Compute new values of $rowscore(i)$ such that vector *rowscore* becomes orthogonal to vector $v(i,k)$: $rowscore(i) = rowscore(i) - (SP \times v(i,k))$.

c) Normalization procedure

Step 6.1: Compute the sum of squares of the object scores: $S^2 = \sum rowscore(i)^2$, and the length $S = \sqrt{S^2}$.

Step 6.2: Compute the normalized object scores: $rowscore(i) = rowscore(i)/S$.

Subscript 1 designates the first iteration. At the end of the iteration process, the column scores will provide estimates of the first column of matrix **U**. The rationale for this operation comes from the basic equation of eigenanalysis (eq. 2.27) applied to matrix **S**:

$$\mathbf{S\,U = U\,\Lambda}$$

Replacing **S** by its value in the definition of the covariance matrix (eq. 4.6), $\mathbf{S} = (n-1)^{-1}\,[y-\bar{y}]'\,[y-\bar{y}]$, one obtains:

$$[y-\bar{y}]'\,[y-\bar{y}]\,\mathbf{U} = (n-1)\,\mathbf{U}\,\mathbf{\Lambda}$$

Since $\mathbf{F} = [y-\bar{y}]\,\mathbf{U}$ (eq. 9.4), it follows that:

$$[y-\bar{y}]'\,\mathbf{F} = (n-1)\,\mathbf{U}\,\mathbf{\Lambda}$$

Hence, the column scores obtained from eq. 9.17 are the values of the first eigenvector (first column of matrix **U**) multiplied by eigenvalue λ_1 (which is the first diagonal element of matrix **Λ**) and by $(n-1)$.

From the first estimate of column scores, a new estimate of row scores is computed using eq. 9.4, $\mathbf{F} = [y-\bar{y}]\,\mathbf{U}$:

$$[\text{row scores}_{i1}] = [y-\bar{y}]\,[u_{i1}] \qquad \textbf{(9.18)}$$

The algorithm alternates between estimating row scores and column scores until convergence. At each step, the row scores (columns called R in Table 9.6) are scaled to length 1 in order to prevent the scores from becoming too large for the computer to handle, which they may easily do. Before this normalization, the length of the row score vector, divided by $(n-1)$, provides the current estimate of the eigenvalue. This length actually measures the amount of "stretching" the row score vector has incurred during an iteration.

This description suggests one of several possible stopping criteria (Table 9.5, step 7): if the estimate of the eigenvalue has not changed, during the previous iteration, by more than a preselected tolerance value, the iteration process is stopped. Tolerance values between 10^{-10} and 10^{-12} produce satisfactory estimates when computing all the eigenvectors of large matrices, whereas values between 10^{-6} and 10^{-8} are sufficient to compute only the first two or three eigenvectors. Another possible stopping criterion would be a minimum percentage of change in the estimate of the eigenvalue.

At the end of the iterative estimation process (Table 9.5, step 8),

- the eigenvector (Table 9.6, line C13) is normalized (i.e. scaled to unit length), and
- the principal component is scaled to length $\sqrt{(n-1)\,\lambda_1}$. This makes its variance equal to its eigenvalue.

Table 9.6 Estimation of axes I (top) and II (bottom) for the centred data of the numerical example (values in boxes), using the "two-way weighted summation" algorithm (TWWS, Table 9.5). Iterations 1 to 13: estimates of the row scores (R1 to R13) and column scores (C1 to C13).

Objects ⇓	Var. 1	Var. 2	R0 (arbitrary)	R1	R1 length=1	R2	R2 length=1	R3	R3 length=1	...	R9	R9 length=1	...	R13	R13 length=1	R13 scaled (var=λ)
$\mathbf{x}_1$	**–3.2**	**–1.6**	1	–64.000	–0.586	–21.103	–0.593	–21.352	–0.595		–21.466	–0.596		–21.466	–0.596	–3.578
$\mathbf{x}_2$	**–2.2**	**1.4**	2	–34.000	–0.311	–9.745	–0.274	–9.037	–0.252		–8.080	–0.224		–8.053	–0.224	–1.342
$\mathbf{x}_3$	**–0.2**	**–2.6**	3	–14.000	–0.128	–6.082	–0.171	–6.977	–0.195		–8.019	–0.223		–8.047	–0.224	–1.341
$\mathbf{x}_4$	**1.8**	**3.4**	4	46.000	0.421	16.633	0.467	17.653	0.492		18.752	0.521		18.780	0.522	3.130
$\mathbf{x}_5$	**3.8**	**–0.6**	5	66.000	0.605	20.297	0.570	19.713	0.550		18.813	0.523		18.786	0.522	3.131
	Estimates of $\lambda_1 \Rightarrow$				27.295		8.895		8.967			9.000			9.000	
C1	18.000	4.000														
C2	5.642	1.905														
C3	5.544	2.257														
C4	5.473	2.449														
C5	5.428	2.554														
C6	5.401	2.612														
C7	5.386	2.644														
C8	5.377	2.661														
C9	5.373	2.671														
C10	5.370	2.677														
C11	5.368	2.680														
C12	5.368	2.681														
C13	5.367	2.682														
C13 length=1	0.895	0.447														

Objects ⇓	Var. 1	Var. 2	R0 (arbitrary)	R1	R1 ortho*	R1 length=1	R2	R2 ortho*	R2 length=1	R2 scaled (var=λ)
$\mathbf{x}_1$	**–3.2**	**–1.6**	1	–64.000	0.001	0.000	0.002	0.001	0.000	0.000
$\mathbf{x}_2$	**–2.2**	**1.4**	2	–34.000	–9.995	–0.500	–9.999	–10.000	–0.500	–2.236
$\mathbf{x}_3$	**–0.2**	**–2.6**	3	–14.000	9.996	0.500	10.001	10.000	0.500	2.236
$\mathbf{x}_4$	**1.8**	**3.4**	4	46.000	–9.996	–0.500	–10.002	–10.001	–0.500	–2.236
$\mathbf{x}_5$	**3.8**	**–0.6**	5	66.000	9.994	0.500	9.998	9.999	0.500	2.236
	Estimates of $\lambda_2 \Rightarrow$					4.998			5.000	
C1	18.000	4.000								
C2	2.000	–4.000								
C2 length=1	0.447	–0.894								

* ortho: scores are made orthogonal to R13 found in the upper portion of the table.

Note that the eigenvalues, eigenvectors, and principal components obtained using this iterative procedure and shown in Table 9.6 are the same as in Subsections 9.1.1 and 9.1.2, except for the signs of the second eigenvector and principal component, which are all changed in this example. One may arbitrarily change all signs of an eigenvector and the corresponding principal component, since signs result from an arbitrary decision made when computing the eigenvectors (Section 2.9). This is equivalent to turning the ordination diagram by 180° if signs are changed on both the first and second principal components, or looking at it from the back of the page, or in a mirror if signs are changed for one axis only.

To estimate the second principal component, eigenvalue, and eigenvector, row scores are again assigned arbitrarily at the beginning of the iterative process. In Table 9.6 (bottom part), the same values were actually chosen as for axis I, as stated in step 2 of the algorithm (Table 9.5). Iterations proceed in the same way as above, with the exception that, during each iteration, the row scores are made orthogonal to the final estimate obtained for the first principal component (column R13 in the upper portion of Table 9.6). This follows from the basic rule that principal components must be linearly independent (i.e. orthogonal) of one another. For the third and following principal axes, the vectors estimating row scores are made orthogonal, in turn, to *all* previously computed principal components.

The algorithm converges fairly rapidly, even with small tolerance values. For the example of Table 9.6, it took 13 iterations to reach convergence for axis I, and 2 iterations only for axis II, using a tolerance value of 10^{-6}. With a tolerance value of 10^{-10}, it took 21 and 2 iterations, respectively. The initial, arbitrary values assigned to the row scores also have an (unpredictable) effect on the number of iterations; e.g. with a different set of initial values [2 5 4 3 1], it took 14 iterations instead of 13 to reach convergence for the first axis (tolerance = 10^{-6}).

Supplementary object and variable

Supplementary objects or variables may easily be incorporated in the calculations using this algorithm. These are objects or variables that have not been used to compute the eigenvalues and eigenvectors of the ordination space, but whose positions are sought with respect to the original set of objects and variables that were used to compute the eigenvalues and eigenvectors. In Ecological application 9.1a for example, where the principal component analysis was computed using the species abundance data, the environmental descriptors used in Ecological application 11.1b could have been added to the analysis as supplementary variables. In addition, since there were 100 pitfall traps of spider observations, the traps that were excluded from the analysis could have been added to the ordination plot as supplementary objects. Preliminary transformations are required: (1) the supplementary variables must be centred on their respective means; (2) for each variable used in the original PCA (e.g. the 12 spider species), supplementary objects must be centred using the mean value of that variable calculated for the original set of objects. When the algorithm has reached convergence for an axis using the original set of objects, it is a simple matter to compute the column scores of the supplementary variables using eq. 9.17 and the row scores of the supplementary objects using eq. 9.18. The final step consists in applying to the

supplementary variable scores the scaling that was applied to the terms of the eigenvector corresponding to the original set of variables and, to the supplementary object scores, the scaling that was applied to the original set of objects.

SVD

3. Another way of computing principal components involves *singular value decomposition* (SVD, Section 2.11). SVD is also a widely used approach to compute correspondence analysis (Section 9.2).

The relationship with principal component analysis is the following. Centre the column vectors of $\mathbf{Y}$ on their respective means, forming matrix $\mathbf{Y}_c$, and compute the covariance matrix $\mathbf{S} = (n-1)^{-1}\mathbf{Y}_c'\mathbf{Y}_c$ (eq. 4.6). Carry out a singular value decomposition: $\mathbf{Y}_c = \mathbf{V}\ \mathbf{W}\ \mathbf{U}_{svd}'$ (eq. 2.31). The following reasoning will show that matrix $\mathbf{U}$ produced by SVD, $\mathbf{U}_{svd}$, is equal to matrix $\mathbf{U}$ computed by eigen-decomposition, $\mathbf{U}_{eigen}$. Use these SVD result to reconstruct $\mathbf{S}$:

$$\mathbf{S} = \frac{1}{n-1}\mathbf{Y}_c'\mathbf{Y}_c = \frac{1}{n-1}(\mathbf{U}_{svd}\mathbf{W}'\ \mathbf{V}')(\mathbf{V}\ \mathbf{W}\ \mathbf{U}_{svd}')$$

Since $\mathbf{V}$ is orthonormal (Section 2.11), $\mathbf{V}'\mathbf{V} = \mathbf{I}$ and one obtains:

$$\mathbf{S} = \frac{1}{n-1}\mathbf{Y}_c'\mathbf{Y}_c = \frac{1}{n-1}\mathbf{U}_{svd}\ \mathbf{W}'\mathbf{W}\ \mathbf{U}_{svd}' \qquad \textbf{(9.19)}$$

In the theory of eigen-decomposition, eq. 2.28 states that $\mathbf{S} = \mathbf{U}_{eigen}\mathbf{\Lambda}\mathbf{U}_{eigen}^{-1}$. Because $\mathbf{U}_{eigen}$ is orthonormal, $\mathbf{U}_{eigen}^{-1} = \mathbf{U}_{eigen}'$ (property 7 of inverses, Section 2.8) and the equation can be rewritten:

$$\mathbf{S} = \mathbf{U}_{eigen}\mathbf{\Lambda}\mathbf{U}_{eigen}'$$

Combining the latter equation with eq. 9.19 shows that

$$\mathbf{U}_{eigen}\mathbf{\Lambda}\mathbf{U}_{eigen}' = \mathbf{U}_{svd}\left(\frac{1}{n-1}\mathbf{W}'\mathbf{W}\right)\ \mathbf{U}_{svd}'$$

hence: $\mathbf{U}_{eigen} = \mathbf{U}_{svd}$ and $\mathbf{\Lambda} = \frac{1}{n-1}\mathbf{W}'\mathbf{W} = \frac{1}{n-1}[w_j^2]$ **(9.20)**

These correspondences can readily be verified, for the numerical example data, by singular value decomposition of the matrix of centred data $\mathbf{Y}_c$:

$$\begin{matrix}\mathbf{Y}_c & = & \mathbf{V} & \mathbf{W} & \mathbf{U}_{svd}'\end{matrix}$$

$$\begin{bmatrix}-3.2 & -1.6\\ -2.2 & 1.4\\ -0.2 & -2.6\\ 1.8 & 3.4\\ 3.8 & -0.6\end{bmatrix} = \begin{bmatrix}-0.5963 & 0.0000\\ -0.2236 & 5.0000\\ -0.2236 & -5.0000\\ 0.5217 & 5.0000\\ 0.5217 & -5.0000\end{bmatrix}\begin{bmatrix}6.0000 & 0\\ 0 & 4.4721\end{bmatrix}\begin{bmatrix}0.8944 & 0.4472\\ -0.4472 & 0.8944\end{bmatrix}$$

(As for eigenvalue decomposition, different SVD functions can revert the signs of some columns of **V** and **U**.) One can check that the squared singular values divided by $(n-1)$ are the eigenvalues, $\lambda_1 = 9$ and $\lambda_2 = 5$, and that $\mathbf{U}_{svd} = \mathbf{U}_{eigen}$ computed in Subsection 9.1.1.

Matrix **G**, which gives the object positions in the correlation biplot (scaling 2), is obtained from **V** as follows:

$$\mathbf{G} = \sqrt{n-1}\ \mathbf{V} \qquad \textbf{(9.21)}$$

Matrix **F**, which gives the object positions in the distance biplot (scaling 1), can be computed from **V** in two different ways:

$$\mathbf{F} = \mathbf{VW} \qquad \text{or} \qquad \mathbf{F} = \sqrt{n-1}\ \mathbf{V}\boldsymbol{\Lambda}^{1/2} \qquad \textbf{(9.22)}$$

When there are as many, or more variables than there are objects (i.e. $p \geq n$, for example in species-rich communities), eigenvalues and eigenvectors can still be computed using any of the three methods described above: Householder reduction, the TWWS algorithm, or singular value decomposition. The covariance matrix is positive semidefinite in such cases, so that null eigenvalues are produced (Table 2.2). When p is much larger than n and all eigenvalues and eigenvectors must be computed, important savings in computer time can be made by applying Householder reduction or singular value decomposition to the cross-product matrix [**YY**'], which is of size $(n \times n)$, instead of [**Y**'**Y**]* which is of size $(p \times p)$ and is thus much larger; **Y** is centred by columns. The eigenvalues of [**YY**'] are the same as the non-zero eigenvalues of [**Y**'**Y**]. Matrix **U** of the eigenvectors of [**Y**'**Y**] can be found from matrix **V** of the eigenvectors of [**YY**'] using the transformation $\mathbf{U} = \mathbf{Y}'\mathbf{V}\boldsymbol{\Lambda}^{-1/2}$. Matrix **F** of the principal components is found from the equation $\mathbf{F} = \mathbf{V}\boldsymbol{\Lambda}^{1/2}$.

Negative eigenvalues may occur in principal component analysis due to the handling of missing values. Pairwise deletion of missing data (Subsection 1.6.2), in particular, creates covariances computed with different numbers of degrees of freedom; this situation can make the covariance matrix indefinite (Table 2.2). A Householder algorithm should be used in such a case because negative eigenvalues come out of SVD as positive singular values (Section 2.11, Application 2).

10 — Metric ordination of community composition data

Different approaches are available to obtain metric ordinations of community composition (species) data (Fig. 9.8): the classical ordination approaches (PCA, this section, and CA, Section 9.2), the transformation-based PCA (tb-PCA), and the

* Matrix **S** differs from the cross-product matrix [**Y**'**Y**] by the division of the cross-products by $(n-1)$ in **S**. The eigenvalues of [**Y**'**Y**] are larger than those of **S** by this factor $(n-1)$, but the eigenvectors are identical.

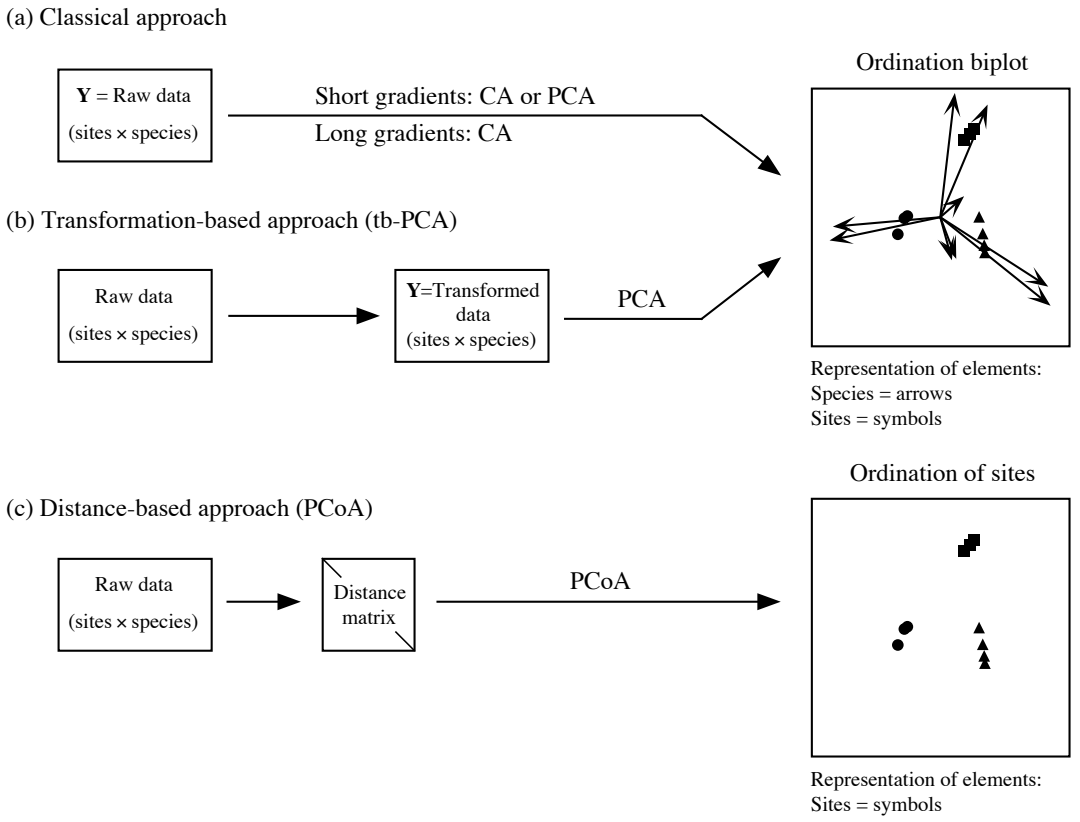


Figure 9.8 Different approaches are available for metric ordination of community composition data: (a) classical PCA and CA, (b) the transformation-based approach, and (c) the distance-based approach (PCoA). Metric ordination methods produce ordinations that fully preserve the distances among sites, as specified in Table 9.1. Modified from Legendre & Gallagher (2001).

distance-based method of principal coordinate analysis (PCoA, Section 9.3). These metric ordination methods produce ordinations that fully preserve the distances among sites. They are discussed here in turn. The distance preserved by each method is specified in Table 9.1. The non-metric method of nMDS (Section 9.4) is not mentioned in Fig. 9.8 because the ordinations that it produces distort the distances among sites.

In the classical approach (Fig. 9.8a), the species-environment relationship is analysed by PCA (this Section) or by CA (Section 9.2). In the early applications of PCA to community ecology, CA was considered preferable to PCA for species data tables sampled in highly diversified regions ("long gradients") because these tables contain many zeros. This is the case, for example, when sampling communities along

extensive spatial or temporal gradients, where the species composition may change greatly along the gradient. For groups of sites that were fairly homogeneous in species composition ("short gradients"), PCA was considered appropriate. A wider array of options is now available.

PCA can be made to preserve some distance that is appropriate for the study of composition data in highly diversified regions, e.g. along gradients, instead of the Euclidean distance D_1 (Fig. 9.8b). Composition data can be transformed using the transformations described in Section 7.7, leading to the transformation-based PCA, or tb-PCA, approach. PCA computed on data transformed using these equations will actually preserve the chord, profile, Hellinger, or chi-square distance, or the chi-square metric among sites, depending on the transformation used. Note that the corresponding distances (D_3 and D_{15} to D_{18} in Table 7.3) have the property of being Euclidean.

tb-PCA

One can also (Fig. 9.8c) compute directly one of the distance functions appropriate for community composition data (Table 7.4) and carry out a principal coordinate analysis (PCoA, Section 9.3) of the distance matrix to obtain an ordination. This is the distance-based approach. PCoA obtains metric ordinations from **D** matrices, whereas nonmetric multidimensional scaling (nMDS, Section 9.4) produces non-metric ordinations that distort the distances among sites. These methods should be used in analyses involving distance functions that cannot be obtained by a data transformation followed by PCA (tb-PCA approach, Fig. 9.8b). Among the distances developed specifically for species data (Table 7.4) are most of the coefficients designed for binary data, e.g. Jaccard ($\sqrt{1 - S_7}$) and Sørensen (D_{13} or $\sqrt{1 - S_8}$), as well as quantitative distance measures like the asymmetric Gower coefficient ($\sqrt{1 - S_{19}}$), the geodesic metric (D_4), Whittaker (D_9), Canberra (D_{10}), Clark (D_{11}), percentage difference (D_{14}), and mean character difference modified for species data D_{19}.

9.2 Correspondence analysis (CA)

Correspondence analysis (CA) was developed independently by several authors. It was first proposed for the analysis of contingency tables by Hirschfeld (1935), Fisher (1940), Benzécri (1969), and others. In a historical review of the subject, Nishisato (1980) traces its origin back to 1933. It was applied in ecology to the analysis of sites × species tables by Roux & Roux (1967), Hatheway (1971), Ibanez & Séguin (1972), Hill (1973b, 1974), Orlóci (1975), and others. Its use was generalized to other types of data tables by Benzécri and his collaborators (Escofier-Cordier, 1969; Benzécri and coll., 1973). Other important books on correspondence analysis are those of Nishisato (1980), Greenacre (1983, 2007), ter Braak (1988), and van Rijckevorsel & de Leeuw (1988). In the course of its history, the method was successively designated under the English names *contingency table analysis* (Fisher, 1940), RQ-*technique* (Hatheway, 1971), *reciprocal averaging* (Hill, 1973b), *correspondence analysis* (Hill, 1974), *reciprocal ordering* (Orlóci, 1975), *dual scaling* (Nishisato,

1980), and *homogeneity analysis* (Meulman, 1982), while it is known in French as *analyse factorielle des correspondances* (Cordier, 1965; Escofier-Cordier, 1969).

Contingency table

Correspondence analysis was first proposed for analysing two-way contingency tables (Section 6.2). In such tables, the states of a first descriptor (rows) are compared to the states of a second descriptor (columns). Data in each cell of the table are

Frequencies

frequencies, i.e. numbers of objects coded with a combination of states of the two descriptors. These frequencies are positive integers or zeros. The most common application of CA in ecology is the analysis of community composition (species presence-absence or abundance values) at sampling sites (Subsection 9.2.4). The rows and columns of the data table then correspond to sites and species, respectively. Such a table is analogous to a contingency table because the data are frequencies.

In general, correspondence analysis can be applied to any data table that is dimensionally homogeneous, meaning that the physical dimensions of all variables are the same (Chapter 3), and that does not contain negative values (i.e. only positive integers or zeros are allowed). The values have to be additive in rows and columns (additivity: see Subsection 1.4.2) to allow computation of row and column sums and transformation of the data table into matrix $\overline{\mathbf{Q}}$ (eq. 9.24). Frequency data have these characteristics. The χ^2 distance (D_{16}, eq. 7.55), which is a coefficient that excludes double-zeros, is used to quantify the relationships among rows and columns in CA (Table 9.1).

Correspondence analysis can also be conducted on contingency tables that compare two *groups* of descriptors. The method is then called *multiple correspondence*

Multiple corresp. analysis

analysis (MCA). For example, the rows of the table could be different species, each divided into a few abundance classes, and the columns, different descriptors of the physical environment with, for each, a number of columns equal to the number of its states. Each site (object) then contributes to several frequencies of the table, but this does not invalidate the results because of the transformations described in the next subsection. The analysis can be done using a standard CA function. Special programs and R functions also exist for MCA, which is not described further in this section. For a table of species × environmental variables, a better way of comparing species to environmental data is canonical correspondence analysis (CCA, Section 11.2), which does not require that the species and environmental data be recoded into a few classes.

Correspondence analysis is primarily an ordination method. As such, it is similar to principal component analysis; it preserves, in the space of the principal axes (i.e. after rotation), the Euclidean distance between *profiles of weighted conditional probabilities*. This is equivalent to preserving the χ^2 distance (D_{16}, eq. 7.55) between the rows or columns of the contingency table. The relationships between correspondence analysis and principal component analysis will be further described in the next subsections.

1 — Computation

This description of correspondence analysis will proceed in three steps. (1) The contingency (or community composition) table will be transformed into a table of contributions to the Pearson chi-square statistic after fitting a null model to the frequency data. (2) The transformed data table will be decomposed to obtain the eigenvalues and eigenvectors, as in PCA. (3) Further matrix operations will lead to the various tables needed for plotting useful diagrams. Besides its role as an ordination method, CA may be used for studying the proximities between the rows (or the columns) of a contingency table, as well as the correspondence between rows and columns as in Section 6.4.

Consider a contingency table with r rows and c columns, as in Section 6.2. Assume that the table is written in such a way that $r \geq c$; the table may be transposed to meet this condition, since the rows and columns of a contingency table play identical roles. The symbolism is as follows:

- Absolute frequencies are represented by f_{ij} and relative frequencies ("probabilities" or "proportions") by p_{ij}.

- p_{ij} is the frequency f_{ij} in cell ij divided by the sum f_{++} of the f_{ij}'s over the whole table. The table containing the relative frequencies p_{ij} is called **Q**; its size is (r×c).

- The weight attached to row i is $p_{i+} = f_{i+}/f_{++}$, where f_{i+} is the sum of the values in row i. Vector $[p_{i+}]$ is of size r = number of rows.

- Likewise, the weight attached to column j is $p_{+j} = f_{+j}/f_{++}$, where f_{+j} is the sum of values in column j. Vector $[p_{+j}]$ is of size c = number of columns.

The computation steps are as follows:

1. *Transform the data table*. — The Pearson chi-square statistic, χ^2_P (eq. 6.5), is a sum of squared χ_{ij} values, computed for every cell ij of the contingency table. Each χ_{ij} value is the standardized residual of a frequency f_{ij} after fitting a null model to the contingency table. The null model states that there is no relationship between the rows and columns of the table (eq. 6.4). Simple algebra shows that the component of χ_{ij} for each cell (eq. 6.26) is:

Component of chi-square

$$\chi_{ij} = \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}}} = \sqrt{f_{++}} \left[\frac{p_{ij} - p_{i+}p_{+j}}{\sqrt{p_{i+}p_{+j}}} \right] \qquad \textbf{(9.23)}$$

Correspondence analysis is based upon a matrix called $\bar{\mathbf{Q}}$ (r×c) in this book:

$$\bar{\mathbf{Q}} = [\bar{q}_{ij}] = \left[\frac{p_{ij} - p_{i+}p_{+j}}{\sqrt{p_{i+}p_{+j}}} \right] \qquad \textbf{(9.24)}$$

Values $\bar{q}_{ij}$, which are at the basis of correspondence analysis, only differ from the χ_{ij} values by the numerical constant $\sqrt{f_{++}}$: $\bar{q}_{ij} = \chi_{ij} / \sqrt{f_{++}}$. This difference causes all the eigenvalues to be smaller than or equal to 1, as shown below. Values $\bar{q}_{ij}$ can also be calculated directly from the f_{ij}'s:

$$\bar{q}_{ij} = \frac{f_{ij} f_{++} - f_{i+} f_{+j}}{f_{++} \sqrt{f_{i+} f_{+j}}} \qquad \textbf{(9.25)}$$

Total inertia

The sum of squares of all values in matrix $\overline{\mathbf{Q}}$, $\sum \bar{q}_{ij}^2$, measures the *total inertia* in $\overline{\mathbf{Q}}$. It is also equal to the sum of all eigenvalues to be extracted by eigenanalysis of $\overline{\mathbf{Q}}$.

SVD

2. *Decomposition of* $\overline{\mathbf{Q}}$. — Singular value decomposition (SVD, eq. 2.31) can be applied to matrix $\overline{\mathbf{Q}}$, with the following result (the symbolism is slightly modified compared to Section 2.11):

$$\overline{\mathbf{Q}}\,(r\times c) = \hat{\mathbf{U}}\,(r\times c)\;\mathbf{W}(\text{diagonal}, c\times c)\;\mathbf{U}'(c\times c) \qquad \textbf{(9.26)}$$

where both $\mathbf{U}$ and $\hat{\mathbf{U}}$ are orthonormal matrices (i.e. matrices containing column vectors that are normalized and orthogonal to one another; Section 4.4) and $\mathbf{W}$ is a diagonal matrix $\mathbf{D}(w_i)$. The diagonal values w_i in $\mathbf{W}$, which are all non-negative, are the singular values of $\overline{\mathbf{Q}}$.

Because $\overline{\mathbf{Q}} = \hat{\mathbf{U}}\,\mathbf{W}\mathbf{U}'$ (eq. 9.26), the multiplication $\overline{\mathbf{Q}}'\overline{\mathbf{Q}}$ gives the following result:

$$\overline{\mathbf{Q}}'\overline{\mathbf{Q}}\,(c\times c) = \mathbf{U}\mathbf{W}'\,(\hat{\mathbf{U}}'\hat{\mathbf{U}})\,\mathbf{W}\,\mathbf{U}' \qquad \textbf{(9.27)}$$

Since $\hat{\mathbf{U}}$ is orthonormal, $\hat{\mathbf{U}}'\hat{\mathbf{U}} = \hat{\mathbf{U}}\hat{\mathbf{U}}' = \mathbf{I}$, hence:

$$\overline{\mathbf{Q}}'\overline{\mathbf{Q}} = \mathbf{U}\,\mathbf{W}'\mathbf{W}\,\mathbf{U}' \qquad \textbf{(9.28)}$$

Equation 2.28 shows that the eigenvalues (forming diagonal matrix $\mathbf{\Lambda}$) and eigenvectors (matrix $\mathbf{U}$) of a square matrix $\mathbf{A}$ obey the relationship:

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{-1}$$

If the vectors in $\mathbf{U}$ are normalized, as they are here, $\mathbf{U}$ is an orthonormal matrix with the property $\mathbf{U}^{-1} = \mathbf{U}'$. As a consequence, eq. 2.28 may be rewritten as

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}' \qquad \textbf{(9.29)}$$

It follows that the diagonal matrix $[\mathbf{W}'\mathbf{W}]$, which contains squared singular values on its diagonal, is the diagonal matrix $\mathbf{\Lambda}(c\times c)$ of the eigenvalues of $\overline{\mathbf{Q}}'\overline{\mathbf{Q}}$. Similarly, the orthonormal matrix $\mathbf{U}$ of eqs. 9.27 and 9.28 is the same as matrix $\mathbf{U}$ of eq. 9.29; it is the matrix of eigenvectors of $\overline{\mathbf{Q}}'\overline{\mathbf{Q}}\,(c\times c)$, containing the loadings of the *columns* of the contingency table. A similar reasoning applied to matrix $\overline{\mathbf{Q}}\overline{\mathbf{Q}}'\,(r\times r)$ shows that the

orthonormal matrix $\hat{\mathbf{U}}$ produced by singular value decomposition is the matrix of eigenvectors of $\overline{\mathbf{Q}}\overline{\mathbf{Q}}'$, containing the loadings of the *rows* of the contingency table.

The relationship between eq. 9.26 and eigenvalue decomposition (eq. 2.22) is the same as in principal component analysis (Subsection 9.1.9). Prior to eigenvalue decomposition, a square matrix of sums of squares and cross products $\overline{\mathbf{Q}}'\overline{\mathbf{Q}}$ is computed. This is similar to using the matrix of sums of squares and cross products $\mathbf{Y}'\mathbf{Y}$ for eigenvalue decomposition in PCA; $\mathbf{Y}'\mathbf{Y}$ is the covariance matrix $\mathbf{S}$ multiplied by the constant $(n-1)$. In PCA, matrix $\mathbf{Y}$ was centred on the column means prior to computing $\mathbf{Y}'\mathbf{Y}$ whereas, in CA, matrix $\overline{\mathbf{Q}}$ is centred by the operation $(O_{ij} - E_{ij})$ (eqs. 9.23 and 9.24). In spite of this centring operation, the sums of the rows and columns of $\overline{\mathbf{Q}}$ are not equal to zero.

Eigen-analysis

Results identical to those of SVD would be obtained by applying eigenvalue decomposition (eqs. 2.22 and 9.1) either to the covariance matrix $\overline{\mathbf{Q}}'\overline{\mathbf{Q}}$, which would produce the matrices of eigenvalues $\mathbf{\Lambda}$ and eigenvectors $\mathbf{U}$, or to matrix $\overline{\mathbf{Q}}\overline{\mathbf{Q}}'$, which would provide the matrices of eigenvalues $\mathbf{\Lambda}$ and eigenvectors $\hat{\mathbf{U}}$. Actually, it is not necessary to repeat the eigenanalysis to obtain $\mathbf{U}$ and $\hat{\mathbf{U}}$, because:

$$\hat{\mathbf{U}}\,(r\times c) = \overline{\mathbf{Q}}\mathbf{U}\,\mathbf{\Lambda}^{-1/2} \qquad \textbf{(9.30)}$$

and

$$\mathbf{U}(c\times c) = \overline{\mathbf{Q}}'\hat{\mathbf{U}}\,\mathbf{\Lambda}^{-1/2} \qquad \textbf{(9.31)}$$

In the sequel, all matrices derived from $\mathbf{U}$ will be without a hat and all matrices derived from $\hat{\mathbf{U}}$ will bear a hat.

Singular value decomposition of matrix $\overline{\mathbf{Q}}$, or eigenvalue analysis of matrix $\overline{\mathbf{Q}}'\overline{\mathbf{Q}}$, always yields one null eigenvalue. This is due to the centring in eq. 9.24, where $(p_{i+}p_{+j})$ is subtracted from each value p_{ij}. The number of positive eigenvalues is $\min(r-1, c-1)$. Hence, when $r \geq c$, there are $(c-1)$ positive eigenvalues. The part of matrix $\mathbf{U}$ that is considered for interpretation is of size $c \times (c-1)$; likewise, the part of $\hat{\mathbf{U}}$ that is considered is of size $r \times (c-1)$.

The analysis, by either SVD or eigenvalue decomposition, is usually performed on matrix $\overline{\mathbf{Q}}$ with $r \geq c$, for convenience. The reason is that not all SVD programs can handle matrices with $r < c$ (function ***svd()*** of R does not present that problem). In addition, when using eigenanalysis, the computation is shorter when performed on the smallest of the two possible covariance matrices, both solutions leading to identical results. If one proceeds from a matrix such that $r < c$, the first $r-1$ eigenvalues are the same as in the analysis of the transposed matrix, the remaining eigenvalues being zero.

Consider now the non-centred matrix $\tilde{\mathbf{Q}}\,(r\times c)$ in which $(p_{i+}p_{+j})$ is not subtracted from each term p_{ij} in the numerator:

$$\tilde{\mathbf{Q}} = [\tilde{q}_{ij}] = \left[\frac{p_{ij}}{\sqrt{p_{i+}p_{+j}}}\right] = \left[\frac{f_{ij}}{\sqrt{f_{i+}f_{+j}}}\right] \qquad \textbf{(9.32)}$$

What happens if the analysis is based on matrix $\tilde{\mathbf{Q}}$ instead of $\overline{\mathbf{Q}}$ (eq. 9.24)? The only difference is that decomposition of $\tilde{\mathbf{Q}}$ produces one extra eigenvalue; all the other results are identical. This extra eigenvalue is easy to recognize because its value is 1 in correspondence analysis. This eigenvalue is meaningless because it only reflects the distance between the centre of mass of the data points in the ordination space and the origin of the system of axes. In other words, it reflects the lack of centring of the scatter of points on the origin (Hill, 1974); it explains none of the dispersion (Lebart & Fénelon, 1971). There are computer programs that do not make the centring; in that case, the first eigenvalue ($\lambda_1 = 1$) and eigenvector must be discarded. All programs that carry out the calculations on matrix $\overline{\mathbf{Q}}$ produce one eigenvalue less than $\min[r, c]$; if the data table $\mathbf{Q}$ is such that $r \geq c$, correspondence analysis yields $(c - 1)$ non-null and positive eigenvalues.

Alternatively, what happens if the analysis is based on the matrix of χ_{ij} values (eq. 9.23) instead of matrix $\overline{\mathbf{Q}}$? Since values $\chi_{ij} = \sqrt{f_{++}}\ \bar{q}_{ij}$, it follows that the total variance in matrix $[\chi_{ij}]$ is larger than that of matrix $\overline{\mathbf{Q}}$ by a factor $(\sqrt{f_{++}})^2 = f_{++}$; hence, all eigenvalues obtained by analysing matrix $[\chi_{ij}]$ are larger than those of $\overline{\mathbf{Q}}$ by a factor f_{++}. The normalized eigenvectors in matrices $\mathbf{U}$ and $\hat{\mathbf{U}}$ remain unaffected. When the analysis is carried out on matrix $\overline{\mathbf{Q}}$, all eigenvalues are smaller than or equal to 1, which is convenient.

Biplot

3. *Compute matrices for biplots*. — Matrices $\mathbf{U}$ and $\hat{\mathbf{U}}$ may be used to plot the positions of the row and column vectors in two separate scatter diagrams. For *biplots*, which are joint plots of the rows and column vectors, various scalings have been proposed. First, matrices $\mathbf{U}$ and $\hat{\mathbf{U}}$ can be weighted by the inverse of the square roots of the column and row weights, written out in diagonal matrices $\mathbf{D}(p_{+j})^{-1/2}$(size $c \times c$) and $\mathbf{D}(p_{i+})^{-1/2}$(size $r \times r$), respectively:

$$\mathbf{V}(c \times c) = \mathbf{D}(p_{+j})^{-1/2}\ \mathbf{U} \qquad \textbf{(9.33)}$$

$$\hat{\mathbf{V}}\ (r \times c) = \mathbf{D}(p_{i+})^{-1/2}\ \hat{\mathbf{U}} \qquad \textbf{(9.34)}$$

Discarding the null eigenvalue, the part of matrix $\mathbf{V}$ to consider for interpretation is of size $c \times (c - 1)$ and the part of matrix $\hat{\mathbf{V}}$ to consider is of size $r \times (c - 1)$.

Matrix $\mathbf{F}$, which gives the positions of the *rows* of the contingency table in the correspondence analysis space, is obtained from the transformed matrix of eigenvectors $\mathbf{V}$, which gives the positions of the *columns* in that space. This is done by applying the usual equation for component scores (eq. 9.4) to data matrix $\mathbf{Q}$, with division by the row weights:

$$\mathbf{F}(r \times c) = \hat{\mathbf{V}}\ \boldsymbol{\Lambda}^{1/2} \qquad \textbf{(9.35a)}$$

or

$$\mathbf{F}(r \times c) = \mathbf{D}(p_{i+})^{-1}\mathbf{Q}\mathbf{V} \qquad \textbf{(9.35b)}$$

In the same way, matrix $\hat{\mathbf{F}}$, which gives the positions of the *columns* of the contingency table in the correspondence analysis space, is obtained from the transformed matrix of eigenvectors $\hat{\mathbf{V}}$, which gives the positions of the *rows* in that space. The equation is the same as above, except that division here is by the column weights:

$$\hat{\mathbf{F}}\,(c\times c) = \mathbf{V}\boldsymbol{\Lambda}^{1/2} \qquad \textbf{(9.36a)}$$

or

$$\hat{\mathbf{F}}\,(c\times c) = \mathbf{D}(\mathrm{p}_{+j})^{-1}\mathbf{Q}'\hat{\mathbf{V}} \qquad \textbf{(9.36b)}$$

Discarding the null eigenvalue, the part of matrix **F** to consider for interpretation is of size $r \times (c-1)$ and the part of matrix $\hat{\mathbf{F}}$ to consider is of size $c \times (c-1)$. With this scaling, matrices **F** and **V** form a pair such that the rows (given by matrix **F**) are at the centroid (also called centre of mass, or "barycentre", from the Greek βαρυς, pronounced "barus", heavy) of the columns in matrix **V**. In the same way, matrices $\hat{\mathbf{F}}$ and $\hat{\mathbf{V}}$ form a pair such that the columns (given by matrix $\hat{\mathbf{F}}$) are at the centroids of the rows in matrix $\hat{\mathbf{V}}$. This property is illustrated in the numerical example below.

Biplots of the rows (e.g. sites) and columns (e.g. species) can be drawn using different combinations of the matrix scalings described above. Scaling types 1 and 2, described below, are the most commonly used by ecologists when analysing community composition data (ter Braak, 1990).

Scalings in CA

• Scaling type 1. — Draw a joint plot with the sites (matrix **F**) at the centroids of the species (matrix **V**). For sites × species data tables, this scaling is the most appropriate if one is primarily interested in representing the distance relationships among the sites because, in matrix **F**, the distances among sites are projections of their χ^2 distances (D_{16}) (ter Braak, 1987c; see Numerical example, Subsection 9.2.2).

• Scaling type 2. — Draw a joint plot with the species (matrix $\hat{\mathbf{F}}$) at the centroids of the sites (matrix $\hat{\mathbf{V}}$). For sites × species data tables, this scaling is the most appropriate if one is primarily interested in representing the distance relationships among the species because, in matrix $\hat{\mathbf{F}}$, the distances among species are projections of their χ^2 distances (see Numerical example, Subsection 9.2.2).

• Scaling type 3. — This is a compromise between scalings 1 and 2. This scaling, called "symmetric" in program CANOCO, does not preserve the chi-square distances among the species or among the site scores. It is obtained by drawing together matrices $\hat{\mathbf{V}}\,\boldsymbol{\Lambda}^{1/4}$ (or $\mathbf{F}\boldsymbol{\Lambda}^{-1/4}$) for sites and $\mathbf{V}\boldsymbol{\Lambda}^{1/4}$ (or $\hat{\mathbf{F}}\,\boldsymbol{\Lambda}^{-1/4}$) for species.

• Scaling type 4. — This scaling is useful in the correspondence analysis of a contingency table crossing two qualitative descriptors or two factors. Draw a joint plot using **F**, which preserves the chi-square distances among the rows, and $\hat{\mathbf{F}}$ which preserves the chi-square distances among the columns of the contingency table. This hybrid scaling correctly represents the chi-square distance relationships among the states of the two qualitative descriptors. In this scaling, the relative positions of the row

and column symbols *along each axis of the plot* are the same as in scaling 3. The range of axis k in scaling 3 multiplied by $\lambda_k^{1/4}$ gives the range of that axis in scaling type 4. The axes in scaling 4 are thus compressed compared to the corresponding axes in scaling 3 because the eigenvalues are always smaller than 1 in CA. The compression is not isotropic, however, because the eigenvalues differ among axes.

Other possible, but less often used scaling methods are discussed by ter Braak (1987c, 1990).

4. *Cumulative fit tables*. — Tables of cumulative fit for columns and rows can be computed in CA, as it was the case in PCA (Subsection 9.1.3.4). The sum of squares of the values in row j of matrix $\hat{\mathbf{F}}$ gives the total variance of column (descriptor) j in the multidimensional ordination space. The relative cumulative fit of descriptor j is found by computing the sum of squared values for axis 1, axes 1 and 2, axes 1, 2 and 3, and so on, and dividing it by the total variance of j. This statistic represents the fit of descriptor j in 1, 2, or more dimensions; it can be interpreted like a coefficient of multiple determination (R^2).

The sum of squared values of row i of matrix **F** gives the squared length of the vector representing object i in the multidimensional ordination space. Use matrix **F** to compute the squared length of each object vector i in 1, 2, 3 … CA dimensions and divide these lengths by the total square length of object vector i. See example in the next subsection. Squared residual lengths can be computed by subtracting from the length of i the sum of squared values for axis 1, axes 1 and 2, axes 1, 2 and 3, etc.

2 — Numerical example

The following numerical example illustrates the calculations involved in correspondence analysis. This example assumes that three species have been observed in three lakes (Table 9.6)*. The justification for analysing community composition data by CA is provided in Subsection 9.2.4. The data table is of small size (3 × 3) to allow readers to follow or repeat the calculations easily.

Matrix **Q** contains the proportions p_{ij}, from which the marginal distributions of the rows and columns, p_{i+} and p_{+j}, are computed. The row and column identifiers are as in Table 9.6:

$$\mathbf{Q} = [p_{ij}] = \begin{array}{c} \\ L1 \\ L2 \\ L3 \end{array} \begin{array}{c} \begin{array}{ccc} Sp1 & Sp2 & Sp3 \end{array} \\ \begin{bmatrix} 0.10 & 0.10 & 0.20 \\ 0.10 & 0.15 & 0.10 \\ 0.15 & 0.05 & 0.05 \end{bmatrix} \end{array} \quad \begin{array}{c} [p_{i+}] \\ \begin{bmatrix} 0.40 \\ 0.35 \\ 0.25 \end{bmatrix} \end{array}$$

$$[p_{+j}] = \begin{bmatrix} 0.35 & 0.30 & 0.35 \end{bmatrix}$$

* Table 9.6 could also represent a contingency table crossing the states of two qualitative descriptors, to illustrate CA of a contingency table. A biplot of the results would use scaling type 4 (Subsection 9.2.1.3).

Table 9.6 Numerical example: a site-by-species data table.

	Species 1 (Sp1)	Species 2 (Sp2)	Species 3 (Sp3)	**Row sums**
Lake 1 (L1)	10	10	20	40
Lake 2 (L2)	10	15	10	35
Lake 3 (L3)	15	5	5	25
Column sums	35	30	35	100

Matrix $\overline{\mathbf{Q}}$ is computed with eq. 9.24:

$$\overline{\mathbf{Q}} = [\bar{q}_{ij}] = \left[\frac{p_{ij} - p_{i+}p_{+j}}{\sqrt{p_{i+}p_{+j}}}\right] = \begin{matrix} \\ L1 \\ L2 \\ L3 \end{matrix} \begin{matrix} \begin{matrix} Sp1 & Sp2 & Sp3 \end{matrix} \\ \begin{bmatrix} -0.10690 & -0.05774 & 0.16036 \\ -0.06429 & 0.13887 & -0.06429 \\ 0.21129 & -0.09129 & -0.12677 \end{bmatrix} \end{matrix}$$

and matrix $\tilde{\mathbf{Q}}$ with eq. 9.32:

$$\tilde{\mathbf{Q}} = [\tilde{q}_{ij}] = \left[\frac{p_{ij}}{\sqrt{p_{i+}p_{+j}}}\right] = \begin{matrix} \\ L1 \\ L2 \\ L3 \end{matrix} \begin{matrix} \begin{matrix} Sp1 & Sp2 & Sp3 \end{matrix} \\ \begin{bmatrix} 0.26726 & 0.28868 & 0.53452 \\ 0.28571 & 0.46291 & 0.28571 \\ 0.50709 & 0.18257 & 0.16903 \end{bmatrix} \end{matrix}$$

The eigenvalues of $\overline{\mathbf{Q}}'\overline{\mathbf{Q}}$ are $\lambda_1 = 0.09613$ (70.1%), $\lambda_2 = 0.04094$ (29.9%), and $\lambda_3 = 0$ (because of the centring). The first two eigenvalues are also eigenvalues of $\tilde{\mathbf{Q}}'\tilde{\mathbf{Q}}$, its third eigenvalue being 1 because $\tilde{\mathbf{Q}}$ is not centred (eq. 9.32). The normalized eigenvectors of $\overline{\mathbf{Q}}'\overline{\mathbf{Q}}$, corresponding to λ_1 and λ_2, are (in columns):

$$\mathbf{U} = \begin{matrix} \\ Sp1 \\ Sp2 \\ Sp3 \end{matrix} \begin{matrix} \begin{matrix} (\lambda_1) & (\lambda_2) \end{matrix} \\ \begin{bmatrix} 0.78016 & -0.20336 \\ -0.20383 & 0.81145 \\ -0.59144 & -0.54790 \end{bmatrix} \end{matrix}$$

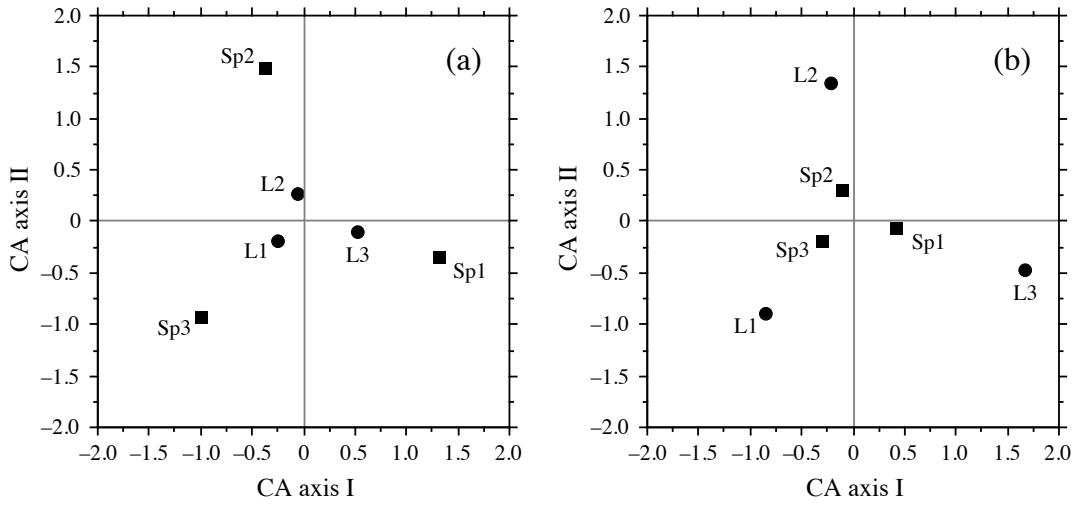


Figure 9.9 Correspondence analysis biplots. (a) Scaling type 1: the rows of the data table (lakes L1 to L3 represented by circles, matrix **F**) are at the centroids (barycentres) of the columns (species Sp1 to Sp3 represented by squares, matrix **V**). (b) Scaling type 2: the species (squares, matrix $\hat{\mathbf{F}}$) are at the centroids (barycentres) of the lakes (circles, matrix $\hat{\mathbf{V}}$).

The normalized eigenvectors of $\overline{\mathbf{Q}}\overline{\mathbf{Q}}'$ are (in columns):

$$\hat{\mathbf{U}} = \begin{array}{c} \\ \text{L1} \\ \text{L2} \\ \text{L3} \end{array} \begin{array}{c} \begin{array}{cc} (\lambda_1) & (\lambda_2) \end{array} \\ \begin{bmatrix} -0.53693 & -0.55831 \\ -0.13043 & 0.79561 \\ 0.83349 & -0.23516 \end{bmatrix} \end{array}$$

The third eigenvector is of no use and is therefore not given. Most programs do not compute it.

In scaling type 1 (Fig. 9.9a), the rows of the data matrix (L1, L2 and L3 in the example), whose coordinates will be stored in matrix **F**, are to be plotted at the centroids of the columns (Sp1, Sp2 and Sp3 in the example). The scaling for the columns is obtained using eq. 9.33:

$$\mathbf{V} = \mathbf{D}(p_{+j})^{-1/2}\,\mathbf{U} = \begin{array}{c} \\ \text{Sp1} \\ \text{Sp2} \\ \text{Sp3} \end{array} \begin{array}{c} \begin{array}{cc} (\lambda_1) & (\lambda_2) \end{array} \\ \begin{bmatrix} 1.31871 & -0.34374 \\ -0.37215 & 1.48150 \\ -0.99972 & -0.92612 \end{bmatrix} \end{array}$$

To put the rows (matrix **F**) at the centroids of the columns (matrix **V**), the position of each row along an ordination axis is computed as the mean of the column positions, weighted by the relative frequencies of the observations in the various columns of that row. Consider the first row of the data table (Table 9.6), for example. The relative frequencies of the three columns in

that row are 0.25, 0.25, and 0.50. Multiplying matrix **V** by that vector provides the coordinates of the first row in the ordination diagram:

$$[0.25 \;\; 0.25 \;\; 0.50] \begin{bmatrix} 1.31871 & -0.34374 \\ -0.37215 & 1.48150 \\ -0.99972 & -0.92612 \end{bmatrix} = [-0.26322 \;\; -0.17862]$$

These coordinates put the first row at the centroid of the columns in Fig. 9.9a; they are stored in the first row of matrix **F**. The row-conditional probabilities for the whole data table are found using the matrix operation $\mathbf{D}(p_{i+})^{-1}\mathbf{Q}$, so that matrix **F** is computed using eq. 9.35b:

$$\mathbf{F} = \mathbf{D}(p_{i+})^{-1}\mathbf{QV} = \begin{array}{c} \\ \text{L1} \\ \text{L2} \\ \text{L3} \end{array} \begin{array}{c} \begin{array}{cc} (\lambda_1) & (\lambda_2) \end{array} \\ \begin{bmatrix} -0.26322 & -0.17862 \\ -0.06835 & 0.27211 \\ 0.51685 & -0.09517 \end{bmatrix} \end{array}$$

Using the formulae for the Euclidean (D_1, eq. 7.32) and χ^2 (D_{16}, eq. 7.55) distances, one can verify that the Euclidean distances among the rows of matrix **F** are equal to the χ^2 distances among the rows of the original data table (Table 9.6):

$$\mathbf{D} = \begin{array}{c} \\ \text{L1} \\ \text{L2} \\ \text{L3} \end{array} \begin{array}{c} \begin{array}{ccc} \text{L1} & \text{L2} & \text{L3} \end{array} \\ \begin{bmatrix} 0 & & \\ 0.49105 & 0 & \\ 0.78452 & 0.69091 & 0 \end{bmatrix} \end{array}$$

Matrix **F** thus provides a proper ordination of the rows of the original data matrix (temperatures in the numerical example).

In scaling type 2 (Fig. 9.9b), the columns, whose coordinates will be stored in matrix $\hat{\mathbf{F}}$, are to be plotted at the centroids of the rows (matrix $\hat{\mathbf{V}}$). The scaling for matrix $\hat{\mathbf{V}}$ is obtained using eq. 9.34:

$$\hat{\mathbf{V}} = \mathbf{D}(p_{i+})^{-1/2}\,\hat{\mathbf{U}} = \begin{array}{c} \\ \text{L1} \\ \text{L2} \\ \text{L3} \end{array} \begin{array}{c} \begin{array}{cc} (\lambda_1) & (\lambda_2) \end{array} \\ \begin{bmatrix} -0.84896 & -0.88276 \\ -0.22046 & 1.34482 \\ 1.66697 & -0.47032 \end{bmatrix} \end{array}$$

To put the columns (matrix $\hat{\mathbf{F}}$) at the centroids of the rows (matrix $\hat{\mathbf{V}}$), the position of each column along an ordination axis is computed as the mean of the row positions, weighted by the relative frequencies of the observations in the various rows of that column. Consider the first column of the data table (Table 9.6), for example. The relative frequencies of the three rows in that column are (10/35 = 0.28571), (10/35 = 0.28571) and (15/35 = 0.42857). Multiplying matrix $\hat{\mathbf{V}}$ by that vector provides the coordinates of the first column in the ordination diagram:

$$[0.28571 \;\; 0.28571 \;\; 0.42857] \begin{bmatrix} -0.84896 & -0.88276 \\ -0.22046 & 1.34482 \\ 1.66697 & -0.47032 \end{bmatrix} = [0.40887 \;\; -0.06955]$$

These coordinates put the first column at the centroid of the rows in Fig. 9.9a; they are stored in the first row of matrix $\hat{\mathbf{F}}$. The column-conditional probabilities for the whole data table are found through matrix operation $\mathbf{D}(p_{+j})^{-1}\mathbf{Q}'$, so that matrix $\hat{\mathbf{F}}$ is computed using eq. 9.36a or 9.36b:

$$\hat{\mathbf{F}} = \mathbf{V}\boldsymbol{\Lambda}^{1/2} = \mathbf{D}(p_{+j})^{-1}\mathbf{Q}'\hat{\mathbf{V}} = \begin{array}{c} \\ \text{Sp1} \\ \text{Sp2} \\ \text{Sp3} \end{array} \begin{array}{c} \begin{array}{cc} (\lambda_1) & (\lambda_2) \end{array} \\ \begin{bmatrix} 0.40887 & -0.06955 \\ -0.11539 & 0.29977 \\ -0.30997 & -0.18739 \end{bmatrix} \end{array}$$

Using the formulae for the Euclidean (D_1, eq. 7.32) and χ^2 (D_{16}, eq. 7.55) distances, one can verify that the Euclidean distances among the rows of matrix $\hat{\mathbf{F}}$ are equal to the χ^2 distances among the columns of the original data table (Table 9.6):

$$\mathbf{D} = \begin{array}{c} \\ \text{Sp1} \\ \text{Sp2} \\ \text{Sp3} \end{array} \begin{array}{c} \begin{array}{ccc} \text{Sp1} & \text{Sp2} & \text{Sp3} \end{array} \\ \begin{bmatrix} 0 & & \\ 0.64128 & 0 & \\ 0.72843 & 0.52458 & 0 \end{bmatrix} \end{array}$$

Matrix $\hat{\mathbf{F}}$ thus provides a proper ordination of the columns of the original data matrix (species abundance classes in the numerical example).

For the numerical example, the table of *Cumulative fit per species* (3 species) is of size (3 × 2) because the CA solution has two dimensions (i.e. two positive eigenvalues) only:

	Cumul. axis 1	**Cumul. axis 2**
Sp1	0.9719	1.0000
Sp2	0.1290	1.0000
Sp3	0.7323	1.0000

In the column that corresponds to the last eigenvalue, the values are always 1 in CA. The table of *Cumulative fit of the objects* (3 lakes) is also of size (3 × 2) in this numerical example:

	Cumul. axis 1	**Cumul. axis 2**
L1	0.6847	1.0000
L2	0.0594	1.0000
L3	0.9672	1.0000

The two tables indicate that species 2 and lake 2 are poorly fitted along axis 1, as can be observed in Fig. 9.9, and that all species and lakes are perfectly represented in 2 dimensions.

3 — Interpretation

The relationship between matrices $\mathbf{V}$ and $\hat{\mathbf{V}}$, which provide the ordinations of the columns and rows of the species data (or contingency) table, respectively, is found by combining eqs. 9.30, 9.33 and 9.34 in the following expression:

$$\hat{\mathbf{V}}\,\mathbf{\Lambda}^{1/2} = \mathbf{D}(p_{i+})^{-1/2}\,\overline{\mathbf{Q}}\,\mathbf{D}(p_{+j})^{1/2}\mathbf{V} \qquad \textbf{(9.37)}$$

This equation means that the ordination of the rows (matrix $\hat{\mathbf{V}}$) is related to the ordination of the columns (matrix $\mathbf{V}$), along principal axis h, by the value $\sqrt{\lambda_h}$ which is a measure of the "correlation" between these two ordinations. The value $(1 - \lambda_h)$ actually measures the difficulty of ordering, along principal axis h, the rows of the contingency table from an ordination of the columns, or the converse (Orlóci, 1978). The highest eigenvalue (0.096 in the above numerical example), or its square root ($\sqrt{\lambda_1} = 0.31$), is thus a measure of dependence between two unordered descriptors, to be added to the measures described in Chapter 6. Williams (1952) discusses different methods for testing the significance of $R^2 = \lambda_h$.

Joint plots (e.g. Fig. 9.9) can be used to draw conclusions about the ecological relationships displayed by the data.

• With scaling type 1, (a) the distances among rows (or sites in the case of a species × sites data table) in reduced space approximate their χ^2 distances, and (b) the rows (sites) are at the centroids of the columns (species). Positions of the centroids are calculated using weights equal to the relative frequencies of the columns (species); columns (species) that are absent from a row (site) have null weights and do not contribute to the position of that row (site). Thus, the ordination of rows (sites) is meaningful. In addition, any row (site) found near the point representing a column (species) is likely to have a high contribution of that column (species); for binary (or species presence-absence) data, the row (site) is more likely to possess the state of that column (or contain that species).

• With scaling type 2, it is the distances among columns (species) in reduced space that approximate their χ^2 distances, whereas columns (species) are at the centroids of the rows (sites). Consequently, (a) the ordination of columns (species) is meaningful, and (b) any column (species) that lies close to the point representing a row (site) is more likely to be found in the state of that row (site), or with higher frequency (abundance) than in rows (sites) that are further away in the joint plot.

For species presence-absence or abundance data, insofar as a species has a unimodal (i.e. bell-shaped) response curve along the axes of ecological variation corresponding to the ordination axes, the optimum for that species should be close to the point representing it in the ordination diagram and its frequency of occurrence or abundance should decrease with the distance from that point. Species that are absent at most sites often appear at the edge of the scatter plot, near the point representing a site where they happen to be present — by chance, or because they are favoured by some

rare condition occurring at that site. Such species have little influence on the analysis because their numerical contributions are small (column sums in Table 9.6). Finally, species that lie near the centre of the ordination diagram may have their optimum in that area of the plot, or have two or several optima (bi- or multi-modal species), or else be unrelated to the pair of ordination axes under consideration. Species of the latter group may express themselves along some other axis or axes; close examination of the raw data table may be required in that case. It is the species found away from the centre of the diagram, but not near the edges, that are the most likely to display clear relationships with the ordination axes (ter Braak, 1987c).

4 — Site × species data tables

Correspondence analysis has been applied to data tables other than contingency tables. Justification is provided by Benzécri and coll. (1973). Notice, however, that the elements of a table to be analysed by correspondence analysis must be *dimensionally homogeneous* (i.e. same physical units, so that they can be added), *non-negative* (≥ 0, so that they can be transformed into probabilities or proportions), and additive so that the sums of rows and columns, f_{i+} and f_{+j}, make sense (additivity: see Subsection 1.4.2). Several types of data possess these characteristics, such as (bio)mass values, concentrations, financial data (in \$, €, ¥, etc.), and species abundances.

Other types of data may be recoded to make the descriptors dimensionally homogeneous and positive; the most widely used data transformations are discussed in Section 1.5. For descriptors with different physical units, the data may, for example, be standardized (which makes them dimensionless; eq. 1.12) and made positive by translation, i.e. by subtracting the most negative value; or they may be divided by the maximum or by the range of values (eqs. 1.10 and 1.11). Data may also be recoded into ordered classes. Regardless of the method, recoding is then a critical step of correspondence analysis. Consult Benzécri and coll. (1973) on this matter.

Several authors, mentioned at the beginning of this section, have applied correspondence analysis to the analysis of site × species matrices containing species presence/absence or abundance data. This generalization of the method is based on the following *sampling model*. If sampling had been designed in such a way as to collect individual organisms (which is usually not the case, the sampled elements being, most often, sampling sites), each organism could be described by two descriptors: the site where it was collected and the taxon to which it belongs. These two descriptors may be recorded in an *inflated data matrix*, which has as many rows as there are individual organism, and two columns identifying the site and the taxon of the individual (qualitative descriptors). The familiar site × species data table is the contingency table resulting from crossing the two descriptors of the inflated data matrix, i.e. the sites and taxa. That table could be analysed using any of the methods applicable to contingency tables. Most methods involving tests of statistical significance cannot be used, however, because the hypothesis of independence of the individual organisms, following the sampling model described above, is not met by species presence-absence

Inflated data matrix

or abundance data collected at sampling sites. An inflated data matrix will be used again in the description of canonical correspondence analysis, Subsection 11.2.1.

Niche theory tells us that species have ecological preferences, meaning that they are found at sites where they encounter favourable conditions. This statement is rooted in the idea that species have unimodal distributions along environmental variables (Fig. 9.10), more individuals being found near some environmental value which is "optimal" for the given species. This has been formalised by Hutchinson (1957) in his *fundamental niche* model. Furthermore, Gause's (1935) competitive exclusion principle suggests that, in their micro-evolution, species should have developed non-overlapping niches. These two principles indicate together that species should be roughly equally spaced in the n-dimensional space of resources. This model has been used by ter Braak (1985) to justify the use of correspondence analysis on presence-absence or abundance data tables; he showed that the χ^2 distance preserved through correspondence analysis (Table 9.1) is an appropriate model for species with unimodal distributions along environmental gradients.

Niche

Let us follow the path travelled by Hill (1973b), who rediscovered correspondence analysis while exploring the analysis of vegetation variation along environmental gradients; he called his method "reciprocal averaging" before realizing that this was correspondence analysis (Hill, 1974). Hill started from the simpler method of *gradient analysis*, proposed by Whittaker (1960, 1967) to analyse site × species data tables. Gradient analysis uses a matrix **Y** (site × species) and an initial vector **v** of values v_j which are ascribed to the various species j as indicators of the physical *gradient* to be evidenced. For example, a score (scale from 1 to 10) could be given to the each species for its preference with respect to soil moisture. These coefficients are used to calculate the positions of the sites along the gradient. The score $\hat{v}_i$ of a site i is calculated as the average score of the species ($j = 1 \ldots p$) present at that site, using the formula:

Reciprocal averaging

$$\hat{v}_i = \frac{\sum_{j=1}^{p} y_{ij} v_j}{y_{i+}} \qquad \textbf{(9.38)}$$

where y_{ij} is the abundance of species j at site i and y_{i+} is the sum of the organisms at this site (i.e. the sum of values in row i of matrix **Y**).

Gradient analysis produces a vector $\hat{\mathbf{v}}$ of the positions of the sites along the gradient under study. Hill (1973b, 1974) suggested to continue the analysis, using now vector $\hat{\mathbf{v}}$ of the ordination of sites to compute a new ordination (**v**) of the species:

$$v_j = \frac{\sum_{i=1}^{n} y_{ij} \hat{v}_i}{y_{+j}} \qquad \textbf{(9.39)}$$

in which y_{+j} is the sum of values in column j of matrix **Y**. Alternating between **v** and $\hat{\mathbf{v}}$ (scaling the vectors at each step as shown in step 6 of Table 9.8) defines an iterative procedure that Hill (1973b) called "reciprocal averaging". This procedure converges towards a unique unidimensional ordination of the species and sites, which is independent of the values initially given to the v_j's; different initial guesses as to the values v_j may however change the number of steps required to reach convergence. Being aware of the work of Clint & Jennings (1970), Hill realized that he had discovered an eigenvector method for gradient analysis, hence the title of his 1973b paper. It so happens that Hill's method produces the barycentred vectors **v** and $\hat{\mathbf{v}}$ for species and sites, which correspond to the first eigenvector of a correspondence analysis. Hill (1973b) showed how to calculate the eigenvalue (λ) corresponding to these ordinations and how to find the other eigenvalues and eigenvectors. Hill thus created a simple algorithm, described in Subsection 9.2.7, for correspondence analysis.

When interpreting the results of correspondence analysis, one should keep in mind that the simultaneous ordination of species and sites aims at determining how useful the ordination of species is, as a whole, for predicting the ordination of the sites. In other words, it seeks the predictive value of one ordination with respect to the other. Subsection 9.2.3 has shown that, for any given dimension h, $(1 - \lambda_h)$ measures the difficulty of ordering, along principal axis h, the row states of the contingency table from an ordination of the column states, or the converse. The interpretation of the relationship between the two ordinations must be done with reference to this statistic.

When it is used as an ordination method, correspondence analysis provides an ordination of the sites which is somewhat similar to that resulting from a principal component analysis of the correlation matrix among species (standardized data). This is to be expected since the first step in the calculation actually consists in weighting each datum by the sums (or the relative frequencies) of the corresponding row and column (eq. 9.24 and 9.25), which eliminates the effects due to the large variances that certain rows or columns may have. In the case of steep gradients (i.e. many zeros in the data matrix), correspondence analysis should produce a better ordination than PCA (Hill, 1973b). This was also shown by Gauch *et al*. (1977) using simulated and observational floristic data. This result logically follows from the fact that the χ^2 distance (D_{16}) is a coefficient that excludes double-zeros from the estimation of resemblance. This is not the case with the Euclidean distance (eq. 7.32), which is the distance preserved in principal component analysis. For this reason, correspondence analysis is one of the methods recommended in Fig. 9.8 for reduced-space ordination of species abundances when the data contain a large number of null values; this situation is encountered when sampling environmental gradients that are long enough for species to replace one another. Data tables with many zeros may contain rare species. Box 9.2 describes a procedure for handling rare species in CA.

As mentioned in Section 8.9, for the clustering of species into associations, correspondence analysis does not seem to escape the problems encountered with principal component analysis (Reyssac & Roux, 1972; Ibanez & Séguin, 1972; Binet *et al*., 1972). The most serious problem is that the species, which are multidimensional

Rare species in CA — Box 9.2

In Chapter 7, it was shown that rare species contribute heavily to the chi-square distance (D_{16}, eq. 7.55), which is the distance preserved in correspondence analysis. The present discussion focuses on the species with *small occurrence values*; they only occur in a small fraction of the study sites. These species generate a large number of zeros in the data matrix. Because zeros have high leverage, they contribute heavily to the total inertia of matrix $\overline{\mathbf{Q}}$ (eq. 9.24). These species contribute very little to the first few CA ordination axes, but they are highly conspicuous in biplots because they are found at the periphery of the graph. Should we keep rare species in CA? If not, which ones should be eliminated?

On the one hand, ecologists who see what they are collecting (e.g. vegetation) may consider rare species as potential indicators of special environmental conditions, but it is not the role of CA to display these conditions. The primary purpose of CA is to display the main axes of variation of the data, not to deal with exceptions. On the other hand, ecologists who sample blindly often consider the occurrence of rare species a chance event which should not be heavily weighted in the analysis. In the case of mobile animals, the presence of an animal at a site is no indication that the site provides favourable conditions for that species.

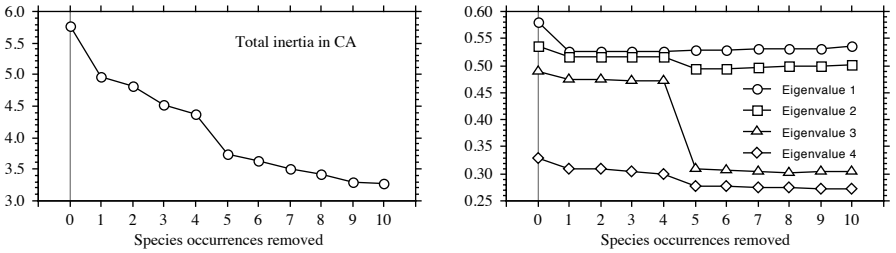
Empirical methods for down-weighting rare species have been proposed and are available in some computer programs, but these methods lack strong ecological foundations. It is better to simply eliminate the rarest species from CA. The following stepwise method has been developed by Daniel Borcard (personal communication):

- For convenience, order the species in the data table in increasing or decreasing occurrences. That will facilitate the stepwise elimination of the species with small occurrence values.
- Carry out a first CA. Note the total inertia as well as the first few eigenvalues (e.g. 4).
- Repeat that step after removing the species with occurrence 1; the species with occurrence 1 and 2; the species with occurrence 1 to 3; and so on. After each analysis, note the total inertia as well as the first few eigenvalues.
- Plot these results. A jump should be observed in total inertia and in some of the eigenvalues. The jump indicates that one has gone too far in removing rare species. [*Continued next page*.]

descriptor-axes, are projected in a low-dimensional space by both PCA and CA. This explains the tendency for the species to form a more or less uniformly dense scatter centred on the origin except in simple situations. It may nevertheless be interesting to superimpose a clustering of species, determined using the methods of Section 8.9, on a reduced-space ordination obtained by correspondence analysis.

Box 9.2 (*continued*)

The following example concerns fish biomass data (47 underwater transects, 156 fish species) collected by researchers Pierre Labrosse and Eric Clua (*Secretariat of the Pacific Community*) near the village of Manuka in the Tonga Islands, under the *DemEcoFish* project funded by the MacArthur Foundation (data used here with permission of the authors).



The curves show that the 61 species with occurrences 1 to 4 can be removed from the analysis with little effects on the first four eigenvalues (right-hand graph). These species generate 24% of the inertia in matrix $\overline{\mathbf{Q}}$ (left-hand graph) subjected to eigenvalue decomposition in CA.

For comparison, the same data were submitted to PCA after Hellinger transformation (Section 7.7). The total variance decreased by only 4.6% after removing the 61 species with occurrences 1 to 4 and the first four eigenvalues were not affected at all by the removal of rare species. The Hellinger transformation was recalculated after each step of species removal.

When sites (objects) and species (descriptors) are plotted together, the joint plot must be interpreted with care. The practice that consists in only associating species with neighbouring species in the plot often gives good results, although it may overlook indications of avoidance of sites by certain species. An interesting complement to correspondence analysis is the direct analysis of the site × species table by the Freeman-Tukey deviates and standardized residuals methods described in Section 6.4. These methods are better at evidencing all the correspondences between sites and species (attraction and avoidance). Applying contingency table analysis to sites × species tables is justified by the same logic that allows correspondence analysis to be applied to such data matrices.

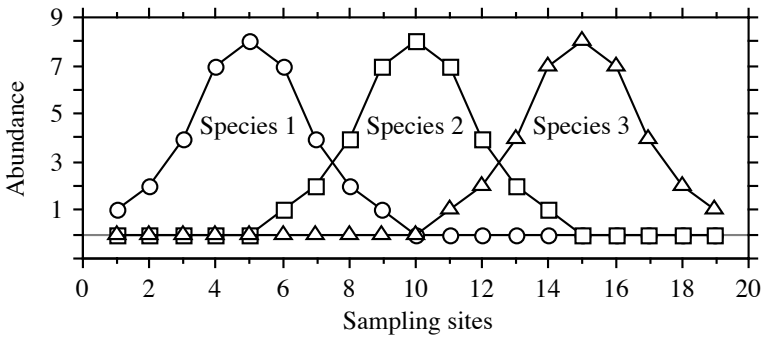


Figure 9.10 Distributions of three species at 19 sampling sites along a hypothetical environmental gradient These artificial data are found in Table 9.7.

5 — Arch effect and detrended correspondence analysis

Environmental gradient

Environmental or temporal gradients often support spatial or temporal succession of species. Since the species that are controlled by environmental factors (*versus* population dynamics, historical events, etc.) generally have unimodal distributions along gradients, the effect of gradients on the distance relationships among sites, calculated on species presence-absence or abundance data, is necessarily nonlinear.

Numerical example 1. A data set was created (Fig. 9.10; Table 9.7) to represent the abundances of three hypothetical species at 19 sites over an environmental gradient along which the species were assumed to have unimodal distributions (Whittaker, 1967).

The three species in Fig. 9.10 have unimodal distributions; each one shows a well-defined mode along the gradient represented by sites 1 to 19. Ordination methods aim at rendering this non-linear phenomenon in a Euclidean space, in particular in two-dimensional plots. In such plots, non-linearities end up being represented by curves,

Table 9.7 Artificial data illustrated in Fig. 9.10.

Sampling sites	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Species 1	1	2	4	7	8	7	4	2	1	0	0	0	0	0	0	0	0	0	0
Species 2	0	0	0	0	0	1	2	4	7	8	7	4	2	1	0	0	0	0	0
Species 3	0	0	0	0	0	0	0	0	0	0	1	2	4	7	8	7	4	2	1

called *arches* or *horseshoes*, described in the next paragraph. While most ecologists are content with interpreting the ordination plots for the information they display about distances among sites, some feel that they should try to reconstruct the original gradient underlying the observed data. Hence their concern with *detrending*, which is an operation carried out on the ordination axes of correspondence analysis. In that operation, the arch is unbent to display the gradient as a linear arrangement of the sites.

Detrending

Euclidean distances calculated on the species data of Fig. 9.10, between site 1 and sites 2, 3, etc., do not increase monotonically from one end of the gradient to the other. These distances form the first row of the Euclidean distance matrix among sites; they are reported on the first row of Table 9.11 in Subsection 9.3.5. Distances from site 1 increase up to site 5, after which they decrease; they increase again up to site 10, then decrease; they increase up to site 15 and decrease again. The other rows of the Euclidean distance matrix display equally complex patterns; they are not shown in Table 9.11 to save space. A PCA algorithm is facing the task of representing these complex patterns in at most three dimensions because PCA ordinations cannot have more axes than the number of original variables (i.e. three species in Fig. 9.10). The result is illustrated in Fig. 9.11, panels a and b. The most dramatic effect is found at the ends of the transect, which are folded inwards along axis I. This is because the Euclidean distance formula considers the extreme sites to be very near each other (small distances due to double-zeros for species 2). This shape is called a *horseshoe*. Figure 9.11b shows that the end sites also go "down" along the third axis. In correspondence analysis on the contrary, extremities of the gradient are, in most instances, not folded inwards in the plot (but see Wartenberg *et al*., 1987, Fig. 3, for a case where this occurs); a bent ordination plot with extremities not folded inwards is called an *arch*, e.g. Fig. 9.11c.

Horseshoe

Arch

The presence in ordination plots of a *bow* (Swan, 1970), *horseshoe* (Kendall, 1971), or *arch* (Gauch, 1982) had already been noted by ecologist Goodall (1954). Benzécri and coll. (1973) discuss the arch under the name *Guttman effect*. Several authors have explained the nature of this mathematical construct, which occurs when the species composition of the sites progressively changes along an environmental gradient. *Detrended correspondence analysis* (DCA; Hill & Gauch, 1980; Gauch, 1982) aims at eliminating the arch effect.

Detrended CA (DCA)

Figure 9.11c helps in understanding the meaning of CA joint plots. This joint plot has been produced using scaling type 1 to preserve the χ^2 distances (D_{16}) among sites; in that respect, this plot is comparable to the PCA ordination shown in Fig. 9.11a. The ordination is two-dimensional since the data set only contains three species. The species (black squares) occupy the edges of a triangle; heavy lines are drawn to materialize their distances to the centre of the plot. Sites 1-5, 10, and 15-19, which only have one species present, occupy the same position as the point representing that species because sites are at the barycentres (centroids) of the species; CA does not spread apart sites that possess a single and same species, even in different amounts. Sites 6-9 and 11-14, which possess two species in various combinations, lie on a line between the two species; their positions along that line depend on the relative

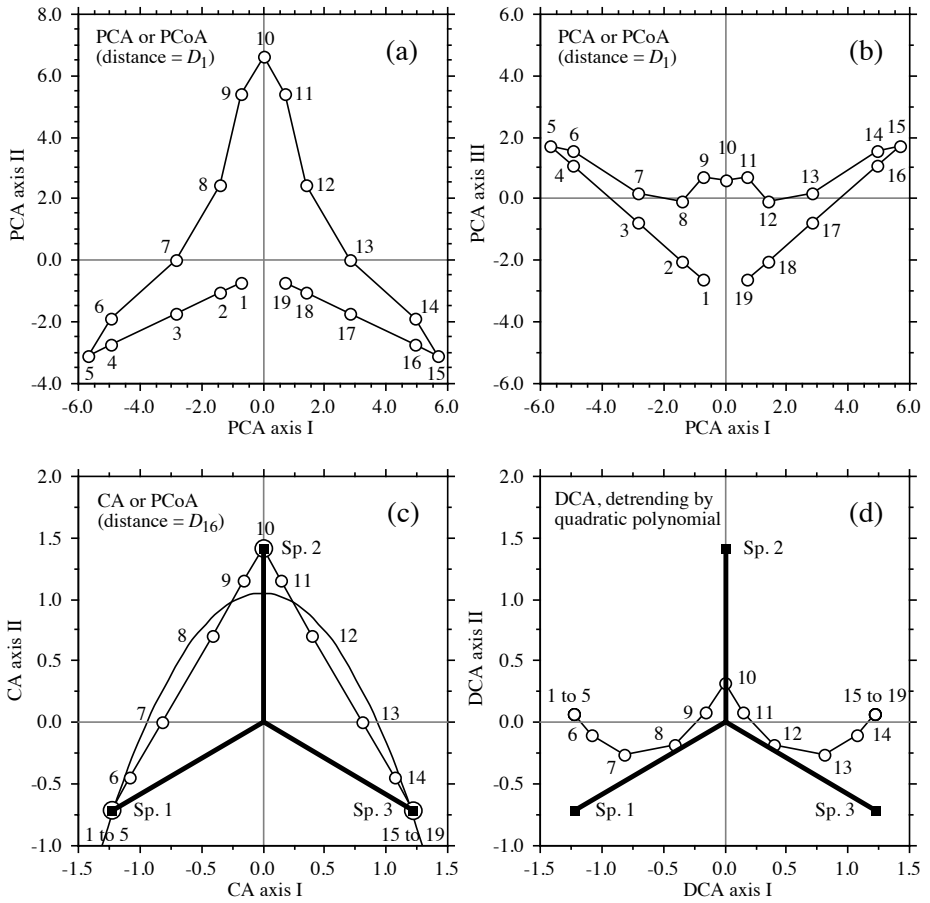


Figure 9.11 Ordinations of the data from Fig. 9.10 and Table 9.7. Circles are sites, and squares in panels c and d are species. Principal component analysis, scaling 1: (a) PCA axes I and II ($\lambda_1 = 50.1\%$, $\lambda_2 = 40.6\%$), (b) axes I and III ($\lambda_1 = 50.1\%$, $\lambda_3 = 9.3\%$). (c) Correspondence analysis, scaling 1, CA axes I and II ($\lambda_1 = 58.1\%$, $\lambda_2 = 41.9\%$). A quadratic polynomial function of axis I is also shown (convex curve): (axis II) = $1.056 - 1.204$ (axis I)^2. (d) Detrended correspondence analysis (scaling type 1, detrending by quadratic polynomial), DCA axes I and II ($\lambda_1 = 58.1\%$, $\lambda_2 = 1.6\%$). (c) and (d) Bold lines drawn from the centres of the plots represent the species axes.

abundances of the two species at each site. No site has three species in this example, so that no point lies inside the triangular shape of the scatter of sites. Considering site 1 (lower left in Fig. 9.11c), examine its distances (D_{16}) to all the other sites in the last row of Table 9.11: they increase from site 6 to 10, after which they remain constant. This corresponds to the relative positions of the sites in the figure. Had the example contained more species, the site points would have displayed a rounded shape.

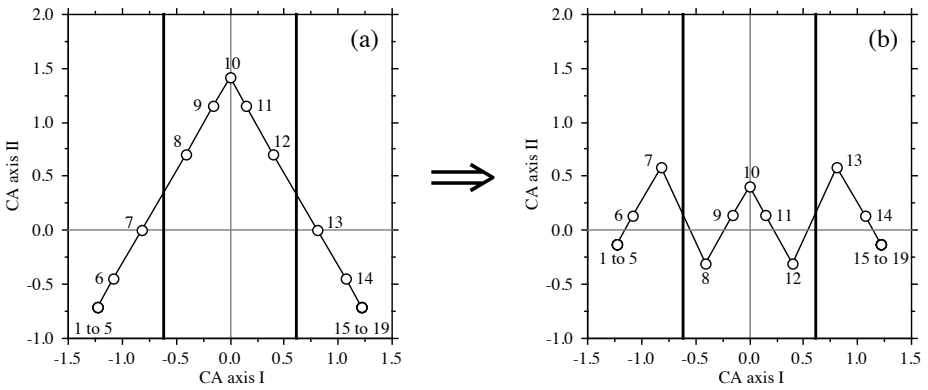


Figure 9.12 Detrending by segments. (a) Three arbitrarily defined segments are delimited by vertical lines in the CA ordination (from Fig. 9.11c). (b) After detrending, the mean of the points in each segment is zero.

Two approaches have been proposed to remove arches in CA, producing detrended correspondence analysis: detrending by segments and by polynomials.

1. When *detrending by segments* (Hill & Gauch, 1980), axis I is divided into a number of *segments* and, within each one, the mean of the scores along axis II is made equal to zero; in other words, data points in each segment are moved along axis II to make their mean coincide with the abscissa. Figure 9.12b shows the result of detrending the ordination of Fig. 9.11c using the three segments defined in Fig. 9.12a. The bottom line is that scores along detrended axis II are meaningless. Proximities among points should in no case be interpreted ecologically, because segmenting generates large differences in scores for points that are near each other in the original ordination but happen to be on either side of segment divisions (Fig. 9.12). The number of segments is arbitrary; different segmentations lead to different ordinations along axis II.

The method is only used with a fairly large number of segments. Programs DECORANA (Hill, 1979b) and CANOCO use a minimum of 10 and a maximum of 46 segments, 26 being the default value that users often take to be the 'recommended' number. This requires a number of data points larger than that in Fig. 9.10.

In order to deal with the contraction of the ends of the gradient when the sites are projected onto the first axis, nonlinear rescaling of the axes is often performed following detrending. An extreme case is represented by Fig. 9.11c where sites 1 to 5 and 15 to 19 each occupy a single point along axis I. To equalize the breadths of the species response curves, the axis is divided into small segments and segments with small within-group variances are expanded, whereas segments with large within-group variances are contracted (Hill, 1979b). Figure 5.5 of ter Braak (1987c) provides a good illustration of the process; ter Braak (1987c) advises *against* the routine use of nonlinear rescaling.

Length of gradient

After detrending by segments and nonlinear rescaling of the axes, the DCA ordination has the interesting property that the axes are scaled in units of the average standard deviation (SD) of species turnover (Gauch, 1982). Along a regular gradient, a species appears, rises to its modal value, and disappears over a distance of about 4 SD; similarly, a complete turnover in species composition occurs, over the sites, in about 4 SD units. A half-change in species composition occurs within about 1 to 1.4 SD units. Thus the length of the first DCA axis is an approximate measure of the length of the ecological gradient, measured in species turnover units. In this respect, DCA with nonlinear rescaling of the axes is a useful method to estimate the lengths of ecological gradients. The length of a gradient revealed by a pilot study may help determine the *extent* (Section 13.0) to be given to a subsequent full-scale study.

2. *Detrending by polynomials* (Hill & Gauch, 1980; ter Braak, 1987c) directly follows from the fact that an arch is produced when a gradient of sufficient length is present in data. When a sufficient number of species are present and replace each other along the gradient, the second CA axis approaches a quadratic function of the first one (i.e. a second-degree polynomial), and so on for the subsequent axes. This is clearly not the case with the data of Table 9.7, which consist of three species only. Figure 9.11c shows that the 'arch' is reduced to a triangular shape in that case.

The arch effect is removed by imposing, in the CA algorithm, the constraint that axis II be uncorrelated not only to axis I (orthogonalization procedure in Table 9.8), but also to its square, its cube, and so on; the degree of the polynomial function is chosen by the user. In the same way, axis III is made uncorrelated to the 1st, 2nd, 3rd … *k*-th degree polynomial of axes I and II. And so forth. When detrending is sought, detrending by polynomial is an attractive method. The result is a continuous function of the previous axes, without the discontinuities generated by detrending-by-segments. However, detrending by polynomials imposes a specific model onto the data, so that the success of the operation depends on how closely the polynomial model corresponds to the data. Detrending by polynomial does not solve the problem of compression of the sites at the ends of the ordination axes.

Detrending by quadratic polynomial was applied to the test data. Figure 9.11c shows the quadratic polynomial (convex curve; among the terms of the quadratic polynomial, only the $(\text{axis I})^2$ term was significant) that was fitted to the CA ordination, which has a triangular shape in the present example. Detrending involves computing and plotting the vertical (residual) distances between the data points and the fitted polynomial. The detrended ordination is shown in Fig. 9.11d. The regression residuals display an elegant but meaningless shape along axis II.

The controversy about detrending raged in the literature for more than 10 years. Key papers are those of Wartenberg *et al*. (1987), Peet *et al*. (1988), and Jackson & Somers (1991b). Wartenberg *et al*. (1987) argued that the arch is an important and inherent attribute of the distances among sites, not a mathematical artifact. The only effect of DCA is to flatten the distribution of points onto axis I without affecting the ordination of sites along that axis. They also pointed out that detrending-by-segments is an arbitrary method for which no theoretical justification has been offered. Similarly, the nonlinear rescaling procedure assumes that, on average, each species appears and disappears at the same rate along the transect and that the parametric variance is an adequate measure of that rate; these assumptions have not been substantiated. Despite these criticisms, Peet *et al*. (1988) still supported DCA on the ground that detrending

and rescaling may facilitate ecological interpretation. Jackson & Somers (1991b) showed that the DCA ordination of sites greatly varied with the number of segments one arbitrarily decides to use, so that the ecological interpretation of the results may vary widely, as do the correlations one can calculate with environmental variables. One should always try different numbers of segments if one decides to use DCA.

Simulation studies involving DCA have been conducted on artificial data representing unimodal species responses to environmental gradients in one (*coenoclines*) or two (*coenoplanes*) dimensions, following the method pioneered by Swan (1970). Kenkel & Orlóci (1986) report that DCA did not perform particularly well in recovering complex gradients. Using Procrustes statistics (Subsection 11.5.2) as measures of structure recovery, Minchin (1987) showed that DCA did not perform well with complex response models and non-regular sampling schemes. Both studies concurred that nMDS (Section 9.4) was a better method than DCA for recovering complex gradients.

Present evidence indicates that detrending should be avoided except for the specific purpose of estimating the lengths of gradients; such estimates remain subject to the condition that the assumptions of the model are true. In particular, DCA should be avoided when analysing data that represent complex ecological gradients. Most ordination techniques are able to recover simple, one-dimensional environmental gradients. When there is a single gradient in the data, detrending is useless since the gradient is well represented by CA axis I.

Satisfactory mathematical solutions to the problem of detrending remain to be found. In the meantime, ordination results should be interpreted with caution and in the light of the type of distance preserved by each method.

6 — Ecological applications

Ecological application 9.2a

The spider data (28 sites × 12 species) of Aart & Smeenk-Enserink (1975) that have been analysed by principal component analysis (PCA) in Ecological application 9.1a are reanalysed here by correspondence analysis (CA). Figure 9.13 presents two CA biplots (scaling types 1 and 2) obtained for these data. Compare the species groups and site ordinations with the PCA biplot presented in Fig. 9.6b (log-transformed data).

Ecological application 9.2b

Cadoret *et al*. (1995) investigated the species composition (presence/absence and abundance) of chaetodontid fish assemblages off Moorea Island, French Polynesia, in order to describe the spatial distribution of the butterflyfishes and determine their relationships with groups of benthic organisms. Sampling was conducted in four areas around the island: (a) Opunohu Bay, (b) Cook Bay, (c) the Tiahura transect across the reef in the northwestern part of the island, and (d) the Afareaitu transect across the reef in the eastern part of the island.

Correspondence analysis (Fig. 9.14) showed that the fish assemblages responded to the main environmental gradients that characterized the sampling sites. For areas c and d (transects across the reef), axis I corresponded to a gradient from the coastline to the ocean; from left to right, in

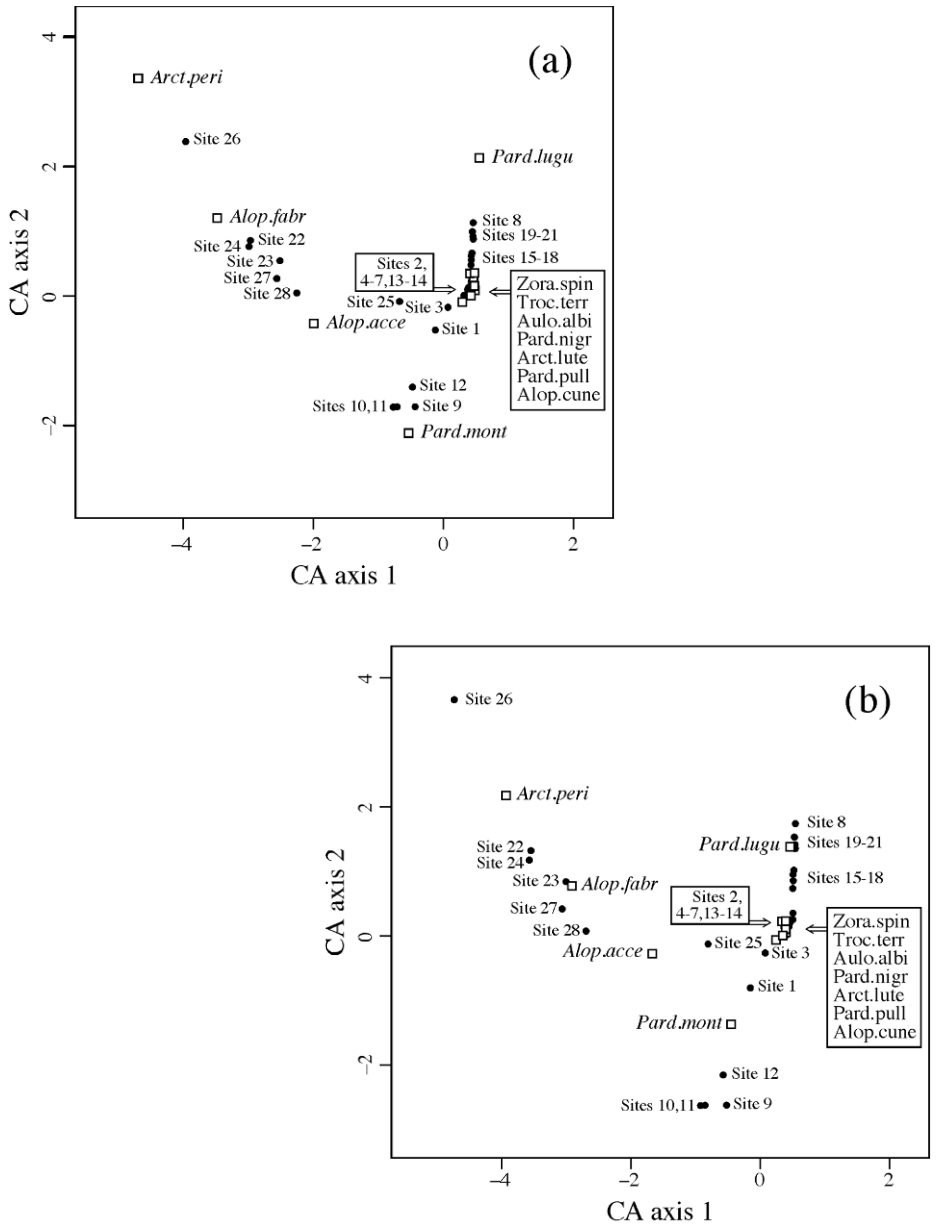


Figure 9.13 Correspondence analysis biplots of the spider data. (a) Scaling type 1: the sites (solid circles) are at the centroids of the species (open squares). (b) Scaling type 2: the species (open squares) are at the centroids of the sites (solid circles). Species abbreviations: see Fig. 9.6.

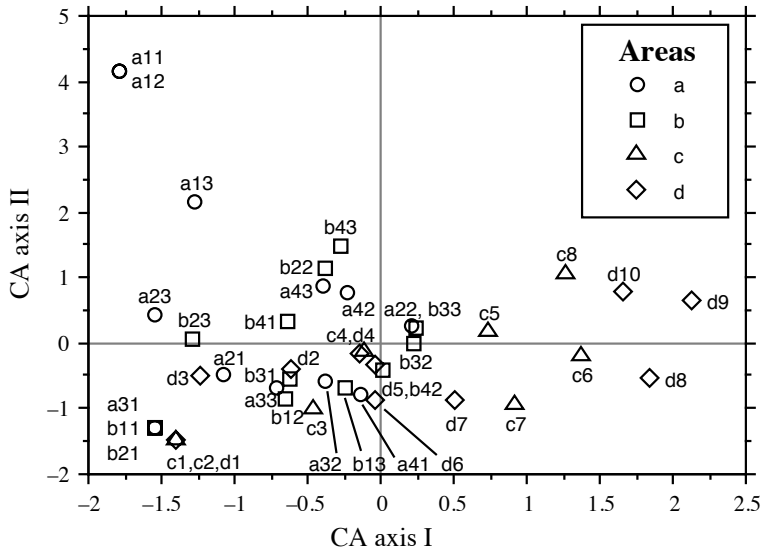


Figure 9.14 Correspondence analysis (CA): ordination of sampling sites with respect to axes I and II from presence/absence observations of butterflyfishes (21 species) in four areas (different symbols) around Moorea Island. Axes I and II explain together 29% of the variation among sites. Species are not drawn; they would have overloaded the plot. Modified from Cadoret *et al*. (1995).

the plot, are the sites of the fringing reef, the shallows (found only in sector c), the barrier reef, and the outer slope. Sites from the bays (areas a and b) are also found in the left-hand part of the graph. Axis II separates the sites located in the upper reaches of Opunohu Bay (a11, a12 and a13, in the upper-left of the plot) from all the others. This application will be further developed, in Subsection 11.2.2, to identify species assemblages and evidence the relationships between species and environmental variables, using canonical correspondence analysis.

Ecological application 9.2c

In a study on the vegetation dynamics of southern Wisconsin, Sharpe *et al*. (1987) undertook a systematic field survey of all forest tracts in two townships. Detrended correspondence analysis was used to display the relationships among stands with respect to species composition. The scores of the first ordination axes were used to construct three-dimensional maps. In the map of the first axis (Fig. 9.15), the scores were generally low in the southern and central portions of the area, and increased towards the west and north. Since the first axis showed a trend from forest tracts dominated by *Acer saccharum* to oak-dominated forests (not shown), Fig. 9.15 indicates that stands dominated by *A. saccharum* were located in the south-central portion of the area, whereas oak-dominated stands were to the west, north and, to a lesser extent, east. Such a mapping, using a 3- or 2-dimensional representation, is often a useful way of displaying synthetic information provided by the scores of objects along the first ordination axes.

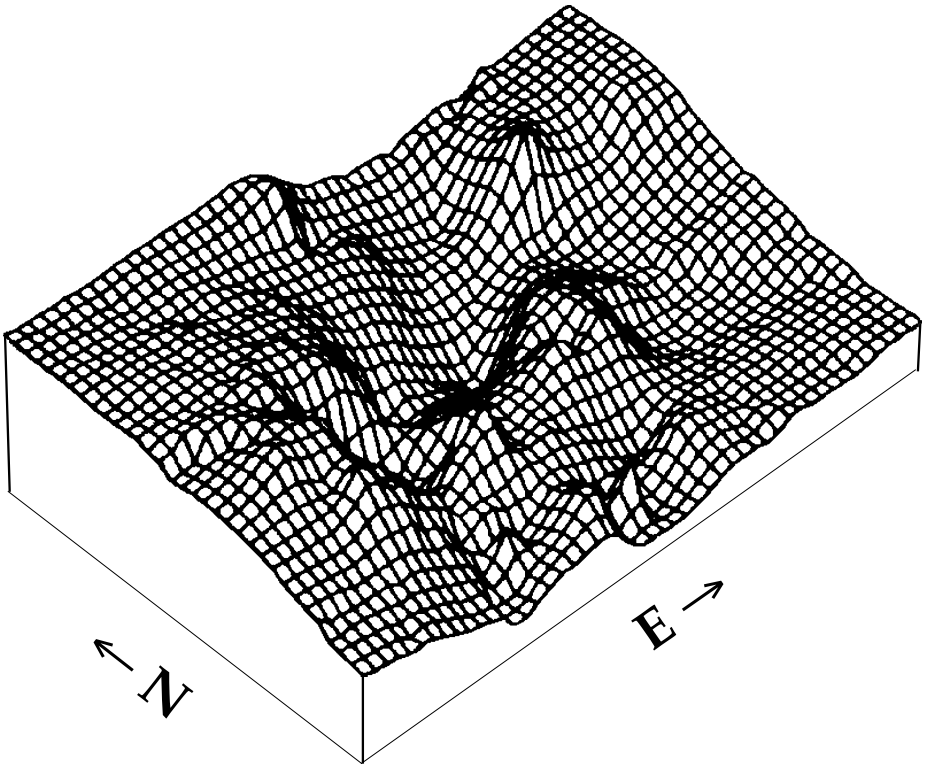


Figure 9.15 Three-dimensional map of the scores of the first ordination axis (detrended correspondence analysis), based on trees observed in 92 forest tracts of southern Wisconsin, U.S.A. (survey area: 11 × 17 km). Modified from Sharpe *et al*. (1987).

Maps like the one displayed in Fig. 9.15 may be produced for the ordination scores computed by any of the methods described in the present chapter; see Section 13.2.

7 — Algorithms

There are several computer programs and R functions available for correspondence analysis; see Section 9.5.

TWWA algorithm

CANOCO (ter Braak, 1988b, 1988c, 1990; ter Braak & Smilauer, 1998) uses Hill's *two-way weighted averaging* (TWWA) algorithm as summarized by ter Braak (1987c). This algorithm is described in Table 9.8. There are three main differences with the TWWS algorithm for PCA presented in Table 9.5: (1) variables are centred in PCA, not in CA; (2) in CA, the centroid of the site scores is not zero and must thus be estimated (step 6.1); (3) in CA, summations are standardized by the row sum, column

Table 9.8 Two-way weighted averaging (TWWA) algorithm for correspondence analysis. From Hill (1973b) and ter Braak (1987c).

a) Iterative estimation procedure

Step 1: Consider a table **Y** with n rows (sites) × p columns (species).
Do NOT centre the columns (species) on their means.

Determine how many eigenvectors are needed. For each one, **DO** the following:

Step 2: Take the row order as the arbitrary initial site scores. (1, 2, …)
Set the initial eigenvalue estimate to 0. In what follows, y_{i+} = row sum for site i, y_{+j} = column sum for species j, and y_{++} = grand total for the data table **Y**.

Iterative procedure begins

Step 3: Compute new species loadings: $colscore(j) = \sum y(i,j) \times rowscore(i)/y_{+j}$

Step 4: Compute new site scores: $rowscore(i) = \sum y(i,j) \times colscore(j)/y_{i+}$

Step 5: For the second and higher-order axes, make the site scores uncorrelated with all previous axes (Gram-Schmidt orthogonalization procedure: see *b* below).

Step 6: Normalize the vector of site scores (procedure *c*, below) and obtain an estimate of the eigenvalue. If this estimate does not differ from the previous one by more than the tolerance set by the user, go to step 7. If the difference is larger than the tolerance, go to step 3.

End of iterative procedure

Step 7: If more eigenvectors are to be computed, go to step 2. If not, continue with step 8.

Step 8: The row (site) scores correspond to matrix $\hat{\mathbf{V}}$. The column scores (species loadings) correspond to matrix $\hat{\mathbf{F}}$. Matrices $\hat{\mathbf{F}}$ and $\hat{\mathbf{V}}$ provide scaling type 2 (Subsection 9.2.1). Scalings 1 or 3 may be calculated if required. Return the eigenvalues, % variance, species loadings, and site scores.

b) Gram-Schmidt orthogonalization procedure

DO the following, in turn, for all previously computed components k:

Step 5.1: Compute the scalar product $SP = \sum (y_{i+} \times rowscore(i) \times v(i,k)/y_{++})$ of the current site score vector estimate with the previous component k. Vector $v(i,k)$ contains the site scores of component k scaled to length 1. This product is between 0 (if the vectors are orthogonal) and 1.

Step 5.2: Compute new values of $rowscore(i)$ such that vector *rowscore* becomes orthogonal to vector $v(i,k)$: $rowscore(i) = rowscore(i) - (SP \times v(i,k))$.

c) Normalization procedure[†]

Step 6.1: Compute the centroid of the site scores: $z = \sum (y_{i+} \times rowscore(i)/y_{++})$.

Step 6.2: Compute the sum of squares of the site scores: $S^2 = \sum (y_{i+} \times (rowscore(i) - z)^2/y_{++})$; $S = \sqrt{S^2}$.

Step 6.3: Compute the normalized site scores: $rowscore(i) = (rowscore(i) - z)/S$.

Step 6.4: At the end of each iteration, S, which measures the amount of shrinking during the iteration, provides an estimate of the eigenvalue. Upon convergence, the eigenvalue is S.

[†] Normalization in CA is such that the *weighted* sum of squares of the elements of the vector is equal to 1.

sum, or grand total, as appropriate, which produces shrinking of the ordination scores at the end of each iteration in CA (step 6.4), instead of stretching as in PCA.

SVD
Householder

R functions for CA use either singular value decomposition (SVD, function ***svd()*** of R) or Householder reduction (function ***eigen()*** of R). SVD and eigen-decomposition were both used to describe the CA algorithm in Subsection 9.2.1; they provide the eigenvalues as well as matrices $\mathbf{U}$ and $\hat{\mathbf{U}}$. The various matrices for the row and column scores used in scalings are then obtained using eqs. 9.33 to 9.36.

9.3 Principal coordinate analysis (PCoA)

Principal component analysis (PCA) is only applicable to data for which the Euclidean distance (D_1) is appropriate, whereas correspondence analysis (CA) is only applicable to frequency-like data for which the χ^2 distance (D_{16}) is appropriate. For other types of data, the relationships among objects are computed with one of the resemblance coefficients described in Chapter 7. The list includes coefficients that can handle binary data (S_1 to S_{14}, S_{24} to S_{27}) and mixtures of quantitative and qualitative descriptors (S_{15}, S_{16}). PCA cannot be applied to these data. CA can be used with presence-absence data for which double zeros must be excluded from object comparisons, but not with mixtures of quantitative and qualitative descriptors.

Euclidean representation

Gower (1966) described a method to obtain a Euclidean representation (i.e. a representation in a Cartesian coordinate system) of a set of objects whose relationships are measured by any distance coefficient chosen by users. This method, known as *principal coordinate analysis* (abbreviated PCoA), *metric multidimensional scaling* (in contrast to the nonmetric method described in Section 9.4), or *classical scaling* by reference to the pioneering work of Torgerson (1958), allows one to position objects in a space of reduced dimensionality while preserving their distance relationships as well as possible; see also Rao (1964).

Mixed precision

The interest of the PCoA method lies in the fact that it may be used with all types of descriptors — even data sets with descriptors of mixed levels of precision, provided that a coefficient appropriate to the data has been used to compute the resemblance matrix (e.g. S_{15} or S_{16}, Chapter 7). It will be shown that, if the distance matrix is metric, i.e. if it contains no violation of the triangle inequality, the relationships among objects can, in most cases, be fully represented in Euclidean space. In the case of violations of the triangle inequality, or when problems of "non-Euclideanarity" occur with metric distances (Gower, 1982; Fig. 9.16), negative eigenvalues are produced. In most cases, this does not impair the quality of the Euclidean representation obtained for the first few principal coordinates. It is also possible to transform the distance matrix, or use an alternative resemblance measure, to eliminate the problem. These topics are discussed in Subsection 9.3.4.

Euclidean model

One may look at principal coordinates as the equivalent of principal components. Principal components, on the one hand, are linear combinations of the original (or standardized) descriptors; *linear* is the key concept. Principal coordinates, on the other hand, are also functions of the original variables, but mediated through the distance function that has been computed among objects. In any case, PCoA can only embed (i.e. fully represent), in Euclidean space, the Euclidean part of a distance matrix. This is not a property of the data, but a result of the Euclidean model, which is forced upon the data because the objective is to draw scatter diagrams on sheets of paper. By doing so, one must accept that whatever is non-Euclidean cannot be drawn on paper. This may be viewed as the problem of drawing points separated by non-Euclidean distances into a Euclidean space.

Like PCoA, the method of nonmetric multidimensional scaling (nMDS, Section 9.4) produces ordinations of objects from any resemblance matrix. It compresses the distance relationships among objects into, say, two or three dimensions in a more efficient way than PCoA. nMDS always obtains a Euclidean representation, even from non-Euclidean-embeddable distances. However, nMDS compresses the distances in a non-linear way and its algorithm is computer-intensive, requiring more computing time than PCoA. The latter is faster for large distance matrices.

1 — Computation

Gower (1966) explained how to compute the principal coordinates of a distance matrix:

- The calculation starts with a distance matrix $\mathbf{D} = [D_{hi}]$. It is also possible to carry out the calculations from a similarity matrix $\mathbf{S} = [S_{hi}]$; the method is detailed in Subsection 9.3.3.

- Matrix **D** is transformed into a new matrix $\mathbf{A} = [a_{hi}]$ by defining:

$$a_{hi} = -\frac{1}{2}D_{hi}^2 \qquad \textbf{(9.40)}$$

The purpose of this transformation is explained in Subsection 9.3.3.

- Matrix **A** is centred to give matrix $\mathbf{\Delta}_1 = [\delta_{hi}]$, using the following equation:

$$\delta_{hi} = a_{hi} - \bar{a}_h - \bar{a}_i + \bar{a} \qquad \textbf{(9.41)}$$

where $\bar{a}_h$ and $\bar{a}_i$ are the means of the row and column corresponding to element a_{hi} of matrix **A**, respectively, and $\bar{a}$ is the mean of all a_{hi}'s in the matrix. The following matrix equation produces the centring described in eq. 9.41:

$$\mathbf{\Delta}_1 = \left(\mathbf{I} - \frac{\mathbf{11}'}{n}\right)\mathbf{A}\left(\mathbf{I} - \frac{\mathbf{11}'}{n}\right) \qquad \textbf{(9.42)}$$

where **I** is an identity matrix of order n and **1** is a column vector of length n containing values "1". Centring has the effect of positioning the origin of the new system of axes at the centroid of the scatter of objects without altering the distances among objects.

Euclidean distance

In the particular case of distances computed using the Euclidean distance coefficient (D_1, eq. 7.32), it is possible to obtain the Gower-centred matrix $\mathbf{\Delta}_1$ directly, i.e. without calculating a matrix **D** of Euclidean distances and going through eqs. 9.40 and 9.41, because $\mathbf{\Delta}_1 = \mathbf{Y}_c\mathbf{Y}_c'$, where $\mathbf{Y}_c$ is **Y** centred by columns. This may be verified using numerical examples. In that particular case, $\mathbf{\Delta}_1$ is always a positive semidefinite matrix (Table 2.2).

• The eigenvalues λ_k and normalized eigenvectors (matrix **U**) are computed and each eigenvector $\mathbf{u}_k$ is multiplied by the square root of its eigenvalue. As a result, the eigenvectors are scaled to lengths equal to the square roots of their eigenvalues:

$$\sqrt{\mathbf{u}'_k\mathbf{u}_k} = \sqrt{\lambda_k}$$

Degenerate **D** matrix

Due to the centring, matrix $\mathbf{\Delta}_1$ always has at least one zero eigenvalue. The reason is that at most $(n - 1)$ real axes are necessary for representing n points in Euclidean space. There may be more than one zero eigenvalue if the distance matrix is degenerate, i.e. if the objects can be represented in fewer than $(n - 1)$ dimensions. In practice, there are c positive eigenvalues and c real axes forming the Euclidean representation of the data, the rule being that $c \leq n - 1$.

With the Euclidean distance (D_1), when there are more objects than descriptors ($n > p$), the maximum value of c is p; when $n \leq p$, then $c \leq n - 1$. Take as example a set of three objects or more, and two descriptors ($n > p$). The objects, as many as they are, may be represented in a two-dimensional space — for example, the scatter diagram of the two descriptors. Consider now the case where there are two objects and two descriptors ($n \leq p$); the two objects only require one dimension for representation.

• After scaling, if the eigenvectors are written as columns (e.g. Table 9.9), the *rows* of the resulting table are the *coordinates of the objects* in the space of principal coordinates, without any further transformation; they form matrix **PC** of the principal coordinates. Plotting the points on, say, the first two principal coordinates produces a reduced-space ordination diagram of the objects in two dimensions.

2 — Numerical example

Readers may get a better feeling of what principal coordinate analysis does by comparing it to principal component analysis. Consider a data matrix **Y** on which a principal component analysis (PCA) has been computed, with resulting eigenvalues, eigenvectors (matrix **U**), and principal components (matrix **F**). If one also computed a Euclidean distance matrix $\mathbf{D} = [D_{hi}]$ for the same n objects, the eigenvectors obtained by principal coordinate analysis would be exactly the same as the principal components. The eigenvalues of the PCoA are equal to the eigenvalues one would

Table 9.9 Principal coordinates of the objects (rows) are obtained by scaling the eigenvectors to $\sqrt{\lambda}$.

	Eigenvalues			
	λ_1	λ_2	$\cdots$	λ_c
Objects	**Eigenvectors**			
$\mathbf{x}_1$	u_{11}	u_{12}	$\cdots$	u_{1c}
$\mathbf{x}_2$	u_{21}	u_{22}	$\cdots$	u_{2c}
$\mathbf{x}_h$	u_{h1}	u_{h2}	$\cdots$	u_{hc}
$\mathbf{x}_i$	u_{i1}	u_{i2}	$\cdots$	u_{ic}
$\mathbf{x}_n$	u_{n1}	u_{n2}	$\cdots$	u_{nc}
Lengths: $\sqrt{\sum_i u_{ik}^2}$ =	$\sqrt{\lambda_1}$	$\sqrt{\lambda_2}$	$\cdots$	$\sqrt{\lambda_c}$
Centroid: $[\bar{u}_k]$ =	0	0	$\cdots$	0

obtain from a PCA conducted on the cross-product matrix $[y - \bar{y}]' [y - \bar{y}]$; these are larger than the eigenvalues of a PCA conducted on the covariance matrix **S** by factor $(n - 1)$ because $\mathbf{S} = (1/(n-1)) [y - \bar{y}]' [y - \bar{y}]$. Since PCA has been defined, in this book, as the eigenanalysis of the covariance matrix **S**, the same PCA eigenvalues can be obtained from a principal coordinate analysis computed on the Euclidean distance matrix among objects, and dividing the resulting PCoA eigenvalues by $(n - 1)$. If one is only interested in the *relative* magnitude of the eigenvalues, this scaling step is not necessary and may be ignored.

The previous paragraph does not mean that principal coordinate analysis is limited to Euclidean distance matrices. It can actually be computed for *any* distance matrix. If the distances cannot readily be embedded in Euclidean space, negative eigenvalues may be obtained, with consequences described in Subsection 9.3.4.

Numerical example 2. The numerical example for principal component analysis (Section 9.1) is used here to illustrate the main steps in the computation of principal coordinates. The example also shows that computing principal coordinates from a matrix of Euclidean distances $\mathbf{D} = [D_{hi}]$ gives the exact same results as a principal component analysis of the raw data, with the exception that the descriptor loadings are not obtained in PCoA. Indeed, information about the original descriptors is not passed on to the PCoA algorithm. Indeed, since PCoA is computed from a distance matrix among objects, it cannot give back the loadings of the descriptors. A method for computing them *a posteriori* is described in Subsection 9.3.3 (eq. 9.45).

1) The matrix of Euclidean distances among the 5 objects of data matrix **Y** used to illustrate Section 9.1 is:

$$\mathbf{D} = \begin{bmatrix} 0.00000 & 3.16228 & 3.16228 & 7.07107 & 7.07107 \\ 3.16228 & 0.00000 & 4.47214 & 4.47214 & 6.32456 \\ 3.16228 & 4.47214 & 0.00000 & 6.32456 & 4.47214 \\ 7.07107 & 4.47214 & 6.32456 & 0.00000 & 4.47214 \\ 7.07107 & 6.32456 & 4.47214 & 4.47214 & 0.00000 \end{bmatrix}$$

2) Matrix $\mathbf{\Delta}_1$ obtained by Gower's centring (eqs. 9.40 and 9.41) is:

$$\mathbf{\Delta}_1 = \begin{bmatrix} 12.8 & 4.8 & 4.8 & -11.2 & -11.2 \\ 4.8 & 6.8 & -3.2 & 0.8 & -9.2 \\ 4.8 & -3.2 & 6.8 & -9.2 & 0.8 \\ -11.2 & 0.8 & -9.2 & 14.8 & 4.8 \\ -11.2 & -9.2 & 0.8 & 4.8 & 14.8 \end{bmatrix}$$

The trace (sum of the diagonal elements) of this matrix is 56. This is $(n - 1) = 4$ times the trace of the covariance matrix computed in PCA, which was 14. The diagonal elements are the squared distances of the points to the multivariate centroid. Note that matrix $\mathbf{\Delta}_1$ could have been obtained directly from data matrix **Y** centred by columns ($\mathbf{Y}_c$), as mentioned in Subsection 9.3.1 for the particular case where **D** is computed using the Euclidean distance coefficient (D_1, eq. 7.32): $\mathbf{\Delta}_1 = \mathbf{Y}_c\mathbf{Y}_c'$. Readers can verify this property numerically for the example.

3) The eigenvalues and eigenvectors of matrix $\mathbf{\Delta}_1$, scaled to $\sqrt{\lambda}$, are shown in Table 9.10. There are only $c = 2$ eigenvalues different from zero; this was to be expected since the distances had been computed from $p = 2$ variables only ($c = p = 2$). The principal coordinates, which are the rescaled eigenvectors of the PCoA, are identical to the principal components (Subsection 9.1.2 and Table 9.6) in this example. Measures of resemblance other than the Euclidean distance may produce a different number of eigenvalues and principal coordinates and they would, of course, position the objects differently.

PCA and PCoA

While the numerical example illustrates the fact that a PCoA computed on a Euclidean distance matrix gives the same results as a PCA conducted on the original data, the converse is also true: taking the coordinates of the objects in the full space (all eigenvectors) obtained from a PCoA and using them as input of a principal component analysis would produce the same PCA eigenvalues as those of the original PCoA (to a

Table 9.10 Principal coordinates computed for the numerical example for PCA developed in Section 9.1. Compare with PCA results in Subsection 9.1.2 and Table 9.6.

	Eigenvalues	
	λ_1	λ_2
Objects	**Eigenvectors**	
$\mathbf{x}_1$	–3.578	0.000
$\mathbf{x}_2$	–1.342	–2.236
$\mathbf{x}_3$	–1.342	2.236
$\mathbf{x}_4$	3.130	–2.236
$\mathbf{x}_5$	3.130	2.236
Eigenvalues of PCoA	36.000	20.000
PCoA eigenvalues/$(n-1)$ = eigenvalues of corresponding PCA	9.000	5.000
Lengths: $\sqrt{\sum_i u_{ik}^2}$ =	6.000 = $\sqrt{36}$	4.472 = $\sqrt{20}$

factor $n-1$), and the principal components will be identical to the principal coordinates. All the signs of any one component may be inverted, though, as explained in Subsection 9.1.9; signs depend on an arbitrary decision made during execution of eigen-decomposition functions (Subsection 2.9.2). Because of this, before presenting their results, users of ordination methods are free to invert all the signs of any principal component or principal coordinate if that suits them better.

3 — Rationale of the method

Gower (1966) has shown that the distance relationships among objects are preserved in the full-dimensional principal coordinate space. His proof is summarized as follows.

• In the total space of the principal coordinates (i.e. all eigenvectors), the distance between objects h and i can be found by computing the Euclidean distance between rows h and i of Table 9.9:

$$D'_{hi} = \left[\sum_{k=1}^{c} (u_{hk} - u_{ik})^2\right]^{1/2} = \left[\sum_{k=1}^{c} u_{hk}^2 + \sum_{k=1}^{c} u_{ik}^2 - 2\sum_{k=1}^{c} u_{hk}u_{ik}\right]^{1/2} \quad \textbf{(9.43)}$$

• Since the eigenvectors are scaled in such a way that their lengths are $\sqrt{\lambda_k}$ (in other words, **U** is scaled here to $\mathbf{\Lambda}^{1/2}$), the eigenvectors have the property that $\mathbf{\Delta}_1 = \mathbf{U}\mathbf{U}'$. One can thus write:

$$\mathbf{\Delta}_1 = [\delta_{hi}] = \mathbf{u}_1\mathbf{u}_1' + \mathbf{u}_2\mathbf{u}_2' + \ldots + \mathbf{u}_c\mathbf{u}_c'$$

from which it can be shown, following eq. 9.43, that:

$$D'_{hi} = [\delta_{hh} + \delta_{ii} - 2\delta_{hi}]^{1/2}$$

Readers can verify this property on the above numerical example.

• Since $\delta_{hi} = a_{hi} - \bar{a}_h - \bar{a}_i + \bar{a}$ (eq. 9.41), replacing the values of δ in the right-hand member of the previous equation gives:

$$\delta_{hh} + \delta_{ii} - 2\delta_{hi} = a_{hh} + a_{ii} - 2a_{hi}$$

hence

$$D'_{hi} = [a_{hh} + a_{ii} - 2a_{hi}]^{1/2}$$

The transformation of **A** into $\mathbf{\Delta}_1$ is not essential. It is simply meant to eliminate one of the eigenvalues, which could be the largest and would only account for the distance between the centroid and the origin.

• The transformation of the matrix of original distances D_{hi} into **A** is such that distances are preserved in the course of the calculations. Actually, one can replace the a_{hi} terms in the previous equation by $-0.5\ D^2_{hi}$ (eq. 9.40), which produces the equation

$$D'_{hi} = \left[-\frac{1}{2}D^2_{hh} - \frac{1}{2}D^2_{ii} + D^2_{hi}\right]^{1/2}$$

and, since $D_{hh} = D_{ii} = 0$ (property of distances),

$$D'_{hi} = [D^2_{hi}]^{1/2}$$

Principal coordinate analysis thus preserves the original distances, regardless of the formula used to compute them. If the distances have been calculated from similarities, $D_{hi} = 1 - S_{hi}$ will be preserved in the full-dimensional principal coordinate space. If the transformation of similarities into distances was done by $D_{hi} = \sqrt{1 - S_{hi}}$ or $D_{hi} = \sqrt{1 - S^2_{hi}}$, then it is these distances that are preserved by the PCoA. As a corollary, these various representations in principal coordinate space should be as different from one another as are the distances themselves.

Gower (1966) has also shown that principal coordinates can be directly computed from a similarity matrix **S** instead of a distance matrix **D**, as follows: (1) make sure that the diagonal of matrix **S** contains 1's and not 0's before centring; (2) centre matrix **S** using eq. 9.41 or 9.42

without applying eq. 9.40 first; (3) compute the eigenvalues and eigenvectors; (4) multiply the elements of each eigenvector k by $\lambda_k^{0.5}$. The distances D'_{hi} among the reconstructed point-objects in the full-dimensional principal coordinate space are not the same as the distances $D_{hi} = (1 - S_{hi})$; they are distorted, being such that $D'_{hi} = \sqrt{2}\sqrt{D_{hi}}$. Looking at it from another viewpoint, the reconstructed distances D'_{hi} are larger than the distances $D_{hi} = (1 - S_{hi})^{0.5}$ by a factor $\sqrt{2}$ without further distortion. These relationships hold only if the centred matrix **S** is positive semidefinite, i.e. if its eigen-decomposition does not produce negative eigenvalues.

To summarize, principal coordinate analysis produces a representation of objects in Euclidean space that preserves the distance relationships computed using any measure selected by users. This is a major difference with PCA, where the distance among objects is always, by definition, the Euclidean distance (Table 9.1). In PCoA, the representation of objects in the reduced space of the first few principal coordinates forms the best possible Euclidean approximation of the original distances, because the sum of squared lengths of the objects in the selected subspace is maximum (Gower, 1982). The quality of a Euclidean representation in a space of principal coordinates can be assessed using a Shepard diagram (Fig. 9.1).

Contrary to principal component analysis, the relationships between the principal coordinates and the original descriptors are not provided by a principal coordinate analysis. Indeed the descriptors, from which distances were initially computed among the objects, do not play any role during the calculation of the PCoA from matrix **D**.

PCoA biplot

However, computing the projections of descriptors in the space of the principal coordinates to produce biplots is fairly simple:

$$\mathbf{S}_{pc} = \frac{1}{(n-1)} \mathbf{Y}_c' \mathbf{U}_{st} \qquad \textbf{(9.44)}$$

$$\mathbf{U}_{proj} = \sqrt{n-1}\ \mathbf{S}_{pc} \mathbf{\Lambda}^{-0.5} \qquad \textbf{(9.45)}$$

$\mathbf{Y}_c$ is the centred matrix of the original descriptors or any other set of explanatory variables that users wish to project in the PCoA biplot. **Y** may need to be transformed before it is centred and used in eq. 9.44; for example, dimensionally heterogeneous physical variables need to be standardized. $\mathbf{U}_{st}$ is the matrix of PCoA eigenvectors ($n \times c$) standardized by columns (eq. 1.12); it may contain a subset of the eigenvectors only, for example the first two. $\mathbf{S}_{pc}$ is the covariance matrix between **Y** and the standardized principal coordinates $\mathbf{U}_{st}$; in the computation of this covariance matrix, eq. 9.44 assumes that the descriptors in matrix **Y** are quantitative. The rows of matrix $\mathbf{U}_{proj}$ correspond to the p descriptors to be added to the biplot, and its columns correspond to the principal coordinates. For a PCoA conducted on a Euclidean distance matrix (D_1) computed from **Y**, the PCoA biplot with matrix **PC** for objects and matrix $\mathbf{U}_{proj}$ for descriptors is identical to a PCA distance biplot of **Y** (Subsection 9.1.4), notwithstanding possible changes of signs along some of the axes between the two analyses.

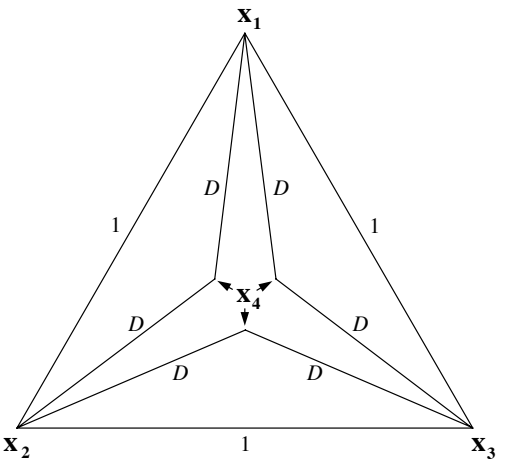


Figure 9.16 This figure, from Gower (1982), illustrates a case where the triangle inequality is not violated, yet no Euclidean representation of the four points ($\mathbf{x}_1$ to $\mathbf{x}_4$) is possible because the distances D, which are all equal, are too small for the three representations of the inner point ($\mathbf{x}_4$) to join in a single point. Assuming that the outer edges are all of lengths 1, the triangle inequality will be violated if D is smaller than 0.5. On the contrary, a two-dimensional Euclidean representation of the four points will be possible with $D = 1/\sqrt{3}$ because then the three representations of $\mathbf{x}_4$ will meet at the centroid. With $D > 1/\sqrt{3}$, the Euclidean representation of the four points, $\mathbf{x}_1$ to $\mathbf{x}_4$, will form a three-dimensional pyramid.

4 — Negative eigenvalues

There are distance matrices that do not allow a full representation of the distance relationships among objects in Euclidean space (i.e. a set of real Cartesian coordinates).

• Problems of Euclidean representation may result from the use of a distance measure that violates the triangle inequality. Such distances are called *semimetric* and *nonmetric* in Tables 7.2 and 7.3.

• Such problems may also result from an imbalance in the distance matrix, due to the handling of missing values. See for instance how missing values are handled in coefficients S_{15}, S_{16}, S_{19}, and S_{20} of Chapter 7, using Kronecker delta functions.

Non-Euclideanarity

• Some *metric* distance matrices present problems of "non-Euclideanarity", as described by Gower (1982, 1985). Figure 9.16 illustrates such a case; the closing of all individual triangles (triangle inequality condition, Section 7.4) is a necessary, but not a sufficient condition to guarantee a full Euclidean representation of a set of objects. This "non–Euclideanarity", when present, translates itself into negative eigenvalues.

For instance, most of the metric distances resulting from the transformation of a similarity coefficient using the formula $D = 1 - S$ are non-Euclidean (Table 7.2). This does not mean that all distance matrices computed using these coefficients are non-Euclidean, but that cases can be found where PCoA produces negative eigenvalues. Among the metric coefficients described in Subsection 7.4.1, several were demonstrated to be Euclidean whereas others are not Euclidean (Table 7.3).

Tables 7.2 and 7.3 show that, for many coefficients, the distances $\sqrt{D}$ or $D = \sqrt{1 - S}$ are Euclidean even though the distances D or $D = (1 - S)$ are not Euclidean. The use of $\sqrt{D}$ or the transformation $D_{hi} = \sqrt{1 - S_{hi}}$ should thus be preferred before computing PCoA using those coefficients. This transformation solves the negative eigenvalue problem even for coefficients that are known to be semimetric. This is the case, for instance, with coefficients S_8, S_{17}, and the percentage difference ($D_{14} = 1 - S_{17}$), which are widely used by ecologists to analyse tables of species presence or abundance data. A square-root transformation of $D_{14} = 1 - S_{17}$, for example, eliminates negative eigenvalues in principal coordinate analysis; see Numerical example 1 (continued) in Subsection 9.3.5. In support of this statement, Gower & Legendre (1986) have shown that coefficient S_8, which is the binary form of S_{17}, is Euclidean when transformed into $D = \sqrt{1 - S_8}$, and simulations have never turned up cases where $D = \sqrt{1 - S_{17}}$ is non-Euclidean.

When one does not wish to apply a square root transformation to the distances, or when negative eigenvalue problems persist in spite of a square root transformation, Gower & Legendre (1986) have shown that the problem of "non–Euclideanarity", and of the negative eigenvalues that come with it, can be solved by adding a (large enough) constant to all values of a distance matrix that would not lend itself to full Euclidean representation. No correction is made along the diagonal, though, because the distance between an object and itself is always zero. Actually, adding some large constant would make the negative eigenvalues disappear and produce a fully Euclidean representation, but it would also create an extra dimensions (and eigenvalue) to express the additional variance so generated. In Fig. 9.17c, for instance, adding a large value, like 0.4, to all six distances among the four points in the graph would create a pyramid, requiring three dimensions for a full Euclidean representation instead of two.

The problem is to add just the right amount to all distances in the matrix to eliminate all negative eigenvalues and produce a Euclidean representation of the distance relationships among objects, without creating unnecessary extra dimensions. Following Gower & Legendre (1986, Theorem 7*), this result can be obtained by adding a constant c to either the squared distances D_{hi}^2 or the original distances D_{hi}. This provides two methods for adjusting the original distances and correcting for their non-Euclidean behaviour.

* The present subsection corrects two misprints in theorem 7 of Gower & Legendre (1986).

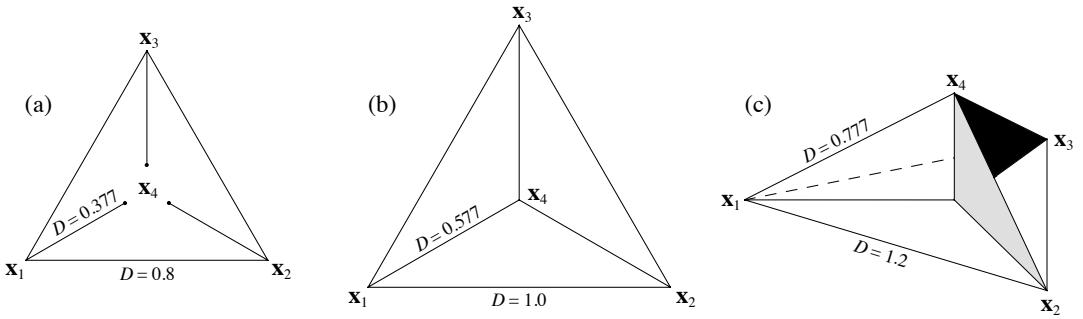


Figure 9.17 (a) Distances among four points constructed in such a way that the system cannot be represented in Euclidean space because the three lines going towards point $\mathbf{x}_4$ do not meet. (b) By adding a constant to all distances ($c_2 = 0.2$ in the present case), correction method 2 makes the system Euclidean; in this example, the distances can be associated with a representation of the points in two-dimensional space. (c) When increasing the distances further (adding again 0.2 to each distance in the present case), the system remains Euclidean but requires more dimensions for representation (three dimensions in this example).

• *Correction method 1* (derived from the work of Lingoes, 1971). — Add a constant to all squared distances D_{hi}^2, except those on the diagonal, creating a new matrix $\hat{\mathbf{D}}$ of distances $\hat{D}_{hi}$ through the following transformation:

$$\hat{D}_{hi} = \sqrt{D_{hi}^2 + 2c_1} \qquad \text{for } h \neq i \tag{9.46}$$

How to obtain c_1 is described a few lines below. Then proceed to the transformation of $\hat{\mathbf{D}}$ into matrix $\hat{\mathbf{A}}$ using eq. 9.40. The two operations may be combined into a single transformation producing the new matrix $\hat{\mathbf{A}} = [\hat{a}_{hi}]$ directly from the original distances D_{hi}:

$$\hat{a}_{hi} = -\frac{1}{2}\hat{D}_{hi}^2 = -\frac{1}{2}(D_{hi}^2 + 2c_1) = -\frac{1}{2}D_{hi}^2 - c_1 \qquad \text{for } h \neq i$$

Then, proceed with eq. 9.41 and recompute the PCoA. The constant to be added, c_1, is the *absolute value of the largest negative eigenvalue* obtained by analysing the original matrix $\mathbf{\Delta}_1$. Constant c_1 is also used in eq. 9.48 below. After correction, all non-zero eigenvalues are augmented by a value equal to c_1, so that the largest negative eigenvalue is now shifted to value 0. As a consequence, the corrected solution has two null eigenvalues (hence a maximum of $n - 2$ dimensions), or more if the matrix is degenerate. The constant c_1 is the smallest value that will produce the desired effect. Any value larger than c_1 would also eliminate all negative eigenvalues and make the system fully Euclidean, but it would also create a solution requiring more dimensions.

• *Correction method 2* (proposed by Cailliez, 1983). — Add a constant c_2 to all elements D_{hi} of matrix **D**, except those on the diagonal, creating a new matrix $\hat{\mathbf{D}}$ of distances $\hat{D}_{hi}$ through the transformation:

$$\hat{D}_{hi} = D_{hi} + c_2 \qquad \text{for } h \neq i \tag{9.47}$$

and proceed to the transformation of $\hat{\mathbf{D}}$ into matrix $\hat{\mathbf{A}}$ using eq. 9.40. The two operations may be combined into a single transformation producing the new matrix $\hat{\mathbf{A}} = [\hat{a}_{hi}]$ directly from the original distances D_{hi}:

$$\hat{a}_{hi} = -\frac{1}{2}(D_{hi} + c_2)^2 \qquad \text{for } h \neq i$$

Then, proceed with eq. 9.41 and recompute the PCoA. The constant to be added, c_2, is equal to the *largest positive eigenvalue* obtained by analysing the following special matrix, which is of order $2n$:

$$\begin{bmatrix} \mathbf{0} & 2\boldsymbol{\Delta}_1 \\ -\mathbf{I} & -4\boldsymbol{\Delta}_2 \end{bmatrix}$$

where **0** is a null matrix, **I** is an identity matrix, $\boldsymbol{\Delta}_1$ is the centred matrix defined by eqs. 9.40 and 9.41, and $\boldsymbol{\Delta}_2$ is a matrix containing values $(-0.5 D_{hi})$ centred using eq. 9.41. The order of each of these matrices is n. Beware: the special matrix is asymmetric. Press *et al*. (2007) describe an algorithm to compute the eigenvalues of such a matrix. Function ***eigen()*** of R can also compute them. The solution has two null eigenvalues (hence a maximum of $n - 2$ dimensions), or more if the matrix is degenerate. The constant c_2 is the smallest value that will produce the desired effect; any value larger than c_2 would also eliminate all negative eigenvalues and make the system fully Euclidean, but the solution would require more dimensions. Figure 9.17a-b shows the effect of adding constant c_2 to a non-Euclidean group of four points, and Fig. 9.17c shows the effect of adding a value larger than c_2.

The two correction methods do not produce the same Euclidean representation. This may be understood by examining the consequences of adding c_2 to the distances in **D**. When $\hat{\mathbf{D}}$ is transformed into $\hat{\mathbf{A}}$ (eq. 9.40), $(D_{hi} + c_2)$ becomes:

$$\hat{a}_{hi} = -0.5 (D_{hi} + c_2)^2 = -0.5 (D_{hi}^2 + 2c_2 D_{hi} + c_2^2) \qquad \text{for } h \neq i$$

The effect on $\hat{a}_{hi}$ does not only depend on the value of c_2 but it also varies with each value D_{hi}. This is clearly not the same as subtracting a constant from all a_{hi} values (i.e. correction method 1). The eigenvectors resulting from one or the other correction also differ from those resulting from a PCoA without correction for negative eigenvalues. The two correction methods, and PCoA without correction, thus correspond to different partitions of the variation because the total variance, given by the trace of centred matrix $\boldsymbol{\Delta}_1$, differs among methods.

How large must constants c_1 and c_2 be for coefficient $D_{14} = 1 - S_{17}$, which is important for the analysis of species abundance data? To answer this question, Legendre & Anderson (1999) simulated species abundance data matrices. After computing distance D_{14}, the correction constants (c_1 for method 1, c_2 for method 2) increased nearly linearly with the ratio (*number of sites : number of species*). In extremely species-poor ecosystems, corrections were the largest; for instance, with a ratio 20:1 (e.g. 200 sites, 10 species), c_1 was near 0.4 and c_2 was near 0.8. When the ratio was near 1:1 (i.e. number of sites ≈ number of species), c_1 was about 0.06 and c_2 was about 0.2. In species-rich ecosystems, corrections were small, becoming smaller as the species richness increased for a constant number of sites; with a ratio 1:2 for example (e.g. 100 sites, 200 species), c_1 was near 0.02 and c_2 was about 0.1. Results also depended to some extent on the data generation parameters.

To summarize, all methods for eliminating negative eigenvalues operate by making the small distances larger, compared to the large distances, in order to allow all triangles to close (Figs. 9.16, 9.17a and b). As explained above, the first approach consists in taking the square root of all distances; this reduces the largest distances more than the small ones. The other two approaches (described above as correction methods 1 and 2) involve adding a constant to all non-diagonal distances; small distances are proportionally more augmented than large distances. In correction method 1, a constant ($2c_1$) is added to the squared distances D^2_{hi} whereas in method 2 a constant (c_2) is added to the distances D_{hi} themselves.*

Numerical example 3. Consider the numerical example used in Subsection 7.4.2 to demonstrate the semimetric nature of the percentage difference (D_{14}). The data matrix contained 3 objects and 5 species. Matrix **D**, matrix $\mathbf{A} = [-0.5 D^2_{hi}]$, and matrix $\mathbf{\Delta}_1$ are:

$$\mathbf{D} = \begin{bmatrix} 0.00000 & 0.05882 & 0.60000 \\ 0.05882 & 0.00000 & 0.53333 \\ 0.60000 & 0.53333 & 0.00000 \end{bmatrix} \quad \mathbf{A} = \begin{bmatrix} 0.00000 & -0.00173 & -0.18000 \\ -0.00173 & 0.00000 & -0.14222 \\ -0.18000 & -0.14222 & 0.00000 \end{bmatrix} \quad \mathbf{\Delta}_1 = \begin{bmatrix} 0.04916 & 0.03484 & -0.08401 \\ 0.03484 & 0.02398 & -0.05882 \\ -0.08401 & -0.05882 & 0.14283 \end{bmatrix}$$

The trace of $\mathbf{\Delta}_1$ is 0.21597. The eigenvalues are: $\lambda_1 = 0.21645$, $\lambda_2 = 0.00000$, and $\lambda_3 = -0.00049$. The sum of the eigenvalues is equal to the trace.

For correction method 1, value $c_1 = 0.00049$ is subtracted from all non-diagonal values of **A** to give $\hat{\mathbf{A}}$, which is then centred (eq. 9.41) to give the corrected matrix $\mathbf{\Delta}_1$:

$$\hat{\mathbf{A}} = \begin{bmatrix} 0.00000 & -0.00222 & -0.18049 \\ -0.00222 & 0.00000 & -0.14271 \\ -0.18049 & -0.14271 & 0.00000 \end{bmatrix} \quad \mathbf{\Delta}_1 = \begin{bmatrix} 0.04949 & 0.03468 & -0.08417 \\ 0.03468 & 0.02430 & -0.05898 \\ -0.08417 & -0.05898 & 0.14315 \end{bmatrix}$$

The trace of the corrected matrix $\mathbf{\Delta}_1$ is 0.21694. The corrected eigenvalues are: $\lambda_1 = 0.21694$, $\lambda_2 = 0.00000$, and $\lambda_3 = 0.00000$. This Euclidean solution is one-dimensional.

* In the R language, function ***pcoa()*** in APE offers these two corrections.

For correction method 2, value $c_2 = 0.00784$, which is the largest eigenvalue of the special matrix, is added to all non-diagonal elements of matrix **D** to obtain $\hat{\mathbf{D}}$, which is then transformed into $\hat{\mathbf{A}}$ (eq. 9.40) and centred (eq. 9.41) to give the corrected matrix $\boldsymbol{\Delta}_1$:

$$\hat{\mathbf{D}} = \begin{bmatrix} 0.00000 & 0.06667 & 0.60784 \\ 0.06667 & 0.00000 & 0.54118 \\ 0.60784 & 0.54118 & 0.00000 \end{bmatrix} \quad \hat{\mathbf{A}} = \begin{bmatrix} 0.00000 & -0.00222 & -0.18474 \\ -0.00222 & 0.00000 & -0.14644 \\ -0.18474 & -0.14644 & 0.00000 \end{bmatrix} \quad \boldsymbol{\Delta}_1 = \begin{bmatrix} 0.05055 & 0.03556 & -0.08611 \\ 0.03556 & 0.02502 & -0.06058 \\ -0.08611 & -0.06058 & 0.14669 \end{bmatrix}$$

The trace of the corrected matrix $\boldsymbol{\Delta}_1$ is 0.22226. The corrected eigenvalues are: $\lambda_1 = 0.22226$, $\lambda_2 = 0.00000$, and $\lambda_3 = 0.00000$. This Euclidean solution is one-dimensional, as was the case with correction method 1.

Using the square root of coefficient D_{14}, matrices **D**, **A** and $\boldsymbol{\Delta}_1$ are:

$$\mathbf{D} = \begin{bmatrix} 0.00000 & 0.24254 & 0.77460 \\ 0.24254 & 0.00000 & 0.73030 \\ 0.77460 & 0.73030 & 0.00000 \end{bmatrix} \quad \mathbf{A} = \begin{bmatrix} 0.00000 & -0.02941 & -0.30000 \\ -0.02941 & 0.00000 & -0.26667 \\ -0.30000 & -0.26667 & 0.00000 \end{bmatrix} \quad \boldsymbol{\Delta}_1 = \begin{bmatrix} 0.08715 & 0.04662 & -0.13377 \\ 0.04662 & 0.06492 & -0.11155 \\ -0.13377 & -0.11155 & 0.24532 \end{bmatrix}$$

The trace of $\boldsymbol{\Delta}_1$ is 0.39739. The eigenvalues are: $\lambda_1 = 0.36906$, $\lambda_2 = 0.02832$, and $\lambda_3 = 0.00000$. No negative eigenvalue is produced using this coefficient. This Euclidean solution is two-dimensional.

If negative eigenvalues are present in a full-dimensional PCoA solution and no correction is made to the distances to eliminate negative eigenvalues, problems of interpretation arise. Since the eigenvectors $\mathbf{u}_k$ are scaled to length $\sqrt{\lambda_k}$, it follows that the axes corresponding to negative eigenvalues are not real, but complex. Indeed, in order for the sum of squares of the u_{ik}'s in an eigenvector $\mathbf{u}_k$ to be negative, the coordinates u_{ik} must be imaginary numbers. When some of the axes of the reference space are complex, the distances cannot be fully represented in Euclidean space, as in the example of Figs. 9.16 and 9.17a.

It is, however, legitimate to investigate whether the Euclidean approximation corresponding to the positive eigenvalues (i.e. the non-imaginary principal coordinates) provides a good representation, when no correction for negative eigenvalues is applied. Cailliez & Pagès (1976) have shown that such a representation is meaningful as long as the largest negative eigenvalue is smaller, in absolute value, than any of the m positive eigenvalues of interest for representation in reduced space (usually, the first two or three).

When there are no negative eigenvalues, the quality of the representation in a reduced Euclidean space with m dimensions can be assessed, as in principal component analysis (eq. 9.5), by the R^2-like ratio:

R^2-like ratio

$$\left(\sum_{k=1}^{m} \lambda_k \right) \Big/ \left(\sum_{k=1}^{c} \lambda_k \right) \qquad (\mathbf{9.5})$$

where c is the number of positive eigenvalues. This comes from the fact that the eigenvalues of a PCoA are the same (to a factor $n - 1$) as those of a PCA performed on the coordinates of the same points in the full-dimensional space of the principal coordinates, e.g. the object coordinates in Table 9.10. Cailliez & Pagès (1976) further suggested that, when negative eigenvalues are present, a correct estimate of the quality of a reduced-space representation can be obtained by the corrected R^2-like ratio:

Corrected R^2-like ratio

$$\frac{\left(\sum_{k=1}^{m} \lambda_k\right) + mc_1}{\left(\sum_{k=1}^{n} \lambda_k\right) + (n-1)\,c_1} \qquad \textbf{(9.48)}$$

where m is the dimensionality of the reduced space, n is the order of the distance matrix (total number of objects), and c_1 is the absolute value of the largest negative eigenvalue; c_1 was found in correction method 1 above. Equation 9.48 gives the same value as if correction method 1 had been applied to the distance matrix, the PCoA had been recomputed, and the quality of the representation had been calculated using eq. 9.5. All non-zero eigenvalues would be augmented by a value equal to $c_1 = |\lambda_n|$, producing the same changes to the numerator and denominator as in eq. 9.48.

5 — Ecological applications

Principal coordinate analysis is an ordination method of great interest to ecologists because the nature of ecological descriptors often makes it necessary to use other measures of resemblance than the Euclidean distance preserved by principal component analysis or the χ^2 distance preserved by correspondence analysis (Table 9.1). Ordination methods such as principal coordinate analysis and nonmetric multidimensional scaling (Section 9.4) provide Euclidean representations of point-objects for any distance measure selected by users.

Numerical example 1 (continued from Subsection 9.2.5). From the data shown in Fig. 9.10 and Table 9.7, four Q-mode distance measures were computed among sites (Table 9.11) to illustrate some properties of principal coordinate analysis.

• Row 1 of Table 9.11 — The Euclidean distance D_1 is a symmetrical coefficient. It is not ideal for species abundance data, and it is only used here for comparison. A principal coordinate analysis of this matrix led to 19 eigenvalues: three positive (accounting for 50, 41, and 9% of the variation, respectively) and 16 null. This was expected since the original data matrix contained three variables.

• Row 2 — Distance D_{14} is often applied to species abundance data. Like its one-complement S_{17}, it excludes double-zeros. Principal coordinate analysis of this distance matrix led to 19 eigenvalues: 11 positive, one null, and 7 negative. The distance matrix was corrected using method 1 of Subsection 9.3.4, which makes use of the largest negative eigenvalue. PCoA produced 17 positive and two null eigenvalues, the largest one accounting for 31% of the variation. The distance matrix was also corrected using method 2 of Subsection 9.3.4, which

Table 9.11 Distance matrices computed from the artificial data in Fig. 9.10 and Table 9.7. Each row in this table corresponds to the first row of a distance matrix, comparing site 1 to itself and to the 18 other sites. The remaining rows of the distance matrices are not shown to save space; readers can compute these matrices from the data in Table 9.7. Values are rounded to a single decimal place.

Sampling sites	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
D_1 (Euclidean)	0.0	1.0	3.0	6.0	7.0	6.1	3.6	4.1	7.0	8.1	7.1	4.6	4.6	7.1	8.1	7.1	4.1	2.2	1.4
$D_{14} = (1 - S_{17})$	0.0	0.3	0.6	0.8	0.8	0.8	0.7	0.7	0.8	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
$\sqrt{D_{14}}$	0.0	0.6	0.8	0.9	0.9	0.9	0.8	0.8	0.9	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
D_{16} (χ^2 distance)	0.0	0.0	0.0	0.0	0.0	0.3	0.8	1.6	2.1	2.4	2.3	2.2	2.2	2.3	2.4	2.4	2.4	2.4	2.4

makes use of the largest eigenvalue of the special matrix. PCoA also produced 17 positive and two null eigenvalues, the largest one accounting for 34% of the variation.

• Row 3 — Principal coordinate analysis was also conducted using the square root of coefficient D_{14}. The analysis led to 18 positive, one null, and no negative eigenvalues, the largest one accounting for 35% of the variation.

• Row 4 — A fourth distance matrix was computed using the χ^2 distance D_{16}, which excludes double-zeros. Principal coordinate analysis produced 19 eigenvalues: two positive (accounting for 4 and 36% of the variation, respectively) and 17 null. The χ^2 distance (D_{16}) is the coefficient preserved in correspondence analysis (CA, Section 9.2), which would also produce two positive eigenvalues with these data. Indeed, CA always produces one dimension less than the original number of species, or fewer in the case of degenerate matrices.

This example shows that different distance measures may lead to very different numbers of dimensions of the Euclidean representations. In the analyses reported here, the numbers of dimensions obtained were 3 for distance D_1, 11 for uncorrected D_{14} (not counting the complex axes corresponding to negative eigenvalues), 17 for D_{14} after correction by the largest negative eigenvalue, 18 for the square root of D_{14}, and 2 for D_{16}.

For the example data, the PCA ordination (Fig. 9.11a, b) is identical to the ordination that would have been obtained from PCoA of a matrix of Euclidean distances among sites, as shown in Numerical example 2 (Subsection 9.3.2). In the same way, the ordination of sites in the CA plot (Fig. 9.11c), which used scaling type 1, is similar to a PCoA ordination that would be obtained from a matrix of χ^2 distances (D_{16}) among sites. The ordinations (Fig. 9.18) obtained from distance coefficients D_{14} and $\sqrt{D_{14}}$ are also of interest because these coefficients are often used to analyse community composition data; they are illustrated in Fig. 9.18a-d. The ordinations produced by these coefficients are quite similar to each other and present horseshoes, which are not as pronounced as in PCA because coefficients D_{14} and $\sqrt{D_{14}}$ exclude double-zeros from the calculations. In Fig. 9.18a (coefficient D_{14}), sites 6 to 14 form an arch depicting the three-species gradient, with arms extending in a perpendicular

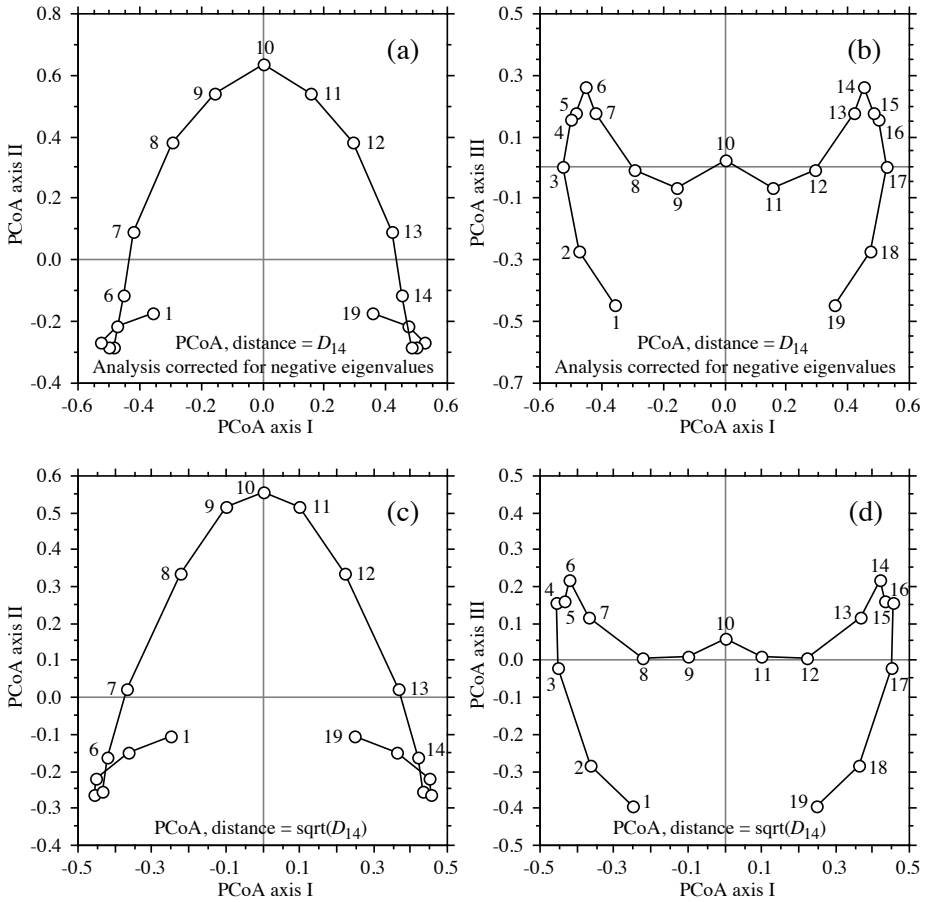


Figure 9.18 Principal coordinate ordinations of the data in Fig. 9.10 and Table 9.7. Distance D_{14}, analysis corrected for negative eigenvalues: (a) PCoA axes I and II ($\lambda_1 = 30.8\%$, $\lambda_2 = 18.6\%$), (b) axes I and III ($\lambda_1 = 30.8\%$, $\lambda_3 = 8.3\%$). Distance $\sqrt{D_{14}}$: (c) PCoA axes I and II ($\lambda_1 = 34.5\%$, $\lambda_2 = 22.9\%$), (d) axes I and III ($\lambda_1 = 34.5\%$, $\lambda_3 = 10.5\%$).

direction (axis III, Fig. 9.18b) to account for the dispersion of sites 1 to 5 and 15 to 19, each group containing one species only, as shown in Table 9.7. The ordination produced by coefficient $\sqrt{D_{14}}$ is very similar to the above (Fig. 9.18c-d). There are two advantages to $\sqrt{D_{14}}$ over D_{14}, though: $\sqrt{D_{14}}$ never produces negative eigenvalues and, in the present case at least, the ordination explains more variation than D_{14} in two or three dimensions.

Only the Euclidean distance and derived coefficients lead to a number of principal axes equal to the original number of descriptors. Other coefficients may produce fewer, or more axes. The dimensionality of a principal coordinate space is a function of the

number of original descriptors, mediated through the distance measure that was selected for the analysis.

There are many applications of principal coordinate analysis in the ecological literature. This method may be used in conjunction with clustering techniques; an example is presented in Ecological application 10.1. Direct applications of the method are summarized in Ecological applications 9.3a and 9.3b. The application of principal coordinate analysis to the clustering of large numbers of objects is discussed in Subsection 8.7.3.

Ecological application 9.3a

Field & Robb (1970) studied the molluscs and barnacles from a rocky shore (21 quadrats) in False Bay, South Africa, in order to determine the influence of factors *emergence* (the height on the shore relative to the mean sea level) and *wave* on these communities. Quadrats 1 to 10, on a transect parallel to the shoreline, differed in their exposure to wave action; quadrats 11 to 21, located on a transect at right angle to the shoreline, covered the spectrum between the mean high and mean low waters of spring tides. 79 species were enumerated, one reaching 10864 individuals in a single quadrat. When going up the shore, quadrats had progressively larger numbers of individuals and smaller numbers of species. This illustrates the principle that increasing environmental stress (here, the *emergence* factor) is accompanied by decreasing diversity. It also shows that the few species that can withstand a high degree of stress do not encounter much interspecific competition and may therefore become very abundant.

The same principal coordinate ordination could have been obtained by estimating species abundances with a lesser degree of precision, e.g. using classes of abundance. Table 7.4 gives the association measures that would have been appropriate for such data.

Species abundances (y'_{ij}) were first normalized by logarithmic transformation $y''_{ij} = \log_e (y'_{ij} + 1)$, and centred ($y_{ij} = y''_{ij} - \bar{y}_i$), to form matrix $\mathbf{Y} = [y_{ij}]$ containing the data to be analysed. Scalar products among quadrat vectors were used as measures of similarity:

$$\mathbf{S}_{n \times n} = \mathbf{Y}_{n \times p} \mathbf{Y}'_{p \times n}$$

Principal coordinates were computed using a variant procedure proposed by Orlóci (1966). Figure 9.19 displays the ordination of quadrats 1 to 19 in the space of the first two principal coordinates. The ordination was also calculated including quadrats 20 and 21 but, since these came from the highest part of the shore, they introduced so much variation in the analysis that the factor *emergence* dominated the first two principal coordinates. For the present ecological application, only the ordination of quadrats 1 to 19 is shown. The authors looked for a relationship between this ordination and the two environmental factors, by calculating Spearman's rank correlation coefficients (eq. 5.3) between the ranks of the quadrats on each principal axis and their ranks on the two environmental factors. This showed that the first principal axis had a significant correlation with elevation with respect to the shoreline (*emergence*), and the second axis was significantly related to *wave action*. The authors concluded that PCoA is well adapted to the study of ecological gradients, provided that the data set is fairly homogeneous. (Correspondence analysis, described in Section 9.2, would have been another appropriate way of obtaining an ordination of these quadrats.)

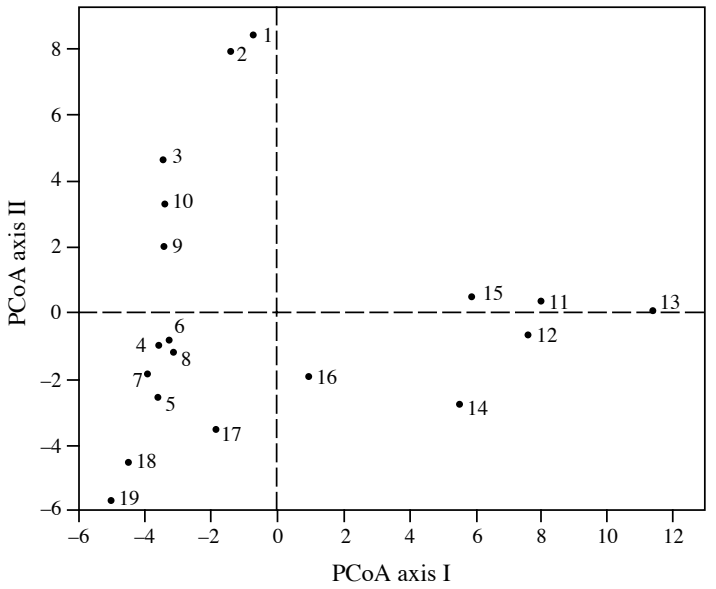


Figure 9.19 Ordination of quadrats 1 to 19 in the space of the first two principal coordinates (PCoA axes I and II). Modified from Field & Robb (1970).

Ecological application 9.3b

Ardisson *et al*. (1990) investigated the spatio-temporal organization of epibenthic communities in the Estuary and Gulf of St. Lawrence, an area ca. 1150 × 300 km. Quantitative data were obtained over 8 years, between 1975 and 1984, from 161 collectors (summer navigation buoys) moored yearly from May through November, and retrieved by the Canadian Coastguard before winter ice formation.

Each year was represented by a data table of 161 sites (buoys) × 5 dominant species (dry biomass). A similarity matrix among sites was computed for each year separately, using the asymmetrical form of the Gower similarity coefficient (S_{19}). The eight yearly matrices were compared to one another using Mantel statistics (Subsection 10.5.1). A principal coordinate analysis (Fig. 9.20) was conducted on the resulting matrix of (1 – Mantel) statistics to determine whether year-to-year differences were random or organized. The among-year pattern of dispersion suggested the existence of a cycle of variation whose length was about equal to the duration of the study. This cycle might represent the response of the Estuary-Gulf system, as an integrated unit, to external inputs of auxiliary energy, although the specific causal process, physical or biotic, remains unknown.

Mantel statistic

The same type of analysis as in this ecological application, i.e. comparing several data matrices about the same objects, could be based on RV coefficients (eq. 11.66) computed between all pairs of data matrices. The corresponding distance-like coefficients (1 – RV) would be assembled in a square distance matrix and PCoA would be computed on that matrix to obtain an ordination of the type illustrated in Fig. 9.20.

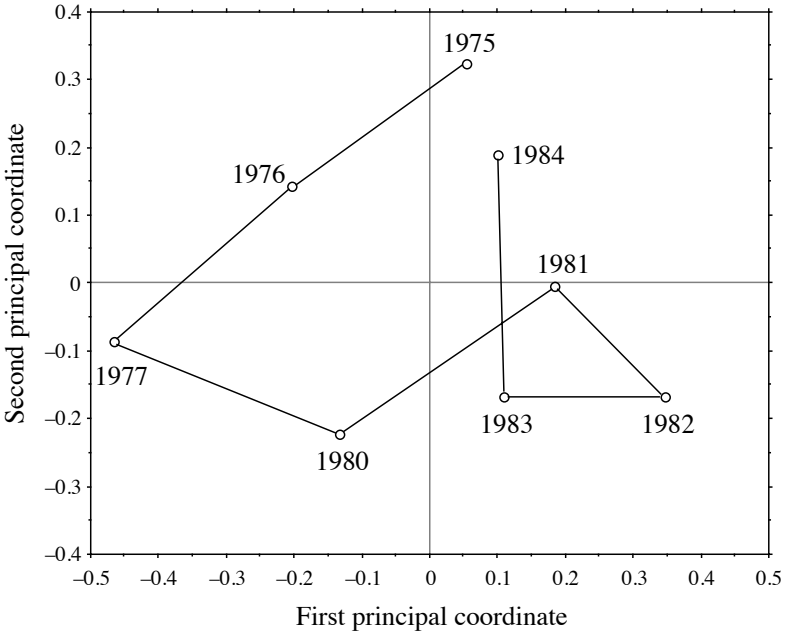


Figure 9.20 Among-year variability illustrated in the space of the first two principal coordinates, obtained from analysing a matrix of Mantel statistics comparing yearly similarity matrices. Recomputed from the Mantel statistic values provided in Fig. 8 of Ardisson *et al*. (1990).

This last ecological application showed that the usefulness of principal coordinate analysis is not limited to projecting, in reduced space, classical resemblance matrices among objects. In that example, the relationships among data tables, as expressed by Mantel statistics, were represented in a Euclidean space using PCoA. The method may actually be applied to any type of symmetric resemblance matrices. This includes cases where the measures of resemblance are obtained directly from observation (e.g. interaction matrices in behavioural studies) or from laboratory work (DNA hybridisation results, serological data, etc.). If the resulting matrix is non-symmetric, it may be decomposed into a symmetric and a skew-symmetric component (Section 2.3), which can be analysed separately by PCoA.

6 — Algorithm

Principal coordinate analysis is easy to compute for any distance matrix, using a standard eigenanalysis function. Follow the steps summarized in Table 9.12.

Table 9.12 Computing principal coordinates.

a) Centre the distance matrix

Transform and centre the distance matrix following Gower's method (eqs. 9.40 and 9.41).

b) Compute the eigenvalues and eigenvectors

Use an eigen-decomposition function†.

c) Final scaling

Scale each eigenvector k to length $\sqrt{\lambda_k}$ to obtain the principal coordinates.

† For a matrix of Euclidean distances D_1, the eigenvalues obtained from PCoA are larger than those of PCA by a factor $(n - 1)$.

9.4 Nonmetric multidimensional scaling (nMDS)

The reduced-space ordination methods of the previous sections produced an ordination (scaling) of the objects in full-dimensional space. Users could then select the first few dimensions and check how well the distance relationships among the objects were preserved in that reduced space. There are cases, however, where the exact preservation of the distances among objects is not of primary importance, the priority being instead the representation of the objects in a small and specified number of dimensions, usually two or three. In such cases, the objective is to plot dissimilar objects far apart in the ordination space and similar objects close to one another. This is called the *preservation of ordering relationships* among objects. The method to do so is called *nonmetric multidimensional scaling* (nMDS, or simply MDS). It was devised by psychometricians Shepard (1962, 1966) and Kruskal (1964a, b). Programs for nMDS were originally distributed by Bell Laboratories in New Jersey, where the method originated; see Carroll (1987) for a review. The method is now available in several major (SPSS, SAS, SYSTAT, etc.) and specialized computer packages as well as in R*. A useful reference is the book of Kruskal & Wish (1978). Relationships between nMDS and other forms of ordination have been described by Gower (1987). Extensions of nMDS to several matrices, weighted models, the analysis of preference data, etc. are discussed by Young (1985) and Carroll (1987). A form of *hybrid scaling*, combining metric and nonmetric scaling criteria, was proposed by Faith *et al*. (1987); it was further explained in Belbin (1991) and is available in packages DECODA (written by Peter R. Minchin) and PATN.

Like principal coordinate analysis (PCoA), nMDS is not limited to Euclidean distance matrices; it can produce ordinations of objects from any distance matrix. The method can also proceed with missing distances — actually, the more missing distances there are, the easier the computations — as long as there are enough measures left to position each object with respect to a few of the others. This feature makes it a method of choice for the analysis of matrices obtained by direct observation (e.g. behaviour studies) or laboratory assays, where missing pairwise distances often occur. Some programs can handle non-symmetric distance matrices, for which they provide a compromise solution between distances in the upper and lower triangular parts of the matrix. Contrary to PCA, PCoA, or CA, which are eigenvector-based methods, nMDS calculations do not maximize the variability associated with individual axes of the ordination; nMDS axes are arbitrary, so that plots may arbitrarily be rotated, centred, or inverted. Reasons for this will become clear from the presentation of the method.

Consider a distance matrix $\mathbf{D}_{n\times n} = [D_{hi}]$ computed using a measure appropriate to the data at hand (Chapter 7). Matrix **D** may also result from direct observations, e.g. affinities among individuals or species found in serological, DNA pairing, or behavioural studies; these matrices may be non-symmetric. Nonmetric multidimensional scaling of matrix **D** may be summarized in the following steps.

1) Specify the number m of dimensions chosen *a priori* for scaling the objects. The output will provide coordinates of the n objects on m axes. If several configurations for different numbers of dimensions are sought — say, 2, 3, 4, and 5 dimensions, they must be computed separately. Several programs actually allow solutions to cascade from high to low numbers of dimensions — for instance from 4 to 3 to 2 to 1.

2) Construct an initial configuration of the objects in m dimensions, to be used as a starting point for the iterative adjustment process of steps 3 to 7. The way this initial configuration is chosen is critical because the solution on which the algorithm eventually converges depends to some extent on the initial positions of the objects. The same problem was encountered with K-means partitioning (Section 8.8); the "space of

---

* nMDS is available in R functions listed in Section 9.5. It is also found in the following commercially available packages (list not exhaustive):

- NTSYSpc. Distribution: see footnote in Section 7.8.
- PATN. Distribution: see footnote in Section 7.8.
- PRIMER. That package was developed by M. R. Carr and K. R. Clarke at the Plymouth Marine Laboratory, Prospect Place, West Hoe, Plymouth PL1 3DH, Great Britain.
- PC-ORD. Distribution: see footnote in Section 11.7. Besides nMDS, PC-ORD contains many other methods of multivariate ecological data analysis.

Local minimum

solutions" may contain several local minima besides the overall minimum (Fig. 8.17). The usual solutions to this problem are the following:

- Run the program several times, starting from different random initial placements of the objects. The solution minimizing the objective function (step 5) is retained.

- Initiate the run from an ordination obtained using another method, e.g. PCoA.

- If the data are thought to be spatially structured and the geographic positions of the objects are known, these geographic positions may be used as the starting configuration for nMDS of a matrix **D** computed from the data.

- Work step by step from higher to lower dimensionality. Compute, for instance, a first nMDS solution in 5 dimensions from a random initial placement of the objects. Note the stress value (eqs. 9.49 to 9.51), which should be low because the high number of dimensions imposes little constraint to the distances. Use 4 of the 5 dimensions so obtained as the initial configuration for a run in 4 dimensions, and so forth until the desired number (m) of ordination dimensions is reached.

3) Calculate a matrix of fitted distances d_{hi} in the ordination space, using one of Minkowski's metrics (D_6, eq. 7.43). Most often, one chooses the second degree of Minkowski's metric, which is the Euclidean distance. (a) In the first iteration, distances d_{hi} are computed from the initial (often random) configuration. (b) In subsequent iterations, the configuration is that obtained in step 6.

Shepard diagram

4) Consider the Shepard diagram (Figs. 9.1 and 9.21b) comparing the fitted distances d_{hi} to the empirical (i.e. original) distances D_{hi}. Regress d_{hi} on D_{hi}. Values forecasted by the regression line are called $\hat{d}_{hi}$. The choice of the type of regression is left to the users, given the choices implemented in the computer program. Usual choices are the linear, polynomial, or monotone regressions (also called "nonparametric", although there are other types of nonparametric regression methods).

Monotone regression is a step-function constrained to always increase from left to right (Fig. 9.21b); this is a common choice in nMDS. A monotone regression is equivalent to a linear regression performed after monotonic transformation of the original distances D_{hi}, so as to maximize the linear relationship between D_{hi} and d_{hi}. The regression is fitted by least squares.

Tied values

If there are tied values among the empirical distances, Kruskal (1964a, b) proposed two approaches that may be followed in monotone regression. Ties are likely to occur when the empirical distances D_{hi} are computed from a table of raw data using one of the coefficients described in Chapter 7; they are less likely to occur when distances result from direct observations. In Fig. 9.21b, for instance, there are ties for several of the values on the abscissa; the largest number of ties is found at $D = D_{max} = 1$.

- In Kruskal's *primary approach*, one accepts the fact that, if an empirical distance D_{hi} corresponds to different fitted values d_{hi}, it also corresponds to different forecasted

values $\hat{d}_{hi}$. Hence the monotone regression line is allowed to go straight up in a column of tied values, subject to the constraint that the regression line is not allowed to decrease compared to the previous values D. The monotone regression line is not a mathematical function in that case, however. In order to insure monotonicity, the only constraint on the $\hat{d}_{hi}$ values is:

$$\text{when } D_{gi} < D_{hi}, \text{ then } \hat{d}_{gi} \leq \hat{d}_{hi}$$

• In the *secondary approach*, the forecasted value $\hat{d}_{hi}$ is the same for all fitted distances d_{hi} that are tied to a given empirical distance value D_{hi}. To insure monotonicity, the constraints on the $\hat{d}_{hi}$ values are:

$$\text{when } D_{gi} < D_{hi}, \text{ then } \hat{d}_{gi} \leq \hat{d}_{hi}$$

$$\text{when } D_{gi} = D_{hi}, \text{ then } \hat{d}_{gi} = \hat{d}_{hi}$$

In this approach, the least-squares solution for $\hat{d}_{hi}$ is the mean of the tied d_{hi}'s when considering a single value D_{hi}. The vertical difference in the diagram between d_{hi} and $\hat{d}_{hi}$ is used as the contribution of that point to the stress formula, below. In Fig. 9.21b, the secondary approach is applied to all tied values found for $D_{hi} < (D_{max} = 1)$, and the primary approach when $D_{hi} = D_{max} = 1$.

Computer programs may differ in the way they handle ties. This may cause major differences between reported stress values corresponding to the final solutions, although the final configurations of points are usually very similar from program to program, except when two programs identify different final solutions having very similar stress values.

A reduced-space scaling would be perfect if all points in the Shepard diagram fell exactly on the regression line (straight line, smooth curve, or step-function); the rank ordering of the fitted distances d_{hi} would be exactly the same as that of the original distances D_{hi} and the value of the objective function (step 5) would be zero.

Objective function

5) Measure the goodness-of-fit of the regression using an objective function. All objective functions used in nMDS are based on the sum of the squared differences between fitted values d_{hi} and the corresponding values $\hat{d}_{hi}$ forecasted by the regression function; this is the usual sum of squared residuals of regression analysis (least-squares criterion, Subsection 10.3.1). Several variants have been proposed and are used in nMDS programs:

$$\textit{Stress} \text{ (formula 1)} = \sqrt{\sum_{h,i} (d_{hi} - \hat{d}_{hi})^2 / \sum_{h,i} d_{hi}^2} \qquad \textbf{(9.49)}$$

$$\textit{Stress} \text{ (formula 2)} = \sqrt{\sum_{h,i} (d_{hi} - \hat{d}_{hi})^2 / \sum_{h,i} (d_{hi} - \bar{d})^2} \qquad \textbf{(9.50)}$$

$$Sstress = \sqrt{\sum_{h,i} (d_{hi}^2 - \hat{d}_{hi}^2)^2} \quad \textbf{(9.51)}$$

The denominators in the two *Stress* formulas (eqs. 9.49 and 9.50) are scaling terms that make the objective functions dimensionless and produce *Stress* values between 0 and 1. These objective functions may apply the square root, or not, without changing the issue; a configuration that minimizes these objective functions would also minimize the non-square-rooted forms. Other objective criteria, such as *Strain*, have been proposed. All objective functions measure how far the reduced-space configuration is from being monotonic to the original distances D_{hi}. Their values are only relative, measuring the decrease in lack-of-fit between iterations of the calculation procedure.

Steepest descent

6) Improve the configuration by moving it slightly in a direction of decreasing stress. This is done by a numerical optimization algorithm called the *method of steepest descent*; the method is explained, for instance, in *Numerical Recipes* (Press *et al*., 2007) and in Kruskal (1964b). The direction of steepest descent is the direction in the space of solutions along which stress is decreasing most rapidly. This direction is found by analysing the partial derivatives of the stress function (Carroll, 1987). The idea is to move points in the ordination plot to new positions that are likely to decrease the stress most rapidly.

7) Repeat steps 3 to 6 until the objective function reaches a small, predetermined value (tolerated lack-of-fit), or until convergence is achieved, i.e. until it reaches a minimum and no further progress can be made. The coordinates calculated at the last passage through step 6 become the coordinates of the *n* objects in the *m* dimensions of the multidimensional scaling ordination.

8) Most nMDS programs rotate the final solution using principal component analysis, for easier interpretation.

In most situations, users of nMDS decide that they want a representation of the objects in two or three dimensions, for illustration or other purpose. In some cases, however, one wonders what the "best" number of dimensions would be for a data set, i.e. what would be the best compromise between a summary of the data and an accurate representation of the distances. As pointed out by Kruskal & Wish (1978), determining the dimensionality of an nMDS ordination is as much a substantive as a statistical question. The substantive aspects concern the interpretability of the axes, ease of use, and stability of the solution. The statistical aspect is easier to approach since stress may be used as a guide to dimensionality. Plot the *stress* values as a function of *dimensionality* of the solutions, using one of the stress formulas above (eqs. 9.49 to 9.51). Since stress decreases as dimensionality increases, choose for the final solution the dimensionality where the change in stress becomes small.

For species count data, Faith *et al*. (1987) have shown, through simulations, that the following strategy yields informative ordination results: (1) standardize the data by

dividing each value by the maximum abundance for that species in the data set; (2) use the Steinhaus (S_{17}) or the Kulczynski (S_{18}) similarity measure; (3) compute the ordination by nMDS.

Besides the advantages mentioned above for the treatment of nonmetric distances or non-symmetric matrices (see also Sections 2.3 and 8.10 on this topic), Gower (1966) pointed out that nMDS can summarize distances in fewer dimensions than principal coordinate analysis (i.e. lower stress in, say, two dimensions). Results of the two methods may be compared by examining Shepard diagrams of the results obtained by PCoA and nMDS, respectively. If the scatter of points in the Shepard diagram for PCoA is narrow, as in Fig. 9.1a or b, the reduced-space ordination is useful in that it correctly reflects the relative positions of the objects. If the scatter is wide or nearly circular (Fig. 9.1c), the ordination diagram is of little use and one may try nMDS to find a more satisfactory solution in a few dimensions. A PCoA solution remains easier to compute in most cases, however, because it does not require multiple runs, and it is obtained using a direct eigenanalysis algorithm instead of an iterative procedure.

Numerical example 1 (continued from Subsections 9.2.5 and 9.3.5)**.** The percentage difference distance matrix (D_{14}) computed in Table 9.11 was subjected to nMDS analysis using the package DECODA written by Peter R. Minchin. This nMDS program uses Stress formula 1 (eq. 9.49). Repeated runs, using $m = 2$ dimensions but different random starting configurations, produced very similar results; the best one had a stress value of 0.0181 (Fig. 9.21a).

Kruskal's secondary approach, explained with computation step 4 above, was used in Fig. 9.21b for all tied values found when $D_{hi} < D_{max}$, while the primary approach was used when $D_{hi} = D_{max} = 1$. The rationale for this follows from the fact that the empirical distances D_{hi} are blocked by an artificial ceiling D_{max} of the distance function, over which they cannot increase. So, pairs of sites tied at distance $D_{max} = 1$, for which d_{hi} is larger than the previous value $\hat{d}$, are not expected to be the same distance apart in the ordination. Hence these values should not contribute to the stress despite their ties.

Using $\sqrt{D_{14}}$ as the distance measure, instead of D_{14}, produced an identical ordination, since nMDS is invariant to monotonic transformations of the distances. The stress value did not change either, because the square root transformation of D_{14} affects only the abscissa of Fig. 9.21b, whereas the stress is computed along the ordinate. The arch effect found in Fig. 9.18a does not appear in Fig. 9.21a. The horizontal axis of the nMDS ordination reproduces the original gradient almost perfectly in this example.

Points in an nMDS plot may be rotated, translated, inverted, or scaled *a posteriori* in any way considered appropriate to achieve maximum interpretability or to illustrate the results. This may be done either by hand or, for example, through canonical analysis of the nMDS axes with respect to a set of explanatory variables (Chapter 11).

With the present data, a one-dimensional ordination (stress = 0.1089) perfectly reconstructed the gradient of sites 1 to 19; the same ordination was always obtained when repeating the run from different random starting configurations and cascading from 3 to 2 to 1 dimensions. This configuration, and the low stress value, were hardly ever obtained when performing the nMDS ordination directly in one dimension, without the cascading procedure.

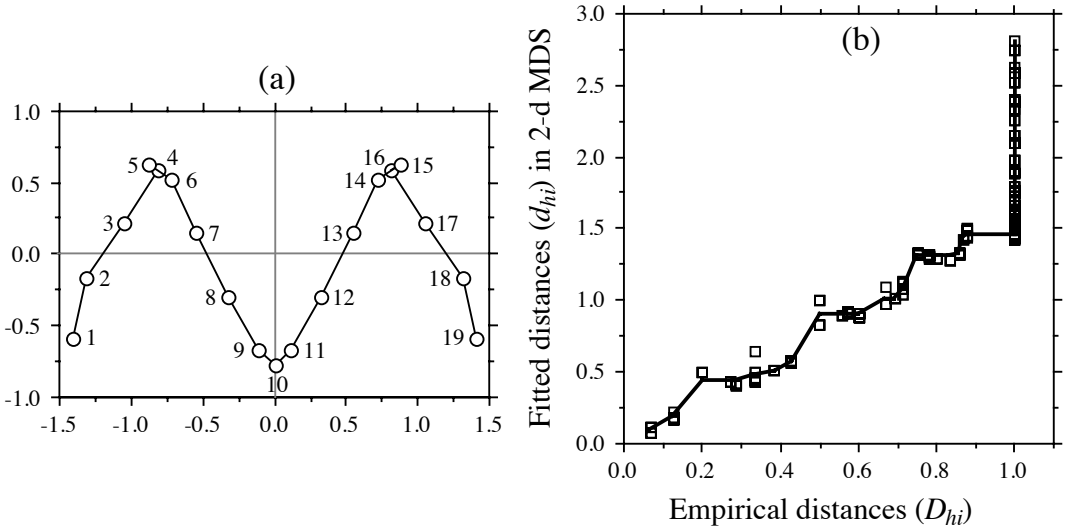


Figure 9.21 (a) nMDS ordination (2-dimensional) of the D_{14} distance matrix in Table 9.11. Sampling sites are numbered as in Fig. 9.10 and Table 9.7. (b) Shepard diagram of the final solution (2-dimensional) showing the monotone regression line fitted by nonparametric regression. The scatter about the line is measured by a stress function (eq. 9.49 to 9.51).

Ecological application 9.4a

Sprules (1980) used nonmetric multidimensional scaling to analyse seasonal changes in zooplankton assemblages at a site located in Lake Blelham, in the Lake District of northern England, and in two experimental enclosures built in that lake. The three sites were surveyed on a weekly basis from June to December 1976. nMDS was preferred to PCA because the responses of the species to environmental gradients could not be assumed to be linear.

For each site, points in the nMDS ordination diagram were connected in chronological order to reflect the seasonal changes in faunal composition. The plot (not reproduced here) is therefore of the same type as Fig. 12.24. In one of the enclosures, the assemblage oscillated about a mean value without any clear cycle; small-size species dominated the assemblage. In the other enclosure and in the lake, changes were more directional; at these sites, predators were more abundant. Based on available evidence, Sprules concluded that the differences observed between the two patterns of seasonal change were related to differences in predation intensity, quality of food available to herbivores, and nutrient dynamics.

Ecological application 9.4b

Redford *et al*. (2010) studied the bacteria living on the surfaces of leaves of 56 tree species found on the University of Colorado campus in Boulder, Colorado, USA. The bacterial communities were characterized by barcoded pyrosequencing. The authors analysed the intra-

and inter-tree-species variation in bacterial community composition. They found the bacterial communities to be more similar within a given tree species and among closely related species than among phylogenetically distant host species. The results were illustrated by nMDS plots computed from matrices of phylogenetic and community ecology distances among bacterial communities. The authors also compared the bacterial communities found on *Pinus ponderosa* needles from different locations with geographic distances ranging from 10 m to more than 10000 km; the bacterial floras of the leaves of three other tree species from Boulder were included in this analysis. The authors found that the bacterial communities were more similar among *P. ponderosa* trees from different locations than among different host species from the University of Colorado campus in Boulder.

Ecological application 9.4c

The relationship between snake community composition (46 sampling sites, 43 species) and environmental variables was investigated by de Fraga *et al*. (2011) in a 25 km^2 portion of Reserva Ducke near Manaus (Amazonas, Brazil) in the rain forest. The authors computed the first nMDS axis of a distance matrix based on the presence-absence of the species (Chao index, Subsection 7.3.4), and related it to the distance from streams in a dispersion diagram. The plot showed that snake community composition was fairly similar among sites close to streams, but varied much more for distances greater than 100 m from streams. Direct ordination of presence and absence data for 26 common snake species across all plots indicated a gradual substitution of species with distance from the streams.

Many ecological applications of nonmetric multidimensional scaling are found in the ecological literature. Two papers are especially interesting: Whittington & Hughes (1972; Ordovician biogeography from the analysis of trilobite fauna), and Fasham (1977; comparison of nonmetric multidimensional scaling, principal coordinate analysis and correspondence analysis for the ordination of simulated coenoclines and coenoplanes). Ecological application 12.6b (Subsection 12.6.5) features a nMDS plot.

9.5 Software

All general-purpose statistical packages offer principal component analysis, but not with options for the scalings used by ecologists. Among the specialized packages recommended for ecological analysis are CANOCO* (ter Braak & Smilauer, 1998 and later editions), PC-ORD and SYN-TAX 2000. For distribution of these programs, see Section 11.7.

* CANOCO was often referred to in this chapter. It was the first program to offer ecologists a whole array of simple and canonical ordination methods and it is still the reference for developers of ordination programs.

Several R-language packages offer functions for ordination of multivariate data.

1. Principal component analysis (PCA). — Functions ***dudi.pca()*** in ADE4, ***PCA()*** in FACTOMINER, ***pca()*** in LABDSV, ***prcomp()*** in STATS, and ***rda()*** in VEGAN compute PCA. Compare the PCA eigenvalues to the broken-stick model (eq. 9.16): ***PCAsignificance()*** in BIODIVERSITYR and ***bstick()*** in VEGAN. A PCA biplot with an equilibrium contribution circle is drawn by ***ordiequilibriumcircle()*** of BIODIVERSITYR. Fuzzy PCA and CA are available in ***dudi.fca()*** of ADE4.

2. Correspondence analysis (CA). — Functions ***dudi.coa()*** of ADE4, ***ca()*** of CA, ***CA()*** of FACTOMINER, ***corresp()*** of MASS and ***cca()*** of VEGAN compute CA. Multiple correspondence analysis (MCA): functions ***dudi.acm()*** of ADE4, ***mjca()*** of CA, ***MCA()*** of FACTOMINER, and ***mca()*** of MASS.

3. Principal coordinate analysis (PCoA). — Functions ***dudi.pco()*** of ADE4, ***pcoa()*** of APE, ***pco()*** of ECODIST, ***pco()*** of LABDSV, ***pcoa()*** and ***pcoa.all()*** of PCNM, ***cmdscale()*** of STATS, and ***wcmdscale()*** of VEGAN compute PCoA. Function ***pcoa.all()*** of package PCNM allows the computation of principal coordinates from distance matrices that have non-zero diagonals, which will be useful in Section 14.2. It also contains an option to output the eigenvectors corresponding to negative eigenvalues; in that case, the eigenvectors are not scaled to lengths of $\sqrt{\lambda_k}$ since that would produce complex vectors, but are kept normalized to lengths 1. Function ***is.euclid()*** of ADE4 checks the Euclidean nature of distance matrices; see Tables 7.2 and 7.3.

4. Nonmetric multidimensional scaling (nMDS). — Functions ***nmds()*** of ECODIST, ***nmds()*** of LABDSV, ***isoMDS()*** of MASS, and ***metaMDS()*** of VEGAN compute nMDS.

R functions for PCA and CA that follow and illustrate the algebra described in Sections 9.1 and 9.2 are available on the page http://numericalecology.com/rcode/.

Chapter

10 *Interpretation of ecological structures*

10.0 Ecological structures

The previous chapters explained how to use the techniques of clustering and ordination to investigate relationships among objects or descriptors. What do these analyses contribute to the understanding of ecological phenomena? Ecological applications in Chapters 8 and 9 have shown how clustering and ordination can synthesize the variability of the data and present it in a format that is easily amenable to interpretation. It often happens, however, that researchers who are using these relatively sophisticated methods do not go beyond the description of the structures of multidimensional data matrices, in terms of clusters or gradients. The descriptive phase must be followed by interpretation, which is conducted using either the descriptors that were used to evidence the structure, or other ecological descriptors that have not yet been involved in the analysis.

Structure

From the previous chapters, it should be clear that the *structure* of a data matrix is the organization of the objects, or descriptors, along gradients in a continuum, or in the form of subsets (clusters). This organization characterizes the data matrix, and it is derived from it. The first phase of multidimensional analysis (i.e. clustering or/and ordination) thus consists in characterizing the data matrix in terms of a simplified structure. In a second phase, ecologists may use this structure to interpret the phenomenon that underlies the data matrix. To do so, analyses are conducted to quantify the relationships between the structure of the data matrix and potentially explanatory descriptors. The methods that are most often used for interpreting ecological structures are described in the present chapter and in Chapter 11.

During the interpretation phase, one must assume that the analysis of the structure has been conducted with care, using measures of association that were appropriate to the objects and/or descriptors of the data matrix (Chapter 7) as well as analytical methods that corresponded to the objectives of the study. Ordination (Chapter 9) is used when gradients are sought, and clustering (Chapter 8) when one is looking for a

partition of the objects or descriptors into subsets. When the gradient is a function of a single or a pair of ordered descriptors, the ordination may be plotted in the original space of the descriptors. When the gradient results from the combined action of several descriptors, the ordination must be carried out in a reduced space using the methods discussed in Chapter 9. It may also happen that an ordination is used as a basis for visual clustering. Section 10.1 discusses the combined use of clustering and ordination to optimize the partition of objects or descriptors.

Explanation
Forecasting
Prediction

The interpretation of structures, in ecology, has three main objectives: (1) *explanation* (often called *discrimination*) of the structure of one or several descriptors, using the descriptors at the origin of the structure or, alternatively, a set of other descriptors that may potentially explain the structure; (2) *forecasting* of one or several descriptors (which are the response, or dependent, variables: Box 1.1), using a number of other descriptors (called the explanatory, or independent, variables); (3) *prediction* of one or several descriptors, using descriptors that can be manipulated experimentally or naturally exhibit environmental variation. The terms *forecasting* and *prediction*, which are not equivalent (Subsection 10.2.2), are often confused in the ecological and statistical literatures. Each of the above objectives covers a large number of numerical methods, which correspond to various levels of precision of the descriptors involved in the analysis.

Section 10.2 reviews the methods available for interpretation. The next sections are devoted to some of the methods introduced in Section 10.2. Regression and other scatterplot smoothing methods are discussed in Section 10.3. Section 10.4 deals with path analysis, which is used to assess causal relationships among quantitative descriptors. Section 10.5 discusses some methods developed to test the relationship between association or data matrices.

10.1 Clustering and ordination

Section 8.2 showed that single linkage clustering accurately accounted for the relationships between highly similar objects. However, due to its tendency to chaining, single linkage agglomeration is not very suitable for investigation of ecological questions. Because ecological data generally form a continuum in A-space (Fig. 7.2), it is often informative to use single linkage clustering in conjunction with an ordination of the objects. In the full multidimensional ordination space of principal component analysis (Section 9.1), Euclidean distances among the main clusters of objects are the same as in the original A-space. Other ordination methods (Sections 9.2 to 9.4) may be more appropriate in other cases. However, when only the first two or three dimensions are considered, ordinations in reduced space may misrepresent the structure by projecting together clusters of objects that are distinct in higher dimensions. Clustering methods allow one to separate clusters whose projections in reduced space may sometimes obscure the relationships between them.

Several authors (e.g. Gower & Ross, 1969; Rohlf, 1970; Schnell, 1970; Jackson & Crovello, 1971; Legendre, 1976) have independently proposed to take advantage of the characteristics of clustering and ordination by combining the results of the two types of analyses on the same diagram. The same similarity or distance matrix (Tables 7.4 to 7.6) is often used for the ordination and cluster analyses. Any clustering method may be used, as long as it is appropriate to the data. If linkage clustering is chosen, it is easy to draw the links between objects onto the ordination diagram, up to a given level of similarity. One may also identify the various similarity levels by using different colours or streaks (for example: solid line for $1.0 \geq S > 0.8$, dashed for $0.8 \geq S > 0.6$, dotted for $0.6 \geq S > 0.4$, etc., or any other convenient combination of codes or levels). If a divisive method or centroid clustering was used, a polygon or envelope may be drawn, on the ordination diagram, around the members of each cluster. This is consistent with the opinion of Sneath & Sokal (1973), who suggested to always simultaneously carry out clustering and ordination on a set of objects. Field *et al*. (1982) expressed the same opinion about marine ecological data. It is therefore recommended, as a routine procedure in ecology, to represent clustering results onto ordination diagrams.

The same approach can be applied to cluster analyses of descriptors. Clustering may be conducted on a dependence matrix among descriptors — especially species (Subsection 8.9.2) — in the same way as for an association matrix among objects. An ordination of species (e.g. Figs. 8.19 and 8.20) or other descriptors can be obtained using one of the ordination methods described in Chapter 9, depending on the measure of dependence among descriptors that is appropriate for the data under study. With quantitative physical or chemical descriptors of the environment, the method of choice is principal component analysis of the correlation matrix (Section 9.1); descriptors are represented by arrows in the ordination diagram. In some cases, before clustering, negative correlations among descriptors can be made positive because they are indicative of resemblance on an inverted scale.

When superimposed onto an ordination, single linkage clustering becomes a most interesting procedure for ecological interpretation. Single linkage clustering is the best complement to an ordination due to its contraction of the clustering space (Table 8.9, Fig. 8.24). Drawing single linkage results onto an ordination diagram provides both the correct positions for the main clusters of objects (from the ordination) and the fine relationships between closely similar objects (from the clustering). It is advisable to only draw the chain of primary connections (Section 8.2) on the ordination diagram because it reflects the changes in the composition of clusters. Otherwise, the groups of highly similar objects may become lost in the multitude of links drawn on the ordination. Ecological application 10.1 provides an example of this procedure.

Jackson & Crovello (1971) suggested to indicate the directions of the links on the ordination diagram (Fig. 10.1). This information may be useful when delineating clusters. In such diagrams, each link of the primary chain is drawn with an arrow. On a link from $\mathbf{x}_1$ to $\mathbf{x}_2$, an arrow pointing towards $\mathbf{x}_2$ indicates that object $\mathbf{x}_1$ has $\mathbf{x}_2$ as its closest neighbour in multidimensional A-space (i.e. in the association matrix among

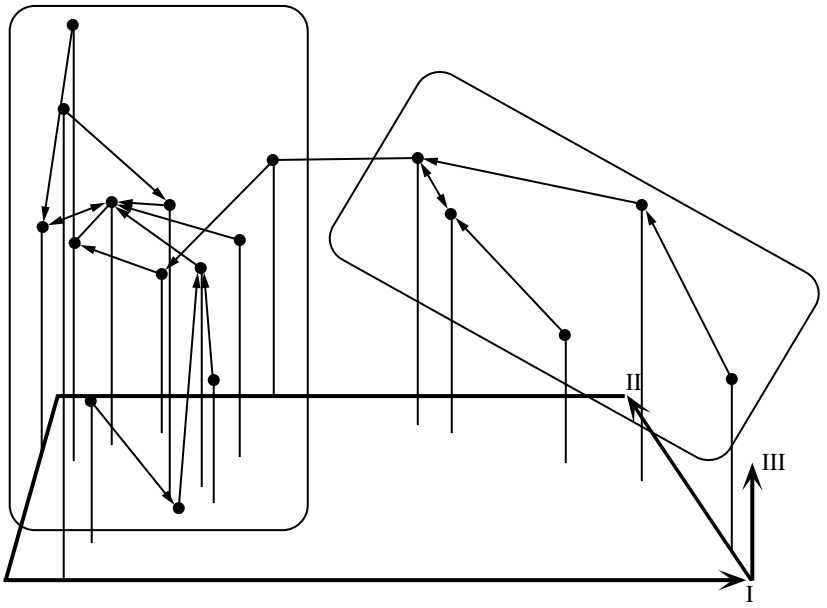


Figure 10.1 Three-dimensional ordination of objects (dots), structured by the primary connections of a single linkage clustering. The arrows (excluding those of the principal axes I to III) specify the directions of the relationships between nearest neighbours; see text. Modified from Jackson & Crovello (1971). Cutting the link without arrows determines two clusters (boxed points).

objects). When $\mathbf{x}_2$ also has $\mathbf{x}_1$ as its closest neighbour, the arrow goes both ways. When $\mathbf{x}_2$ has $\mathbf{x}_3$ as its closest neighbour, the arrow from $\mathbf{x}_2$ points towards $\mathbf{x}_3$. New links formed between objects that are already members of clusters do not receive arrows. These links may be removed to separate the clusters.

Ecological application 10.1

Single linkage clustering was illustrated by Ecological application 8.2 taken from a study of a group of ponds, based upon zooplankton. The same example (Legendre & Chodorowski, 1977) is used again here. Twenty ponds were sampled on islands of the St. Lawrence River, east and south of Montréal (Québec). Similarity coefficient S_{20} (eq. 7.27) was computed with $k = 2$. The matrix of similarities among ponds was used to compute both single linkage clustering and an ordination in reduced space by principal coordinate analysis. In Fig. 10.2, the chain of primary connections is superimposed onto the ordination, in order to evidence the clustering structure. The ponds are divided between a cluster of periodic ponds, which are dry during part of the year (encircled), and a cluster of permanent ponds. Ponds with identification numbers beginning with the same digit (which indicates the region) tend to be close to one another and to cluster first with one another. The second digit refers to the island on which a pond was located.

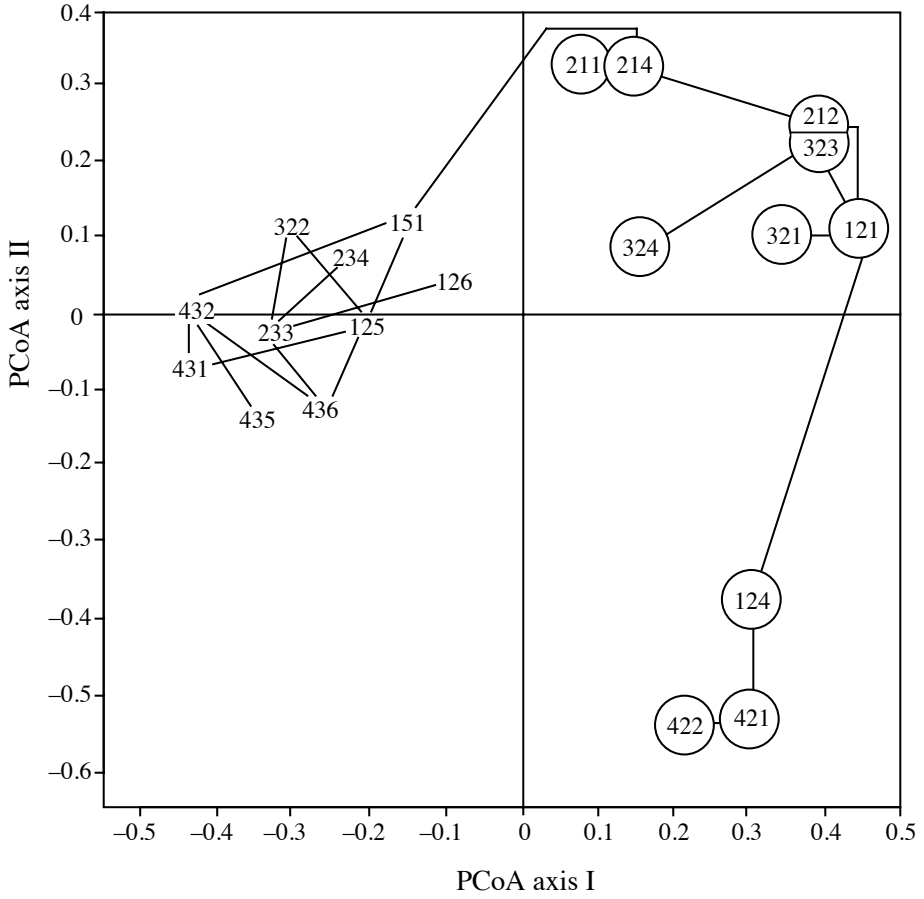


Figure 10.2 Comparison of 20 ponds on the basis of their zooplankton fauna. Ordination in a space of principal coordinates (principal axes I and II), and superimposition of the chain of primary connections obtained by single linkage clustering. The encircled ponds are periodic; the others are permanent. Adapted from Legendre & Chodorowski (l977).

When no clear clustering structure is present in the data but groups are still needed, for management purpose for instance, arbitrary groups may be delineated by drawing a regular grid on the reduced-space ordination diagram. This grid may be orthogonal (i.e. square or rectangular) or polar (division into triangles from the point of origin of the graph coordinates). Another method is to divide the objects according to the quadrants of the ordination in reduced space (in 2^d groups for a d-dimensional space); the result is the hierarchic classification scheme of Lefkovitch (1976) described in Subsection 8.7.3.

Figure 10.3 summarizes the steps involved in producing a cluster analysis and an ordination from a resemblance matrix. Description of the data structure is clearer when the clustering results are drawn onto the ordination. In order to assess to what extent the clustering and the ordination correspond to the resemblance matrix from which they originate, these representations may be compared to the original resemblance matrix using matrix correlation or related methods (Subsection 8.12.2).

10.2 The mathematics of ecological interpretation

The present section summarizes the numerical methods available for the interpretation of ecological structures. The most widely used of these techniques (regression, path analysis, matrix comparison, the fourth-corner method, and canonical analysis) are discussed in Sections 10.3 to 10.6 and in Chapter 11. A few other methods are briefly described in the present section.

The numerical methods presented in this section are grouped into three subsections, which correspond to the three main objectives of ecological interpretation, set in Section 10.0: explanation, forecasting, and prediction. For each of these objectives, there is a summary table (Tables 10.1 to 10.3) intended to facilitate the choice of methods best suited to the researchers' ecological objectives and the nature of their data.

Ecological interpretation, and especially the *explanation* and *forecasting* of the structure of several descriptors (i.e. multivariate data), may be conducted following two approaches, which are the indirect and direct comparison schemes (Fig. 10.4).

Indirect comparison

Indirect comparison proceeds in two steps. The structure (ordination axes, or clusters) is first identified from a set of descriptors (response data) of prime interest in the study. In a second step, the structure is interpreted using either (a) the descriptors that were analysed in the first step to identify the structure, or (b) another set of descriptors that may help explain the structure. In his chapter on ordination analysis, ter Braak (1987c) referred to this form of analysis as *indirect gradient analysis* because he was mostly concerned with the study of environmental gradients.

Direct comparison

In *direct comparison*, one simultaneously analyses the response and explanatory data matrices in order to identify how they are related. Direct comparison is done by the asymmetric methods of canonical analysis (Sections 11.1 and 11.2), which allow one to bring out the ordination structure of a response data set that is explained by another data set; ter Braak (1987c) refers to this approach as *direct gradient analysis*.

Other forms of direct comparison analysis are available. One can compare similarity or distance matrices, derived from the original data matrices, using the techniques of matrix comparison (Section 10.5); this type of comparison should, however, be restricted to test hypotheses that concern similarities or distances, not raw data (Subsection 10.5.1).

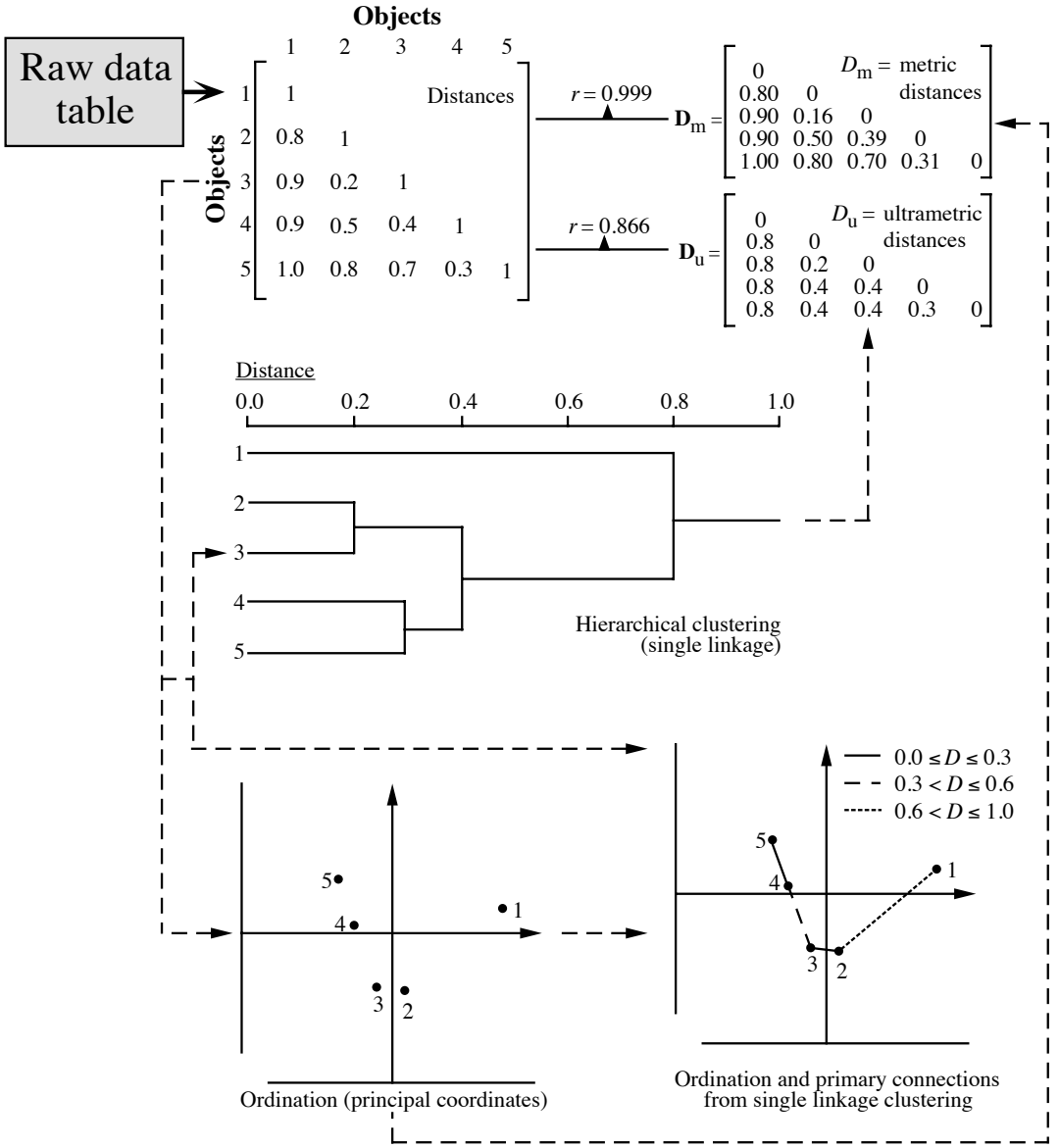


Figure 10.3 Identification of the structure of five objects, using clustering and ordination. Bottom right: the chain of primary connections is superimposed on a 2-dimensional ordination, as in Figs. 10.1 and 10.2. Top: the reduced-space ordination and the clustering results are compared to the resemblance matrix from which they originate. Upper right (top): a matrix of metric distances $\mathbf{D}_m$ is computed from the reduced-space ordination, and compared to the original distances using matrix correlation; $r = 0.999$ is a rather high score. Upper right (below): a cophenetic $\mathbf{D}_u$ matrix (Section 8.3) is computed from the dendrogram, and compared to the original distances using matrix correlation ($r = 0.866$).

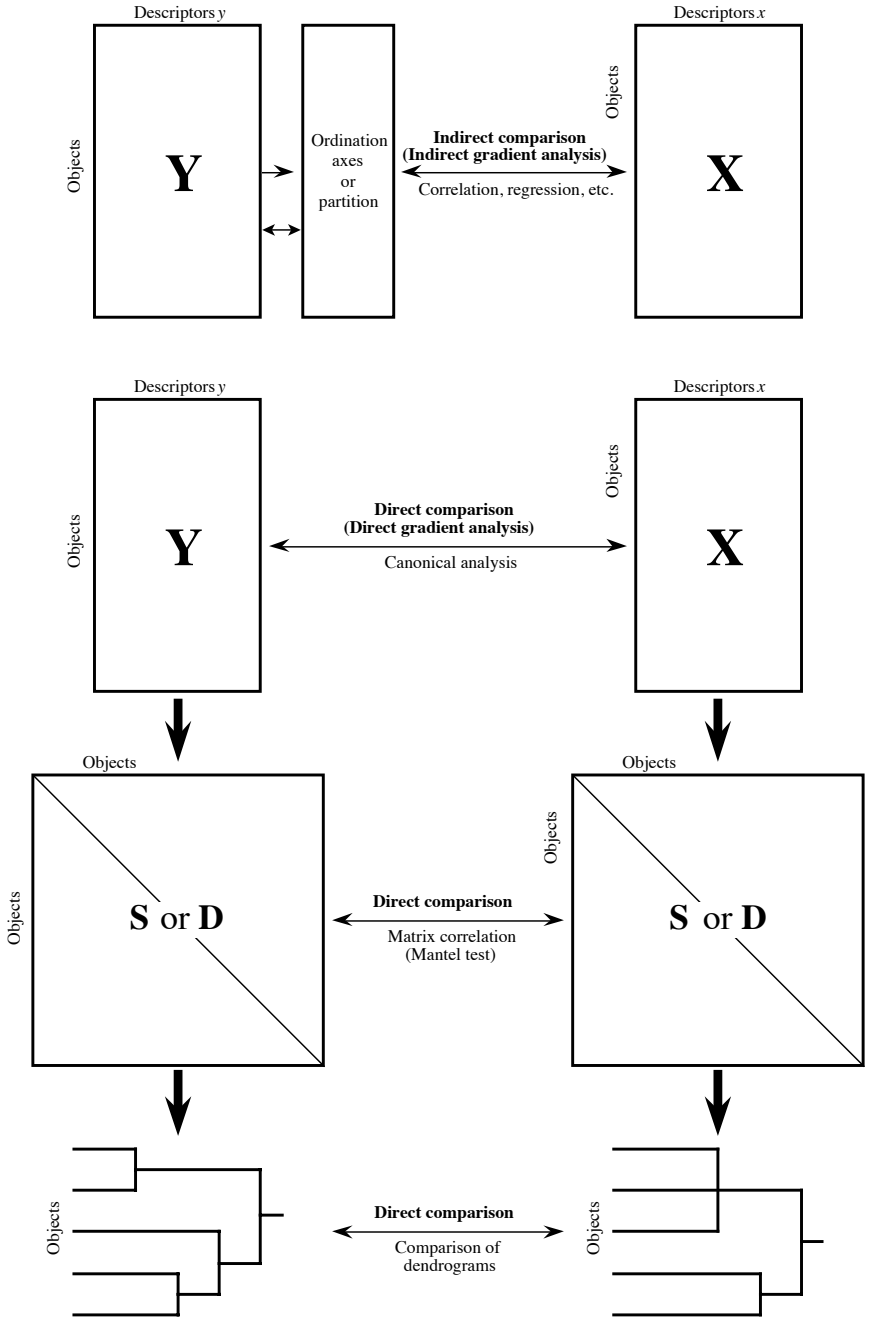


Figure 10.4 Indirect and direct comparison approaches for analysing and interpreting the structure of ecological data. Single thin arrow: inference of structure. Double arrow: interpretation strategy.

Consensus index

Permutation test

One can also compare dendrograms derived from resemblance matrices, using *consensus indices*; this approach should be restricted to test hypotheses that concern dendrograms. Two main approaches have been developed to test the significance of consensus statistics: (1) a probability distribution derived for a given consensus statistic may be used, or (2) a specific test may be carried out to assess the significance of the consensus statistic, in which the reference distribution is found by permuting the two dendrograms under study in an appropriate way (Lapointe & Legendre, 1995). Readers are referred to the papers of Day (1983, 1986), Shao & Rohlf (1983), Shao & Sokal (1986), Lapointe & Legendre (1990, 1991, 1992a, 1992b, 1995), and Steel & Penny (1993), where these methods are described. Lapointe & Legendre (1994) used the three forms of direct comparison analysis (i.e. comparison of raw data, distance matrices, and dendrograms; Fig. 10.4) on five data sets describing the same objects. In that study, all methods essentially led to similar conclusions, with minor differences.

The interpretation of a structure using the descriptors from which it originates makes it possible to identify which of these descriptors mainly account for the structuring of the objects. In some ordination methods (i.e. principal component and correspondence analysis), the eigenvectors readily identify the important descriptors. Other types of ordination, or the clustering techniques, do not directly provide this information, which must be found *a posteriori* using methods of indirect comparison. This type of interpretation does not allow one to perform formal tests of significance. The reason is that the structure under study is derived from the very same descriptors that are now used to interpret it; it is thus not independent of them.

Interpretation of a structure using external information (data matrix **X** in Fig. 10.4) is central to numerical ecology. This approach is used, for example, to diagnose abiotic conditions (response data matrix **Y**) from the available biological descriptors (explanatory data matrix **X**) or, alternatively, to forecast the responses of species assemblages (matrix **Y**) using available environmental descriptors (matrix **X**). In the same way, it is possible to compare two groups of biological descriptors or two matrices of environmental data. Until the mid-1980's, the indirect comparison scheme was favoured because of methodological problems with canonical correlation analysis, which was then the only method available in computer packages to analyse two sets of descriptors. When new methods and computer programs (including R functions) were made available, direct comparison became widely used in the ecological literature.

In the indirect comparison approach, the first set of descriptors is reduced to a single or a few one-dimensional variables, i.e. a partition resulting from clustering, or one or several ordination axes, the latter being generally interpreted one at a time. It follows that the methods of interpretation for univariate descriptors (e.g. correlation, regression) can also be used for indirect comparisons. This is the approach used in Tables 10.1 and 10.2.

1 — Explaining ecological structures

Table 10.1 summarizes the methods available to *explain* the patterns found in one or several ecological descriptors. *Explaining* is taken here in the sense of looking for correlations and using them to formulate hypotheses. The purpose is data exploration, not hypothesis testing. The first dichotomy of the table separates methods for univariate descriptors (used also in the indirect comparison approach) from those for multivariate data.

Methods used for explaining the structure of *univariate descriptors* belong to three major groups: (1) measures of dependence, (2) discriminant analysis, (3) and methods for qualitative descriptors. Methods used for explaining the structure of *multivariate descriptors* belong to two major types: (4) asymmetric canonical analysis using a response and an explanatory matrix, and (5) symmetric canonical analysis comparing two interchangeable data matrices. (6) Supplementary data associated with the sites and the species of a community composition data matrix can be related in fourth-corner analysis. The following paragraphs briefly review these groups of methods.

1. Various coefficients have been described in Chapters 4 and 5 to measure the *dependence* between two descriptors exhibiting linear or monotonic relationships (i.e. the parametric and nonparametric *correlation coefficients*). When there are more than two descriptors, one may use the *coefficients of partial correlation* (Section 4.5) or the *coefficient of concordance* (Section 5.4). The *coefficient of multiple determination* (R^2), computed in *multiple linear regression*, may be used to assess the dependence of a quantitative response descriptor on an explanatory matrix containing quantitative or mixed-level descriptors. *Dummy variable regression* is a special case of multiple regression where the explanatory matrix contains qualitative descriptors recoded into dummy variables, as explained in Subsection 1.5.7. These different types of regression are briefly discussed in Subsection 10.2.2, in relation with Table 10.2, and in more detail in Section 10.3.

2. Explaining the structure of a qualitative descriptor is often called *discrimination*, when the aim of the analysis is to identify explanatory descriptors that would allow one to discriminate among the various states of the qualitative descriptor. *Linear discriminant analysis* may be used when (1) the explanatory (or *discriminant*) descriptors are quantitative, (2) the distributions of the within-group residuals are not too far from normal, and (3) the within-group dispersion matrices are reasonably homogeneous. Linear discriminant analysis (LDA) is described in Section 11.3. Its use with species data is discussed in Section 11.6, where alternative strategies are proposed.

3. When both the descriptor to be explained and the explanatory descriptors are qualitative, one may use *multidimensional contingency table analysis*. It is then imperative to follow the rules, given in Section 6.3, concerning the models to use when a distinction is made between the explained and explanatory descriptors. When the response variable is binary, *logistic regression* is a better choice than multidimensional

Table 10.1 Numerical methods to *explain* the structure of descriptors, using either the descriptors from which the structure originates, or other, potentially explanatory descriptors. In parentheses, identification of the section where a method is discussed. Tests of significance cannot be performed when the structure of a descriptor is explained by the descriptors at the origin of that structure.

1) Explanation of the structure of a *single* descriptor, or *indirect comparison* . . . . . **see 2**
   2) Structure of a quantitative or a semiquantitative descriptor . . . . . **see 3**
      3) Explanatory descriptors are quantitative or semiquantitative . . . . . **see 4**
         4) To *measure* the dependence between descriptors . . . . . **see 5**
            5) Pairs of descriptors: *Pearson r*, for quantitative descriptors exhibiting linear relationships (4.2); *Kendall* τ or *Spearman r*, for quantitative or semiquantitative descriptors exhibiting monotonic relationships (5.3)
            5) A single quantitative descriptor as a function of several others: *coefficient of determination* R^2 of *multiple regression* (4.5, 10.3.3)
            5) Several descriptors exhibiting monotonic relationships: *coefficient of concordance W* (5.4)
         4) To *interpret* the structure of a single descriptor: *partial Pearson r*, for quantitative descriptors exhibiting linear relationships (4.5); *partial Kendall* τ, for descriptors exhibiting monotonic relationships (5.3)
      3) Explanatory descriptors are qualitative: R^2 of *dummy variable regression* (10.3)
      3) Estimation of the dependence between descriptors of the sites and descriptors of the species (any precision level): *the fourth-corner method* (10.6)
   2) Structure of a qualitative descriptor (*or* of a classification) . . . . . **see 6**
      6) Explanatory descriptors are quantitative: *linear discriminant analysis* (LDA, 11.3)
      6) Explanatory descriptors are qualitative: *multidimensional contingency table analysis* (6.3); *discrete discriminant analysis* (10.2)
      6) Explanatory descriptors are of mixed precision: *logistic regression* (in most cases, the explained descriptor is binary; 10.3)

1) Explanation of the structure of a *multivariate* data matrix . . . . . **see 7**
   7) *Direct comparison* . . . . . **see 8**
      8) Asymmetric analysis of a response matrix by an explanatory matrix: *redundancy analysis* (RDA, 11.1); *canonical correspondence analysis* (CCA, 11.2); multivariate regression tree analysis (MRT, 8.11). Basic statistic in RDA: *canonical* R^2
      8) Symmetric comparison of two data matrices: *canonical correlation analysis* (CCorA, 11.4), *co-inertia analysis* (CoIA, 11.5.1), *Procrustes analysis* (Proc, 11.5.2). Statistics: *RV* (11.5.1), Trace**W** and $m_{12}{}^2$ (10.5.4, 11.5.2)
      8) Compare classifications computed from two data matrices: *contingency table analysis* (6.2), *modified Rand index* (8.12)
   7) *Indirect comparison* . . . . . **see 10**
      10) Ordination in reduced space: each axis is treated in the same way as a single quantitative descriptor . . . . . **see 2**
      10) Clustering: each partition is treated as a qualitative descriptor . . . . . **see 2**

contingency table analysis. An additional advantage is that logistic regression allows one to use explanatory variables presenting a mixture of precision levels. For qualitative variables, the equivalent of discriminant analysis is called *discrete discriminant analysis*. Goldstein & Dillon (1978) describe this form of analysis.

4. The standard approach for comparing two sets of descriptors is *canonical analysis* (Chapter 11). In ecology, the asymmetric forms of canonical analysis, where the two data matrices do not play the same role, are the most widely used. Asymmetric analyses involve a response matrix **Y** and an explanatory matrix **X**. The methods are called *redundancy analysis* (RDA, Section 11.1) and *canonical correspondence analysis* (CCA, Section 11.2). The difference between these two methods is the same as between principal component and correspondence analyses (Table 9.1). An alternative method of asymmetric analysis is multivariate regression tree analysis (MRT, Section 8.11), which looks for cutting points in the explanatory descriptors **X** that create compact groups in the response data **Y**.

5. It is also possible to compare two matrices that play the same role and can be interchanged in the analysis. These symmetric analyses are carried out by *canonical correlation analysis* (CCorA, 11.4), *co-inertia analysis* (CoIA, 11.5.1), and *Procrustes analysis* (Proc, 11.5.2).

6. Consider a (site × species) matrix containing community composition data (presence-absence or abundance), for which supplementary variables are known for the sites (e.g. habitat characteristics, spatial data) and for the species (e.g. biological or behavioural traits). The *fourth-corner method*, described in Section 10.6, offers a way of estimating the dependence between the supplementary variables of the rows and those of the columns and testing the resulting correlation-like statistics for significance.

2 — Forecasting ecological structures

Forecasting model
Predictive model

A distinction has to be made between *forecasting* and *prediction* in ecology. Forecasting models extend, into the future or to different situations, structural relationships among descriptors that have been quantified for a given data set. A set of relationships among variables, which simply describe the changes in one or several descriptors in response to changes in others as computed from a "training set", make up a *forecasting* model. In contrast, when the relationships are considered causal and to describe a mechanistic process, the model is *predictive*. A condition to successful forecasting is that the values of all important variables that have not been observed (or controlled, in the case of an experiment) be about the same in the new situation as they were during the survey or experiment. In addition, forecasting does not allow extrapolation beyond the observed range of the explanatory variables. *Forecasting models* (also called *correlative models*) are frequently used in ecology, where they are sometimes misleadingly called "predictive models". Forecasting models are useful provided that the above conditions are fulfilled. In contrast, predictive models describe known or assumed causal relationships. They allow one to estimate the effects, on

Table 10.2 Numerical methods to *forecast* one or several descriptors (response or dependent variables) using other descriptors (explanatory or independent variables). In parentheses, identification of the section where a method is discussed.

1) Forecasting the structure of a *single* descriptor, or *indirect comparison* . . . . . **see 2**
   2) The response variable is quantitative . . . . . **see 3**
      3) The explanatory variables are quantitative . . . . . **see 4**
         4) Null or low correlations among explanatory variables: *multiple linear regression* (10.3); *nonlinear regression* (10.3)
         4) High correlations among explanatory variables (collinearity): *ridge regression* (10.3); *regression on principal components* (10.3)
      3) The explanatory variables are qualitative: *dummy variable regression* (10.3)
   2) The response variable is qualitative (*or* a classification) . . . . . **see 5**
      5) Response: two or more groups; explanatory variables are quantitative (but qualitative variables may be recoded into dummy variables): *identification functions in discriminant analysis* (11.3)
      5) Response: binary (presence-absence); explanatory variables are quantitative (but qualitative variables may be recoded into dummy var.): *logistic regression* (10.3)
   2) The response and explanatory variables are quantitative, but they display a nonlinear relationship: *nonlinear regression* (10.3)

1) Forecasting the structure of a *multivariate* data matrix . . . . . **see 6**
   6) *Direct comparison* . . . . . **see 7**
      7) Linear modelling: *redundancy analysis* (RDA, 11.1); *canonical correspondence analysis* (CCA, 11.2)
      7) Find a tree-like decision model: *multivariate regression tree analysis* (MRT, 8.11)
   6) *Indirect comparison* . . . . . **see 8**
      8) Ordination in reduced space: each axis is treated in the same way as a single quantitative descriptor . . . . . **see 2**
      8) Clustering: each partition is treated as a qualitative descriptor . . . . . **see 2**

some variables, of changes in other variables; they will be briefly discussed at the beginning of the next subsection.

Methods in Table 10.2 are used to *forecast* descriptors. As in Table 10.1, the first dichotomy in the table distinguishes the methods that allow one to forecast values of a single descriptor (*response* or *dependent* variable) from those that may be used to simultaneously forecast several descriptors. Forecasting methods belong to four major groups: (1) regression models, (2) identification functions, (3) asymmetric canonical analysis methods, and (4) multivariate regression trees.

1. Methods belonging to *regression* models are numerous. Several regression methods include measures of dependence that have already been mentioned in the discussion of Table 10.1: *multiple linear regression* (the explanatory variables are quantitative or mixed), *dummy variable regression* (a special case of multiple regression where the explanatory matrix contains qualitative descriptors (e.g. ANOVA factors) recoded into dummy variables, as explained in Subsection 1.5.7), and *logistic regression* (the explanatory variables may be of mixed levels of precision; the response variable is qualitative). Section 10.3 provides a detailed description of several regression methods.

2. *Identification functions* are part of linear discriminant analysis (Section 11.3), which was briefly described in the previous subsection. These functions allow the assignment of any object to one of the states of a qualitative descriptor, using the values taken by several quantitative variables (i.e. the explanatory or discriminant variables). As mentioned in the previous subsection, the distributions of the discriminant variables must not be too far from normality, and their within-group dispersion matrices must be reasonably homogeneous (i.e. about the same in all groups).

3. Canonical analysis, and especially *redundancy analysis* and *canonical correspondence analysis*, which were briefly discussed in the previous subsection (and in more detail in Sections 11.1 and 11.2), allow one to model a data matrix from the descriptors of a second data matrix; these two data matrices form the "training set". Using the resulting model, it is possible to forecast the position of any new observation among those of the "training set", for example along environmental gradients. The new observation may represent some condition that may occur in the future, or at a different but comparable location.

4. An alternative forecasting method of analysis is multivariate regression tree analysis (MRT, Section 8.11). This method produces a decision tree in which the response data **Y** are divided into groups, whereas the bifurcations of the tree correspond to splits in the explanatory variables **X** that can be used for forecasting the positions of new observations.

3 — Ecological prediction

As explained in the Preface, predictive modelling does not belong to numerical ecology *sensu stricto*. However, some methods of numerical ecology may be used to analyse causal relationships among a small number of descriptors, thus linking numerical ecology to predictive modelling. Contrary to the *forecasting* or *correlative models* (previous subsection), *predictive models* allow one to foresee how some variables of interest would be affected by changes in other variables. Prediction is possible when the model is based on causal relationships among descriptors (i.e. not only correlative evidence). Causal relationships are stated as hypotheses (theory) for modelling; they may also be validated through experiments in the laboratory or in the field. In *manipulative experiments*, one observes the responses of some descriptors to

Predictive model

Experiment

Table 10.3 Numerical methods for analysing causal relationships among ecological descriptors, with the purpose of *predicting* one or several descriptors using other descriptors. In parentheses, identification of the section where the methods are discussed. In addition, forecasting methods (Table 10.2) may be used for prediction when there are reasons to believe that the relationships between the explanatory and response variables are of causal nature.

1)	The causal relationships among descriptors are given by hypothesis . . . . . **see 2**
	2) Quantitative descriptors; linear causal relationships: *causal modelling using correlations* (4.5); *path analysis* (10.4)
	2) Qualitative descriptors: *logit* and *log-linear models* (6.3)
1)	Hidden variables (latent variables, factors) are assumed to cause the observed structure of the descriptors: *confirmatory factor analysis* (not discussed in this book)

user-determined changes in other descriptors, by reference to a *control*. Besides manipulative experiments, which involve two or more treatments, Hurlbert (1984) recognizes *mensurative experiments*, which involve measurements made at one or more points in space or time and allow one to test hypotheses about patterns in space (Chapters 13 and 14) and/or time (Chapter 12). The numerical methods in Table 10.3 allow one to explore a network of causal hypotheses, using the observed relationships among descriptors. The design of experiments and analysis of experimental results are discussed by Mead (1988) who offers a statistically-oriented presentation, and by Underwood (1997) in a book emphasizing ecological experiments.

One may hypothesize that there exist causal relationships among the observed descriptors or, alternatively, that the observed descriptors are caused by underlying hidden variables. Depending on the hypothesis, the methods for analysing causal relationships are not the same (Table 10.3). Methods appropriate to the first case belong to the family of *path analysis* (Section 10.4). The second case leads to *confirmatory factor analysis*, which is not discussed in this book; see e.g. Brown (2006) or Harrington (2009) on this subject. The present chapter only discusses the former. In addition to these methods, techniques of forecasting (Table 10.2) may be used for predictive purposes when there are reasons to believe that the relationships between explanatory and response variables are of causal nature.

Fundamentals of *path analysis* are presented in Section 10.4. Path analysis is an extension of multiple linear regression and is thus limited to quantitative or binary descriptors (including qualitative descriptors recoded as dummy variables: Subsection 1.5.7). In summary, path analysis is used to decompose and interpret the relationships among a small number of descriptors, assuming that (a) there is a *(weak) causal order* among descriptors, and (b) the relationships among descriptors are *causally closed*. *Causal order* means, for example, that y_2 possibly (but not necessarily) affects y_3 but that, under no circumstance, y_3 would affect y_2 through the

same process. Double causal "arrows" are allowed in a model only if different mechanisms may be hypothesized for the reciprocal relationships. Using this assumption, it is possible to set a causal order between y_2 and y_3. The assumption of *causal closure* implies independence of the residual causalities, which are the unknown factors responsible for the residual variance (i.e. the variance not accounted for by the observed descriptors). Path analysis is restricted to a small number of descriptors. This is not due to computational problems, but to the fact that the interpretation becomes complex when the number of descriptors in a model becomes large. When the analysis involves three descriptors only, the simple method of *causal modelling using correlations* may be used (Subsection 4.5.4).

For qualitative descriptors, Fienberg (1980; his Chapter 7) explains how to use *logit* or *log-linear models* (Section 6.3) to determine the signs of causal relationships among such descriptors, by reference to diagrams similar to the path diagrams of Section 10.4.

10.3 Regression

Random variable

The purpose of regression analysis is to describe the relationship between a *dependent* (or *response*) *random** *variable* (y) and a set of *independent* (or *explanatory*) *variables*, in order to forecast or predict the values of y for given values of the independent variables $x_1, x_2, \ldots, x_m$. Box 1.1 gives the terminology used to refer to the dependent and independent variables of a regression model in an empirical or causal framework. The explanatory variables may be either random*, or controlled (and, consequently, known *a priori*). On the contrary, the response variable must of necessity be a random variable. That the explanatory variables be random or controlled will be important when choosing the appropriate computation method (model I or II).

Model

A *mathematical model* is simply a mathematical formulation (algebraic, in the case of regression models) of a relationship or a set of relationships among variables, whose parameters have to be estimated or tested against a hypothesis; in other words, it is a simplified mathematical description of a real-life system. Regression, with its many variants, is the first type of modelling method presented in this chapter for the analysis of ecological structures. It is also used as a platform to help introduce the principles of structure analysis. The same principles will apply to more mathematically advanced forms, collectively referred to as canonical analysis, which are discussed in Chapter 11.

* A random variable is a variable whose values are assumed to result from some random process (Section 1.0); these values are not known before observations are made. A random variable is *not* a variable consisting of numbers drawn at random; such variables, usually generated with the help of a pseudo-random number generator, are used by statisticians to assess the properties of statistical methods under some distribution hypotheses.

Regression modelling may be used for description, inference, or forecasting/prediction:

1. Description aims at finding the best functional relationship among variables in the model, and estimating its parameters, based on available data. In mathematics, a function $y = f(x)$ is a rule of correspondence, often written as an equation, that associates with each value of x one and only one value of y. A well-known functional relationship in physics is Einstein's equation $E = mc^2$, which describes the amount of energy E associated with given amounts of mass m; the scalar value c^2 is the parameter of the model, where c is the speed of light in vacuum.

2. Inference means generalizing the results of a set of observations to the whole target population, as represented by a sample drawn from that population. Inference may consist in estimating the confidence intervals within which the true values of the statistical population parameters are likely to be found, or testing *a priori* hypotheses about the values of model parameters in the statistical population. (1) The ecological hypotheses may simply concern the *existence* of a relationship, e.g. the slope or the intercept are different from 0. The test consists in finding the *two-tailed* probability of observing the slope (b_1) or intercept (b_0) values that have been estimated from the sample data, given the null hypothesis (H_0) stating that the slope (β_1) or intercept (β_0) parameters are zero in the statistical population. These tests are described in manuals of elementary statistics. (2) In other instances, the ecological hypothesis concerns the sign that the relationship should have. One then tests the *one-tailed* null statistical hypotheses (H_0) that the intercept or slope parameters in the statistical population are zero, against alternative hypotheses (H_1) that they have the signs (positive or negative) stated in the ecological hypotheses. For example, one might want to test Bergmann's law (1847), that the body mass of homeotherms, within species or groups of closely related species, *increases* with latitude. (3) There are also cases where the ecological hypothesis states specific values for the parameters. Consider for instance the isometric relationship specifying that mass should increase as the cube of the length in animals, or in log form: $\log(mass) = b_0 + 3 \log(length)$. Length-to-mass relationships found in nature are most often allometric, especially when considering a multi-species group of organisms. Reviewing the literature, Peters (1983) reported allometric slope values from 1.9 (algae) to 3.64 (salamanders).

3. Forecasting (or prediction) consists in calculating values of the response variable using a regression equation. Forecasting (or prediction) is sometimes described as *the* purpose of ecology. In any case, ecologists agree that empirical or hypothesis-based regression equations are helpful tools for management. This objective is achieved by using the equation that minimizes the residual mean square error, or maximizes the coefficient of determination (r^2 in simple regression; R^2 in multiple regression).

A study may focus on one or two of the above objectives, but not necessarily all three. Satisfying two or all three objectives may call upon different methods for computing the regressions. In any case, these objectives differ from that of correlation

Correlation or regression analysis? Box 10.1

Regression analysis is a type of modelling. Its purpose is either to find the best functional model relating a response variable to one or several explanatory variables, in order to test hypotheses about the model parameters, or to forecast or predict values of the response variable.

The purpose of correlation analysis is quite different. It aims at establishing whether there is *interdependence*, in the sense of the coefficients of dependence of Chapter 7, between two random variables, without assuming any functional or explanatory-response or causal link between them.

In model I simple linear regression, where the explanatory variable of the model is controlled, the distinction is easy to make; in that case, a correlation hypothesis (i.e. interdependence) is meaningless. Confusion comes from the fact that the coefficient of determination, r^2, which is essential to estimate the forecasting value of a regression equation and is automatically reported by most regression programs, happens to be the square of the coefficient of linear correlation.

When the two variables are random (i.e. not controlled), the distinction is more tenuous and depends on the intent of the investigator. If the purpose is modelling (as broadly defined in the first paragraph of this Box), model II regression is the appropriate type of analysis; otherwise, correlation should be used to measure the interdependence between such variables. In Sections 4.5 and 10.4, the same confusion is rampant, since correlation coefficients are used as an *algebraic tool* for choosing among causal models or for estimating path coefficients.

analysis, which is to support the existence of a relationship between two random variables, without reference to any functional or causal link between them (Box 10.1).

This section does not attempt to present regression analysis in a comprehensive way. Interested readers are referred to general texts of (bio)statistics such as Sokal & Rohlf (1995), specialized texts on regression analysis (e.g. Draper & Smith, 1981; Neter *et al*., 1996), or textbooks such as those of Ratkowski (1983) or Ross (1990) for nonlinear estimation. The purpose here is to survey the main principles of regression analysis and, in the light of these principles, explain the differences among the regression models most commonly used by ecologists: simple linear (model I and model II), multiple linear, polynomial, partial, nonlinear, and logistic. Some smoothing methods will also be described. Several other types of regression will be

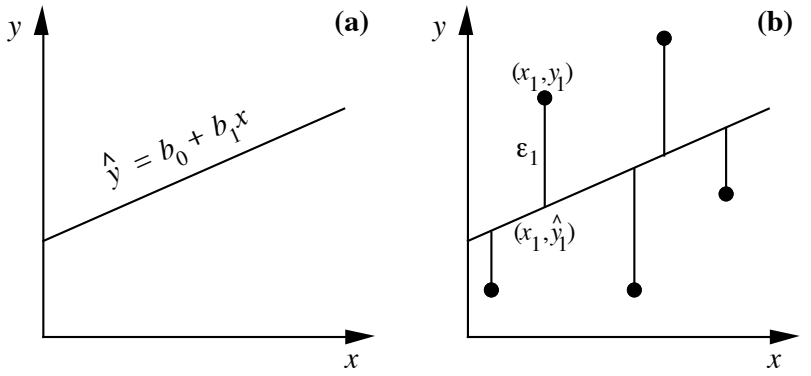


Figure 10.5 (a) Linear regression line, of equation $\hat{y} = b_0 + b_1 x$, fitted to the scatter of points shown in b. (b) Graphical representation of regression residuals ε_i (vertical lines); ε_1 is the residual for point 1 with coordinates (x_1, y_1).

mentioned, such as ridge regression, multivariate linear regression, and monotone or nonparametric regression.

Regression

Incidentally, the term *regression* has a curious origin. It was coined by the anthropologist Francis Galton (1889, pp. 95-99), a cousin of Charles Darwin, who was studying the relationship between the heights of parents and children. Galton observed "that the Stature of the adult offspring … [is] … more *mediocre* than the stature of their Parents", or in other words, closer to the population mean; so, Galton said, they *regressed* (meaning *going back*) towards the population mean. He called the slope of this relationship "the ratio of 'Filial Regression' ". For this historical reason, the slope parameter is now known as the regression coefficient.

1 — Simple linear regression: model I

Linear regression is used to compute the parameters of a first-degree equation relating variables y and x. The expression *simple linear regression* applies to cases where there is a single explanatory variable x. The equation (or model) for simple linear regression has the form:

$$\hat{y} = b_0 + b_1 x \qquad \textbf{(10.1)}$$

This corresponds to the equation of a straight line (hence the name *linear*) that crosses the scatter of points in some optimal way and allows the computation of an estimated value $\hat{y}$ (along the ordinate scale of the scatter diagram) for any value of x (abscissa; Fig. 10.5a). Parameter b_0 is the estimate of the intercept of the regression line with the *ordinate*; it is also called the y-intercept. Parameter b_1 is the slope of the regression line; it is also called the *regression coefficient*. In Subsection 10.3.4 on polynomial

Intercept
Slope

regression, a distinction will be made between linearity in parameters and linearity in response to the explanatory variables.

The intercept b_0 has the same physical dimensions as y, whereas the regression coefficient b_1 has the physical dimensions of $[y]/[x]$ (Section 3.1) so that b_1x has the same physical dimensions as y. As a consequence, the regression equation (eq. 10.1) is dimensionally homogeneous (Section 3.2).

When using this type of regression, one must be aware of the fact that a *linear model* is imposed on the data. In other words, one assumes that the relationship between variables may be adequately described by a straight line and that the vertical dispersion of observed values above and below the line is the result of a random process. The difference between the observed and estimated values along y, noted $\varepsilon_i = (y_i - \hat{y}_i)$ for every observation i, may be either positive or negative since the observed data points lie above and below the regression line. ε_i is called the *residual* value of observation y_i after fitting the regression line (Fig. 10.5b). Including ε_i in the equation allows one to describe exactly the ordinate value y_i of each point (x_i, y_i) in the data set; y_i is equal to the value $\hat{y}_i$ predicted by the regression equation plus the residual ε_i:

$$y_i = \hat{y}_i + \varepsilon_i = b_0 + b_1 x_i + \varepsilon_i \qquad \textbf{(10.2)}$$

This equation is the *linear model* of the relationship. $\hat{y}_i$ is the predicted, or *fitted* value corresponding to each observation i. The model assumes that the only deviations from the linear functional relationship $y = b_0 + b_1 x$ are vertical differences ("errors") ε_i on values y_i of the response variable, and that there is no "error" associated with the estimation of x. "Error" is the traditional term used by statisticians for deviations of all kind due to random processes, and not only measurement error. In practice, when it is known by hypothesis — or found by studying a scatter diagram — that the relationship between two variables is not linear, one may either try to linearise it (Section 1.5), or else use polynomial or nonlinear regression methods to model the relationship (Subsections 10.3.4 and 10.3.6, below).

Model I

Besides the supposition that the variables under study are linearly related, *model I regression* makes the following additional assumptions about the data:

1. The explanatory variable x is controlled, or it is measured without error. (The concepts of random and controlled variables have been briefly explained above.)

2. For any given value x_i of x, the values y in the statistical population are independently and normally distributed. This does not mean that the response variable y must be normally distributed, but instead that the "errors" ε_i are normally distributed about a mean of zero. One also assumes that the ε_i's have the same variance for all values of x in the range of the observed data (homoscedasticity: Box 1.3).

So, model I regression is appropriate to analyse results of controlled experiments, and also the many cases of field data where a response random variable y is to be related to sampling variables under the control of the researcher (e.g. location in time and space, volume of water filtered). The next subsection will show how to use model II regression to analyse situations where these assumptions are not met.

Least squares

In simple linear regression, one is looking for the straight line with equation $\hat{y} = b_0 + b_1 x$ that minimizes the sum of squares of the vertical residuals, ε_i, between the observed values and the regression line. This is the *principle of least squares*, first proposed by the mathematician Adrien Marie Le Gendre from France, in 1805, and later by Karl Friedrich Gauss from Germany, in 1809; these two mathematicians were interested in estimation problems of astronomy. This sum of squared residuals, $\Sigma (y_i - \hat{y}_i)^2$, offers the advantage of providing a unique solution, which would not be the case if one chose to minimize another function — for example $\Sigma |y_i - \hat{y}_i|$. It can also be shown that the straight line that meets the *ordinary least-squares* (OLS) criterion passes through the centroid, or centre of mass $(\bar{x}, \bar{y})$ of the scatter of points, whose coordinates are the means $\bar{x}$ and $\bar{y}$. The formulae for parameters b_0 and b_1 of the line meeting the least-squares criterion are found using partial derivatives. The solution is:

OLS

$$b_1 = s_{xy}/s_x^2 \quad \text{and} \quad b_0 = \bar{y} - b_1\bar{x} \qquad \textbf{(10.3)}$$

where s_{xy} and s_x^2 are estimates of covariance and variance, respectively (Section 4.1). These formulae, written in full, are found in textbooks of introductory statistics. Least-squares estimates of b_0 and b_1 can also be computed directly from the **x** and **y** data vectors using eq. 2.19. Least-squares estimation provides the line of best fit for parameter estimation and forecasting when the explanatory variable is controlled.

Regressing y on x does not lead to the same least-squares equation as regressing x on y. Figure 10.6a illustrates this for two random variables, which would represent a case for model II regression discussed in the next subsection. Even when x is a random variable, the variables will continue to be called x and y (instead of y_1 and y_2) to keep the notation simple. Although the covariance s_{xy} is the same for the calculation of the regression coefficient of y on x ($b_{1(y \cdot x)}$) and that of x on y ($c_{1(x \cdot y)}$), the denominator of the slope equation (eq. 10.3) is s_x^2 when regressing y on x, whereas it is s_y^2 when regressing x on y. Furthermore, the means $\bar{x}$ and $\bar{y}$ play inverted roles when estimating the two intercepts, $b_{0(y \cdot x)}$ and $c_{0(x \cdot y)}$. This emphasizes the importance of clearly defining the explanatory and response variables when performing regression.

The two least-squares regression lines come together only when all observation points fall on the same line (correlation = 1). According to eq. 4.7, $r_{xy} = s_{xy}/s_x s_y$. So, when $r = 1$, $s_{xy} = s_x s_y$ and, since $b_{1(y \cdot x)} = s_{xy}/s_x^2$ (eq. 10.3), then $b_{1(y \cdot x)} = s_x s_y / s_x^2 = s_y / s_x$. Similarly, the slope $c_{1(x \cdot y)}$, which describes the same line in the transposed graph, is $s_x/s_y = 1/b_{1(y \cdot x)}$. In the more general case where r is not equal to 1, $c_{1(x \cdot y)} = r_{xy}^2 / b_{1(y \cdot x)}$. When the two regression lines are drawn on the same graph, assuming that the variables have been standardized prior to the

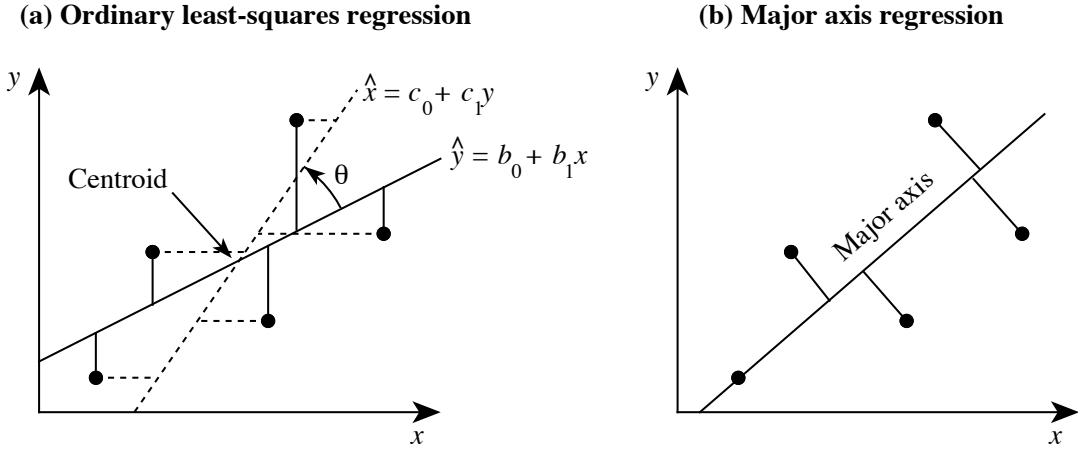


Figure 10.6 (a) Two least-squares regression equations are possible in the case of two random variables (called x and y here, for simplicity). When regressing y on x, the sum of *vertical* squared deviations is minimized (full lines); when regressing x on y, the sum of *horizontal* squared deviations is minimized (dashed lines). Angle θ between the two regression lines is computed using eq. 10.5. (b) In major axis regression, the sum of the squared Euclidean distances to the regression line is minimized.

computations, there is a direct relationship between the Pearson correlation coefficient r_{xy} and angle θ between the two regression lines:

$$\theta = 90° - 2 \tan^{-1} r, \quad \text{or} \quad r = \tan\left(\frac{90° - \theta}{2}\right) \qquad \textbf{(10.4)}$$

If $r = 0$, the scatter of points is circular and angle $\theta = 90°$, so that the two regression lines are at a right angle; if $r = 1$, the angle is 0°. Computing angle θ for non-standardized variables, as in Fig. 10.6a, is a bit more complicated:

$$\theta = 90° - \text{sign}(r) \times [\tan^{-1}(r\, s_x / s_y) + \tan^{-1}(r\, s_y / s_x)] \qquad \textbf{(10.5)}$$

where sign(r) is the sign of the correlation coefficient.

Coefficient of determination

The *coefficient of determination* r^2 measures how much of the variance of each variable is explained by the other. This coefficient has the same value for the two regression lines. The amount of explained variance for y is the variance of the fitted values $\hat{y}_i$, calculated as:

$$s^2_{\hat{y}} = \Sigma (\hat{y}_i - \bar{y})^2 / (n - 1) \qquad \textbf{(10.6)}$$

whereas the total amount of variation in variable y is

$$s_y^2 = \Sigma (y_i - \bar{y})^2 / (n-1)$$

It can be shown that the coefficient of determination, which is the ratio of these two values (the two denominators $(n-1)$ cancel out), is equal to the square of the Pearson correlation coefficient r. It is thus designated by r^2:

$$r^2 = s_{\hat{y}}^2 / s_y^2 \qquad \textbf{(10.7)}$$

With two random variables, the regression of y on x makes as much sense as the regression of x on y. In that case, the coefficient of determination may be computed as the product of the two regression coefficients: $r^2 = b_{1\,(y \cdot x)} c_{1\,(x \cdot y)}$. The coefficient of correlation is then the geometric mean of the coefficients of linear regression of each variable on the other, to which the sign of one of the regression coefficients is imposed: $r = \text{sign}\,(b_{1\,(y \cdot x)}) \times (b_{1\,(y \cdot x)} c_{1\,(x \cdot y)})^{1/2}$; function sign() is described after eq. 10.5. It may also be computed as the square of r in eq. 4.7:

$$r^2 = \frac{(s_{xy})^2}{s_x^2 \, s_y^2} \qquad \textbf{(10.8)}$$

Coefficient of non-determination

A value $r^2 = 0.81$, for instance, means that 81% of the variation in y is explained by x, and vice versa. In Section 10.4, the quantity $(1 - r^2)$ will be called the *coefficient of nondetermination*; it measures the proportion of the variance of a response variable that is not explained by the explanatory variable(s) of the model.

When x is a controlled variable, one must be careful not to interpret the coefficient of determination in terms of interdependence, as one would for a coefficient of correlation, in spite of their algebraic closeness and the fact that one coefficient can, indeed, be calculated directly from the other (Box 10.1).

2 — Simple linear regression: model II

Model II

When both the response and explanatory variables of the model are random (i.e. *not* controlled by the researcher), there are errors associated with the measurements of both x and y. Such situations call for methods that are referred to as *model II regression*. As a parallel to model II ANOVA, which is concerned with the analysis of the effect of a random factor on a random variable (Sokal & Rohlf, 1995, Section 8.7), model II regression is concerned with the analysis of two random variables. In model II regression, different computational procedures are required for description and inference, as opposed to forecasting; these three objectives of regression analysis were described at the beginning of Section 10.3.

1. Model II regression can be used for description and inference, that is, to estimate the slope of a process (parametric estimation) corresponding to the linear relationship

between the measured variables, and compute confidence intervals around the slope or test its significance. Examples:

- In aquatic ecology, *in vivo* fluorescence is routinely used to estimate the amount of chlorophyll *a* in phytoplankton. These variables, which are both random and measured with error, must be related by model II regression to establish their functional relationship (slope). The slope can also be tested for significance. If the objective is to forecast chlorophyll *a* from fluorescence values, see point 2 below.

- In freshwater sediment, one may be interested in comparing the rate of microbial anaerobic methane production to total particulate carbon in two environments (e.g. two lakes) where several sites have been studied. Since total particulate carbon and methane production have been measured with error in the field, rates are given by the slopes of model II regression equations computed on the data from the two lakes separately; the confidence intervals of these slopes may serve to compare the two environments.

Model II regression can be used with the more simple purpose of drawing a line in a graph of two random variables. This can be done for the above examples.

2. Model II regression can also be used for forecasting, that is, for computing fitted values about one variable from the values of the other. The method to be used in that case is ordinary least squares (OLS). The reason is simple: OLS is the method that produces fitted values with the smallest error, defined as $\Sigma (y_i - \hat{y}_i)^2$ (Subsection 10.3.1). Hence, OLS is also one of the methods that can be used in model II situations, when the purpose is forecasting. Example:

- In microbial ecology, the concentrations of two substances produced by bacterial metabolism have been measured. One is of economical interest, but difficult to measure with accuracy, whereas the other is easy to measure. Determining their relationship by regression may allow ecologists to use the second substance as a proxy for the first. An OLS regression model can be used to estimate the concentrations of the first substance from the concentrations of the second.

3. Another application of model II regression concerns deterministic models, which are often used to describe ecological processes. In order to test how good a model is at describing reality, one can run the model with observed values of the control variables and compare the values predicted by the model to the observed values of the response variable. Since both sets of variables (control, response) are random, the values predicted by the model are just as random as the values of the response variables, so that they should be related and compared using model II regression. The hypothesis is one-tailed in this case; indeed, a model accurately reflects the field process only if its predictions are *positively* correlated with the field observations. Theory and examples are provided by Mesplé *et al.* (1996).

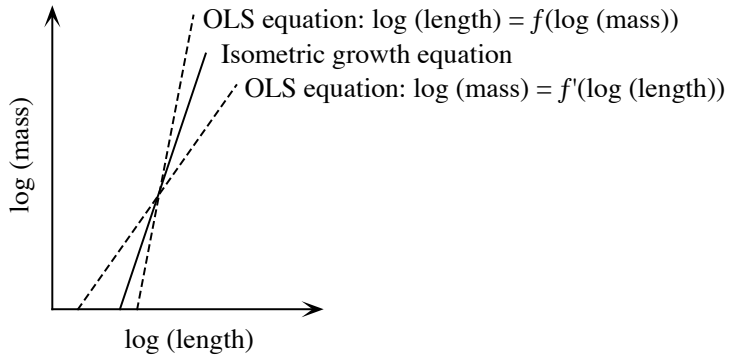


Figure 10.7 Isometric growth is depicted by the functional relationship $\log(mass) = b_0 + 3\log(length)$. The ordinary least-squares (OLS) regression line of $\log(mass)$ on $\log(length)$ would suggest allometric growth of one type, while the OLS regression line of $\log(length)$ on $\log(mass)$ would suggest allometric growth of the opposite type.

In the descriptive examples above, one was interested in estimating the parameters of the equation that describes the functional relationship between pairs of random variables in order to quantify underlying physiological or ecological processes. When both variables are random, as in these examples, model II regression should be used for parameter estimation since the slope found by ordinary least squares (OLS) is too small in absolute value, due to the presence of measurement error in the explanatory variable. OLS regression should only be used when x is fixed by experiment (model I regression) or is a random variable measured with little error compared to y (see recommendation 1 at the end of this subsection). OLS regression (model I or II) should also be used when the objective of the study is forecasting (see recommendation 6).

To better understand the above assertion, let us consider the relationship between length and mass of adult animals of a given species. Let us further assume that the relationship is isometric ($mass = c \times length^3$) for the species under study; this equation would correspond to the case where all individuals, short or long, have the same shape (fatness). The same functional equation, in log form, is $\log(mass) = b_0 + 3\log(length)$, where b_0 is the log of parameter c. Since individual measurements are each subject to a large number of small genetic and environmental influences, presumably additive in their effects and uncorrelated among individuals, it is expected that both length and mass include random deviations from the functional equation; measurement errors must be added to this inherent variability. In such a system, the slope of the OLS regression line of $\log(mass)$ on $\log(length)$ would be smaller than 3 (Fig. 10.7; Ecological application 10.3a), which would lead one to conclude that the species displays allometric growth, with longer individuals thinner than short ones. On the contrary, the slope of the regression line of $\log(length)$ on $\log(mass)$, computed in the transposed space, would produce a slope smaller than 1/3; its inverse, drawn in Fig. 10.7, is larger than 3; this slope would lead to the opposite conclusion, i.e. that shorter individuals are thinner than long ones. This apparent paradox is simply due to the fact that OLS regression is inappropriate to describe the functional relationship between these variables.

Several methods have been proposed to estimate model II regression parameters, and a controversy has raged in the literature about which method was the best. The following methods are the most popular — although, surprisingly, the major statistical packages, except R, are still ignoring them (except method 4, OLS). For methods 1 to 3 described below, slope estimates can easily be calculated with a pocket calculator, from values of the means, variances, and covariance, computed with standard statistical software.

Methods 1, 2, and 4 are special cases of the *structural relationship*, which assumes that there is error ε_i on y and δ_i on x, ε_i and δ_i being independent of each other. As stated above, "error" means deviation of any kind due to a random process, not only measurement error. The maximum likelihood (ML) estimate of the slope for such data is (Madansky, 1959; Kendall & Stuart, 1966):

ML slope formula

$$b_{\mathrm{ML}} = \frac{s_y^2 - \lambda s_x^2 + \sqrt{(s_y^2 - \lambda s_x^2)^2 + 4\lambda s_{xy}^2}}{2 s_{xy}} \qquad \textbf{(10.9)}$$

where s_y^2 and s_x^2 are the estimated variances of y and x, respectively, s_{xy} is their covariance, and λ is the ratio $\sigma_\varepsilon^2 / \sigma_\delta^2$ of the variances of the two error terms.

When λ is large or s_{xy} is very small, another equation form may provide greater computational accuracy than eq. 10.9. It is derived from the property that the slope of the regression line of y on x is the inverse of the slope of the regression of x on y in the case of symmetric regression lines. After the proper substitutions, eq. 10.9 becomes:

$$b_{\mathrm{ML}} = \frac{2 s_{xy}}{s_x^2 - (s_y^2/\lambda) + \sqrt{[s_x^2 - (s_y^2/\lambda)]^2 + (4 s_{xy}^2/\lambda)}} \qquad \textbf{(10.10)}$$

The model II regression methods are derived from eq. 10.9 or eq. 10.10.

Major axis

1. *Major axis regression (MA).* — In this method, the estimated regression line is the first principal component of the scatter of points (see principal component analysis, Section 9.1). The quantity that is minimized is the sum, over all points, of the squared *Euclidean distances* between the points and the regression line (Fig. 10.6b), instead of *vertical distances* as in OLS (Fig. 10.6a). In this method, one assumes that the two error variances (σ_ε^2 on y and σ_δ^2 on x) are equal, so that their ratio $\lambda = 1$. This assumption is strictly met, for example, when both variables have been measured using the same instrument and all of the error is measurement error (McArdle, 1988). The slope of the major axis is estimated by the following formula (Pearson, 1901; Jolicoeur, 1973; Sokal & Rohlf, 1995), which is a special case of eq. 10.9 for $\lambda = 1$:

$$b_{\mathrm{MA}} = \frac{s_y^2 - s_x^2 + \sqrt{(s_y^2 - s_x^2)^2 + 4 (s_{xy})^2}}{2 s_{xy}} \qquad \textbf{(10.11)}$$

The positive square root is used in the numerator. A second equation is obtained by using the negative square root; it estimates the slope of the minor axis, which is the second principal component, of the bivariate scatter of points. When the covariance is near 0, b_{MA} is estimated using eq. 10.10 (with $\lambda = 1$) instead of eq. 10.11, in order to avoid numerical indetermination.

The slope of the major axis may also be calculated using estimates of the slope of the OLS regression line, b_{OLS}, and of the correlation coefficient, r_{xy}:

$$b_{MA} = \frac{d \pm \sqrt{d^2 + 4}}{2} \quad \text{where} \quad d = \frac{(b_{OLS})^2 - r_{xy}^2}{r_{xy}^2 \times b_{OLS}}$$

The positive root of the radical is used when the correlation coefficient is positive, and conversely when it is negative.

Just as with principal component analysis, this method is useful in situations where both variables are expressed in the same physical units or are dimensionless (naturally, or after standardization or ranging). Many natural ecological variables are not in the same physical units. Major axis regression has been criticized because, in that case, the slope estimated by major axis regression is not invariant under an arbitrary change of scale such as expansion (Section 1.5) and, after a change of scale, b_{MA} cannot be directly calculated using the change-of-scale factor. In these conditions, the actual *value* of the slope may be meaningless (Teissier, 1948; Kermack and Haldane, 1950; Ricker, 1973; McArdle, 1988) or difficult to interpret. By comparison, the slopes of the OLS, SMA, and RMA (described below) regression lines are not invariant either to change-of-scale transformations, but the slopes of the transformed data can easily be calculated using the change-of-scale factor. For example, after regressing a mass variable in g onto a length variable in cm, if the OLS slope is b_1 (in g/cm), then after rescaling the explanatory variable from cm to m, the OLS slope becomes $b'_1 = b_1 \times 100$.

Permutation test

Significance of b_{MA} estimates can be tested by permutation (Section 1.2); the values of one or the other variable (i.e. x or y) are permuted a large number of times and slope estimates are computed using eq. 10.11. The test should be carried out on the lesser of the two slopes in absolute value: b_1 of y on x, or $b'_1 = 1/b_1$ of x on y. If the objective is simply to assess the relationship between the two variables under study, the correlation coefficient should be tested for significance instead of the slope of a model II regression line.

When the variances s_y^2 and s_x^2 are equal, the slope estimated by eq. 10.11 is ±1, the sign being that of the covariance, whatever the value of s_{xy}. As in the case of SMA (below), permutations produce slope estimates of +1 or –1 in equal numbers, with a resulting probability near 0.5 whatever the value of the correlation. This result is meaningless. The practical consequence is that, if the slope estimate b_{MA} is to be tested by permutations, variables should not be standardized (eq. 1.12).

C.I. of MA slope

Alternatively, one may compute the confidence interval of the slope at a predetermined confidence level and check whether the value 0 (or, for that matter, any other value of interest) lies inside or outside the confidence interval. Computation of the confidence interval involves several steps; the formulae are given in Jolicoeur & Mosimann (1968), Jolicoeur (1990), and Sokal & Rohlf (1995, pp. 589-591), among others. When both n and the ratio of the eigenvalues of the bivariate distribution (see principal component analysis, Section 9.1) are small, limits of the confidence interval cannot be computed because it covers all 360° of the plane. Such a confidence interval always includes slope 0, as well as any other value. For example, when $n = 10$, the ratio of the eigenvalues must be larger than 2.21 for the 95% confidence interval to be real; for $n = 20$, the ratio must be larger than 1.63; and so on.

It frequently happens in ecology that a scatter plot displays a bivariate lognormal distribution; the univariate frequency distributions of such variables are positively skewed, with longer tails in the direction of the higher values. Such distributions may be normalized by applying a log transformation (Subsection 1.5.6; Fig. 1.11). This transformation also solves the problem of dimensionally heterogeneous variables and makes the estimate of the major axis slope invariant over expansion (multiplication or division by a constant: Section 1.5) — but not over translation. One should verify, of course, that the log-transformed data conform to a bivariate normal distribution before proceeding with major axis regression.

This property can easily be demonstrated as follows. Consider a model II functional equation describing the linear relationship between two log-transformed variables x and y:

$$\log(y) = b_0 + b_1 \log(x)$$

If x and y are divided by constants c_1 and c_2 respectively (expansion), one obtains new variables $x' = x/c_1$ and $y' = y/c_2$, so that $x = c_1x'$ and $y = c_2y'$. The functional equation becomes:

$$\log(c_2y') = b_0 + b_1\log(c_1x')$$

$$\log(y') + \log(c_2) = b_0 + b_1\log(c_1) + b_1\log(x')$$

$$\log(y') = [b_0 + b_1\log(c_1) - \log(c_2)] + b_1\log(x')$$

which may be rewritten as

$$\log(y') = b_0' + b_1\log(x')$$

where $b_0' = [b_0 + b_1\log(c_1) - \log(c_2)]$ is the new intercept, while the slope of $\log(x')$ is still b_1. So, under log transformation, the slope b_1 is invariant for any values of expansion coefficients c_1 and c_2; it differs, of course, from the major axis regression coefficient (slope) of the untransformed variables.

Dividing x and y by their respective standard deviations, s_x and s_y, is an expansion which makes the two variables dimensionless. It thus follows that the major axis slope of the original log-transformed data is the same as that of the log of the standardized (dimensionless) data. This

also applies to other standardization methods such as division by the maximum value or the range (eqs. 1.10 and 1.11).

Readers who prefer numerical examples can easily check the above derivation by computing a principal component analysis on a small data set containing two log-transformed variables only, with or without expansion (multiplication or division by a constant prior to the log transformation). The angles between the original variables and the first principal component are easily computed as the $\cos^{-1}$ of the values in the first normalized eigenvector (Subsection 9.1.3); the slopes of the major axis regression coefficients of $y = f(x)$ and $x = f(y)$, which are the tangents (tan) of these angles, remain the same over such a transformation.

Standard major axis

2. *Standard major axis (SMA).* — Regression using variables that are not dimensionally homogeneous produces results that vary with the scales of the variables. If the physical dimensions are arbitrary (e.g. length measurements that may indifferently be recorded in mm, cm, m, or km), the slope estimate is also arbitrary. In ordinary least-squares regression (OLS), the slope and confidence interval values change proportionally to the measurement units. For example, multiplying all y values by 10 produces a slope estimate ten times larger, whereas multiplying all x values by 10 produces a slope estimate 10 times smaller. This is not the case with MA; the major axis slope does not scale proportionally to the units of measurement. For that reason, it may be desirable to make the variables dimensionally homogeneous prior to model II regression.

Standard major axis regression is MA regression performed on standardized variables, which are thus dimensionally homogeneous. It is computed as follows:

- Standardize variables x and y using eq. 1.12.
- Compute MA regression on the standardized variables. The slope estimate is always +1 or –1; the sign is that of the covariance s_{xy} or correlation coefficient r_{xy}.
- Back-transform the slope estimate to the original units by multiplying it by (s_y/s_x).

As a consequence, the slope of the *standard major axis* (SMA), or *reduced major axis*, is computed as the ratio (Teissier, 1948):

$$b_{\mathrm{SMA}} = \sqrt{s_y^2/s_x^2} = \mathrm{sign}\,(r) \times (s_y/s_x) \qquad \textbf{(10.12)}$$

where sign(r) is the sign of the correlation coefficient. This formula is obtained from eq. 10.9 by assuming that the error variances σ_ε^2 and σ_δ^2 of y and x, respectively, are identically proportional to their respective variances σ_y^2 and σ_x^2; in other words, $\sigma_\varepsilon^2/\sigma_y^2 = \sigma_\delta^2/\sigma_x^2$. This assumption is unlikely to be strictly true with real data, except in cases where both variables are counts (e.g. numbers of organisms), raw or log-

transformed (McArdle, 1988). Replacing variances σ_y^2 and σ_x^2 by their unbiased estimates s_y^2 and s_x^2 gives the following value to λ in eq. 10.9:

$$\lambda = \sigma_\varepsilon^2 / \sigma_\delta^2 = \sigma_y^2 / \sigma_x^2 = s_y^2 / s_x^2$$

Equation 10.9 then simplifies to eq. 10.12. Since the square root $\sqrt{s_y^2 / s_x^2}$ is either positive or negative, the slope estimate receives the sign of the Pearson correlation coefficient, which is the same as that of the covariance s_{xy} in the denominator of eq. 10.9 or that of the OLS slope estimate. The b_{SMA} estimate is also the geometric mean of the OLS regression coefficient of y on x and the *reciprocal* of the regression coefficient of x on y; this is why the method is also called *geometric mean regression*, besides a variety of other names.

From equations 4.7 (Pearson r), 10.3 (b_{OLS}) and 10.12 (b_{SMA}), one can show that

$$b_{SMA} = |b_{OLS}| / r_{xy} \quad \text{when} \quad r_{xy} \neq 0 \qquad \textbf{(10.13)}$$

So, in addition to eq. 10.12, one can easily compute b_{SMA} from eq. 10.13 using values of b_{OLS} and r_{xy} provided by an OLS regression program. This equation also shows that, when the variables are highly correlated ($r \to 1$), $b_{SMA} \to b_{OLS}$. When they are not, b_{SMA} is always larger than b_{OLS} for positive values of r, and smaller for negative values of r; in other words, b_{OLS} is always closer to 0 than b_{SMA}.

When $r_{xy} = 0$, the b_{SMA} estimate obtained from eq. 10.12, which is the ratio of the standard deviations, is meaningless. It does not fall to zero when the correlation is zero, except in the trivial case where s_y is zero (Jolicoeur, 1975, 1990). Since the b_{SMA} estimate is independent of the presence of a significant covariance between x and y (eq. 10.12), users should always compute a Pearson correlation coefficient and test it for significance prior to computing the slope of a standard major axis regression line. If r is not significantly different from zero, b_{SMA} should not be computed.

The slope of the standard major axis cannot be tested for significance by a regular permutation test. There are two reasons for this.

Permutation test

- Consider permutation testing. The b_{SMA} slope estimate is $\pm s_y / s_x$ but, for all permuted data, s_y / s_x is a constant. Giving the signs of the permuted covariances to the permuted slope estimates inevitably produces a probability near 0.5 of obtaining, by permutation, a value as extreme as or more extreme than the estimate b_{SMA}.

- The confidence interval of the slope b_{SMA}, described below, is inappropriate to test the null hypothesis $\beta = 0$ because the ratio s_y / s_x cannot be zero unless s_y is equal to zero. This is a trivial case, unsuitable for regression analysis (Sokal & Rohlf, 1995).

McArdle (1988) suggests that the solution to this problem is to test the correlation coefficient r_{xy} for significance instead of testing b_{SMA}. Warton *et al.* (2006, Appendix F) describe a permutation test of the SMA slope based on residuals.

C.I. of SMA slope

When needed, an approximate confidence interval $[b_1, b_2]$ can be computed for b_{SMA} as follows (Jolicoeur & Mosimann, 1968):

$$b_1 = b_{SMA}\left[\sqrt{(B+1)} - \sqrt{B}\right]$$

$$b_2 = b_{SMA}\left[\sqrt{(B+1)} + \sqrt{B}\right]$$

where $$B = t^2\,(1 - r^2) \,/\, (n-2)$$

and t is a two-tailed Student's $t_{\alpha/2}$ value for significance level α and $(n-2)$ degrees of freedom.

Ranged major axis

3. *Ranged major axis regression (RMA).* — An alternative transformation to make the variables dimensionally homogeneous is *ranging* (eqs. 1.10 and 1.11). This transformation does not make the variances equal and thus does not lead to the problems encountered with SMA regression. It leads to RMA, which proceeds as follows:

- Transform the y and x variables into y' and x', respectively, using eq. 1.11. For relative-scale variables (Subsection 1.4.1), which have zero as their natural minimum, the ranging transformation is carried out using eq. 1.10.

- Compute MA regression between the ranged variables y' and x'. Test by permutation if a test is required.

- Back-transform the estimated slope and confidence interval limits to the original units by multiplying them by the ratio of the ranges, $(y_{max} - y_{min})/(x_{max} - x_{min})$.

The RMA slope estimator has several desirable properties when variables x and y are not expressed in the same units. The slope estimator scales proportionally to the units of x and y. The estimator is not insensitive to the covariance, as is the case for SMA. Finally, it is possible to test the hypothesis that an RMA slope estimate is equal to a stated value, in particular 0 or 1. As in MA, this may be done either by permutations, or by comparing the confidence interval of the slope to the hypothetical value of interest. Thus, whenever MA regression cannot be used because of incommensurable units, RMA regression can be used. There is no reason, however, to use RMA when the variables are expressed in the same units.

Prior to RMA, one should check for the presence of outliers, using a scatter diagram of the objects. Outliers cause important changes to the estimates of the ranges of the variables. Outliers that are not aligned with the bulk of the objects may thus have an undesirable influence on the slope estimate. RMA should not be used in the presence of such outliers.

OLS method

4. *Ordinary least squares (OLS).* — The OLS method is derived from eq. 10.10 by assuming that there is no error on x, so that the error variance on x, σ^2_δ, is zero and thus $\lambda = \sigma^2_\varepsilon / \sigma^2_\delta = \infty$. After simplification, the OLS slope is equal to (eq. 10.3)

$$b_{\mathrm{OLS}} = s_{xy} / s^2_x$$

The remainder of the subsection is devoted to the description of general properties and the comparison of model II regression methods.

C.I. of intercept

With all methods of model II regression, an estimate of the intercept, b_0, can be computed from b_1 and the centroid of the scatter of points $(\bar{x}, \bar{y})$, using eq. 10.3. The same equation can be used to calculate *approximate estimates* of the confidence limits of the intercept. Warton *et al.* (2006) describe more precise estimates of these confidence limits.

The first three methods (MA, SMA, RMA) have the property that the slope of the regression $y = f(x)$ is the reciprocal of the slope of $x = f(y)$. This property of symmetry is desirable here since there is no functional distinction between x and y in a model II situation. OLS regression does not have that property (Fig. 10.6a).

Users of model II regression techniques are never certain that the assumptions of the various methods are met by the variables in the data sets (i.e. MA: $\sigma^2_\varepsilon = \sigma^2_\delta$ so that $\lambda = \sigma^2_\varepsilon / \sigma^2_\delta = 1$; SMA: $\lambda = \sigma^2_y / \sigma^2_x$; OLS: $\sigma^2_\delta = 0$ so that $\lambda = \sigma^2_\varepsilon / \sigma^2_\delta = \infty$). For that reason, McArdle (1988) carried out an extensive simulation study to investigate the influence of the error variances, σ^2_ε for y and σ^2_δ for x, on the efficiency (i.e. precision of the estimation) of the MA, SMA and OLS methods, measuring how variable the estimated slopes were under various conditions. Likewise, Jolicoeur (1990) used simulations to investigate the effects of small sample sizes and low correlations on the slope estimates obtained by MA and SMA. D. J. Currie, P. Legendre and A. Vaudor (unpublished study) also used numerical simulations to investigate the relationship between slope estimate formulas. They compared MA to OLS and MA to SMA in the *correlation situation*, defined as that where researchers are interested in describing the slope of the bivariate relationship displayed by two correlated random variables, i.e. variables that are not controlled or error-free. The results of all these simulations lead to the following recommendations for the estimation of parameters of functional linear relationships between variables that are random (i.e. not controlled) and measured with error (Table 10.4). They were first presented in a guide (Legendre, 2008b) distributed with the R package LMODEL2.

Recom-mendations

1. If the magnitude of the random variation (i.e. the error variance*) on the response variable y is much larger (i.e. more than three times) than that on the explanatory variable x, use OLS as the model II regression method. Otherwise, proceed as follows.

* Contrary to the sample variance, the error variance on x or y cannot be estimated from the data. It can only be estimated from knowledge of the way the variables were measured.

Table 10.4 Recommendations for the application of the model II regression methods. The numbers refer to the corresponding recommendation paragraphs (recom.) in the text.

- The error on y is much larger than the error on x: use OLS (recom. 1)
- The data distribution is close to bivariate normal (recom. 2)
  - The variables are in the same physical units or dimensionless, the error variance is about the same for x and y: use MA (recom. 3)
  - The variables are not dimensionally homogeneous. The error variance along each axis is proportional to the variance of the corresponding variable (recom. 4)
    - There are no outliers in the scatter diagram: RMA can be used (recom. 4.1)
    - The Pearson correlation coefficient r is significant: SMA can be used (recom. 4.2)
- The data distribution is clearly not bivariate normal (recom. 2)
  - The relationship between x and y is linear: use OLS (recom. 5)
- The objective is to compute forecasted (i.e. fitted) values $\hat{y}$: use OLS (recom. 6)
- The objective is to compare observations to model predictions: use MA (recom. 7)

2. Check whether the data are approximately bivariate normal, either by examining a scatter diagram or by performing a formal test of significance. If they are not, attempt transformations to make the distribution bivariate normal. For data that are or can be made to be reasonably bivariate normal, consider recommendations 3 and 4. If not, see recommendation 5.

3. For bivariate normal data, if the two variables are expressed in the same physical units (untransformed variables that were originally measured in the same units) or are dimensionless (e.g. log-transformed variables), and if it can reasonably be assumed that the error variances of the variables are approximately equal, use major axis (MA) regression.

When no information is available on the ratio of the error variances and there is no reason to believe that it may differ from 1, MA may be used provided that the results are interpreted with caution. MA produces unbiased slope estimates and accurate confidence intervals (Jolicoeur, 1990).

MA can be used with dimensionally heterogeneous variables (1) when the purpose of the analysis is to compare slopes computed from these variables measured in an identical way in different systems (e.g. at two or more sampling sites). It may also be useful (2) when the objective of the study is to test the hypothesis that the slope of the major axis of the empirical data does not differ from a value given by theory.

4. For bivariate normal data, if MA cannot be used because the variables are not expressed in the same physical units or the error variances on the two axes differ, two methods are available to estimate the parameters of the functional linear relationship if it can be assumed that the error variance on each axis is proportional to the variance of the corresponding variable, i.e. (error variance of y / sample variance of y) ≈ (error variance of x / sample variance of x). This condition is often met with counts (e.g. number of plants or animals) or log-transformed data (McArdle, 1988). The two following methods can be used if their specific conditions are met by the data.

4.1. Ranged major axis regression (RMA) can be used if there are no outliers in the scatter of points. Prior to RMA, one should check for the presence of outliers, using a scatter diagram of the objects.

4.2. Standard major axis regression (SMA) can be used if the coefficient of linear correlation (Pearson r) is significant. SMA regression should not be computed when this condition is not met.

The SMA slope cannot be tested by a standard permutation test, but the correlation coefficient r can. See also Warton *et al*. (2006, Appendix F) for a permutation test of the SMA slope based on residuals. Confidence intervals should be used with caution: simulations have shown that, as the slope departs from ±1, the SMA slope estimate is increasingly biased and the confidence interval includes the true value less and less often. Even when the slope is near ±1, the confidence interval is too narrow if n is very small or if the correlation is weak.

5. If the distribution is not bivariate normal and the data cannot be transformed to satisfy that condition (e.g. if the distribution possesses two or several modes), one should wonder whether the slope of a regression line is really an adequate model to describe the functional relationship between the two variables. Since the distribution is not bivariate normal, there seems little reason to apply models such as MA, SMA or RMA, which primarily describe the first principal component of a bivariate normal distribution. So, (1) if the relationship is linear, OLS is recommended to estimate the parameters of the regression line. The significance of the slope should be tested by permutation, however, because the distributional assumptions of the parametric test are not satisfied. (2) If a straight line is not an appropriate model, polynomial or nonlinear regression should be considered.

6. When the purpose of the study is not to estimate the parameters of a functional relationship, but simply to forecast or predict values of y for given x's, use OLS in all cases. OLS is the only method that minimizes the squared residuals in y. The OLS regression line itself is meaningless. Do not use the OLS standard error and confidence bands unless x is known to be free of error (Sokal and Rohlf, 1995: 545, Table 14.3).

7. Observations may be compared to the predictions of a statistical or deterministic model (e.g simulation model) in order to assess the quality of the model. If the model contains random variables measured with error, use MA for the comparison when the observations and model predictions are in the same units.

If the model fits the data well, the MA slope is expected to be 1 and the intercept 0. A slope that significantly differs from 1 indicates a *difference* between observed and simulated values that is proportional to the observed values. For relative-scale variables, a MA intercept that significantly differs from 0 suggests the existence of a systematic difference between observations and simulations (Mesplé *et al*., 1996).

8. With all methods, the confidence intervals are large when n is small; they become smaller as n goes up to about 60, after which they change much more slowly. Model II regression should ideally be applied to data sets containing 60 observations or more.

Numerical examples illustrating the cases found in Table 10.4 are described in Legendre (2008b). The data and R script are found in the help file of the ***lmodel2()*** function (Section 10.7). Other interesting examples are found in Warton *et al*. (2006).

Ecological application 10.3a

Laws & Archie (1981) re-analysed data published in two previous papers that had quantified the relationships between the log of respiration rates and the log of biomass for zooplankton under various temperature conditions. The authors of the original papers had computed OLS slopes and confidence intervals (model I regression) of the biomass-respiration relationships for each temperature condition. They had come to the conclusions (1) that the *surface law*, which states that the slope of the log-log relationship should fall between 0.66 and 1.00, was not verified by the data, and (2) that the slope significantly varied as a function of temperature. Based on the same data, Laws & Archie recomputed the slopes using the standard major axis method. They found that all slopes were larger than estimated by OLS (same phenomenon as in Fig. 10.7) and that none of them was significantly outside the 0.66 to 1.00 interval predicted by the surface law. Furthermore, comparing the slopes of the different temperature data sets at $\alpha = 0.02$, they found that they did not differ significantly from one another.

3 — Multiple linear regression

When there are several explanatory variables $x_1, x_2, \ldots, x_m$, it is possible to compute a regression equation where the response variable y is a linear function of all explanatory variables x_j. The multiple linear regression model is a direct extension of simple linear regression:

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_m x_{im} + \varepsilon_i \qquad \textbf{(10.14)}$$

for object i. Equation 10.14 leads to the well-known formula for the fitted values:

$$\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_m x_{im} \qquad \textbf{(10.15)}$$

Using ordinary least squares (OLS), the vector of regression parameters $\mathbf{b} = [b_j]$ is easily computed from matrix eq. 2.19: $\mathbf{b} = [\mathbf{X}'\mathbf{X}]^{-1}\,[\mathbf{X}'\mathbf{y}]$. If an intercept ($b_0$) must be estimated, a column of 1's is added to matrix $\mathbf{X}$ of the explanatory variables. QR decomposition (Section 10.7) is an alternative, computer-efficient method for the computation of regression coefficients in univariate or multivariate regression.

Equation 10.15 provides a model I estimation, which is valid when the x_j variables have been measured without error. This is the only method presently available in commercial statistical packages and, for this reason, it is the multiple regression model most widely used by ecologists. McArdle (1988) proposed a multiple regression method, the *standard minor axis*, to be used when the explanatory variables of the model are random (i.e. with measurement error or natural variability). McArdle's standard minor axis is the multivariate equivalent of the standard major axis (SMA) method described in the previous subsection.

Standard minor axis

Another approach is *orthogonal distance regression* (ODR), computed through generalized least squares. The method minimizes the sum of the squares of the orthogonal distances between each data point and the curve described by the model equation; this is the multivariate equivalent of the major axis regression (MA) method described in the previous subsection. ODR is used extensively in econometrics. Boggs & Rogers (1990) give entry points to the numerous papers that have been published on the subject in the computer science and econometric literature and they propose an extension of the method to nonlinear regression modelling. They also give references to ODRPACK*, a public-domain collection of FORTRAN subprograms for *weighted orthogonal distance regression*, which allows estimation of the parameters that minimize the sum of squared weighted orthogonal distances from a set of observations to the curve or surface determined by the parameters.

Orthogonal distance regression

When the same multiple regression model is to be computed for several response variables $y_1, \ldots, y_i, \ldots, y_p$, regression coefficients can be estimated by ordinary least squares for all response variables simultaneously, using a single matrix expression:

$$\hat{\mathbf{B}} = [\mathbf{X}'\mathbf{X}]^{-1}[\mathbf{X}'\mathbf{Y}]$$

The procedure is called *multivariate linear regression* (Finn, 1974). In this expression, which is the multivariate equivalent of eq. 2.19, $\mathbf{X}$ is the matrix of explanatory variables, $\mathbf{Y}$ is the matrix of the p response variables, and $\hat{\mathbf{B}}$ is the matrix of regression coefficients. The coefficients found using this equation are the same as those obtained from multiple regressions computed in separate runs for each response variable. The multivariate matrix of fitted values is obtained by the following matrix expression, which will serve as the basis for redundancy analysis (eq. 11.3) in Section 11.1:

Multivariate linear regression

$$\hat{\mathbf{Y}} = \mathbf{X}\,[\mathbf{X}'\mathbf{X}]^{-1}\,\mathbf{X}'\mathbf{Y} \qquad \textbf{(10.16)}$$

Two types of regression coefficients can be computed in regression analysis.

- *Ordinary regression coefficients*, represented by symbols b, are computed on the original variables. The physical dimension of coefficient b_j associated with explanatory variable x_j is (dimension of y / dimension of x_j). These regression

* ODRPACK is available from the following Web site: <http://www.netlib.org/odrpack/>.

coefficients are useful when the regression equation is to be used to compute estimated values of y for objects that have not been involved in the estimation of the regression parameters, and for which y and x values are available. This is the case, for instance, when a regression model is validated using a new set of observations: estimates $\hat{y}$ are computed from the regression equation to be validated, using the observed values of the explanatory variables x_j, and they are compared to the corresponding observed y's, to assess how efficient the regression model is at calculating y for new data.

• In contrast, *standard regression coefficients*, often represented by symbols b', are computed on standardized variables **X** and **y**. Standard regression coefficients are dimensionless. These regression coefficients are useful as a means of assessing the relative importance of each explanatory variables x_j included in the regression model: the variables with the highest standard regression coefficients (in absolute values) are those that contribute the most to the estimated $\hat{y}$ values. The relationship between coefficients b and b' obtained by ordinary least-squares estimation is: $b'_{yx_j} = b_{yx_j} s_{x_j} / s_y$, where b_{yx_j} is the partial regression coefficient for explanatory variable x_j.

It is interesting and important to note that, for the objects that were used to estimate the regression parameters, the fitted values $\hat{y}$ computed from the ordinary regression coefficients (hence from the original variables) are identical to the fitted values computed from standard regression coefficients (i.e. from the standardized variables).

Partial regression coefficient

Both the ordinary and standard regression coefficients in multiple regression are *partial regression coefficients*. The term *partial* means that each regression coefficient is a measure, standardized or not, of the rate of change that variable y would have per unit of variable x_j, if all the other explanatory variables in the study were held constant. The concept of partial regression is further developed in Subsection 10.3.5. Partial regression coefficients can be tested by permutation using methods similar to those described in Subsection 11.1.8 for canonical redundancy analysis (RDA).

Collinearity

When the explanatory variables x_j of the model are uncorrelated, multiple regression is a straightforward extension of simple linear regression. In experimental work, controlled variables may satisfy this condition if the experiment has been planned with care and the design is balanced. With observational data, however, the explanatory variables used in multiple regression models are most often collinear (i.e. correlated to one another), and it will be seen that strong collinearity may affect the ability to correctly estimate the regression parameters. How to deal with this problem will depend on the purpose of the analysis. If one is primarily interested in forecasting, the objective is to maximize the coefficient of multiple determination (called R^2 in multiple regression); collinearity of the explanatory variables is not a concern. For description or inference, however, the primary interest is to correctly estimate the parameters of the model; the effect of multicollinearity on the estimates of the model parameters must then be minimized.

Identify collinear variables

Prior to regression, different methods can be used to identify fully or highly collinear variables.

- One can check if the group of explanatory variables is of full rank. This can be done by singular value decomposition (SVD) of the data matrix (Section 2.11, Application 1): the matrix is not of full rank if one or more of the singular values are 0. Alternatively, one can compute the determinant of the covariance matrix of a group of variables: the determinant is 0 if the group includes variables that are linearly dependent on other variables in the group (Section 2.6, property 5; Section 2.7).

- If the rank of the matrix is smaller than its order, check subgroups of explanatory variables. Place the variables in an order that seems suitable; for example, put the most ecologically informative or easy-to-measure variables first. Compute SVD of the matrix containing the first two variables, then the first three, and so on. SVD produces a singular value of zero when a variable that is fully collinear with the previous ones is included in the group. When identified, remove the fully collinear variable from the set of explanatory variables and resume the exploration of the remaining variables.

- For variables that are not fully collinear, compute the extent to which each variable is collinear with the other variables in the group. This is done by computing *variance inflation factors* (*VIF*; Neter *et al.,* 1996, their Sections 9.5 and 10.2). Each variable *j* is regressed, in turn, on all the other variables in the group and the coefficient of determination (R_j^2, eq. 10.20) of that regression model is noted. The *VIF* for variable *j* is computed as follows:

VIF

$$VIF_j = \frac{1}{1 - R_j^2} \qquad \textbf{(10.17)}$$

All *VIF* coefficients can actually be found in a single operation by computing the inverse of the correlation matrix, $\mathbf{R}^{-1}$, among the variables in the group under study; the diagonal elements of that inverse matrix are the *VIF* coefficients. VIF_j is 1 for a variable *j* that has correlations of 0 with all the other variables in the group, and is larger than 1 when the correlations between *j* and some or all the other variables differ from 0 (positive or negative correlation values). Variables that have high *VIF* coefficients can be scrutinized and considered as candidates for elimination from the group of explanatory variables. Different cut-off values have been proposed to identify highly collinear variables: $VIF > 5$, or > 10 (Neter *et al.,* 1996), or > 20 (ter Braak & Smilauer, 2002).

The effect of collinearity on the estimates of regression parameters may be described as follows. Let us assume that one is regressing *y* on two explanatory variables x_1 and x_2. If x_1 is uncorrelated to x_2, the variables form a well-defined Cartesian plane. If *y* is represented as an axis orthogonal to that plane, a multiple linear regression equation corresponds to a plane in the three-dimensional space; this plane represents the variation of *y* as a linear function of x_1 and x_2. If x_1 is strongly correlated (i.e. collinear) to x_2, the axes of the base plane form an acute angle instead of being at

right angle. In the limit situation where $r(x_1, x_2) = 1$, they become a single axis. With such correlated explanatory variables, the angles determined by the slope coefficients (b_1 and b_2), which set the position of the regression plane in the x_1–x_2–y space, are more likely to be unstable; their values may change depending on the random component ε_i in y_i. In other words, two samples drawn from the same statistical population may be modelled by regression equations with very different parameters — even to the point that the signs of the regression coefficients may change.

Simulation is the easiest way to illustrate the effect of collinearity on the estimation of regression parameters. Vectors $\mathbf{x_1}$ and $\mathbf{x_2}$ were generated, each containing 100 random normal deviates N(0,1), and assembled into an explanatory matrix $\mathbf{X}$. Because the data were generated at random, vectors $\mathbf{x_1}$ and $\mathbf{x_2}$ should be uncorrelated. Actually, the correlation between them was –0.002. The *control data set* was completed by computing a response variable $\mathbf{y_1}$ as the sum of $\mathbf{x_1}$ and $\mathbf{x_2}$, to which a random component was added in the form of an error term ε composed of random normal deviates N(0,2):

$$y_{1i} = x_{i1} + x_{i2} + \varepsilon_i$$

For the *test data set*, two correlated explanatory variables $\mathbf{w_1}$ and $\mathbf{w_2}$ were created by multiplying matrix $\mathbf{X}$ by the square root of a correlation matrix stating that the correlation between $\mathbf{x_1}$ and $\mathbf{x_2}$ should be 0.8:

$$\mathbf{W} = [\mathbf{w_1}, \mathbf{w_2}] = \mathbf{X}\mathbf{R}^{0.5} \quad \text{where} \quad \mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}] \text{ and } \mathbf{R}^{0.5} = \begin{bmatrix} 1.0 & 0.8 \\ 0.8 & 1.0 \end{bmatrix}^{0.5} = \begin{bmatrix} \sqrt{0.8} & \sqrt{0.2} \\ \sqrt{0.2} & \sqrt{0.8} \end{bmatrix}$$

$\mathbf{R}^{0.5}$ is computed using eq. 2.29; Cholesky factorization (Section 2.12) of $\mathbf{R}$ may be used instead of square root decomposition. Since $\mathbf{x_1}$ and $\mathbf{x_2}$ are N(0,1) random deviates, they have expected values of 0 and are orthogonal for large n. The covariance matrix of $\mathbf{W}$ can be developed as follows, which shows that its expected value is equal to the imposed correlation matrix $\mathbf{R}$:

$$\frac{1}{n-1}\mathbf{W}'\mathbf{W} = \frac{1}{n-1}\mathbf{R}^{0.5}\mathbf{X}'\mathbf{X}\mathbf{R}^{0.5} = \mathbf{R}^{0.5}\left(\frac{1}{n-1}\mathbf{X}'\mathbf{X}\right)\mathbf{R}^{0.5} = \mathbf{R}^{0.5}\mathbf{I}\mathbf{R}^{0.5} = \mathbf{R}$$

For the simulated data, the correlation between $\mathbf{w_1}$ and $\mathbf{w_2}$ turned out to be 0.801, which is very close to 0.8. The test data set was completed by computing a variable $\mathbf{y_2}$ from $[\mathbf{w_1}, \mathbf{w_2}]$ with the same error term ε as in the equation for $\mathbf{y}_1$ above:

$$y_{2i} = w_{i1} + w_{i2} + \varepsilon_i$$

Each data matrix was divided into five independent groups of 20 observations each, and multiple regression equations were computed; the groups were independent of one another since the generated data were not autocorrelated. Results are shown in Table 10.5. Note the high variability of the slope estimates obtained for the test data groups (lower panel, with collinearity in the explanatory variables) compared to the control data groups (upper panel). In two cases, the signs of the regression coefficients were changed: for b_1 in group 5 and for b_2 in group 1.

Parsimony

When trying to find the 'best' possible model describing an ecological process, another important aspect is the principle of *parsimony*, also called *Ockham's razor*.

Table 10.5 Parameters of the multiple regression equations for two data sets, each divided into five groups of 20 objects. Top: control data where variables x_1 and x_2 are uncorrelated. Bottom: test data with $r(\mathbf{x_1}, \mathbf{x_2}) \approx 0.8$. Note how the range and standard deviation statistics indicate higher slope variability among the test groups (lower panel). The intercepts are the same in the two panels.

$\hat{y}_1 = b_0 + b_1 x_1 + b_2 x_2 \quad \Rightarrow$	b_0	b_1	b_2
Group 1	0.922	1.457	0.247
Group 2	0.002	–0.033	1.032
Group 3	0.494	1.264	1.206
Group 4	0.343	0.614	0.339
Group 5	0.209	0.410	1.410
Mean	0.394	0.742	0.847
Range of slope estimates = Max – Min		1.491	1.163
Standard deviation of slope estimates		0.615	0.524

$\hat{y}_2 = b_0 + b_1 w_1 + b_2 w_2 \quad \Rightarrow$	b_0	b_1	b_2
Group 1	0.922	1.988	–0.718
Group 2	0.002	–0.819	1.563
Group 3	0.494	0.985	0.855
Group 4	0.343	0.663	0.048
Group 5	0.209	–0.440	1.796
Mean	0.394	0.475	0.709
Range of slope estimates = Max – Min		2.807	2.514
Standard deviation of slope estimates		1.129	1.050

Ockham's razor

This principle, formulated by the English logician and philosopher William Ockham (1290-1349), professor at Oxford University, states that

Pluralites non est ponenda sine necessitate

which literally translates: "Multiplicity should not be posited without necessity". In other words, unnecessary assumptions should be avoided (i.e. "shaved away") when formulating hypotheses. Following this principle, parameters should be used with parsimony in modelling, so that any parameter that does not significantly contribute to the model (e.g. by increasing the R^2 coefficient in an important way, or by decreasing *AIC*) should be eliminated. Indeed, any model containing as many parameters as the number of data points can be adjusted to perfectly fit the data. The corresponding 'cost' is that there is no degree of freedom left to test its significance, hence the 'model' cannot be extended to any other situation.

When the explanatory variables of the model are orthogonal to one another (no collinearity, for example among the controlled factors of well-planned and balanced factorial experiments), applying Ockham's razor is easy: one can remove from the model any variable whose contribution (slope parameter) is not statistically significant. Tests of significance for the partial regression coefficients (i.e. the individual b's) are described in standard textbooks of statistics. The task is not that simple, however, with observational data, because these often display various degrees of collinearity. The problem is that significance may get 'diluted' among collinear variables contributing in the same way to the explanation of a response variable y. Consider a data set where an explanatory variable x_1 makes a significant contribution to a regression model; introducing a highly correlated copy of x_1 in the calculation is usually enough to make the contribution of each copy non-significant, simply as the result of the collinearity that exists between copies (if the second copy is a perfect copy of x_1, the regression coefficients must be computed using a generalized inverse; see Section 2.11, Application 3). Linear dependence (or full collinearity) in a group of explanatory variables is easy to detect; see *Identify collinear variables* in the margin a few pages above. Multicollinearity (without full collinearity) among explanatory variables is measured by *VIF* coefficients (eq. 10.17). Hocking (1976) compared a number of methods proposed for selecting variables in linear regression exhibiting collinearity.

Some statistical programs offer procedures that allow one to compute and compare all possible regression submodels for a small set of k explanatory variables. When such a procedure is not available and one does not want to manually test all possible models, heuristic methods that have been developed for selecting the 'best' subset of explanatory variables may be used, although with caution. The explanatory variables with the strongest contributions may be chosen by backward elimination, forward selection, or stepwise procedure. The three strategies do not necessarily lead to the same selection of explanatory variables.

Backward elimination

- The *backward elimination procedure* is easy to understand. All variables are initially included and, at each step, the variable that contributes the least to explaining the response variable (usually that with the smallest partial correlation) is removed, until all explanatory variables remaining in the model have a significant partial regression coefficient. Some programs express the selection criterion in terms of a F-to-remove (F-statistic for testing the significance of the partial regression coefficient) or a p-to-remove criterion (same, but expressed in terms of probability), instead of the value of the partial correlation, or else in terms of AIC or AIC_c (eqs. 10.22 and 10.23, below).

Forward selection

- The *forward selection procedure* starts with no explanatory variable in the model. The variable entered is the one that produces the largest increase in R^2, provided this increase is significantly different from zero using a predetermined significance level. The procedure is iteratively repeated until no more explanatory variable can be found that produces a significant increase in R^2. Calculations may be simplified by computing partial correlations for all variables not yet in the model, and only testing the significance of the largest partial correlation. Again, some programs base the final decision for including an explanatory variable on a F-to-enter value, which is

equivalent to using the actual probability values, or on *AIC* or AIC_c (eqs. 10.22 and 10.23, below). The major problem with forward selection is that all variables included at previous steps are kept in the model, even though some of them may finally contribute little to the R^2 after incorporation of some other variables.

Stepwise procedure

• The latter problem may be alleviated by the *stepwise procedure*, which alternates between forward selection and backward elimination. After each step of forward inclusion, the significance of all the variables in the model is tested, and those that are not significant are excluded before the next forward selection step.

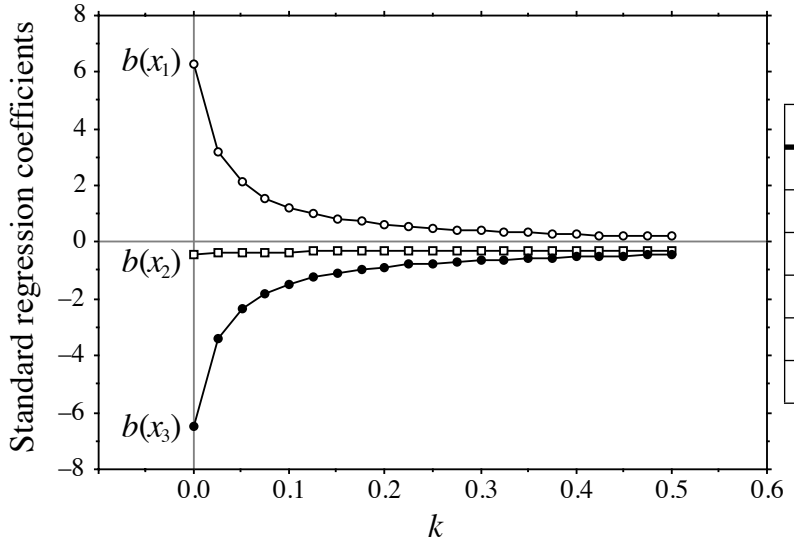
In any case, a problem common to all stepwise inclusion procedures remains: when a model with, say, k explanatory variables has been selected, the procedure offers no guarantee that there does not exist another subset of k explanatory variables, with significant partial correlations, that would explain together more of the variation of y (larger R^2) than the subset selected by stepwise procedure. Furthermore, Sokal & Rohlf (1995) warn users that, after doing repeated tests, the probability of type I error is far greater than the nominal significance value α. The stepwise approach to regression can only be recommended in empirical studies, where one must reduce the number of explanatory variables in order to simplify data collection during the next phase of field study.

There are other ways to counter the effects of multicollinearity in multiple regression. Table 10.5 shows that collinearity has the effect of inflating the variance of regression coefficients, with the exception of the intercept b_0. When the objective is forecasting or prediction, one can use regression on principal components or ridge regression, described below. These methods reduce the variance of the regression coefficients, which leads in turn to better predictions of the response variable. However, the regression coefficients they produce are biased; despite of that, they are still better estimates of the 'true' regression coefficients than those obtained by ordinary multiple regression for collinear variables. In other words, the price to pay for reducing the inflation of variance is some bias in the estimates of the regression coefficients. This may provide better forecasting or prediction than the ordinary multiple regression solution since, as a consequence of the larger variance in the regression coefficients, multicollinearity tends to increase the variance of the forecasted or predicted values (Freund & Minton, 1979).

Regression on principal components

• *Regression on principal components* consists of the following steps: (1) perform a principal component analysis on the matrix of the explanatory variables **X**, (2) compute the multiple regression of y on the principal components (matrix **F**, eq. 9.4) of **X** instead of the original explanatory variables, and (3) find back the contributions of the explanatory variables by multiplying matrix **U** of the eigenvectors with the vector of regression coefficients **c** of y regressed on the selected principal components (without including the intercept). One obtains a new vector **b** of contributions of the original variables to the regression equation as follows:

$$\mathbf{b}_{(m \times 1)} = \mathbf{U}_{(m \times k)} \mathbf{c}_{(k \times 1)} \qquad \textbf{(10.18)}$$



k	R^2
0.0	0.346
0.1	0.275
0.2	0.257
0.3	0.249
0.4	0.245
0.5	0.243

Figure 10.8 'Ridge trace' diagram showing the estimates of the standard regression coefficients $b(x_1)$ to $b(x_3)$ for explanatory variables x_1 to x_3 as a function of k. Table on the right: decrease of R^2 as a function of k.

where m is the number of explanatory variables in the analysis and k is the number of principal components retained for step 3. This procedure does not necessarily resolve the problem of multicollinearity, although it is true that the regression is performed on principal components, which are not correlated to one another by definition. Consider the following case: if all m eigenvectors are kept in matrix **U** for step 3, one obtains exactly the same regression coefficients as in ordinary multiple regression. When **X** contains collinear variables, there is a gain in stability of the regression coefficients only if some of the principal components are eliminated from the computation of eq. 10.18. One may either eliminate the eigenvectors with the smallest eigenvalues or, better, use only in eq. 10.18 the principal components that significantly contribute to explain the variation of y. By doing so, the regression coefficient estimates become biased, of course. In problems involving a small number of explanatory variables, regression on principal components may be difficult to use because the number of principal components is small, so that eliminating one of them from the analysis may result in a large drop in R^2. Ecological application 12.7 provides an example of regression on principal components.

Ridge regression

• *Ridge regression*, developed by Hoerl (1962) and Hoerl & Kennard (1970a, b), approaches the problem in a different way; another important paper on the subject is

Marquardt & Snee (1975). Instead of the usual matrix eq. 2.19 $\mathbf{b} = [\mathbf{X}'\mathbf{X}]^{-1}[\mathbf{X}'\mathbf{y}]$, the regression coefficients are estimated using a modified equation,

$$\mathbf{b} = [\mathbf{X}'\mathbf{X} + k\mathbf{I}]^{-1}[\mathbf{X}'\mathbf{y}], \quad \text{where} \quad k > 0. \qquad \textbf{(10.19)}$$

Hence, the method consists in increasing the diagonal terms (variances) of the covariance matrix $[\mathbf{X}'\mathbf{X}]$ by a constant positive quantity k. This reduces the variance of the regression coefficients while creating a bias in the resulting estimates. So, users are left with the practical problem of choosing a value for k that is optimal in some sense. This is accomplished by computing regression coefficient estimates for a series of values of k, and plotting them (ordinate) as a function of k (abscissa); this plot is called the 'ridge trace', for historical reasons (Hoerl, 1962). After studying the plot, one chooses a value of k which is as small as possible, but large enough that the regression coefficient estimates change little after it. Since ridge regression is usually computed on standardized variables, no intercept is estimated. A number of criteria have been proposed by Obenchain (1977) to help choose the value of k. These criteria must be used with caution, however, since they often do not select the same value of k as the optimal one.

An example of a 'ridge trace' diagram is presented in Fig. 10.8. The data set consists of a response variable y and three collinear explanatory variables x_1 to x_3; their empirical correlation matrix is the following:

	y	x_1	x_2	x_3
y	1			
x_1	–0.40	1		
x_2	–0.44	0.57	1	
x_3	–0.41	**0.99**	0.56	1

Variables x_1 and x_3 are highly correlated. The leftmost regression coefficient estimates in Fig. 10.8 (for $k = 0$) are the standardized OLS multiple regression coefficients. Going from left to right in the figure, the regression coefficients stabilize after a sharp decrease or increase. One may decide that setting the cut-off point at $k = 0.2$ would be an appropriate compromise between small k and stable regression coefficients. Boudoux & Ung (1979) and Bare & Hann (1981) provide applications of ridge regression to forestry; in both papers, some regression coefficients change signs with increasing k. An application of ridge regression to modelling heterotrophic bacteria in a sewage lagoon ecosystem is presented by Troussellier *et al.* (1986, followed-up by Troussellier & Legendre, 1989).

Coefficient of determination, R^2

The coefficient of multiple determination R^2, also called the *unadjusted* coefficient of multiple determination, is the square of the multiple correlation coefficient R of Section 4.5; it varies between 0 and 1. $R^2_{\mathbf{y}|\mathbf{X}}$ measures the proportion of the variation of variable y about its mean that is explained by the linear model of the variables

included in explanatory matrix **X**. As in simple linear regression, where the coefficient of determination is r^2 (eq. 10.7), $R^2_{\mathbf{y}|\mathbf{X}}$ is the regression sum of squares (SS) divided by the total sum of squares (total SS, TSS), or the one-complement of the ratio of the sum of squared residuals (residual sum of squares, RSS) to the total sum of squares (TSS):

$$R^2_{\mathbf{y}|\mathbf{X}} = \frac{\text{regression SS}}{\text{total SS}} = \frac{\Sigma(\hat{y}_i - \bar{y})^2}{\Sigma(y_i - \bar{y})^2} = 1 - \frac{\text{RSS}}{\text{TSS}} \qquad \textbf{(10.20)}$$

The expected value of R^2 in a regression involving m random predictors is not 0 but $m/(n-1)$, as explained below. As a consequence, if **X** contains $m = (n-1)$ predictors that are linearly unrelated to the response variable y, for example m columns of random numbers, $R^2 = 1$ even though the explanatory variables explain none of the variation of y. For that reason, R^2 cannot be interpreted as a correct (i.e. unbiased) estimate of the proportion of variation of y explained by **X**.

Three useful statistics can, however, be derived from R^2. They serve distinct purposes in regression analysis.

Adjusted R^2

1. The *adjusted coefficient of multiple determination* R^2_a or *adjusted* R^2 (Ezekiel, 1930), provides an unbiased estimate of the proportion of variation of y explained by **X**. The formula takes into account the numbers of degrees of freedom (d.f.) of the numerator and denominator portions of R^2:

$$R^2_a = 1 - \frac{\text{residual mean square}}{\text{total mean square}} = 1 - (1 - R^2_{\mathbf{y}|\mathbf{X}})\left(\frac{\text{total d.f.}}{\text{residual d.f.}}\right) \qquad \textbf{(10.21)}$$

- In ordinary multiple regression, the total degrees of freedom of the F-statistic are $(n-1)$ and the residual d.f. are $(n-m-1)$, where n is the number of observations and m is the number of explanatory variables in the model (eq. 4.40).

- In multiple regression through the origin, where the intercept is forced to zero, the total degrees of freedom of the F-statistic are n and the residual d.f. are $(n-m)$.

These same degrees of freedom are used in eq. 10.21. The logic of this adjustment is the following: in ordinary multiple regression, a random predictor explains on average a proportion $1/(n-1)$ of the response's variation, so that m random predictors explain together, on average, $m/(n-1)$ of the response's variation; in other words, the expected value of R^2 is $E(R^2) = m/(n-1)$. Applying eq. 10.21 to that value, where all predictors are random, gives $R^2_a = 0$. In regression through the origin, a random predictor explains on average a proportion $1/n$ of the response's variation, so that m random predictors explain together, on average, m/n of the response's variation, and $R^2 = m/n$. Applying eq. 10.21 to that case gives, again, $R^2_a = 0$.

R^2_a is a suitable measure of goodness of fit for comparing the success of regression equations fitted to different data sets, with different numbers of objects and explanatory variables. Using simulated data with normal error, Ohtani (2000) has shown that R^2_a is an unbiased estimator of the contribution of a set of random

predictors **X** to the explanation of *y*. This adjustment may be too conservative when $m > n/2$ (Borcard *et al*., 2011); this is a rule of thumb rather than a statistical principle.

With real matrices of random variables (defined at the beginning of Section 10.3), when the explanatory variables explain no more of the response's variation than the same number of variables containing random numbers, the value of R^2_a is near zero; it can be negative on occasion. Contrary to R^2, R^2_a does not necessarily increase with the addition of explanatory variables to the regression model if these explanatory variables are linearly unrelated to *y*. R^2_a is a better estimate of the population coefficient of determination ρ^2 than R^2 (Zar, 1999, Section 20.3) because it is unbiased.

Healy (1984) pointed out that Ezekiel's (1930) adjusted R^2 equation (R^2_a, eq. 10.21) makes sense and should be used when **X** contains observed values of random variables. That is not the case for ANOVA fixed factors, which can be used in a multiple regression equation when they are recoded into binary dummy variables or Helmert contrasts (Subsection 1.5.7).

In canonical analysis (Chapter 11), the canonical R^2 is called the *bimultivariate redundancy statistic* (Miller & Farr, 1971), *canonical coefficient of determination*, or *canonical* R^2. Using numerical simulations, Peres-Neto *et al*. (2006) have shown that, in redundancy analysis (RDA, Section 11.1), for normally distributed data or Hellinger-transformed species abundances, the adjusted canonical R^2 (R^2_a, eq. 11.5), obtained by applying eq. 10.21 to the canonical R^2 ($R^2_{\mathbf{Y}|\mathbf{X}}$, eq. 11.4), produces unbiased estimates of the contributions of the variables in **X** to the explanation of a response matrix **Y**, just as in multiple regression. With simulated data, they also showed the artificial increase of R^2 as the number of unrelated explanatory variables in explanatory matrix **X** increases.

AIC, AIC$_c$

2. The *Akaike Information Criterion* (*AIC*) is a measure of the goodness of fit of the data to an estimated statistical model (Akaike, 1974). When comparing linear regression models, *AIC* is computed as follows (RSS, TSS: see eq. 10.20):

$$AIC = n \log_e\left(\frac{\text{RSS}}{n}\right) + 2k \qquad \textbf{(10.22)}$$

where k is the number of parameters, including the intercept, in the regression equation. Independence of the observations is assumed in the calculation of *AIC*, as well as normality of the residuals and homogeneity of their variances. The following formula is also found in the literature: $AIC = n \log_e((1 - R^2)/n) + 2k$. A constant, $n\log_e(\text{TSS})$, must be added to this formula to obtain eq. 10.22. Since *AIC* is used to compare different models of the same response data, either formula will identify the same model as the one that minimizes *AIC*.

The corrected form of *AIC*, abbreviated AIC_c (Hurvich & Tsai, 1993), is *AIC* with a second-order correction for small sample size:

$$AIC_c = AIC + \frac{2k(k+1)}{n-k-1} \qquad \textbf{(10.23)}$$

Burnham & Anderson (2002) strongly recommend using AIC_c rather than AIC when n is small or k is large. Because AIC_c converges towards AIC when n is large, AIC_c should be used with all sample sizes.

The AIC_c statistic is not the basis for a test of significance. It plays a different role than the F-test (below): it is used to compare models. For a given data set, several competing models may be ranked by AIC_c. The model with the *smallest* value of AIC_c is the best-fitting one, i.e. the most likely for the data. For example, in selection of explanatory variables, the model for which AIC_c is minimum is retained.

F-statistic

3. The F-statistic (see eq. 4.40) serves as the basis for the test of significance of the coefficient of multiple determination, R^2. A parametric test can be used if the regression residuals are normal. Otherwise, a permutation test should be used.

F-statistic for nested models

There is another way of comparing models statistically, but it is limited to nested models of the same response data. A model is nested in another if it contains one or several variables less than the reference model. The method consists in calculating the R^2 of the two linear models and computing a F-statistic to test the difference in R^2 between them. The F-statistic is computed as follows for two nested models, the most inclusive containing m_2 variables and the model nested into it containing m_1 variables:

$$F = \frac{(R^2_{y.1\ldots m_2} - R^2_{y.1\ldots m_1}) / (m_2 - m_1)}{(1 - R^2_{y.1\ldots m_2}) / (n - m_2 - 1)}$$

The difference in R^2 is tested for significance parametrically with $\nu_1 = (m_2 - m_1)$ and $\nu_2 = (n - m_2 - 1)$ degrees of freedom, or by permutation. This method can be used in forward selection or backward elimination. It is implemented, for example, in functions ***ordiR2step()*** of VEGAN and ***forward.sel()*** of PACKFOR (Subsection 11.1.10, paragraph 7), which can be used in models involving a single response variable y.

As a final note, it is useful to remember that several types of explanatory variables can be used in multiple regression:

Dummy variable regression

• Binary descriptors can be used as explanatory variables in multiple regression, together with quantitative variables. This means that multistate qualitative variables can also be used, insofar as they are recoded into binary dummy variables, as described in Subsection 1.5.7*. This case is referred to as *dummy variable regression*.

• Geographic information may be used in multiple regression models in different ways. On the one hand, latitude (Y) and longitude (X) information form perfectly valid quantitative descriptors if they are recorded as axes of a Cartesian plane. Geographic data in the form of degrees-minutes-seconds should, however, be recoded to decimal

* In R, qualitative multistate descriptors used as explanatory variables are automatically recoded into dummy variables by function ***lm()*** if they are identified as *factors* in the data frame.

form before they are used as explanatory variables in regression. The X and Y coordinates may be used either alone, or in the form of a polynomial (X, Y, X^2, XY, Y^2, etc.). Regression using such explanatory variables is referred to as *trend surface analysis* in Chapter 13. Spatial eigenfunctions, described in Chapter 14, are more sophisticated descriptions of geographic relationships among study sites; they can also be used as explanatory variables in regression.

• If replicate observations are available for each site, the grouping of observations, which is also a kind of geographic information, may be used in multiple regression as a qualitative multistate descriptor, recoded into a set of dummy variables.

• Finally, any analysis of variance may be reformulated as a linear regression analysis; actually, linear regression and ANOVA both belonging to the General Linear Model. Consider one-way ANOVA for instance: the classification criterion can be written as a multistate qualitative variable and, as such, recoded as a set of dummy variables (Subsection 1.5.7) on which multiple regression may be performed. The analysis of variance table obtained by multiple regression is identical to that produced by ANOVA. This equivalence is discussed in more detail by ter Braak & Looman (1987) in an ecological framework. Draper & Smith (1981) and Searle (1987) discuss in some detail how to apply multiple regression to various analysis of variance configurations. ANOVA by regression can be extended to cross-factor (two-way or multiway) ANOVA. How to carry out these analyses is described in Subsection 11.1.10, point 4, for the more general analysis of multivariate response data **Y** (MANOVA).

4 — Polynomial regression

Several solutions have been proposed to the problem of fitting, to a response variable y, a nonlinear function of a single explanatory variable x. An elegant and easy solution is to use a polynomial of x, whose terms are treated as so many explanatory variables in a multiple regression procedure. In this approach, y is modelled as a polynomial function of x:

Polynomial model

$$\hat{y} = b_0 + b_1 x + b_2 x^2 + \ldots + b_k x^k \qquad \textbf{(10.24)}$$

Such an equation is linear in its parameters (if one considers the terms $x^2, \ldots, x^k$ as so many explanatory variables), although the modelled response of y to the explanatory variable x is nonlinear. The degree of the equation, which is its highest exponent, determines the shape of the curve: each degree above 1 (straight line) and 2 (concave up or down) adds an inflexion point to the curve. Increasing the degree of the equation always increases its adjustment to the data (R^2). If one uses as many parameters b (including the intercept b_0) as there are data points, one can fit the data perfectly ($R^2 = 1$). However, the cost of that perfect fit is that there are no degrees of freedom left to test the relationship and, therefore, the "model" cannot be extended to other situations. Hence, a perfectly fitted model is useless. In any case, a high-degree polynomial would be of little interest in view of the principle of parsimony (Ockham's razor) discussed in Subsection 10.3.3, which states that the best model is the simplest

one that adequately describes the relationship. Each term of a polynomial expression is called a *monomial*.

Monomial

So, the problem left to ecologists is to find the most parsimonious polynomial equation that adequately fits the data. The methods for selecting variables, described above for multiple regression, may be used to profit here.

One can start with a polynomial equation of degree k (e.g. $k = 4$) and use a selection procedure, based on *AIC*, to determine which subset of the monomials produces the most parsimonious model. Backward, forward or stepwise procedures can be applied. One could add the following constraint: that all monomials in the final model be significant, e.g. at level $\alpha = 0.05$. It may turn out that some higher-degree monomials are retained by the selection procedure, and are significant, whereas some of the lower-order monomials are excluded; this is entirely permissible. Beware: in some statistical packages, selection of monomials in polynomial regression only removes higher-degree monomials; monomials of degrees lower than k cannot be removed if x^k is retained in the model. These procedures do not produce a parsimonious model in cases where some lower-degree terms should be eliminated.

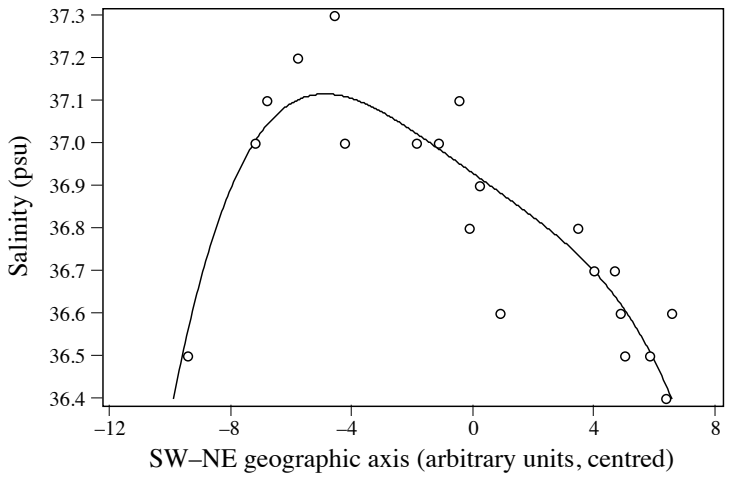
The successive terms of an ordinary polynomial expression are collinear. Starting for instance with a variable x made of the successive integers 1 to 10, variables x^2, x^3, and x^4 computed from it display the following correlations:

	x	x^2	x^3	x^4
x	1			
x^2	0.975	1		
x^3	0.928	0.987	1	
x^4	0.882	0.961	0.993	1

The problem of multicollinearity is severe with such data. Centring variable x on its mean before computing the polynomial is good practice. It reduces the linear dependency of x^2 on x (it actually eliminates it when the x values are at perfectly regular intervals, as in the present example), and somewhat alleviates the problem for the higher terms of the polynomial. This may be enough when the objective is descriptive. If, however, it is important to estimate the exact contribution (standard regression coefficient) of each term of the polynomial in the final equation, the various monomials (x, x^2, etc.) should be made orthogonal to one another before computing the regression equation. Orthogonal monomials may be obtained, for example, through the Gram-Schmidt procedure described in Table 9.5 and in textbooks of linear algebra for instance Lipschutz (2009); see function ***poly()*** in Section 10.7.

Orthogonal monomials

Numerical example. Data from the ECOTHAU program (Ecology of the Thau lagoon, southern France; Amanieu *et al.*, 1989) are used to illustrate polynomial regression. Salinity (response variable y) was measured at 20 sites in the brackish Thau lagoon (Mediterranean Sea) on 25 October 1988. The lagoon is elongated in a SW-NE direction. The explanatory variable x



Regression equation:

	Regr. coeff.	Standard coeff.
b_0	36.929	0
$b_1(x)$	–0.0509	–0.9699
$b_2(x^4)$	–0.0001	–0.7261

Figure 10.9 Polynomial regression line describing the structure of salinity (psu: practical salinity units) in the Thau lagoon (Mediterranean Sea) along its main geographic axis on 25 October 1988.

is the projection of the positions of the sampling sites on the long axis of the lagoon, as determined by principal component analysis of the site coordinates. Being a principal component, variable x is centred. The other terms of an ordinary 6th-degree polynomial were computed from it. After stepwise selection, the model with the lowest AIC_c contained variables x, x^4 and x^5 ($AIC_c = –80.845$, $R^2_a = 0.815$); the regression parameters for x^4 and x^5 were not significant at the 0.05 level. Then, all possible models involving x, x^2, x^3, x^4 and x^5 were computed. The model with the largest number of significant regression coefficients contained variables x and x^4 ($AIC_c = –80.374$, $R^2_a = 0.792$). These results indicate that the model with three monomials (x, x^4 and x^5) is slightly better in terms of AIC_c and is thus the best-fitting model for the data. The line fitted to the second model, which is more parsimonious with only two explanatory variables (x and x^4), is shown in Fig. 10.9).

5 — Partial linear regression and variation partitioning

There are situations where two or more complementary sets of hypotheses may be invoked to explain the variation of an ecological variable. For example, the abundance of a species could vary as a function of biotic and abiotic factors. Regression modelling may be used to study one set of factors, or the other, or the two sets together. In most if not all cases involving field data (by opposition to experimental designs), there are correlations among variables across the two (or more) explanatory data sets. Partial regression is a way of estimating how much of the variation of the response variable can be attributed exclusively to one set once the effect of the other has been taken into account and controlled for. The purpose may be to estimate the amount of variation that can be attributed exclusively to one or the other set of explanatory variables and the amount explained jointly by the two explanatory data sets, or else to

estimate the vector of fitted values corresponding to the exclusive effect of one set of variables. When the objective is simply to assess the unique contribution of each explanatory variable, there is no need for partial regression analysis: the coefficients of multiple regression of the standardized variables already provide that information since they are *standard partial regression coefficients*.

Matrix of covariables

Consider three data sets. Vector **y** is the response variable whereas matrices **X** and **W** contain the explanatory variables. Assume that one wishes to model the relationship between **y** and **X**, while controlling for the effects of the variables in matrix **W**, which is called the *matrix of covariables*. The roles of **X** and **W** could of course be inverted.

Variation partitioning

Variation partitioning consists in apportioning the variation* of variable **y** among two or more explanatory data sets. This approach was first proposed by Mood (1969, 1971) and further developed by Borcard *et al*. (1992) and Peres-Neto *et al*. (2006). The method is described here for two explanatory data sets, **X** and **W**, but it can be extended to more explanatory matrices. When **X** and **W** contain random variables (defined at the beginning of Section 10.3), adjusted coefficients of determination (R^2_a, eq. 10.21) are used to compute the fractions following the method described below. Ordinary R^2 (eq. 10.20) are used instead of R^2_a when **X** and **W** represent ANOVA fixed factors coded into binary dummy variables or Helmert contrasts (Subsection 1.5.7).

Figure 10.10 sets a nomenclature, [a] to [d], for the fractions of variation that can be identified in **y**. Kerlinger & Pedhazur (1973) called this form of analysis "commonality analysis" by reference to the common fraction of variation (fraction [b] in Fig. 10.10) that two sets of explanatory variables may explain jointly. Partial regression assumes that the effects are linear and additive. There are two ways of carrying out the partitioning computations, depending on whether one wishes to obtain vectors of fitted values corresponding to fractions of variation, or simply estimate the amounts of variation corresponding to the fractions. In the description that follows, the fractions of variation are computed from R^2_a statistics.

(1) If one is interested in obtaining a partial regression equation and computing a vector of partial fitted values, one first computes the residuals of **y** on **W** (noted $\mathbf{y}_{\text{res}|\mathbf{W}}$) and the residuals of **X** on **W** (noted $\mathbf{X}_{\text{res}|\mathbf{W}}$):

Residuals of **y** on **W**: $$\mathbf{y}_{\text{res}|\mathbf{W}} = \mathbf{y} - \mathbf{W}\,[\mathbf{W}'\mathbf{W}]^{-1}\mathbf{W}'\,\mathbf{y}$$

Residuals of **X** on **W**: $$\mathbf{X}_{\text{res}|\mathbf{W}} = \mathbf{X} - \mathbf{W}\,[\mathbf{W}'\mathbf{W}]^{-1}\mathbf{W}'\,\mathbf{X}$$

In both cases, the regression coefficients are computed here through eq. 2.19 in which **X** is replaced by **W**. QR decomposition (see Section 10.7), which is also used in some

* The term *variation*, a less technical and looser term than *variance*, is used because one is partitioning the total sum of squared deviations of **y** from its mean (total SS). In variation partitioning, there is no need to divide the total SS of **y** by its degrees of freedom to obtain the variance s^2_y (eq. 4.3).

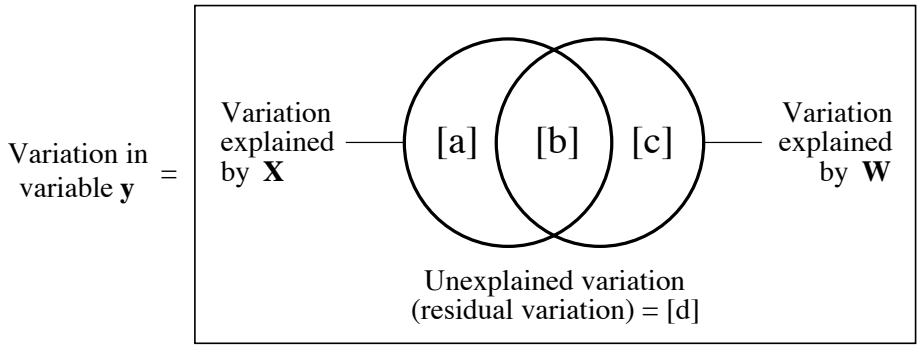


Figure 10.10 Partition of the variation of a response variable **y** among two sets of explanatory variables, **X** and **W**. The rectangle represents 100% of the variation in **y**. Fraction [b] is the intersection (*not* the interaction) of the variation explained by linear models of **X** and **W**. Adapted from Legendre (1993).

situations in Subsection 11.1, e.g. in Table 11.5, offers another way of computing regression equations.

Then, two computation methods are available: one can either

(1.1) regress $\mathbf{y}_{\text{resl}\mathbf{W}}$ on $\mathbf{X}_{\text{resl}\mathbf{W}}$,

(1.2) or regress **y** on $\mathbf{X}_{\text{resl}\mathbf{W}}$. The same partial regression coefficients are obtained in both cases, as will be verified in the numerical example below. Between calculation methods, the vectors of fitted values only differ by the values of the intercepts. The R^2 of analysis 1.1 is the partial R^2 whereas that of analysis 1.2 is the semipartial R^2; their square roots are the partial and semipartial correlation coefficients (Box 4.1).

(2) If one is interested in estimating the fractions resulting from partitioning the variation of vector **y** among the explanatory data sets **X** and **W**, there is a simple way to obtain the information, considering the ease with which multiple regressions can be computed using R or commercial statistical packages:

• Compute the multiple regression of **y** against **X** and **W** together. The corresponding R^2_a measures the fraction of information [a + b + c], which is the sum of the fractions of variation [a], [b], and [c] defined in Fig. 10.10. For the example data set (below), $R^2 = 0.5835$, so $R^2_a = 0.3913 = [a + b + c]$. The vector of fitted values corresponding to fraction [a + b + c], which is required to plot Fig. 10.13 (below), is also computed.

• Compute the multiple regression of **y** against **X**. The corresponding R^2_a measures [a + b], which is the sum of the fractions of variation [a] and [b]. For the example data,

$R^2 = 0.4793$, so $R^2_a = 0.3817 = [a + b]$. The vector of fitted values corresponding to fraction [a + b], which is required to plot Fig. 10.13, is also computed.

• Compute the multiple regression of **y** against **W**. The corresponding R^2_a measures [b + c], which is the sum of the fractions of variation [b] and [c]. For the example data, $R^2 = 0.3878$, so $R^2_a = 0.2731 = [b + c]$. The vector of fitted values corresponding to fraction [b + c], which is required to plot Fig. 10.13, is also computed.

• If needed, fraction [d] may be computed by subtraction. For the example, it is equal to $1 - [a + b + c]$, or $1 - 0.3913 = 0.6087$.

As explained in Subsection 10.3.3, the adjusted R-square, R^2_a (eq. 10.21), is an unbiased estimator of the real contribution of a set of random variables **X** to the explanation of **y**. Following Peres-Neto *et al*. (2006), the values of the individual fractions [a], [b], and [c] must be computed by combining the R^2_a values obtained from the three multiple regressions that produced fractions [a + b + c], [a + b], and [b + c]:

• fraction [a] is computed by subtraction, using the R^2_a values: $[a] = [a + b + c] - [b + c]$;

• likewise, fraction [c] is computed by subtraction, using the R^2_a values: $[c] = [a + b + c] - [a + b]$;

• fraction [b] is also obtained by subtraction, using the R^2_a values, in the same way as the quantity B used for comparing two qualitative descriptors in Section 6.2:

$$[b] = [a + b] + [b + c] - [a + b + c] \quad or \quad [b] = [a + b] - [a] \quad or \quad [b] = [b + c] - [c]$$

Negative [b]

Fraction [b] may be negative. As such, it is not a rightful measure of variance; this is another reason why it is referred to by the looser term *variation*. A negative fraction [b] indicates that two variables (or groups of variables **X** and **W**), together, explain **y** better than the sum of the individual effects of these variables. This can happen: see Numerical examples 2 and 3. Fraction [b] is the *intersection* of the variation explained by linear models of **X** and **W**. It is *not an interaction* in the ANOVA sense.

Vectors of fitted values corresponding to fractions [a] and [c] can be computed using partial regression, as explained above, while the vector of residuals of the regression equation that uses all predictors corresponds to fraction [d]. No fitted vector can be estimated for fraction [b], however, because no partial regression model can be written for that fraction. No degrees of freedom are attached to fraction [b]; hence [b] cannot be tested for significance.

Selection of explanatory variables

If a selection procedure (backward, forward, stepwise; Subsection 10.3.3) is used, it must be applied to data matrices **X** and **W** separately, before partitioning, in order to preserve fraction [b] of the partition. Applying the selection to matrices **X** and **W** combined could result in the elimination of variables from one or both matrices because they are correlated with variables in the other matrix, thereby reducing or eliminating fraction [b].

Table 10.6 Data collected at 20 sites in the Thau coastal lagoon on 25 October 1988. There are two bacterial response variables (Bna and Ma), three environmental variables (NH_4, phaeopigments, and bacterial production), and three spatial variables (the X and Y geographic coordinates measured with respect to arbitrary axes and centred on their respective means, plus the quadratic monomial X^2). The variables are further described in the text. The code names of these variables in the present section are y, x_1 to x_3, and w_1 to w_3, respectively.

Site No.	Bna	Ma y	NH_4 x_1	Phaeo. *a* x_2	Prod. x_3	X w_1	Y w_2	X^2 w_3
1	4.615	10.003	0.307	0.184	0.274	–8.75	3.7	76.5625
2	5.226	9.999	0.207	0.212	0.213	–6.75	2.7	45.5625
3	5.081	9.636	0.140	0.229	0.134	–5.75	1.7	33.0625
4	5.278	8.331	1.371	0.287	0.177	–5.75	3.7	33.0625
5	5.756	8.929	1.447	0.242	0.091	–3.75	2.7	14.0625
6	5.328	8.839	0.668	0.531	0.272	–2.75	3.7	7.5625
7	4.263	7.784	0.300	0.948	0.460	–1.75	0.7	3.0625
8	5.442	8.023	0.329	1.389	0.253	–0.75	–0.3	0.5625
9	5.328	8.294	0.207	0.765	0.235	0.25	–1.3	0.0625
10	4.663	7.883	0.223	0.737	0.362	0.25	0.7	0.0625
11	6.775	9.741	0.788	0.454	0.824	0.25	2.7	0.0625
12	5.442	8.657	1.112	0.395	0.419	1.25	1.7	1.5625
13	5.421	8.117	1.273	0.247	0.398	3.25	–4.3	10.5625
14	5.602	8.117	0.956	0.449	0.172	3.25	–2.3	10.5625
15	5.442	8.487	0.708	0.457	0.141	3.25	–1.3	10.5625
16	5.303	7.955	0.637	0.386	0.360	4.25	–5.3	18.0625
17	5.602	10.545	0.519	0.481	0.261	4.25	–4.3	18.0625
18	5.505	9.687	0.247	0.468	0.450	4.25	–2.3	18.0625
19	6.019	8.700	1.664	0.321	0.287	5.25	–0.3	27.5625
20	5.464	10.240	0.182	0.380	0.510	6.25	–2.3	39.0625

Numerical example 1. The example data set (Table 10.6) is from the ECOTHAU research program mentioned in the numerical example of Subsection 10.3.4 (Amanieu *et al*., 1989). It contains two bacterial variables (Bna, the concentration of colony-forming units of aerobic heterotrophs growing on bioMérieux nutrient agar, with low NaCl concentration; and Ma, the concentration of aerobic heterotrophs growing on marine agar with a salt content of 34 gL^{-1}); three environmental variables (NH_4 in the water column, in $\mu molL^{-1}$; phaeopigments from degraded chlorophyll *a*, in μgL^{-1}; and bacterial production, determined by incorporation of tritiated thymidine in bacterial DNA, in $nmolL^{-1}d^{-1}$); and three spatial variables of the sampling sites on the nodes of an arbitrarily located grid (the X and Y geographic coordinates, in km, each centred on its mean, and the quadratic monomial X^2, which was found to be important for explaining the response variables). All bacterial and environmental variables were log-transformed using $\log_e(x + 1)$. One of the bacterial variables, Ma, is used here as the response variable **y**; the three environmental variables form the matrix of explanatory variables **X**; the three spatial variables make up matrix **W** of the covariables. Table 10.6 will be used again in

Section 13.4. A multiple regression of **y** against **X** and **W** together was computed first as a reference. The regression equation was the following:

$$\hat{y} = 9.64 - 0.90x_1 - 1.34x_2 + 0.54x_3 + 0.10w_1 + 0.14w_2 + 0.02w_3$$
$$(R^2 = 0.5835;\ R^2_a = 0.3913 = [\text{a} + \text{b} + \text{c}])$$

The adjusted coefficient of determination (R^2_a) is an unbiased estimate of the proportion of the variation of **y** explained by the regression model containing the 6 explanatory variables; it corresponds to fraction [a+b+c] in the partitioning table below and to the sum of fractions [a], [b] and [c] in Fig. 10.10. The vector of fitted values was also computed; after centring, this vector will be plotted as fraction [a + b + c] in Fig. 10.13. Since the total sum of squares in **y** is 14.9276 [$SS = s^2_y \times (n-1)$], the R^2 allowed the computation of the sum of squares corresponding to the vector of fitted values: $SS(\hat{\mathbf{y}}) = 14.9276 \times 0.5835 = 8.7109$. This value can also be obtained by computing directly the sum of squared deviations about the mean of the values in the fitted vector $\hat{\mathbf{y}}$.

For calculation of the partial regression equation using method 1.1, the residuals* of the regression of **y** on **W** were computed. One way is to use the following equation, which requires adding a column of "1" to matrix **W** in order to estimate the regression intercept:

$$\mathbf{y}_{\text{resl}\mathbf{W}} = \mathbf{y} - \mathbf{W}\,[\mathbf{W}'\mathbf{W}]^{-1}\mathbf{W}'\,\mathbf{y}$$

The residuals of the regressions of **X** on **W** were computed in the same way:

$$\mathbf{X}_{\text{resl}\mathbf{W}} = \mathbf{X} - \mathbf{W}\,[\mathbf{W}'\mathbf{W}]^{-1}\mathbf{W}'\,\mathbf{X}$$

Then, vector $\mathbf{y}_{\text{resl}\mathbf{W}}$ was regressed on matrix $\mathbf{X}_{\text{resl}\mathbf{W}}$ with the following result:

regression equation: $\hat{y} = 0 - 0.90x_{r(\mathbf{W})1} - 1.34x_{r(\mathbf{W})2} + 0.54x_{r(\mathbf{W})3}$ $(R^2 = 0.3197)$

The value $R^2 = 0.3197$ is the partial R^2. In its calculation, the denominator is the sum of squares corresponding to fractions [a] and [d], as shown for the partial correlation coefficient in Box 4.1.

For calculation through method 1.2, **y** was regressed on matrix $\mathbf{X}_{\text{resl}\mathbf{W}}$ with the following result:

regression equation: $\hat{y} = 8.90 - 0.90x_{r(\mathbf{W})1} - 1.34x_{r(\mathbf{W})2} + 0.54x_{r(\mathbf{W})3}$ $(R^2 = 0.1957)$

The value $R^2 = 0.1957$ is the semipartial R^2. The semipartial R^2 is the square of the semipartial correlation defined in Box 4.1. It represents the fraction of the total variation of **y** explained by the partial regression equation because, in its calculation, the denominator is the total sum of squares of the response variable **y**, [a+b+c+d]. That value is shown in the variation partitioning table below, but it will not be used to compute the individual fractions of variation.

Note that the three regression coefficients for the three *x* variables in the last equation are exactly the same as in the two previous equations; only the intercepts differ. This gives substance to the statement of Subsection 10.3.3 that regression coefficients obtained in multiple

* In the R language, regression residuals can be computed using ***residuals(lm())***.

linear regression are *partial regression coefficients* in the sense of the present subsection. Between calculation methods, the vectors of fitted values only differ by the value of the intercept of the regression of **y** on $\mathbf{X}_{\text{res}|\mathbf{W}}$, 8.90, which is also the mean of **y**. The centred vector of fitted values will be plotted as fraction [a] in Fig. 10.13.

The calculation of partial regression can be done in the opposite way, regressing **y** on **W** while controlling for the effects of **X**. First, $\mathbf{y}_{\text{res}|\mathbf{X}}$ and $\mathbf{W}_{\text{res}|\mathbf{X}}$ were computed. Then, for method 1.1, $\mathbf{y}_{\text{res}|\mathbf{X}}$ was regressed on $\mathbf{W}_{\text{res}|\mathbf{X}}$ with the following result:

regression equation: $\hat{y} = 0 - 0.10 w_{r(\mathbf{X})1} - 0.14 w_{r(\mathbf{X})2} + 0.02 w_{r(\mathbf{X})3}$ $(R^2 = 0.2002)$

where $R^2 = 0.2002$ is the partial R^2. For method 1.2, **y** was regressed on $\mathbf{W}_{\text{res}|\mathbf{X}}$ with the following result:

regression equation: $\hat{y} = 8.90 - 0.10 w_{r(\mathbf{X})1} - 0.14 w_{r(\mathbf{X})2} + 0.02 w_{r(\mathbf{X})3}$ $(R^2 = 0.1043)$

where $R^2 = 0.1043$ is the semipartial R^2, shown in the variation partitioning table below, but not used to compute the individual fractions of variation.

Again, the three regression coefficients in these partial regression equations are exactly the same as in the first regression equation of this example; only the intercepts differ. Between calculation methods, the vectors of fitted values only differ by the value of the intercept of the regression of **y** on $\mathbf{X}_{\text{res}|\mathbf{W}}$, 8.90, which is also the mean of **y**. The centred vector of fitted values will be plotted as fraction [c] in Fig. 10.13.

To estimate fraction [a + b] of Fig. 10.10, the multiple regression of **y** on the three original (non-residualized) variables in **X** was computed. The regression equation was:

$$\hat{y} = 10.20 - 0.93 x_1 - 2.02 x_2 + 0.89 x_3 \qquad (R^2 = 0.4793;\ R^2_a = 0.3817 = [\text{a} + \text{b}])$$

The value $R^2_a = 0.3817$ is an unbiased estimate of the fraction of the variation of **y** accounted for by the linear model of the three explanatory variables **X**. The vector of fitted values was computed; after centring, this vector will be plotted as fraction [a + b] in Fig. 10.13.

To obtain fraction [b + c] of Fig. 10.10, the multiple regression of **y** on the three original (non-residualized) variables in **W** was computed. The regression equation was:

$$\hat{y} = 8.32 + 0.09 w_1 + 0.10 w_2 + 0.03 w_3 \qquad (R^2 = 0.3878;\ R^2_a = 0.2731 = [\text{b} + \text{c}])$$

The value $R^2_a = 0.2731$ is the unbiased estimation of the fraction of the variation of **y** accounted for by the linear model of the three explanatory variables **W**. The vector of fitted values was computed; after centring, this vector will be plotted as fraction [b + c] in Fig. 10.13.

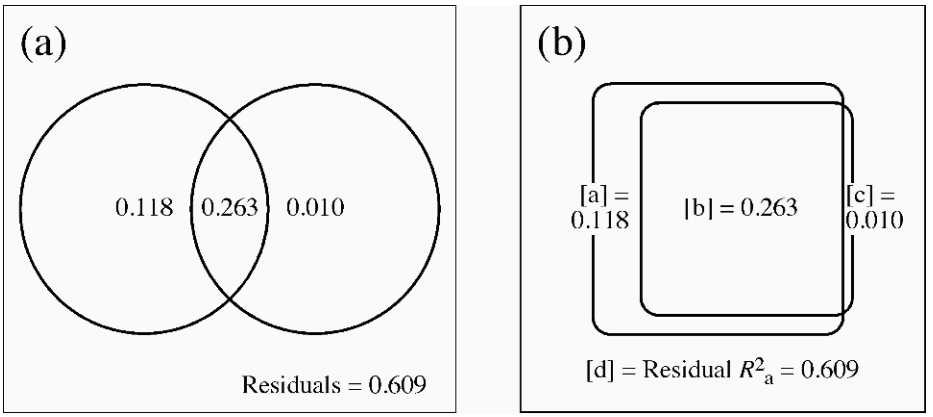


Figure 10.11 Venn diagram illustrating the results of variation partitioning of the numerical example. (a) Diagram drawn by the plotting function ***plot.varpart()*** of the VEGAN package. The circles are of equal sizes despite differences in the corresponding R^2_a. (b) Prior to publication of the partitioning results, the diagram can be redrawn, here using rounded rectangles, to better represent the relative fraction sizes with respect to the size of the outer rectangle, which represents the total variation in the response data. The fractions are identified by letters [a] to [d]; the value next to each identifier is the adjusted R^2 (R^2_a). Rectangle sizes are approximate.

Following the fraction nomenclature convention set in Fig. 10.10, the variation partitioning results were assembled in the following table (rounded values):

Fractions of variation	Sums of squares (SS)	Proportions of variation of y (R^2)	Adjusted R^2 (R^2_a)
[a + b]	7.1547	0.4793	0.3817
[b + c]	5.7895	0.3878	0.2731
[a + b + c]	8.7109	0.5835	0.3913
[a]	2.9213	*0.1957*	0.1183
[b]	4.2333	*0.2836*	0.2634
[c]	1.5562	*0.1043*	0.0097
Residuals = [d]	6.2167	*0.4165*	0.6087
[a + b + c + d]	14.9276	1.0000	1.0000

The partitioning results are illustrated as Venn diagrams in Fig. 10.11*. In Chapter 11, Fig. 11.6 shows partitioning results for multivariate response data involving three explanatory matrices.

* A Venn diagram with proportional circle and intersection sizes can be obtained using function ***venneuler()*** of the same-name package (Section 10.7).

As mentioned at the beginning of this subsection, and following Peres-Neto *et al*. (2006), when **X** and **W** contain random variables, R^2_a values corresponding to [a + b + c], [a + b], and [b + c] are used to compute, by subtraction, the fractions [a] to [d] shown in column 4 of the table. R^2_a provides unbiased estimates of the contributions of the explanatory data sets **X** and **W** to **y** when **X** and **W** contain random variables. The adjusted fractions [a], [b], and [c] cannot be directly computed using the non-adjusted fractions computed from non-adjusted R^2 coefficients, shown in italics in the 3*rd* column. When *n* is small as in this example, the estimated fractions computed from R^2_a may be very different from the fractions computed from R^2 values.

Ordinary R^2 (3*rd* column) are used to compute the fractions (values in italics) when **X** and **W** represent ANOVA fixed factors coded into dummy variables. When these values are required, they can be calculated by subtraction from the R^2 values in the first three rows of the table: R^2[a] = R^2[a+b+c] – R^2[b+c] = 0.1957 (which is equal to the R^2 of the partial regression equation computed above through method 1.2); R^2[c] = R^2[a+b+c] – R^2[a+b] = 0.1043 (which is equal to the R^2 of the partial regression equation computed above through method 1.2); R^2[b] = R^2[a+b] + R^2[b+c] – R^2[a+b+c] = 0.2836 (this value can only be obtained by subtraction). The sums of squares in the 2nd column of the table are obtained by multiplying these R^2 values by the total sum of squares in **y**, which is 14.9276.

The *partial correlation coefficient* between **y** and matrix **X** while controlling for the effect of **W** can be obtained from the values [a] and [d] in the column "Sums of squares" of the table, as explained in Box 4.1 of Section 4.5:

$$r_{\mathbf{yX.W}} = \sqrt{\frac{[a]}{[a+d]}} = \sqrt{\frac{2.9213}{2.9213 + 6.2167}} = 0.5654$$

This value is not the same as the *semipartial* R^2, which is computed as follows (Box 4.1):

$$r_{\mathbf{y(X.W)}} = \sqrt{\frac{[a]}{[a+b+c+d]}} = \sqrt{\frac{2.9213}{14.9276}} = 0.4424$$

Tests of significance of the fractions

If the conditions of homoscedasticity and normality of the residuals are satisfied, the fractions (with the exception of [b]) can be tested for significance through parametric tests. For fractions [a + b + c], [a + b], and [b + c], one can use the results of the parametric tests produced by the statistical software. For fractions [a] and [c], one must construct a *F*-statistic as in eq. 11.22, using the sum of squares corresponding to fraction [a] (symbol: SS[a]) or [c] (symbol: SS[c]) in the numerator, and the residual sum of squares corresponding to [d] (symbol: SS[d]) in the denominator, together with appropriate numbers of degrees of freedom. The test statistic for fraction [a], for example, is constructed as follows:

$$F_{[a]} = \frac{SS[a]/m}{SS[d]/(n-m-q-1)}$$

where *m* is the number of explanatory variables in set **X** and *q* is the number of covariables in set **W**. In the parametric framework, the statistic is tested against the

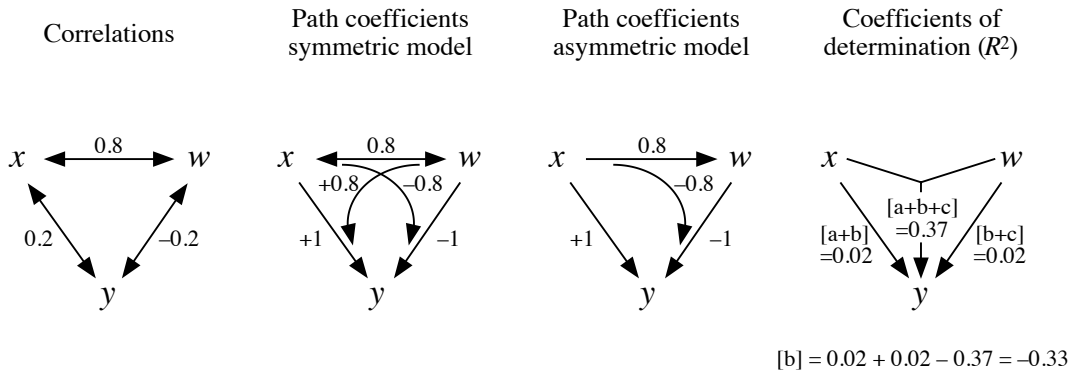


Figure 10.12 Correlations, path coefficients, and coefficients of determination for Numerical example 2.

F–distribution with m and $(n - m - q - 1)$ degrees of freedom. An example of that *F*-statistic for the test of partial or semipartial correlation coefficients is given in Box 4.1 for the simple case where there is a single variable in **X** and **W**.

If the conditions of homoscedasticity or normality of the residuals are not satisfied, one can use permutation tests to obtain p-values. Permutation of the raw data is used to test fractions [a + b + c], [a + b], and [b + c]. To test fractions [a] and [c], permutation of the residuals of a null or full model should be used (Anderson & Legendre, 1999). These permutation methods are described in Subsection 11.1.8.

Numerical example 2. This example illustrates the appearance of a negative fraction [b] when there are strong direct effects of opposite signs of x and w on y and a strong correlation between x and w (non-orthogonality). For three variables measured over 50 objects, the following correlations are obtained: $r(x, w) = 0.8$, $r(y, x) = 0.2$ and $r(y, w) = -0.2$; y, x, and w have the same meaning as in the previous numerical example. $r(y, x)$ and $r(y, w)$ are not statistically significant at the $\alpha = 0.05$ level. Referring to Section 10.4, one may use path analysis to compute the direct and indirect causal covariation relating the explanatory variables x and w to the response variable y. One can also compute the coefficient of determination of the model $y = f(x, w)$; its value is $R^2 = 0.40$. From these values, the partition of the variation of y can be achieved: R^2 of the whole model = 0.40, R^2_a = [a + b + c] = 0.37447; $r^2(w, y) = 0.04$, R^2_a = [a + b] = 0.02; $r^2(x, y) = 0.04$, R^2_a = [b + c] = 0.02. Hence, [b] = [a + b] + [b + c] – [a + b + c] = –0.33447, [a] = [a + b] – [b] = 0.35447, and [c] = [b + c] – [b] = 0.35447. How is that possible?

Carrying out path analysis (Fig. 10.12), and assuming a symmetric model of relationships (i.e. w affects x and x affects w), the direct effect of x on y, $p_{xy} = 1.0$, is positive and highly significant, but it is counterbalanced by a strong negative indirect covariation of –0.8 going through w. In the same way, $p_{wy} = -1.0$ (which is highly significant), but this direct effect is counterbalanced by a strong positive indirect covariation of +0.8 going through x. As a result, and although they both have the maximum possible value of 1.0 for direct effects on the

response variable y, both w and x turn out to have non-significant total correlations with y. In the present variation partitioning model, this translates into small adjusted amounts of explained variation $[a + b] = 0.02$ and $[b + c] = 0.02$, and a negative value for fraction [b]. If an asymmetric model of relationship had been assumed (e.g. w affects x but x does not affect w), essentially the same conclusion would have been reached from path analysis.

Numerical example 3. Another situation can give rise to a negative fraction [b], i.e. when there is no linear correlation between y and one of the explanatory variables, e.g. $r(y, x) = 0.0$, but the other two correlations differ from 0, e.g. $r(y, w) = 0.5$ and $r(x, w) = 0.5$. For this example, assuming again $n = 50$, we find $[a + b + c] = 0.30497$, $[a + b] = -0.02083$, and $[b + c] = 0.23438$ (computed from the R^2_a coefficients), so that $[b] = -0.09142$. The partial explanation of the variation of y provided by x, estimated by the partial regression or partial correlation coefficient, is not zero and may be significant in the statistical sense: using path analysis (Section 10.4) for this example, the direct effect of x on y is $p_{xy} = -0.33333$ ($p = 0.019$, which is significant) and the indirect effect is 0.33333, these two effects summing to zero. The direct effect of w on y is $p_{wy} = 0.66667$ and its indirect effect is -0.16667. The negative [b] fraction indicates that x and w, together, explain the variation of y better than the sum of the individual effects of these variables. The signs of the regression coefficients (path coefficients) actually vary depending on the signs of the correlations $r(y, w)$ and $r(x, w)$.

The above decomposition of the variation of a response vector **y** between two sets of explanatory variables **X** and **W** was described by Whittaker (1984) for the simple case where there is a single regressor in each set **X** and **W**. Whittaker showed that the various fractions of variation may be represented as vectors in space, and that the value of fraction [b] [noted G(12:) by Whittaker, 1984] is related to the angle θ between the two regressors through the following formula:

$$1 - 2\cos^2(\theta/2) \leq [b] \leq 2\cos^2(\theta/2) - 1 \qquad \textbf{(10.25)}$$

Fraction [b] for orthogonal regressors

θ is related to the coefficient of linear correlation (eq. 10.4). This formula has three interesting properties. (1) If the two regressors are orthogonal ($r = 0$), then $2\cos^2(\theta/2) = 1$, so that $0 \leq [b] \leq 0$ and consequently $[b] = 0$. Turning the argument around, the presence of a non-zero fraction [b] indicates that the two explanatory variables are not orthogonal. There are also instances where [b] is zero with two non-orthogonal regressors; a simple example is when the two regressors are uncorrelated with **y** and explain none of its variation. (2) If the two regressors are identical, or at least pointing in the same direction ($\theta = 0°$), then $-1 \leq [b] \leq 1$. It follows that the proportion of variation of **y** that is accounted for by either regressor (fraction [b]) may be, in some cases, as large as 1, i.e. 100%. (3) The formula allows for negative values of [b], as shown in Numerical example 2.

In conclusion, fraction [b] represents the fraction of variation of **y** that may indifferently be attributed to **X** or **W**. The interpretation of a negative [b] is that the two processes, represented in the analysis by data sets **X** and **W**, are competitive; in other words, they have opposite effects, one process hindering the contribution of the other in the joint regression model. One could use eq. 6.15, $S = [b]/[a + b + c]$, to quantify how similar **X** and **W** are in explaining **y**. Whittaker (1984) also suggested that if **X** and **W** represent two factors of an experimental design, [b] may be construed as a

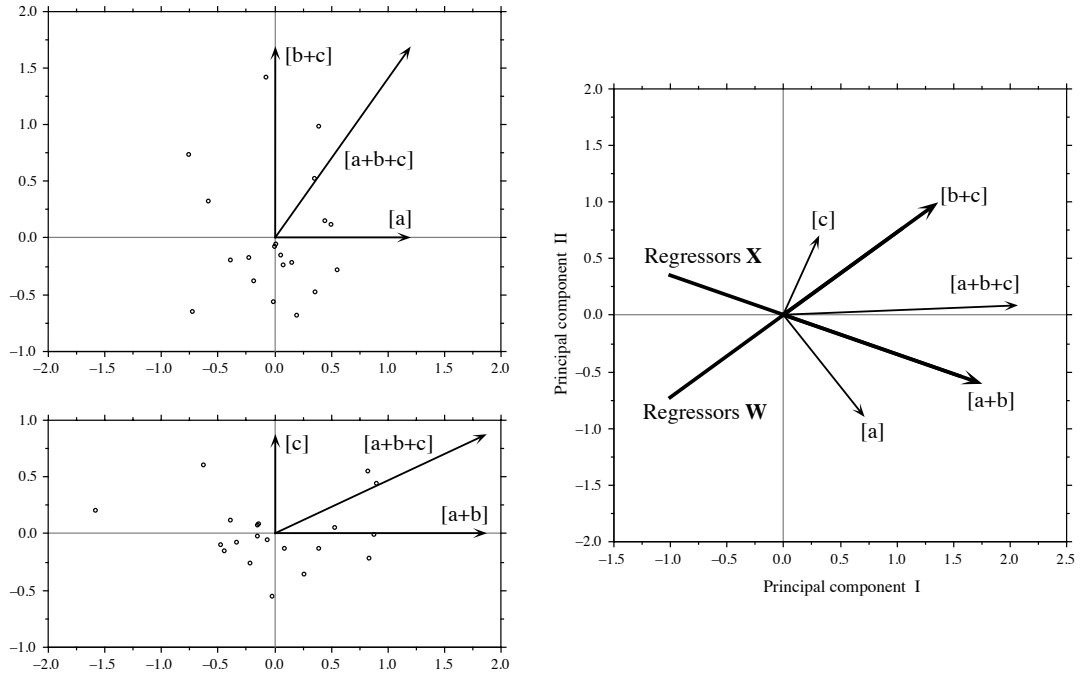


Figure 10.13 Numerical example of partial regression analysis: representation of the fitted vectors in regression space. Vectors are represented with lengths proportional to their standard deviations. Upper left: scatter diagram of objects along orthogonal vectors [a] and [b + c]. Vector [a + b + c], also shown, is obtained by adding vectors [a] and [b + c]. Lower left: same for orthogonal vectors [c] and [a + b]. Right: all five fitted vectors are represented in a compromise plane obtained by principal component analysis (PCA axes I and II, which explain 96.7% of the variation). [a] is still orthogonal to [b + c] in three-dimensional space, and [c] to [a + b]; these orthogonal relationships are slightly deformed by the projection in two dimensions.

measure of the effective balance (i.e. orthogonality) of the design; [b] is 0 in a balanced crossed design.

Whittaker's representation may be used even when regressors **X** and **W** are multivariate data sets. Figure 10.13 illustrates the angular relationships among the fitted vectors corresponding to the fractions of variation of Numerical example 1. One plane is needed for vectors {[a], [b + c], and [a + b + c]} in which [a] is orthogonal and additive to [b + c]; another plane is needed for vectors {[c], [a + b], and [a + b + c]} where [c] is orthogonal and additive to [a + b]. However, the sets {[a], [b + c]} and {[c], [a + b]} belong to different planes, which intersect along vector [a + b + c]; so, the whole set of fitted vectors is embedded in a three-dimensional space when there are two explanatory data sets; this is independent of the number of variables in each set. The vector of residuals corresponding to fraction [d] is orthogonal to all the fitted

vectors and lies in a fourth dimension. Whittaker (1984) gives examples involving more than two explanatory data sets. The graphical representation of the partitioned fitted vectors in such cases requires spaces with correspondingly more dimensions.

Ecological application 10.3b

Birks (1996) used partial regressions to analyse the mountain plant species richness in 75 grid squares covering Norway (109 species in total), in order to test whether the nunatak hypothesis was necessary to explain the present distribution of these plants. The nunatak, or refugial hypothesis, holds that apparent anomalies in present-day species distributions are explained by survival through glaciations on ice-free mountain peaks or rocky outcrops, called 'nunataks' (from Inuit *nunataq*), projecting above continental glaciers. Implicit in this hypothesis is that the presumed refugial species have poor dispersal ability. According to the nunatak hypothesis, one would expect a concentration of rare plants in the glacial refuges or their vicinity. Hence, a variable describing unglaciated areas (3 abundance classes for occurrence of presumed unglaciated areas) was introduced in the analysis to represent "history". The alternative hypothesis, called "tabula rasa", holds that present-day distributions are well-explained by the environmental control model (Whittaker, 1956; Bray & Curtis, 1957). To materialise this hypothesis in the analysis, Birks used 10 explanatory variables that described bedrock geology, geography, topography, and climate. "Geography" was introduced in the analysis in the form of a third-degree polynomial of the geographic coordinates, which allowed a representation of the geographic variation of species richness by a cubic trend surface of latitude and longitude, as explained in Subsection 13.2.1; the terms of the polynomial representing latitude and longitude^2 were retained by a forward selection procedure.

(1) Birks (1996) first used a form of stepwise multiple regression, adding variables in a specified order, to determine the importance of unglaciated areas in explaining mountain plant species richness. In the "ecology first" analysis, history (i.e. variable "unglaciated areas") was introduced last in the analysis; it added about 0.1% to the explained variation, whereas the environmental variables explained together 84.9% of the variation. In the "history first" analysis, history was entered first; it only explained 7.6% of the variation, which was not a significant contribution. (2) The contribution of "history" did not improve in partial regression analyses, when controlling for either land area per grid square alone, or land area, latitude and longitude. Modern ecological variables such as bedrock geology, climate, topography, and geography were considerably more effective explanatory variables of species richness than "history". (3) In order to find out whether "history" made a unique statistically significant contribution to the variation of the species richness when the effects of the other variables were controlled for, Birks computed variation partitioning, described above, after partial regression analyses, using non-adjusted R^2 coefficients. Fraction [a], corresponding to the influence of all environmental variables independent of "history", explained 77.4% of the variation of species richness; fraction [b], in which "environment" covaried with "history", explained 7.5%; fraction [c], "history" independent of "environment", explained 0.1%; the unexplained variation, fraction [d], was 15.0%. Fraction [b] is likely to result from the spatial coincidence of unglaciated areas with high elevation, western coastal areas, and certain types of bedrock, all these being included among the environmental variables.

In another paper, Birks (1993) used partial canonical correspondence analysis, instead of partial regression analysis, to carry out the same type of analysis (including variation decomposition) on a matrix of grid cells × species presence/absence. Again, the results suggested that there was no statistically significant contribution from unglaciated areas in

explaining present-day distribution patterns when the effects of modern topography, climate, and geology were considered first.

These two papers (Birks, 1993, 1996) show that the hypothesis of survival in glacial nunataks is unnecessary to explain the present-day patterns of species distribution and richness of Norwegian mountain plants. Following Ockham's razor principle (Subsection 10.3.3), this unnecessary assumption should be avoided when formulating hypotheses intended to explain present-day species distributions.

6 — Nonlinear regression

Logistic equation

In some applications, ecologists know from existing theory the algebraic form of the nonlinear relationship between a response variable and one or several explanatory variables. An example is the logistic equation, which describes population growth in population dynamics:

$$N_t = \frac{K}{1 + e^{(a - rt)}} \qquad \textbf{(10.26)}$$

This equation gives the population size (N_t) of a species at time t as a function of time (t). The equation contains three parameters a, r, and K, which are adjusted to the data; r is the Malthus parameter describing the natural rate of increase of the population, and K is the support capacity of the ecosystem. Nonlinear regression allows one to estimate the parameters (a, r, and K in this example) of the curve that best fits the data, for a user-selected function. This type of modelling does not assume linear relationships among the variables; the equation to be fitted is provided by the user. The algorithm for nonlinear parameter estimation tries to minimize an objective function.

The most usual objective functions to minimize are (1) the usual least-squares criterion $\Sigma (y_i - \hat{y}_i)^2$ and (2) the sum of squared Euclidean distances of the points to the regression function. These two criteria are illustrated in Fig. 10.6. The parameters of the best-fitting equation are found by iterative adjustment; users usually have the choice among a variety of rules for stopping the iterative search process. Common choices are: when the improvement in R^2 becomes smaller than some preselected value, when some preselected maximum number of iterations is reached, or when the change in all parameters becomes smaller than a given value. Useful references on this topic are Hollander & Wolfe (1973), Ratkowsky (1983), Ross (1990), Huet *et al.* (1992), and Bates & Chambers (1992). Nonlinear regression is available in several statistical packages, including R (see Section 10.7).

Consider the Taylor equation relating the means $\bar{y}$ and variances s_y^2 of several groups of data:

$$s_{y_k}^2 = a \bar{y}_k^b \qquad \textbf{(1.17)}$$

One must decide whether the equation should be fitted to the data by nonlinear regression, or to the corresponding logarithmic form (eq. 1.18) by linear regression. Look at the data in the original mean-variance space and in the transformed log(mean)-log(variance) space, and choose the form for which the data are homoscedastic.

Other often-encountered functions are the exponential, hyperbolic, Gaussian, and trigonometric (for periodic phenomena; see Subsection 12.4.5), and other growth models for individuals or populations.

Monotone regression

As an alternative to linear or nonlinear regression, Conover (1980, his Section 5.6) proposed *monotone regression* which may be used when (1) the relationship is monotonic (increasing or decreasing), (2) the purpose is forecasting or prediction rather than parameter estimation, and (3) one does not wish to carefully model the functional relationship; see also Iman & Conover (1983, their Section 12.6). Monotone regression consists in assigning ranks to the x and y observations and computing a linear regression on these ranks. Simple, natural rules are proposed to reassign real-number values to the forecasted/predicted values obtained from the rank-based equation for given values of x. Monotone regression is sometimes called *nonparametric regression*. A specialized form of monotone regression is used in nMDS algorithms (Section 9.4).

7 — Logistic regression

Binary variables form an important category of response variables that ecologists may wish to model. In process studies, one may wonder whether a given effect will be present under a variety of circumstances. Population ecologists are also often interested in determining the factors responsible for the presence or absence of a species. When the explanatory variables of the model are qualitative, modelling may call upon log-linear models computed on multiway contingency tables (Section 6.3). When the explanatory variables are quantitative, or represent a mixture of quantitative and qualitative data, logistic regression is the approach of choice.

In logistic regression, the response variable is binary (presence-absence, or 1-0; see example below). A linear model of quantitative explanatory variables would necessarily produce some forecasted/predicted values larger than 1 and some values smaller than 0. Consider Fig. 10.14, which illustrates the example developed below. A linear regression line fitting the data points would have a positive slope and would span outside the vertical [0, 1] interval, so that the equation would forecast ordinate values smaller than 0 (for small x) and larger than 1 (for large x); these would not make sense since the response variable can only be 0 or 1.

If one tries to predict the *probability* of occurrence of an event (for example the presence of a species), instead of the event itself (0 or 1 response), the model should be able to produce real-number values in the range [0, 1]. The logistic equation (eq. 10.26) described in Subsection 10.3.6 provides a sigmoid model for such a

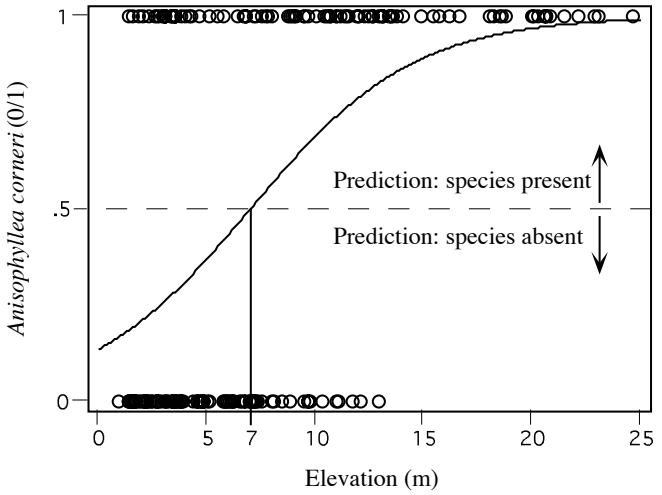


Figure 10.14 Logistic regression equation fitted to presence/absence of *Anisophyllea corneri*, as a function of elevation, in 200 forest quadrats.

response between limit values (Fig. 10.14). It is known to adequately model several ecological, physiological and chemical phenomena. Since the extreme values of the probabilistic response to be modelled are 0 and 1, then $K = 1$, so that eq. 10.26 becomes:

$$p = \frac{1}{1 + e^{-z}} \qquad \textbf{(10.27)}$$

where p is the probability of occurrence of the event. z is a linear function of the explanatory variable(s):

$$z = b_0 + b_1 x \qquad \text{for a single predictor } x \qquad \textbf{(10.28a)}$$

or $$z = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p \qquad \text{for several predictors} \qquad \textbf{(10.28b)}$$

Note that there are other, equivalent algebraic forms for the logistic equation. A form equivalent to eq. 10.27 is: $p = e^z/(1 + e^z)$.

For the error part of the model, the ε_i values cannot be assumed to be normally distributed and homoscedastic, as it is the case in linear regression, since the response variable can only take two values (presence or absence). The binomial distribution is the proper model in such a case, or the multinomial distribution for multistate qualitative response variables, as allowed in some computer software (e.g. CATMOD in

SAS). The parameters of the model cannot be estimated by ordinary least-squares since the error term is not normally distributed. This is done instead by maximum likelihood. Logistic regression is a special case of the generalized linear model (GLM: McCullagh & Nelder, 1983; Section 10.7); least-squares regression is another special case of GLM. According to the *maximum likelihood principle*, the best values for the parameters of a model are those for which the likelihood is maximum. The likelihood L of a set of parameter estimates is defined as the probability of observing the values that have actually been observed, given the model and the parameter estimates. This probability, which is not the same as p in eq. 10.27, is expressed as a function of the parameters:

Maximum likelihood

$$L = \mathrm{p}(\text{observed data} \mid \text{model, parameters})$$

So, one iteratively searches for parameter estimates that maximize the likelihood function.

Numerical example. Data describing the structure of a tree community, sampled over a 50–ha plot in the Pasoh forest*, Malaysia, were studied by He *et al*. (1994, 1996, 1997). The plot was established to monitor long-term changes in a primary tropical forest. The precise locations of the 334 077 individual trees and shrubs at least 1 cm in diameter at breast height (dbh) were determined (825 species in total) and a few environmental variables were recorded at the centres of 20 × 20 m quadrats. The present example uses the presence or absence of a species, *Anisophyllea corneri* Ding Hou (Cucurbitales), in each quadrat. One hundred quadrats were selected at random in the plot among those where *A. corneri* was present, and 100 among the quadrats where it was absent, for a total of 200 quadrats. Results of the logistic regression study presented below were reported by He *et al*. (1997).

Stepwise logistic regression was used to model the presence or absence of the species with respect to *slope* and *elevation* (i.e. altitude in metres measured by reference to the lowest part of the forest plot floor), using the SPSS software package. Following the calculations, *elevation* was included in the model for its significant contribution, whereas *slope* was left out. The linear part of the fitted model (eq. 10.28a) was:

$$z = -1.8532 + 0.2646 \times \mathit{elevation}$$

Significance of the regression coefficients was tested using the Wald statistic, which is the square of the ratio of a regression coefficient to its standard error; this statistic is distributed like χ^2. Both the intercept and slope coefficients of the model were significant ($p < 0.001$).

As explained above, the probability of the observed values of the response variable, for given values of the parameters, is called the likelihood. Since a probability is in the range [0, 1], its natural logarithm is a negative number. It is customary to multiply it by –2 to obtain $-2 \log_e(L)$, noted –2LL, a positive number that measures of how poorly a model fits the data; –2LL = 0 represents a perfect fit. This value presents the advantage of being distributed like χ^2,

* The Pasoh forest is one of the CTFS permanent forest plots. See note on these forest plots in Subsection 6.5.3.

so that it can be tested for significance. The significance of the model was tested using the following table:

	χ^2	ν	p (χ^2)
Intercept only	277.259	199	0.0002
Difference	59.709	1	< 0.0001
Intercept + *elevation*	217.549	198	0.1623
Difference	1.616	1	0.2057
Intercept + *elevation* + *slope*	215.933	197	0.1690
Goodness of fit	183.790	198	0.7575

Parameters were added to the model, one by one, as long as they improved the fit. The procedure is the same as in log-linear models (e.g. Table 6.6).

• For a model with an intercept only, –2LL = 277.259. The hypothesis to be tested was that –2LL = 277.259 was not significantly different from 0, which would be the value of –2LL for a model fitting the data perfectly. Degrees of freedom were computed as the number of observations (200) minus the number of fitted parameters (a single one up this point). The significant χ^2 statistic ($p < 0.05$) indicated that the model did not fit the data well.

• Inclusion of *elevation* added a second parameter to the model; this parameter was fitted iteratively and the resulting value of –2LL was 217.549 at convergence, i.e. when –2LL did not change by more than a small preselected value. Since the probability associated with the χ^2 statistic was large, the null hypothesis that the model fitted the data could not be rejected. The difference in χ^2 between the two models (277.259 – 217.549 = 59.709) was tested with 1 degree of freedom. The significant probability ($p < 0.05$) showed that *elevation* brought a significant contribution to the likelihood of the model.

• Inclusion of *slope* added a third parameter to the model. The resulting model also fitted the data well ($p > 0.05$), but the difference in χ^2 between the two models (217.549 – 215.933 = 1.616) was not significant ($p = 0.2057$), indicating that *slope* did not significantly contribute to increase the likelihood of the model. Hence, *slope* was left out of the final model.

The last row of the table tested a goodness-of-fit statistic that compared the observed values (0 or 1 in logistic regression) to the probabilities forecasted by the model, which included the intercept and *elevation* in this example (Norusis, 1990, p. 52). The statistic (183.790) is distributed like χ^2 and has the same number of degrees of freedom as the χ^2 statistic for the complete model. In the present example, this statistic was not significant ($p > 0.05$), which led to conclude that there was no significant discrepancy between the forecasted values and the data.

Putting back the observed values of the explanatory variable(s) into the model (eq. 10.28a) provided estimates of z. For instance, one of the quadrats in the example data had *elevation* = 9.5 m, so that

$$z = -1.8532 + 0.2646 \times 9.5 = 0.6605$$

Incorporating this value into eq. 10.27 provided the following probability that *A. corneri* would be present in the quadrat:

$$p = \frac{1}{1 + e^{-0.6605}} = 0.659$$

Since $p > 0.5$, the forecast was that the species should be found in this quadrat. In general, if $p < 0.5$, the event is unlikely to occur whereas it is likely to occur if $p > 0.5$. (Flip a coin if a forecasted value is required in a case where $p = 0.5$ exactly.) With the present equation, the breaking point between forecasted values of 0 and 1 (i.e. the point where $p = 0.5$) corresponded to an *elevation* of 7 m. The logistic curve fitted to the *A. corneri* data is shown in Fig. 10.14.

Classification table

Forecasted values may be used to produce a classification (or "confusion") table, as in linear discriminant analysis (Section 11.3), in which the forecasted values are compared to observations. For the example data, the classification table was:

Observed	*Forecasted*		*Percent*
	0	1	*correct*
0	78	22	78%
1	35	65	65%
Total correct classification			71.5%

Since most values are in the diagonal cells of the table, one concludes that the logistic regression equation based solely on elevation was successful at forecasting the presence of *A. corneri* in the quadrats.

Gaussian logistic model

A Gaussian logistic equation may be used to model the unimodal response of a species to an environmental gradient. Fit the logistic equation with a quadratic response function $z = b_0 + b_1 x + b_2 x^2$, instead of eq. 10.28a, to obtain a Gaussian logistic model; the response function for several predictors (eq. 10.28b) may be modified in the same way. See ter Braak & Looman (1987) for details.

Linear discriminant analysis (Section 11.3) has often been used by ecologists to study niches of plants or animals, before logistic regression became widely available in computer packages. Williams (1983) gives examples of such works. The problem with discriminant analysis is that it constructs a linear model of the explanatory variables, so that the forecasted values are not limited to the [0, 1] range. Negative values and values higher than 1 can be produced, which are ecologically unrealistic for presence-absence data. This problem does not appear with logistic regression, which is available in major statistical packages as well as in S-PLUS®, MATLAB® and R. This question is further discussed in Section 11.6.

In procedure CATMOD of SAS, the concept of logistic regression is extended to multi-state qualitative response variables. Trexler & Travis (1993) provide an application of logistic regression to an actual ecological problem, including selection of the most parsimonious model; they also discuss the relative merits of various alternatives to the logistic model.

8 — Splines and LOWESS smoothing

There are instances where one is only interested in estimating an empirical relationship between two variables, without formally modelling the relationship in an equation and estimating its parameters. In such instances, smoothing methods may be the most appropriate, since they provide an empirical representation of the relationship, efficiently and at little cost in terms of time spent specifying a model. Since they fit the data locally (i.e. within small windows), smoothing methods are useful when the relationship greatly varies in shape along the abscissa. This is the opposite of the parametric regression methods, where a single set of parameters is used to adjust the same function to all data points (global fit). Smoothing methods are far less sensitive to exceptional values and outliers than regression, including polynomial regression. Several numerical methods are available for smoothing.

Moving average

A simple way to visualize an empirical relationship is the method of moving averages, described in more detail in Section 12.2. Define a 'window' of a given width, position it at one of the margins of the scatter diagram, and compute the mean ordinate value (y) of all the observations in the window. Move the window by small steps along the abscissa, recomputing the mean every time, until the window reaches the opposite margin of the scatter diagram. Plot the window means as a function of the positions of the window centres along the abscissa. Link the mean estimates by line segments. This empirical line may be used to estimate y as a function of x.

Piecewise polynomial fitting by "splines" is a more advanced form of local smoothing. In its basic form, spline estimation consists in dividing the range of the explanatory variable x (which is also the width of the scatter diagram) into a number of intervals, which are generally of equal widths and separated by *knots*, and adjusting a polynomial of order k to the data points within each segments using polynomial regression (Subsection 10.3.4). To make sure that the transitions between spline segments are smooth at the junction points (knots), one imposes two constraints: (1) that the values of the function be equal on the left and right of the knots, and (2) that the (k–1) first derivatives of the curves be also equal on the left and right of the knots. Users of the method have to make arbitrary decisions about (1) the level k of the polynomials to be used for regression (a usual choice is cubic splines) and (2) the number of segments along the abscissa. If a large enough number of intervals is used, the spline function can be made to fit every data point. A smoother curve is obtained by using fewer knots. It is recommended to choose the interval width in such a way as to have at least 5 or 6 data points per segment (Wold, 1974). Knots should be positioned at or near inflexion points, where the behaviour of the curve changes (see example below). A large body of literature exists about splines. Good introductory texts are Chambers (1977), de Boor (1978), Eubank (1988), and Wegman & Wright (1983). The simplest text is Montgomery & Peck (1982, Section 5.2.2); it inspired the explanation of the method that follows.

When the positions of the knots are known (i.e. decided by users), a cubic spline model *with no continuity restriction* is written as:

$$\hat{y} = \sum_{j=0}^{3} b_{0j} x^j + \sum_{k=1}^{h} \sum_{j=0}^{3} b_{kj} (x - t_k)_+^j \qquad \textbf{(10.29)}$$

In this equation, the parameters b_{0j} in the first sum correspond to a cubic polynomial equation in x. The parameters b_{kj} in the second sum allow the curve segments to be disconnected at the positions of the knots. There are h knots, and their positions along the abscissa are represented by t_k; the knots are ordered in such a way that $t_1 < t_2 < \ldots < t_h$. This equation, written out in full, is the following for a single knot (i.e. $h = 1$) located at position t:

$$\hat{y} = b_{00} + b_{01}x + b_{02}x^2 + b_{03}x^3 + b_{10}(x - t)_+^0 + b_{11}(x - t)_+^1 + b_{12}(x - t)_+^2 + b_{13}(x - t)_+^3$$

The expression $(x - t_k)_+$ takes the value $(x - t_k)$ when $x - t_k > 0$ (i.e. if the given value x is to the right of the knot), and 0 when $x - t_k \leq 0$ (for values of x on the knot or to the left of the knot). The constraint of continuity is implemented by giving the value zero to all terms b_{kj}, except the last one. In eq. 10.29, it is these parameters that allow the relationship to be described by discontinuous curves; by removing them, eq. 10.29 becomes a cubic splines equation with continuity constraint:

Cubic splines

$$\hat{y} = \sum_{j=0}^{3} b_{0j} x^j + \sum_{k=1}^{h} b_k (x - t_k)_+^3 \qquad \textbf{(10.30)}$$

which has a single parameter b_k for each knot. Written in full, eq. 10.30 is the following for two knots (i.e. $h = 2$) located at positions $t_1 = -5$ and $t_2 = +4$, as in the numerical example below:

$$\hat{y} = b_{00} + b_{01}x + b_{02}x^2 + b_{03}x^3 + b_1(x + 5)_+^3 + b_2(x - 4)_+^3$$

This approach is not the one used in advanced spline smoothing packages because it has some numerical drawbacks, especially when the number of knots is large. It is, however, the most didactic, because it shows spline smoothing to be an extension of OLS polynomial regression. Montgomery & Peck (1982) give detailed computational examples and show how to test the significance of the difference in R^2 between models with decreasing numbers of knots, or between a spline model and a simple polynomial regression model. They finally show that *piecewise linear regression* — that is, fitting a continuous series of straight lines through a scatter of points — is a natural extension of the spline eq. 10.30 in which the exponent is limited to 1.

LOWESS

LOWESS refers to *Locally Weighted Scatterplot Smoothing* (Cleveland, 1979). This method is an extension of moving averages in the sense that, for each value x_i along the abscissa, a value $\hat{y}_i$ is estimated from the data present in a window around x_i. The

number of data points included in the moving window is a proportion f, determined by users, of the total number of observations; a commonly-used first approximation for f is 0.5. The higher this proportion, the smoother the line of fitted values will be. For the end values, all observed points in the window come from the same side of x_i; this prevents the lines from becoming flat near the ends. Estimation proceeds in two steps:

- First, a weighted simple linear regression is computed for the points within the window and an estimate $\hat{y}_i$ is obtained. Weights, given to the observation points by a 'tricube' formula, decrease from the focal point x_i outwards. Points outside the window receive a zero weight. This regression procedure is repeated for all values x_i for which estimates are sought.

- The second step is to make these first estimates more robust, by reducing the influence of exceptional values and outliers. Residuals are computed from the fitted values and, from these, new weights are calculated that give more importance to the points with low residuals. Weighted linear regression is repeated, using as weights the products of the new weights with the original neighbourhood weights. This second step may be repeated until the recomputed weights display no more changes.

Trexler & Travis (1993) give a detailed account of the LOWESS method, together with a full example, and details on two techniques for choosing the most appropriate value for f. The simplest approach is to start with a (low) initial value, and increase it until a non-random pattern along x appears in the residuals; at that point, f is too large. Other important references are Chambers *et al*. (1983) and Cleveland (1985).

Numerical example. Consider again the dependence of salinity on the position along a transect, as modelled in Fig. 10.9. This same relationship may be studied using cubic splines and LOWESS (Fig. 10.15). For splines smoothing, the arbitrary rule stated above (5 or 6 points at least per interval) leads to 3 or 4 intervals. Figure 10.9 indicates, on the other hand, that there are at least three regions in the scatter of points, which can be delimited by knots located at approximately –5 and +4 along the abscissa. The computed spline regression equation which follows has $R^2 = 0.841$:

$$\hat{y} = 37.500 + 0.291x + 0.072x^2 + 0.006x^3 - 0.005\,(x+5)_+^3 - 0.007\,(x-4)_+^3$$

The difference in explained variation between this spline model and a cubic polynomial model ($R^2 = 0.81$, Fig. 10.9) is not significant.

The LOWESS curve also clearly suggests the presence of three distinct physical processes which determine the values of salinity along the long axis of the lagoon, i.e. from abscissa –10 to about –5, the central portion, and the right-hand portion from abscissa 4 and on.

Other smoothing methods are available in computer software, such as negative exponentially weighted smoothing (the influence of neighbouring points decreases exponentially with distance); inverse squared distance smoothing, described in Subsection 13.2.2 (eq. 13.21 with $k = 2$); distance-weighted least-squares smoothing (the surface is allowed to bend locally to fit the data); and step smoothing (a step function is fitted to the data).

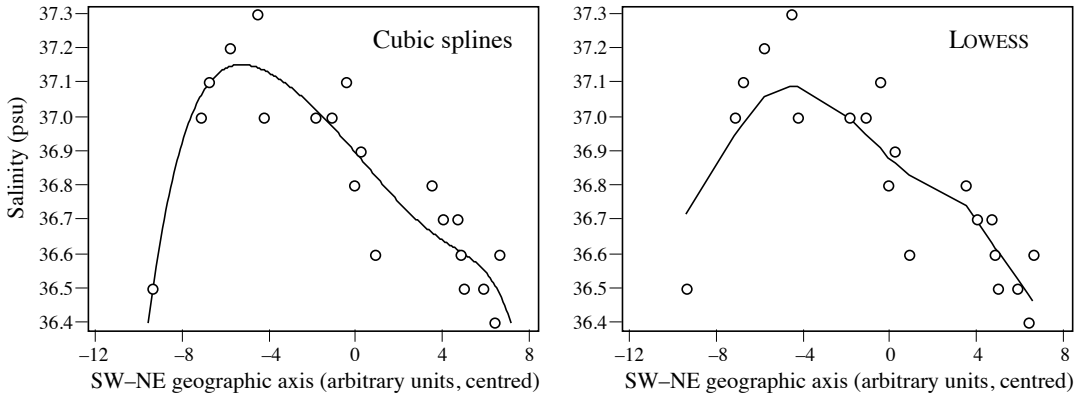


Figure 10.15 Cubic splines and LOWESS scatter diagrams describing the relationship of salinity with the position of the sites along the main geographic axis of the Thau lagoon, on 25 October 1988. Cubic splines were computed with knots at –5 and +4 on abscissa. For LOWESS (computed using SYSTAT), the proportion of the points included in each smoothing window was $f = 0.5$.

10.4 Path analysis

Subsection 4.5.4 showed that causal relationships among descriptors cannot be unambiguously derived from the sole examination of correlation coefficients, whether simple, multiple, or partial. Several causal models may account for the same correlation coefficients. In the case of *prediction* (*versus forecasting*, see Subsection 10.2.2), however, causal (and not only correlative) relationships among descriptors must be established with reasonable certainty. *Path analysis* is an extension of *multiple linear regression* (Subsection 10.3.3) that allows the decomposition and interpretation of *linear* relationships among a (small) number of descriptors. It is thus possible to formally state *a priori hypotheses* concerning the causal relationships among descriptors and, using path analysis, examine their consequences given the coefficients of regression and correlation computed among these descriptors.

Structural equation modelling

Path analysis was developed by Wright (1921, 1960). It is now recognized as a special case of a more general method called *structural equation modelling* (SEM), which includes latent variables (unmeasured, but estimated in the model by several measured variables) in addition to the measured variables. Structural equation models allow both exploratory and confirmatory modelling, meaning that the method is suited to develop as well as test theories. There are many interesting applications of path analysis and SEM in ecology, evolution, population genetics, and the social sciences. An introductory presentation of path analysis is found in Sokal & Rohlf (1995). The present section only provides a summary of path analysis showing its link with linear

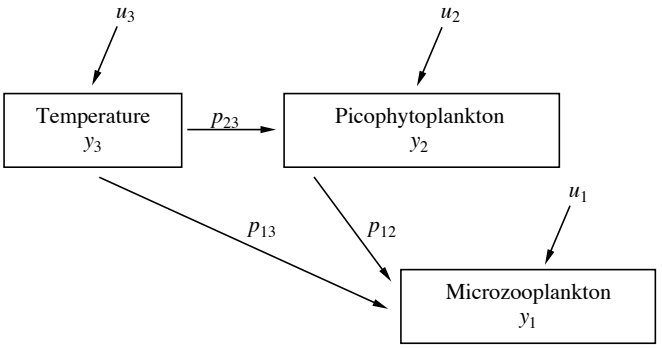


Figure 10.16 Path diagram for three linearly related descriptors. Adapted from Nie *et al*. (1975).

regression, and concludes with an ecological application. More complete and detailed presentations of path analysis and structural equation modelling are found in the books of Shipley (2002), Pugesek *et al*. (2003) and Grace (2006) written for ecologists, as well as books written for the social sciences, e.g. Kaplan (2009) and Kline (2011).

Causal order

Path diagram

Causal closure

As mentioned in Section 10.2, path analysis is based on two fundamental assumptions. (1) There exists a *causal order* among the variables. This causal order, which must be defined by the researchers, may be derived from ecological theory, or established experimentally (for a brief discussion of experiments, see Subsection 10.2.3). The assumption is that of *weak causal* ordering, e.g. y_1 *may* affect y_2 but y_2 *cannot* affect y_1. In *path diagrams* (Figs. 10.16 to 10.18), the causal ordering is represented by arrows, e.g. $y_1 \rightarrow y_2$. (2) No model can account for all the observed variance. Path models thus include *residual variables* u_i, which represent the unknown factors responsible for the residual variance (i.e. the variance not accounted for by the observed descriptors). The assumption of *causal closure* implies the independence of the residual causal variables; in other words, one assumes the existence of residual variables such that $u_1 \rightarrow y_1$ and $u_2 \rightarrow y_2$, whereas $u_1 \rightarrow y_2$ or $u_2 \rightarrow y_1$ is not allowed.

Numeral example. A simple example, with three variables exhibiting causal relationships, is used to illustrate the main features of path analysis. It is adapted from Nie *et al*. (1975, p. 386 *et seq*.). The example considers hypothesized relationships among water temperature, picophytoplankton (algae < 2 μm), and microzooplankton (e.g. ciliates) grazing on the picophytoplankton. In the model, it is assumed that water temperature (y_3) directly affects the growth of microzooplankton (y_1) and picophytoplankton (y_2), whose abundance, in turn, affects that of microzooplankton. Following the terminology of Sokal & Rohlf (1995, Section 16.3), y_2 and y_3 are *predictor* (or explanatory) variables while y_1 is the *criterion* (or response) variable. Figure 10.16 illustrates this hypothetical network of causal relationships in schematic form. Since the three variables probably do not explain all the observed variance, the model also

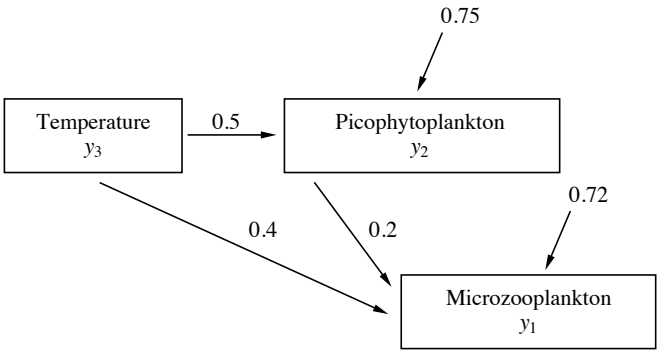


Figure 10.17 Results of path analysis for the example of Fig. 10.16. See text.

includes residual variables u_1 to u_3. The *causal ordering* of Fig. 10.16 is summarized in the following system of linear equations:

$$y_3 = u_3$$

$$y_2 = p_{23}y_3 + u_2$$

$$y_1 = p_{13}y_3 + p_{12}y_2 + u_1$$

Path coefficient

where parameters p_{ij} are the *path coefficients*. All variables are centred on their respective means. The hypothesis of causal closure implies that:

$$s(u_1,u_2) = s(u_1,u_3) = s(u_2,u_3) = 0$$

because the residual causes are independent; s represents covariances.

The path coefficients are estimated using multiple linear regression (Subsection 10.3.3):

$$\hat{y}_2 = p_{23}y_3$$

$$\hat{y}_1 = p_{13}y_3 + p_{12}y_2$$

There are no intercepts (coefficients p_0) in the regression equations because the data are centred. For a model with n descriptors, one can estimate all path coefficients using at most $(n-1)$ regression equations. Each descriptor is predicted from the descriptors with immediately higher causal order. Two regression equations are needed to calculate the three path coefficients in Fig. 10.16. Let us use the following values for the path coefficients (Fig. 10.17) and coefficients of determination (R^2) of the numerical example:

$$\hat{y}_2 = 0.5y_3 \qquad R^2 = 0.25$$

$$\hat{y}_1 = 0.4y_3 + 0.2y_2 \qquad R^2 = 0.28$$

and the following correlation coefficients among the descriptors:

$$r_{12} = 0.4 \qquad r_{13} = 0.5 \qquad r_{23} = 0.5$$

The correlation r_{13} depends on both the direct relationship between y_1 and y_3 and the indirect relationship *via* y_2 (Figs. 10.16 and 10.17). Path analysis makes it possible to interpret the correlation r_{13} within the framework of the above model of causal relationships. Because the regressions that provide the estimates of the path coefficients are computed using *standardized* variables (eq. 1.12), it follows (Sokal & Rohlf, 1995, eq. 16.6) that

$$\begin{aligned} r_{13} &= p_{13} + r_{23}p_{12} \\ &= 0.4 + 0.5 \times 0.2 \\ &= 0.4 + 0.1 = 0.5 \end{aligned}$$

The correlation between y_3 (predictor variable) and y_1 (criterion variable) includes the direct contribution of y_3 to y_1 (path coefficient p_{13}), and also the common causes behind the correlations between y_3 and y_1. More generally, the correlation between a predictor variable y_i and a criterion variable y_1 includes the direct contribution of y_i to y_1, plus the common causes behind the correlations between y_i and any other variable that has a direct effect on y_1. These various contributions may either increase (as in the present example) or decrease the correlation between the predictor and criterion variables. The correlation coefficient r_{13} thus includes both a direct (0.4) and an indirect component (0.1).

Coefficient of nondetermination

Coefficients of nondetermination[*] are used to estimate the fractions of the variance that are not explained by the models (Fig. 10.17):

$$\begin{aligned} r^2(u_2, y_2) &= 1 - R^2_{2.3} = 1 - 0.25 = 0.75 \\ r^2(u_1, y_1) &= 1 - R^2_{1.23} = 1 - 0.28 = 0.72 \end{aligned}$$

One concludes that 75% of the variance of picophytoplankton (y_2) and 72% of the variance of microzooplankton (y_1) are not explained by the causal relationships stated in the model. The same results are obtained using the following general formula (Sokal & Rohlf, 1995):

$$\begin{aligned} r^2(u_1, y_1) &= 1 - \left[\sum_i p^2_{1i} + 2 \sum_{ij} p_{1i} p_{1j} r_{ij} \right] \\ &= 1 - [(p^2_{12} + p^2_{13}) + 2(p_{12}p_{13}r_{23}) \\ &= 1 - [(0.04 + 0.16) + 2(0.2 \times 0.4 \times 0.5)] \\ &= 1 - [0.20 + 2(0.04)] \\ &= 1 - 0.28 = 0.72 \end{aligned}$$

The above results may be summarized in a single table. In the numerical example (Table 10.7), 0.1/0.5 = 20% of the covariation between microzooplankton (y_1) and temperature (y_3) is through picophytoplankton (y_2). In addition, 0.2/0.4 = 50% of the observed relationship between microzooplankton (y_1) and picophytoplankton (y_2) is not causal, and thus spurious

[*] The *coefficient of nondetermination* is $(1 - R^2)$; $\sqrt{1 - R^2}$ is called the *coefficient of alienation*.

Table 10.7 Decomposition of bivariate covariation among the (standardized) variables of Fig. 10.17. Adapted from Nie *et al*. (1975).

Bivariate relationships	Total covariation	Causal covariation			Noncausal covariation
		Direct	Indirect	Total	
	(A)	(B)	(C)	(D = B+C)	(A–D)
y_2y_3	$r_{23} = 0.5$	0.5	0.0	0.5	0.0
y_1y_3	$r_{13} = 0.5$	0.4	0.1	0.5	0.0
y_1y_2	$r_{12} = 0.4$	0.2	0.0	0.2	0.2

according to the path model of Figs. 10.16 and 10.17. Such spurious correlations occur when two descriptors are caused by a third one (e.g. Fig. 4.11, Model 2) whose values have not been observed in the study.

Path analysis can be applied to more than three variables. As the number of variables increases, interpretation of the results becomes more complex and the number of possible models increases rapidly. In practice, path analysis is restricted to exploring the causal structure of relatively simple systems. This type of analysis is very useful in many ecological situations, if only because it forces researchers to explicitly state their hypotheses about the causal relationships among descriptors. The method helps assess the consequences of hypotheses, given the observed covariation among descriptors. Other methods, mentioned in Table 10.3, must be used when the descriptors do not exhibit linear relationships, or when they are not quantitative.

The following Ecological application 10.4 concerns freshwater ecology. Other applications of path analysis may be found, for example, in the fields of bacterial ecology (Troussellier *et al*., 1986), biological oceanography (Gosselin *et al*., 1986; Legendre *et al*., 1991), and plant ecology (Hermy, 1987; Kuusipalo, 1987).

Ecological application 10.4

Harris & Charleston (1977) used path analysis to compare the microhabitats of two pulmonate snails, *Lymnaea tomentosa* and *L. columella*. The two species live in freshwater marshes; there are no obvious differences in the physical or chemical features of their respective habitats. Path analysis was used to examine, for each of the two species, the hypothetical model of causal relationships represented in schematic form in Fig. 10.18. In this model, water was assumed to affect snail numbers directly, and also *via* mud and flocculence, since both factors are partly determined by the amount of water present. The amount of mud was also expected to influence

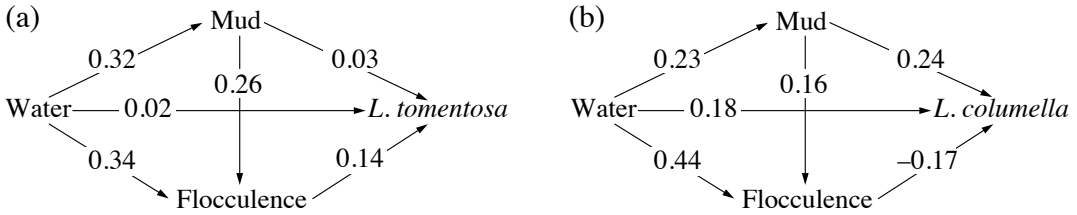


Figure 10.18 Path diagrams of the hypothesized effects of water, mud and flocculence on population densities of two pulmonate snails in marsh microhabitats. After Harris & Charleston (1977).

snails directly; however, larger mud areas are less likely to contain vegetation and are thus more likely to be flocculent, hence the indirect path from mud to snails *via* flocculence.

Results of path analysis (Fig. 10.18) suggest major differences between the microhabitats of the two species. Overall, increasing water cover has a direct (positive) effect on *L. columella*; in addition, flocculent mud appears to favour *L. tomentosa* whereas *L. columella* seem to prefer firm mud. The effects of water and mud on *L. columella* are thus direct, whereas they are indirect on *L. tomentosa* (i.e. *via* flocculence). The tentative hypothesis generated by the path diagrams must be further tested by observations and experiments. However, designing experiments to test the role played by the consistency of mud, while controlling for other (confounding) variables, would require considerable ingenuity.

10.5 Matrix comparisons

Regression and path analysis are restricted to the interpretation of univariate response variables. Other methods are required to perform direct comparison analyses when the descriptors form multivariate data tables. As shown in Fig. 10.4, canonical methods (Chapter 11) analyse the relationship between two rectangular data tables, whereas Mantel tests and derived forms, described in the present section, relate similarity or distance matrices derived from rectangular data tables.

Three main approaches are discussed in the present section. The Mantel test (Subsection 10.5.1) and derived forms (partial Mantel test, multiple regression on distance matrices, Subsection 10.5.2) are used to test relationships between association matrices, not between the rectangular date tables from which they originate. This is also the case of the analysis of similarities (ANOSIM, Subsection 10.5.3). The Procrustes test (Subsection 10.5.4) is different: it assesses the relationship between two rectangular data tables, not between association matrices. That test, derived from Procrustes analysis (Subsection 11.5.2), is presented in the present section to indicate that there are alternatives to the Mantel test to relate data matrices. Chapter 11 describes several other methods for the comparison of raw data matrices.

1 — Two association matrices: Mantel test

The Mantel (1967) test is a method to compare two similarity (**S**) or distance matrices (**D**), computed for the same objects, and test a hypothesis about the relationship between these matrices. For simplicity, the presentation will focus on distance matrices **D**. *Mantel tests should not be used to test hypotheses about the relationships between the original data tables, for reasons explained at the end of this subsection*. The data tables used to compute the two distance matrices must have been obtained independently of each other (i.e. different variables). One of the matrices may actually represent a hypothesis instead of real data, as shown below.

Ecological theory sometimes predicts relationships between resemblance matrices (**S** or **D**). This is the case with neutral theory, which predicts a monotonic relationship to appear in similarity decay plots where community composition similarity is expected to decreases with geographic distance (Nekola & White, 1999; Hubbell, 2001). In genetics, the theory of isolation by distance (Wright, 1943) is based on the fact that in sexually reproducing organisms, individuals tend to find mates in nearby rather than distant populations; for sessile organisms, this theory applies to species with short-range dispersal. As a consequence, populations living near each other tend to be more genetically similar than distant populations. In both cases, the theoretical predictions can be tested by analysing matrices of ecological or genetic distances $\mathbf{D_Y}$ versus geographic distances $\mathbf{D_X}$ using the Mantel test or regression on distance matrices (Subsection 10.5.2). Matrices $\mathbf{D_Y}$ and $\mathbf{D_X}$ must be computed for the same n objects *listed in the same order*. For ecological data, the choice of an appropriate resemblance measure is discussed in Chapter 7. In the two examples of the present paragraph, one of the matrices contains geographic distances among sites and Mantel tests may be used to test the predictions of these theories concerning distances. Other statistical methods can and should be used to test other predictions of these theories.

z_M statistic

The basic form of the Mantel statistic, called z_M, is the scalar product (Section 2.5) of the (*unstandardized*) values in the two resemblance matrices, excluding the main diagonal, which only contains trivial values (1's for similarities, 0's for distances) for which no estimate has been computed (Fig. 10.19). A second approach is to *standardize* the values in each of the two vectors of resemblance before computing the Mantel statistic. The cross-product statistic, divided by the number of distances in each half-matrix minus 1 [i.e. $(n(n-1)/2) - 1$], is bounded between –1 and +1; it behaves like a correlation coefficient and is called r_M. A third approach is to transform the distances into ranks (Dietz, 1983) before computing the standardized Mantel statistic; this is equivalent to computing a Spearman correlation coefficient (Section 5.3) between the corresponding values of matrices $\mathbf{D_Y}$ and $\mathbf{D_X}$.

r_M statistic

Permutation test

Mantel statistics are tested by permutation (Section 1.2). The n objects forming the rows and columns of the similarity or distance matrices are the permutable units, so that the permutations actually concern the n objects, not the $[n(n-1)/2]$ values in each half-matrix of distances. The testing procedure for Mantel statistics is summarized in Box 10.2.

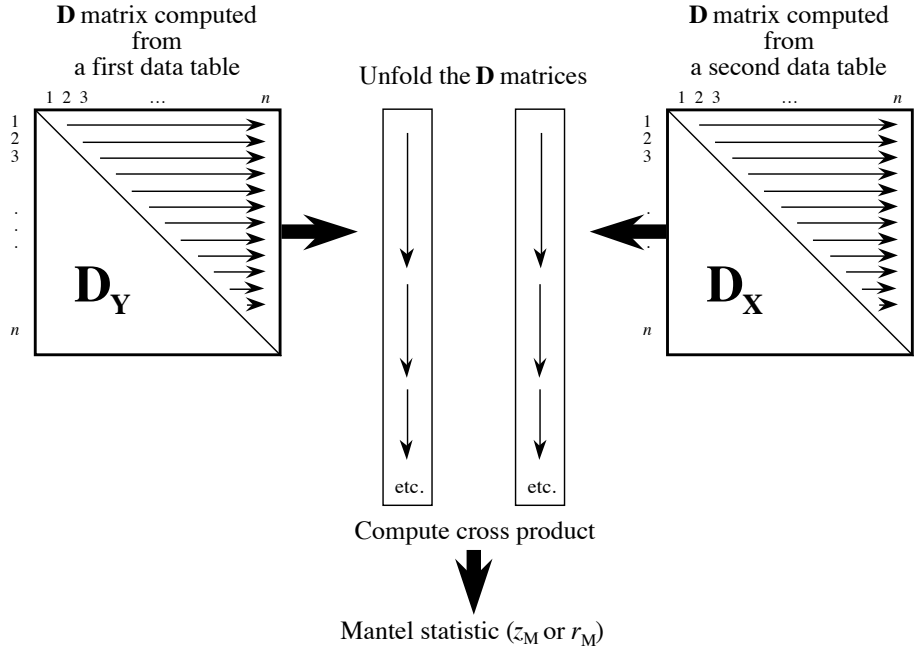


Figure 10.19 The Mantel statistic is the scalar product (sum of cross products) of the corresponding values in two distance matrices (**D**). Values in the vectors representing the unfolded matrices (i.e. written out as vectors) may be standardized before computing the statistic (r_M), or not (z_M), or transformed into ranks.

The permutation test leads to the same p-value with statistics z_M or r_M because all cross-product results, permuted or not, are affected in the same way by linear transformations such as standardization (eq. 1.12) of one or both vectors of distances. This is a most important property of the Mantel test. Thanks to it, the arbitrary values used in *model matrices* (below) are not an issue because any pair of chosen contrasting values leads to the same p-value.

Mantel tests are usually one-tailed since, in most cases, ecologists have a strong hypothesis about the sign of the correlation between the two matrices being compared. The hypothesis may be that the two distance matrices are positively related, which leads to a test of significance in the upper tail of the reference distribution. This is certainly the case when testing a hypothesis of isolation by distance in genetics. When comparing a similarity to a distance matrix, as in similarity decay plots, one generally expects a negative relationship to be found, if any; the test is then in the lower tail of the reference distribution.

Theory of the Mantel test — Box 10.2

Hypotheses

H_0: The distances among objects in matrix $\mathbf{D_Y}$ are not (linearly or monotonically) related to the corresponding distances in $\mathbf{D_X}$.

H_1: The distances among points in matrix $\mathbf{D_Y}$ are related to the distances in $\mathbf{D_X}$.

Test statistics

• Mantel (1967) statistic: $z_M = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbf{D_Y}_{ij} \mathbf{D_X}_{ij}$ where i and j are row and column indices of the $\mathbf{D}$ matrices.

• Standardized Mantel statistic: $r_M = \frac{1}{d-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \text{stand}(\mathbf{D_Y})_{ij} \, \text{stand}(\mathbf{D_X})_{ij}$

where stand($\mathbf{D_Y}$) stand($\mathbf{D_X}$) contain standardized distances in their upper-triangular portions and $d = [n(n-1)/2]$ is the number of distances in the upper-triangular portion of each matrix.

Distribution of the test statistic

• According to H_0, the vector of values observed for any one object could have been observed for any other object; in other words, the objects are the permutable units. A realization of H_0 is obtained by permuting the objects (rows) in one of the original data matrices, bringing with them their vectors of values for the observed variables, and recomputing the distance matrix.

• An equivalent result is obtained by permuting at random the rows and corresponding columns of matrix $\mathbf{D_Y}$. Either $\mathbf{D_Y}$ or $\mathbf{D_X}$ can be permuted at random, with the same net effect.

• Repeating the above operation, the different permutations produce a set of values of the Mantel statistic, z_M or r_M, obtained under H_0. These values estimate the sampling distribution of the Mantel statistic under H_0.

Statistical decision

As in any other statistical test, the decision to reject H_0 or not is made by comparing the actual value of the auxiliary variable (z_M or r_M) to the reference distribution obtained under H_0. If the actual value of the Mantel statistic is one likely to have been obtained under the null hypothesis (no relationship between $\mathbf{D_Y}$ and $\mathbf{D_X}$), H_0 is not rejected; if it is too extreme to be considered a likely result under H_0, H_0 is rejected. See Section 1.2 for details.

Remarks

• The z_M or the r_M statistics may be transformed into another statistic, called t by Mantel (1967), which is asymptotically normal. It is tested by referring to a table of the standard normal distribution. It provides a good approximation of the probability **when *n* is large**.

• Like the Pearson correlation coefficient, the Mantel statistic formula is a linear model that brings out the linear component of the relationship between the values in two distance matrices. Strong nonlinearities may prevent the identification of relationships in Mantel tests. This led Mantel (1967) and Dietz (1983) to suggest the use of the Spearman or Kendall nonparametric correlation coefficients, instead of Pearson's r, as the statistic in Mantel tests.

Examples of Mantel tests are found in Upton & Fingleton (1985), Legendre & Fortin (1989), Sokal & Rohlf (1995), and elsewhere. The Mantel test is the statistical basis for the Mantel correlogram described in Subsection 13.1.6.

The Mantel test is only valid if matrix $\mathbf{D_X}$ is independent of the resemblance measures in $\mathbf{D_Y}$, i.e. $\mathbf{D_X}$ should not be derived in any way from $\mathbf{D_Y}$ nor from the data that were used to compute $\mathbf{D_Y}$. The Mantel test has two chief domains of application in community ecology:

1. It may be used to compare two resemblance matrices computed from empirical data and test a hypothesis about the relationship between the distances, as in the similarity decay plot example described above. For the test to be valid, $\mathbf{D_X}$ must be computed from the same objects but a different set of variables than those used to compute $\mathbf{D_Y}$.

Model matrix

2. The Mantel test may also be used to assess the goodness-of-fit of data to an *a priori* distance model. The test compares the empirical distance matrix to a *model matrix* (also called a *pattern* or *design matrix*). This matrix is constructed to represent the model to be tested; in other words, it depicts the alternative hypothesis of the test. For example, in the Mantel correlogram (Subsection 13.1.6), the model is a classification of the distances in two groups, e.g. the distances smaller than a given value of interest and the larger distances. Figure 13.14 (Chapter 13) shows two matrices, $\mathbf{X}(1)$ and $\mathbf{X}(2)$, representing such models. Other examples are given by Sokal & Rohlf (1995, Section 18.3).

The Mantel test cannot be used to check the conformity to a matrix $\mathbf{D_Y}$ of a model derived from the same data, e.g. to test the conformity of $\mathbf{D_Y}$ to a group structure obtained by clustering matrix $\mathbf{D_Y}$. In such a case, the model matrix $\mathbf{D_X}$, which depicts the *alternative hypothesis* of the test, would describe a structure made to fit the very data that would now be used to test the null hypothesis. The hypothesis ($\mathbf{D_X}$) would not be independent of the data ($\mathbf{D_Y}$) used to test it. Such a test would be incorrect; it would almost always reject the null hypothesis and support the conformity of $\mathbf{D_Y}$ to $\mathbf{D_X}$. This point has been mentioned in Subsection 8.12.2.

Goodness-of-fit Mantel tests have been used in vegetation studies to investigate hypotheses related to questions like the concept of climax (McCune & Allen, 1985) and the environmental control model (Burgman, 1987, 1988). Hypotheses of niche segregation have been tested for trees by Legendre & Fortin (1989), and for animals by Hudon & Lamarche (1989). Somers & Green (1993) used Mantel tests based on Spearman correlation coefficients (see Box 10.2, Remarks) to assess the relationship between crayfish catches in six Ontario lakes and five model matrices corresponding to different ecological hypotheses. Considering what is now known about the properties of the Mantel test (see Box 10.3), in all these applications, the Mantel test should be replaced by canonical analysis (Chapter 11), which provides more powerful tests of significance.

Further developments, Mantel test — Box 10.3

Mantel tests should be restricted to test hypotheses concerning distances. A Mantel test between two distance matrices $\mathbf{D_Y}$ and $\mathbf{D_X}$ derived from raw data tables **Y** and **X** is not equivalent to (1) the test of a correlation coefficient computed between two vectors of raw data, (2) a test computed by linear regression between a response vector **y** and an explanatory matrix **X**, or (3) a test in canonical analysis between a multivariate response matrix **Y** and an explanatory matrix **X**. This statement is supported by the following observations.

1. The sum of squares of the distances *is not* the sum of squares of the raw data (Legendre *et al*., 2005; Legendre & Fortin, 2010). On the one hand,

$$\mathrm{SS}(\mathbf{Y}) = \sum_{j=1}^{p} \sum_{i=1}^{n} (y_{ij} - \bar{y}_j)^2 = \left(\sum_{i \neq h} D_{ih}^2 \right) / n$$

as shown in Box 6.1, eqs. 6.55 and 6.56. On the other hand, the sum of squares of the distances in a distance matrix **D** is computed as follows:

$$\mathrm{SS}(\mathbf{D}) = \sum_{i \neq h} (D_{ih} - \bar{D})^2 = \sum_{i \neq h} D_{ih}^2 - \frac{\left(\sum_{i \neq h} D_{ih} \right)^2}{n(n-1)/2}$$

The last equation is for symmetric distance matrices, where only the D_{ih} values in the upper or lower-triangular portion of **D** are used. The right-hand parts of the two equations above are irreductible to one another. Consider the numbers 1 to 10 for example: their total sum of squares SS(**Y**) in the first equation is 82.5 and the sum of squares of the Euclidean distances among these values (SS(**D**), second equation) is 220. SS(**Y**) is the denominator of R^2 in multiple regression (eq. 10.20) and canonical analysis (eq. 11.4) whereas SS(**D**) is the denominator of the R^2 in regression on distance matrices (Subsection 10.5.2); in simple Mantel tests, the square root of this R^2 is the Mantel statistic r_M. As a consequence, the R^2 computed by regressing $\mathbf{D_Y}$ on $\mathbf{D_X}$ has nothing to do with the canonical R^2 obtained by analysing the variation of **Y** with respect to the variation of **X**.

2. An example given by Legendre *et al*. (2005) concerns a group of four sites that have one species in common; in addition, each site harbours one species that is not present in any of the three other sites (Fig. 10.20). This group of sites clearly displays spatial variation in community structure, or beta diversity (Subsection 6.5.3). The total sum of squares of the species data, SS(**Y**), is 3.0; it is positive, as expected for a group of sites showing beta diversity (Box 6.1). However, the sum of squared distances in the upper (or lower) triangular portion of matrix **D**, SS(**D**), is zero. Because SS(**D**) is the denominator of the R^2 in regression on distance matrices and the square root of this R^2 is the Mantel statistic r_M, the variation in the data shown in Fig. 10.20a cannot be analysed by a Mantel test because the Mantel correlation r_M would be indeterminate. This example also shows that SS(**D**) is not a measure of beta diversity.

(a) Data

	Sp.1	Sp.2	Sp.3	Sp.4	Sp.5
Site 1	1	1	0	0	0
Site 2	1	0	1	0	0
Site 3	1	0	0	1	0
Site 4	1	0	0	0	1

(b) **D** = [1 – Jaccard similarity]

	Site 1	Site 2	Site 3	Site 4
Site 1	0	0.667	0.667	0.667
Site 2	0.667	0	0.667	0.667
Site 3	0.667	0.667	0	0.667
Site 4	0.667	0.667	0.667	0

Figure 10.20 Illustrative example. (a) Community composition data table and (b) derived distance matrix, $D_{ij} = (1 - S_{ij})$, based on the Jaccard similarity index (S_7). Redrawn from Legendre *et al*. (2005).

Box 10.3 (*continued*)

3. Numerical simulations involving two variables were carried out by Legendre & Fortin (2010, Table 2) to demonstrate the difference in power between tests of significance of correlation coefficients between two variables and Mantel tests carried out between distance matrices computed from these same variables. A population correlation value was imposed between two vectors of random variables, as in Table 10.5 of Subsection 10.3.3. When the correlation value was 0, all tests (the parametric and permutation test of the correlation coefficient, as well as the Mantel test) had correct levels of type I error, i.e. all tests rejected H_0 at the α level in a proportion of the cases approximately equal to α. When the population correlation was $\rho = 0.5$, the mean of the Pearson correlations computed on samples of $n = 10$ to 100 data (10000 repetitions for each value of n) was approximately 0.5; the mean of the Mantel r_M statistics was near 0.2. Tests of the Pearson correlations increased in power as n increased, from a rejection rate of H_0 of 0.455 for $n = 10$ to 1.000 for $n = 100$; Mantel tests had a rejection rate of 0.279 for $n = 10$ to 0.968 for $n = 100$. When the population correlation was negative ($\rho = -0.5$), the mean of the Pearson correlations was approximately –0.5; the mean of the Mantel r_M statistics was near 0.2; note the positive sign. Powers for the two statistics were the same as when the population correlation was $\rho = 0.5$. To summarize, these simulations showed the following: when it detects a correlation in the original data, the Mantel test may not correctly estimate the sign of the correlation coefficient, and it produces tests with lower power than the test of Pearson's r. Conclusion: the Mantel test is inappropriate to test hypotheses concerning correlations in raw data.

4. Dutilleul *et al*. (2000) described cases where the values of the Mantel statistics were negative whereas the Pearson correlation was strictly 0; their Table 4 also showed cases, for real bivariate data, where the signs of the Mantel statistics varied but were unrelated to the signs of the Pearson correlations. Again, the Mantel test seemed inappropriate to test hypotheses concerning correlations in raw data.

2 — More than two association matrices

Partial Mantel test

Smouse *et al*. (1986) proposed to compute partial correlations involving similarity or distance matrices. Consider distance matrices $\mathbf{D}_1$, $\mathbf{D}_2$, and $\mathbf{D}_3$ computed from three multivariate data tables, using a distance measure appropriate to each case. The partial Mantel statistic, $r_M(\mathbf{D}_1\mathbf{D}_2.\mathbf{D}_3)$, estimating the correlation between matrices $\mathbf{D}_1$ and $\mathbf{D}_2$ while controlling for the effect of $\mathbf{D}_3$, is computed in the same way as a partial correlation coefficient (eq. 4.36), except that the calculation is based here on standardized Mantel statistics r_M (Box 10.2) instead of Pearson correlations r. For symmetric distance matrices, only the upper (or lower) triangular portions are used in the calculations. The tests of significance applicable to partial Mantel statistics (permutation tests) are described in Legendre (2000) and in Appendix 4 of Legendre & Fortin (2010).

Like Mantel tests, partial Mantel tests are only applicable to questions that concern relationships among three distance matrices, not raw data. Another method described in the present section, multiple regression on distance matrices, is applicable to questions involving more than three distance matrices,

Ecological application 10.5

This application analyses the microgeographic morphological differentiation of muskrats (*Ondatra zibethicus*) in the upper basin of River La Houille in southern Belgium. Muskrats were introduced into Bohemia (now part of the Czech Republic) from North America in 1905 for breeding and fur production. In later years, the species was introduced into other European countries, including Belgium, where individuals were released to the wild in 1928 (Le Boulengé, 1972). After their release from breeding farms, muskrats colonized ponds and waterways throughout Europe.

Muskrats were captured during a government-sponsored trapping campaign conducted in 1971-1972 to eradicate rats from the ponds of the upper La Houille basin (approximately 150 km^2) where the river forms a broad, 15 km long loop, before flowing towards the Ardennes Department of France where it becomes a tributary of River Meuse (Fig. 10.21a). Muskrats were captured in nine local population zones, seven of which are included in the part of the study of Le Boulengé *et al*. (1996) reported here. Age and sex of the captured specimens were determined and measurements of the skull were taken. Mahalanobis distances based on 10 age-adjusted skull measurements were computed among the muskrat population zones.

Despite the absence of environmental heterogeneity across the study region, significant skull morphological differences were identified among the local populations by ANOVA and MANOVA. These differences were possibly due to founder effects and/or colonization of the tributaries by animals from different origins, coupled with a spatial pattern of genetic relatedness among the zones. The question addressed by the authors was: how can the relatedness of the populations in the different zones be explained? Are geographically closer populations more similar in their skull morphology? And then, what is it to be "geographically closer" for muskrats, which are semi-aquatic mammals? The populations are genetically interconnected by the migration of the young which, after weaning, may disperse to other population zones. In this study, the relationships to be tested clearly concerned distances (morphological and geographic), so Mantel tests were appropriate.

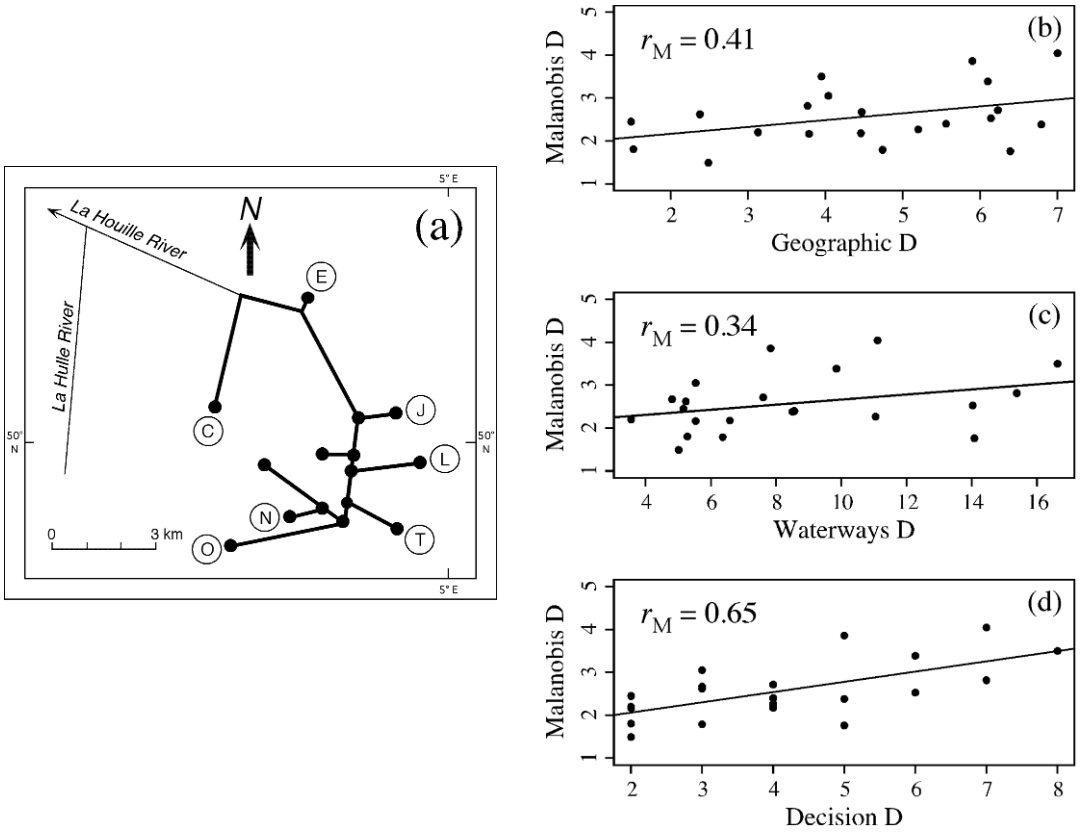


Figure 10.21 (a) Schematic representation of the upper La Houille River network in Belgium, showing the seven muskrat local population zones, identified by letters C to T. (b) to (d) Distance comparison diagrams and simple Mantel statistics (r_M) for three types of geographic separation distances. An OLS regression line indicates the trend in each graph.

The authors measured the geographic distance between the zones in two ways: straight-line geographic distance and distance along the waterways. Muskrats are fast-moving animals when they travel, so perhaps the actual distance is not really the determining factor. For that reason, the authors also devised a "Decision distance", which is the number of furcations of the river network separating two zones. At these points, a migrating muskrat must decide to go either left or right along the river network. The relationship between the three types of geographic separation distances and the morphometric (Mahalanobis) distances are shown in Fig. 10.21b-d, where simple Mantel correlations (r_M) are shown. The graphs also show that the different types of geographic distances are linearly related to the morphometric distances, so that the Mantel statistic based on the Pearson correlation coefficient was appropriate in this study.

The authors formulated a sociobiological hypothesis called "isolation by distance along corridors". This hypothesis states that the decision distances explain the morphometric differentiation of the local populations best. Partial Mantel tests were computed to compare the ability of different distances to explain the morphological distances. Comparisons of the geographic and decision distances produced the following results:

r_M(Mahalanobis Geographic.Decision) = 0.03, p = 0.455

r_M(Mahalanobis Decision.Geographic) = 0.55, p = 0.010

showing that the decision distance matrix explained the morphometric distances significantly better than the geographic distances. Likewise, the decision distance matrix explained the morphometric distances significantly better than the waterways distances; there was no significant difference in explanation of the morphometric distances between the geographic and waterways distances (results computed from the distance matrices found in the paper). Hence the results were consistent with the hypothesis of "isolation by distance along corridors".

Causal modelling

Partial Mantel tests are not always easy to interpret. Legendre & Troussellier (1988) have shown the consequences of all possible three-matrix causal models on the significance of Mantel and partial Mantel statistics. The models (and their predictions) are the same as those illustrated in Fig. 4.11 for three simple variables. This approach leads to a form of *causal modelling on resemblance matrices* (Legendre, 1993). It should only be used to analyse questions that require the modelling of distances, not raw data.

In ecology, this type of analysis has been used mostly to study the distribution of organisms (matrix $\mathbf{D}_1$) with respect to environmental variables (matrix $\mathbf{D}_2$) while considering the spatial locations of the sampling sites (matrix $\mathbf{D}_3$). We now know that in all these applications, including Legendre & Trousselllier (1988), the Mantel test should be replaced by canonical analysis (Chapter 11), which provides much more powerful tests of significance (see Box 10.4).

Regression on distance matrices

Permutation test

One may also want to model the variation in a first distance matrix as a function of the variation in other distance matrices about the same objects. *Multiple regression on resemblance matrices* has been suggested by several authors (Hubert & Golledge, 1981; Smouse *et al*., 1986; Manly, 1986; Krackhardt, 1988) to address research questions formulated in terms of distances. Legendre *et al*. (1994) described appropriate testing procedures for evolutionary studies where the response data was a dendrogram or an evolutionary tree. The parameters of the multiple regression model are obtained using a procedure similar to that of the Mantel test (Fig. 10.22). The response distance matrix $\mathbf{D_Y}$, which represents the evolutionary tree, is unfolded into a vector **y**; likewise, each explanatory distance matrix $\mathbf{D_X}$ is unfolded into a vector **x**. A multiple regression is computed in which **y** is a function of vectors $\mathbf{x}_j$. The parameters of that regression (the coefficient of multiple determination R^2_j and the partial regression coefficients) are tested by permutations, as follows. When the response distance matrix $\mathbf{D_Y}$ is an ordinary distance or similarity matrix, the permutations of the corresponding vector **y** are carried out in the way of the Mantel permutational test (Subsection 10.5.1). When it is an ultrametric matrix representing a dendrogram (Subsection 8.3.1), the double-permutation method of Lapointe and Legendre (1990,

Further developments, partial Mantel test — Box 10.4

During the past 15 years, partial Mantel tests and regression on distance matrices have been used in many ecological papers that had for objective to analyse the spatial variation of community composition (raw data, not distances) among sites, i.e. beta diversity (Subsection 6.5.3). Some of these papers were listed as examples by Legendre *et al*. (2005). To demonstrate that the Mantel test should not be used for that type of objective, that paper presented simulation results involving multivariate, spatially correlated data. The simulations compared canonical analysis (RDA, Section 11.1) to Mantel tests to detect the effect of environmental variables **X** on species-like response data **Y**, as well as the presence of spatial structures in the species-like data (10 simulated species, $n = 100$). The results found in Table 1 and Fig. 3 of Legendre *et al*. (2005) showed the following:

• The two testing methods had correct levels of type I error. They were thus statistically valid.

• When **Y** was related to the environmental variables **X** (plus random error in **Y**), RDA detected a significant relationship in 97% of the simulations whereas the Mantel test detected it in 49% of the cases.

• Using the distance-based Moran's eigenvector map method of spatial analysis (dbMEM, Section 14.1) in RDA, significant spatial structures were detected in the simulated data in 99% of the cases, compared to 8 to 22% of the cases detected by Mantel tests.

These findings support the conclusion that the Mantel test is inappropriate to test hypotheses concerning correlations in raw data. Other simulation results, where community composition data were simulated according to Hubbell's (2001) neutral model, led to the same conclusions about the difference in power between the two types of tests when applied to raw data (Legendre *et al*., 2008).

Not everyone agrees about the questions that can be answered by Mantel tests. See the controversy raised by Tuomisto & Ruokolainen (2006) and the exchanges that followed in the ecological literature (Pélissier *et al*., 2008; Laliberté, 2008; Legendre *et al*., 2008; Tuomisto & Ruokolainen, 2008). Everyone now seems to agree, however, that Mantel tests should be limited to questions about relationships between distance matrices.

1991) is used. When it is a path-length matrix representing an additive tree (i.e. a cladogram in phylogenetic studies), a triple-permutation method (Lapointe and Legendre, 1992a) is used. Vectors $\mathbf{x}_j$ representing the explanatory matrices are kept fixed with respect to one another during the permutations. Selection of explanatory matrices may be done by forward selection, backward elimination, or a stepwise procedure, which are described in Legendre *et al*. (1994). For research questions that do not strictly concern distances, the method of multiple factor analysis (MFA, briefly described at the end of Subsection 11.5.1) should be used for analysis.

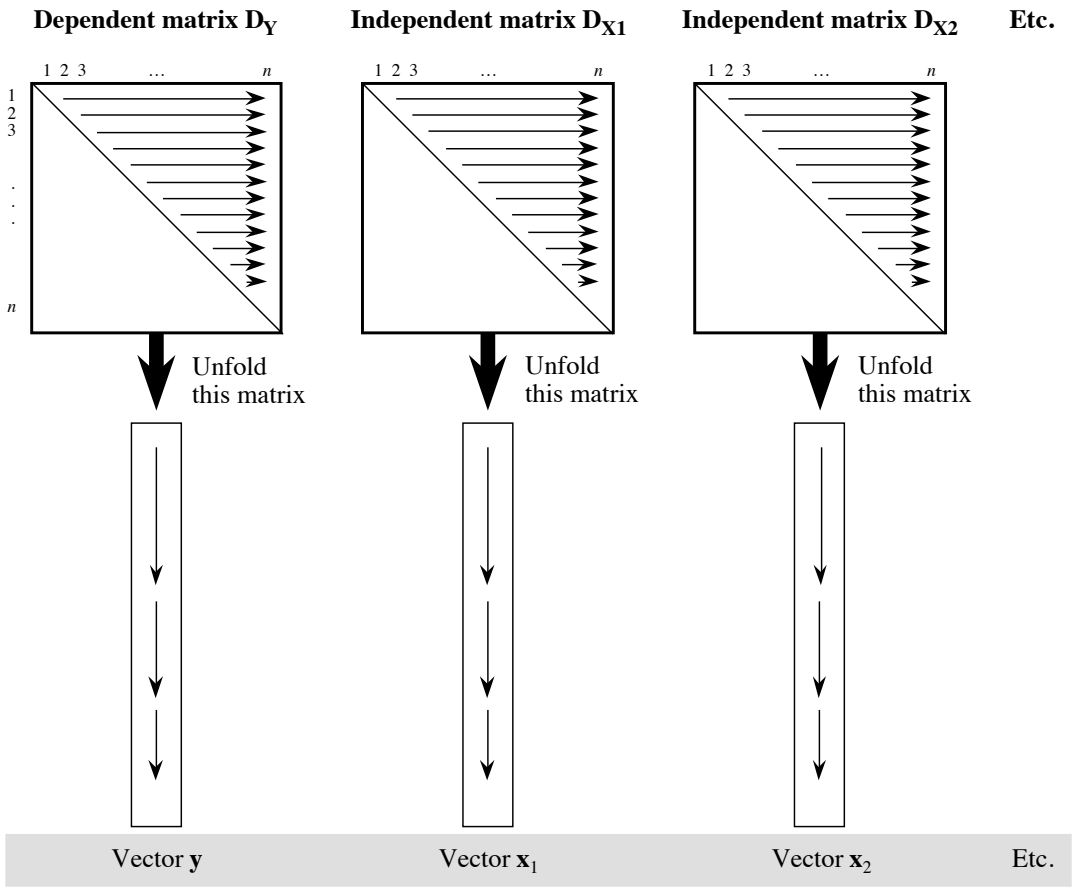


Figure 10.22 Multiple regression is computed on the vectors resulting from unfolding matrices $\mathbf{D_Y}$ (response) and $\mathbf{D_{X1}}$, $\mathbf{D_{X2}}$, etc. (explanatory).

The CADM method to test the congruence among distance matrices, described in Subsection 5.4.3, is another extension of the Mantel test to several distance matrices.

3 — ANOSIM test

Focusing on problems of analysis of variance that involved community composition data, Clarke (1988, 1993) developed a parallel approach to the goodness-of-fit Mantel tests. Clarke's method, called ANOSIM (*ANalysis Of SIMilarities*), is implemented in the PRIMER package, referred to in Section 9.4, and in R. In PRIMER, program ANOSIM includes one-way and two-way analyses (crossed or nested) for replicated data, whereas program ANOSIM2 covers two-way analyses without replication (Clarke &

(a) **X** = ranked distances

D	Group 1 5	Group 1 6	Group 2 7	Group 2 8	Group 2 9
5					
6	2				
7	4	8.5			
8	6.5	5	1		
9	8.5	10	3	6.5	

(b) **Y** = model matrix

D	Group 1 5	Group 1 6	Group 2 7	Group 2 8	Group 2 9
5					
6	–1/20				
7	1/30	1/30			
8	1/30	1/30	–1/20		
9	1/30	1/30	–1/20	–1/20	

Figure 10.23 (a) Distances from the numerical example in Fig. 12.22a are transformed into ranks, the most similar pair receiving rank 1. (b) Weighting required to compute the ANOSIM statistic as a Mantel statistic.

Warwick, 1994). After a brief presentation of Clarke's statistic, below, the similarities and differences between the ANOSIM and Mantel approaches will be shown.

Consider the situation illustrated in Fig. 10.23a. The distances shown in Fig. 12.22a were transformed into ranks, the least dissimilar pair (i.e. the most similar) receiving rank 1. Tied values in Fig. 12.22a were given mean rank values, as usual in nonparametric statistics. Objects are arbitrarily numbered 5, 6, 7, 8, 9. The objects are assumed to form two groups, defined here on *a priori* bases; the two groups are not supposed to result from clustering as in Fig. 12.22a. The two *a priori* groups are (5, 6) and (7, 8, 9). The null hypothesis is of the ANOVA type:

H_0: There is no difference between the two (or more) groups.

In Fig. 10.23a, does one find the kind of variation among distance values that one might expect if the data corresponded to the null hypothesis? Clarke (1988, 1993) proposed the following statistic to assess the differences among groups:

$$R = \frac{\bar{r}_B - \bar{r}_W}{n\,(n-1)\,/\,4} \qquad \textbf{(10.31)}$$

where $\bar{r}_B$ is the mean of the ranks in the *between*-group submatrix (i.e. in Fig. 10.23a, the rectangle crossing groups 1 and 2), $\bar{r}_W$ is the mean of the ranks in all *within*-group submatrices (i.e. the two triangles in the figure), and n is the total number of objects. In the present example, $\bar{r}_B = 7.083$ and $\bar{r}_W = 3.125$, so that $R = 0.79167$ (eq. 10.31).

Using ranks instead of the original distances is not a fundamental requirement of the method. It comes from a (reasonable) recommendation, by Clarke and co-authors, that the test statistic should reflect the patterns formed among sites represented by

Permutation test

multidimensional scaling plots (nMDS, Section 9.4), which preserve rank-transformations of distances. The R statistic is tested by permutations of the objects, as explained in Box 10.2. The denominator of eq. 10.31 is chosen in such a way that $R = 1$ if all the lowest ranks are in the "within-group" submatrices, and $R = 0$ if the high and low ranks are perfectly mixed across the "within" and "between" submatrices. R is unlikely to be substantially smaller than 0; this would indicate that the similarities within groups are systematically lower than among groups.

Clarke (1988, 1993) actually applied the method to the analysis of several groups. This is also the case in the nonparametric ANOVA-like example of the Mantel test in the Sokal & Rohlf (1995) book. The statistic (eq. 10.31) can readily handle the more-than-two-group case: $\bar{r}_B$ is then the mean of the ranks in *all* between-group submatrices, whereas $\bar{r}_W$ is the mean of the ranks in *all* within-group submatrices.

Equation 10.31 may be reformulated as a Mantel cross-product statistic z_M (Box 10.2). To achieve this, define a model matrix containing positive constants in the "between-group" portion and negative constants in the "within-group" parts:

- the "between" values (shaded area in Fig. 10.23b) are chosen to be the inverse of the number of between-group distances (1/6 in this example), divided by the denominator of eq. 10.31, i.e. $[n(n-1)/4]$ (which is 5 in the present example);

- similarly, the "within" values in Fig. 10.23b are chosen to be the inverse, with negative signs, of the number of distances in all within-group submatrices (–1/4 in the example), also divided by $[n(n-1)/4]$ (= 5 in the present example).

The coding is such that the sum of values in the half-matrix is zero. The unstandardized Mantel statistic (Box 10.2), computed between matrices $\mathbf{D_X}$ and $\mathbf{D_Y}$ of Fig. 10.23, is $z_M = 0.79167$. This result is identical to Clarke's ANOSIM statistic.

Since the permutation method is the same in the Mantel and ANOSIM procedures, the tests should produce similar p-values. They may differ slightly in practice because different programs, and even different runs of the same program, may produce different sequences of permutations of the objects. As shown in Subsection 10.5.1, *any binary coding* of the "within" and "between" submatrices of the model matrix leads to the same probabilities. Of course, interchanging the small and large values produces a change of sign of the statistic and turns an upper-tail test into a lower-tail test. The only substantial difference between the Mantel goodness-of-fit and ANOSIM tests is one of tradition: Clarke (1988, 1993) and the ANOSIM function in the PRIMER package (Clarke & Warwick, 1994) and in R (Section 10.7) transform the distances into ranks before computing eq. 10.31. Since Clarke's R is equivalent to a Mantel statistic computed on ranked distances, it is thus analogous to a Spearman correlation coefficient (eqs. 5.1 and 5.3).

The Mann-Whitney U statistic could also be used for analysis-of-variance-like tests of significance performed on distance matrices. This has been suggested by

Gordon (1994) in a different context, i.e. as a way of measuring the differentiation of clusters produced by clustering procedures (internal validation criterion), as reported in Section 8.13. In Gordon's method, distances are divided in two subsets, i.e. the within-group (W) and between-group (B) distances — just like in Clarke's method. A U statistic is computed between the two subsets. U is closely related to the Spearman rank correlation coefficient (eqs. 5.1 and 5.3); a U test of a variable against a dummy variable representing a classification in two groups is equivalent to a Spearman correlation test (same probability). Since Clarke's statistic is also equivalent to a Spearman correlation coefficient, the Mann-Whitney U statistic should lead to the exact same probability as the Clarke or Mantel statistics, if U was used as the statistic in a Mantel-like permutation test. [Using the U statistic as an internal validation criterion, as proposed by Gordon (1994), is different. On the one hand, the grouping of data into clusters is obtained from the distance matrix that is also used for testing; this is not authorized in an analysis-of-variance approach. On the other hand, Gordon's Monte Carlo testing procedure differs from the Mantel permutation test.]

4 — Procrustes test

In Greek mythology, Procrustes was a son of Poseidon and a rogue. He invited travellers to spend the night with him, then tied them down to an iron bed and either cut off their limbs if they were taller than the bed, or stretched the victims if they were too short, till they fitted in.

Procrustes analysis, proposed by Gower (1971b, 1975, 1987), is primarily a canonical ordination method; it is described in Subsection 11.5.2. The Procrustes test (PROTEST) is presented here as a statistical method for comparing two rectangular data matrices about the same objects. It is appropriate to answer questions about the relationship between the original data sets (i.e. raw data), which is not the case of the Mantel test. Another statistic that can be used in the same situation is the RV coefficient (eqs. 11.65 and 11.66) described with co-inertia analysis (Subsection 11.5.1).

The purpose of Procrustes analysis is to find a compromise ordination for two raw data matrices with the same objects in rows, using a rotational-fit algorithm that minimizes the sum of squared distances between corresponding points of the two matrices in a joint ordination. In that ordination, each object has two representations, one from each matrix, so that the scatter diagram allows one to visualize the differences between the two original matrices. In *orthogonal Procrustes*, two matrices are considered and fitted using rigid-body motions (translation, rotation, and mirror reflection). *Generalized Procrustes analysis* is the extension of the method to more than two matrices. Details are found in the references given above.

The present subsection focuses on the residual sum-of-squares statistic of orthogonal Procrustes analysis, which is a goodness-of-fit statistic. It was called m^2 by Gower and is computed as follows:

Procrustes statistic m^2

$$m_{12}^2 = \text{Trace}(\mathbf{Y}_1\mathbf{Y}_1') - \frac{(\text{Trace}\mathbf{W})^2}{\text{Trace}(\mathbf{Y}_2\mathbf{Y}_2')} \quad \textbf{(10.32)}$$

where $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are the two rectangular matrices of raw data to be analysed, with column vectors centred on their respective means, and $\mathbf{W}$ is a diagonal matrix of singular values found by the singular value decomposition $\mathbf{Y}_1'\mathbf{Y}_2 = \mathbf{VWU}'$ (SVD, eq. 2.31).

Equation 10.32 is not symmetric; indeed, the m_{12}^2 value resulting from fitting $\mathbf{Y}_2$ to $\mathbf{Y}_1$ differs from m_{21}^2 obtained by fitting $\mathbf{Y}_1$ to $\mathbf{Y}_2$. To solve that problem, transform the column-centred matrix $\mathbf{Y}_1$ to $\mathbf{Y}_{1.\text{tr}}$ by dividing each value of $\mathbf{Y}_1$ by the square root of the trace of $\mathbf{Y}_1\mathbf{Y}'_1$, which is the same as the trace of $\dot{\mathbf{Y}}_1'\mathbf{Y}_1$; that trace is easily computed as the sum of squares of all values in $\mathbf{Y}_1$. Using the same method, transform the centred matrix $\mathbf{Y}_2$ to $\mathbf{Y}_{2.\text{tr}}$. For $\mathbf{Y}_{1.\text{tr}}$ and $\mathbf{Y}_{2.\text{tr}}$, the two Procrustes statistics are now identical:

$$m_{12}^2 = m_{21}^2 = 1 - (\text{Trace}\mathbf{W})^2 \quad \textbf{(10.33)}$$

Jackson (1995) suggested using the symmetric orthogonal Procrustes statistic m_{12}^2 (eq. 10.33) as a measure of concordance, or similarity, between two data matrices representing, in particular, species abundances and environmental variables. The statistic is tested by permutation. Jackson (1995) called this procedure the *Procrustean randomization test* (PROTEST). He provided examples of applications to ecological data: benthic invertebrates, lake morphometry, lake water chemistry, and geographic coordinates, for 19 lakes in Ontario, Canada. What Jackson actually compared in that paper were, for each data set, the first two ordination axes from correspondence analysis (CA, for benthic invertebrates) or principal component analysis (PCA, for lake morphometry and chemistry); the geographic coordinates were left untransformed. The PROTEST method, as re-described by Peres-Neto and Jackson (2001), can actually be used to test the significance of the relationship between data matrices in all situations where co-inertia and orthogonal Procrustes analyses are applicable (Section 11.5). In R, function ***protest()*** of VEGAN uses Trace$\mathbf{W}$ instead of m_{12}^2 or m_{21}^2 as the test statistic in the Procrustean permutation test.

PROTEST

Numerical simulations carried out by Peres-Neto & Jackson (2001) showed that PROTEST was more powerful than the Mantel test to identify correlations generated between raw data matrices. This finding is in accordance with the conclusions of other authors, reported in Box 10.3, that the Mantel test should not be used to test hypotheses concerning correlations between raw data matrices.

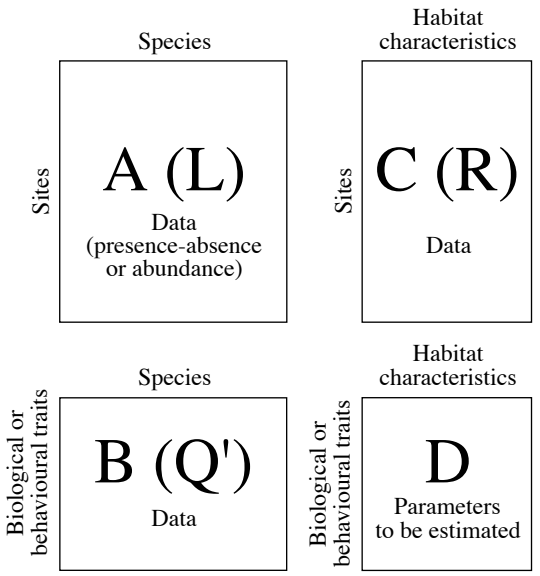


Figure 10.24 Given the information in matrices **A**, **B**, and **C**, the fourth-corner problem is to estimate the parameters in the fourth-corner matrix **D** that crosses the habitat characteristics with the biological or behavioural traits of the species. In Dray & Legendre (2008), matrix **A** is called **L**, **B** is called **Q'**, and **C** is called **R**.

10.6 The fourth-corner problem

How do the biological and behavioural characteristics of species determine the niches they occupy or their geographic locations in an ecosystem?

This question, which stems from niche theory, has long been neglected by ecologists because they lacked an appropriate method of analysis. Observation of species in nature helps ecologists formulate hypotheses in that respect. Testing such hypotheses requires (1) a way of detecting relationships between species traits and habitat characteristics, and (2) of testing the significance of these relationships. A method of analysis for this problem was proposed by Legendre *et al*. (1997a) and the statistical theory was completed by Dray & Legendre (2008) and by ter Braak *et al*. (2012).

Consider a matrix **A** ($n \times p$) containing data on the presence-absence or abundance of p species at n sites (Fig. 10.24)[*]. A second matrix **B** ($q \times p$) describes q biological or behavioural traits of the same p species. A third matrix **C** ($n \times m$) contains information about m habitat characteristics (environmental variables) at the n sites. How does one

[*] Matrices **A** to **D** are transposed compared to the presentation in Legendre *et al*. (1997).

go about associating the q biological and behavioural traits to the m habitat characteristics? To help find a solution, let us translate the problem into matrix form:

$$\begin{bmatrix} \mathbf{A}_{(n \times p)} & \mathbf{C}_{(n \times m)} \\ \mathbf{B}_{(q \times p)} & \mathbf{D}_{(q \times m)} \end{bmatrix} \qquad \textbf{(10.34)}$$

Using this representation, the problem may now be stated as follows:

• How does one go about estimating the parameters in matrix **D** ($q \times m$) where the q biological and behavioural traits are related to the m habitat characteristics?

• Are these parameters significant in some sense, i.e. are they different from 0 (no relationship) or from the value they could take in a randomly organized environment?

The statistical problem of estimating the parameters in matrix **D** is referred to as the *fourth-corner problem* because matrix **D** lies in the fourth corner of the matrix arrangement shown in eq. 10.34. Data in matrix **A** belong to the presence/absence or abundance types (only presence-absence data were considered by Legendre *et al*., 1997a). Matrices **B** and **C** may contain quantitative or qualitative (nominal) data. The papers referenced at the beginning of the section describe solutions to accommodate the different types of variables. The relationship between **B** and **C** mediated by **A** can also be analysed by a related method called RLQ analysis (Dolédec *et al*., 1996).

1 — Comparing two qualitative variables

The first situation considered here concerns two qualitative variables, one from matrix **B** (behaviour), the other from matrix **C** (habitat). Any qualitative variable can be expanded into a series of binary variables, one for each state (Subsection 1.5.7).

Numerical example. In test cases 1 and 2 (Table 10.8), **A** is a matrix of presence-absence of species at two sites; **B** and **C** contain supplementary variables (qualitative, two states) for the rows and columns of **A**, respectively. To fix ideas, let us assume that the variable in **B** describes two feeding habits (herbivorous, carnivorous) and **C** is the nature of the substrate at the study sites on a coral reef (live coral, turf). This example describes the approach for qualitative variables (Subsection 10.6.1) and introduces the method for significance testing (Subsection 10.6.2).

Matrices **A**, **B** and **C** (or **L**, **Q** and **R**) are all needed to estimate the parameters in the fourth-corner matrix **D**. The three matrices can be combined by multiplication around the set of four matrices while preserving matrix compatibility:

clockwise: $\mathbf{D} = \mathbf{B}\,\mathbf{A}'\,\mathbf{C}$ or $\mathbf{D} = \mathbf{Q}'\,\mathbf{L}'\,\mathbf{R}$ **(10.35)**

or counter-clockwise: $\mathbf{D}' = \mathbf{C}'\,\mathbf{A}\,\mathbf{B}'$ or $\mathbf{D}' = \mathbf{R}'\,\mathbf{L}\,\mathbf{Q}$ **(10.36)**

For the two test cases of the numerical example, matrix **D** is shown in Table 10.8. Equations 10.35 and 10.36 have an equivalent in traditional statistics. If the data in **A**,

Table 10.8 Test cases for qualitative variables. Matrices are transposed to reduce their widths in the page. **A'** is (10 species × 2 sites), **B'** is (10 species × 2 feeding habits), and **C'** is (2 habitat types × 2 sites). So, **D'** is (2 habitat types × 2 feeding habits). Probabilities (p) are one-tailed, assuming that H_1 states the sign of the relationship. H_1 is indicated by a sign in each cell of **D'**, + meaning that the actual value is larger than the expected value and is tested in the upper tail, and – in the opposite case. Probabilities computed after 9999 permutations. *E* = exact probabilities; see text.

Test case 1

A':	Site 1	Site 2
Sp. 1	1	0
Sp. 2	0	1
Sp. 3	1	0
Sp. 4	1	0
Sp. 5	1	0
Sp. 6	0	1
Sp. 7	0	1
Sp. 8	0	1
Sp. 9	0	1
Sp. 10	0	1

B':	Herbiv.	Carniv.
	0	1
	0	1
	0	1
	0	1
	0	1
	1	0
	1	0
	1	0
	1	0
	1	0

C':	Site 1	Site 2
Live coral	1	0
Turf	0	1

D':	Herbiv.	Carniv.
Live coral	0 – p = 0.029 *E* = 0.031	4 + p = 0.189 *E* = 0.188
Turf	5 + p = 0.029 *E* = 0.031	1 – p = 0.189 *E* = 0.188

Contingency statistic:

G = 8.4562, p (9999 permutations) = 0.021

Test case 2

A':	Site 1	Site 2
Sp. 1	1	1
Sp. 2	1	1
Sp. 3	1	1
Sp. 4	1	1
Sp. 5	1	1
Sp. 6	1	1
Sp. 7	1	1
Sp. 8	1	1
Sp. 9	1	1
Sp. 10	1	1

B':	Herbiv.	Carniv.
	0	1
	0	1
	0	1
	0	1
	0	1
	1	0
	1	0
	1	0
	1	0
	1	0

C':	Site 1	Site 2
Live coral	1	0
Turf	0	1

D':	Herbiv.	Carniv.
Live coral	5 p = 1.000 *E* = 1.000	5 p = 1.000 *E* = 1.000
Turf	5 p = 1.000 *E* = 1.000	5 p = 1.000 *E* = 1.000

Contingency statistic:

G = 0.0000, p (9999 permutations) = 1.000

Inflated data matrix

B, and **C** are frequencies, they can be combined to form an "inflated data matrix". Matrix **D**, which results from crossing the two columns of the inflated matrix, is a contingency table as shown in Table 10.9; values d_{ij} in matrix **D** are frequencies or pseudo-frequencies (see Ecological application 10.6). So, a solution that naturally comes to mind for significance testing is to compute a χ^2 statistic, using either Pearson's (eq. 6.5) or Wilks' formula (eq. 6.6, also called the *G* statistic). The *G* statistic is used here; it is the first type of fourth-corner statistic.

Table 10.9 Inflated data matrix (left); there is one row in this matrix for each species "presence" (value 1) in matrix **A'** of test case 1 (Table 10.8). The contingency table (matrix **D'**, right) is constructed from the inflated matrix.

Inflated data matrix

Occurrences in test case 1	Feeding habits from **B**	Habitat types from **C**
Sp. 1 @ Site 1	Carnivorous	Live coral
Sp. 2 @ Site 2	Carnivorous	Turf
Sp. 3 @ Site 1	Carnivorous	Live coral
Sp. 4 @ Site 1	Carnivorous	Live coral
Sp. 5 @ Site 1	Carnivorous	Live coral
Sp. 6 @ Site 2	Herbivorous	Turf
Sp. 7 @ Site 2	Herbivorous	Turf
Sp. 8 @ Site 2	Herbivorous	Turf
Sp. 9 @ Site 2	Herbivorous	Turf
Sp. 10 @ Site 2	Herbivorous	Turf

Contingency table

D':	Herbivorous	Carnivorous
Live coral	0	4
Turf	5	1

Dray & Legendre (2008) have shown that species abundance data can be used as well as presence-absence data in the calculation of fourth-corner statistics and in the permutation tests described in the next two subsections.

For large contingency tables **D**, relationships among descriptor states could be visualized in a correspondence analysis (CA) biplot (Subsection 9.2.1). Consider matrix **D** shown in Table 10.10 (below) as an example. It may be simplified as follows before CA: for the cells where d_{ij} is significant, code those that are above the expected value (sign + in the matrix) with +1 and those that are below the expected value (sign – in the matrix) with –1. Code the non-significant cells with 0. After coding, add 1 to all cells because CA requires that the values in the matrix subjected to the analysis be non-negative. Carry out CA of the coded matrix and use scaling type 4 for the biplot. A CA biplot remains a simplified summary; it contains less precise information than the original matrix **D**.

2 — Test of statistical significance

Permutation test

In fourth-corner problems, one cannot test the G statistics in the usual manner because, in the general case (although not in test case 1 of Table 10.8), several species are observed at any one sampling site so that the rows of the inflated matrix are not independent of one another; several rows of that matrix result from observations at a single site. To solve the problem, G is tested by permutations (Section 1.2). The procedure is as follows.

Hypotheses

• H_0: the species traits (matrix **B**) are unrelated to the characteristics of the sites (matrix **C**), their relationships (links) being mediated by the species presence-absence or abundance data (matrix **A**). Different permutation null models are detailed in the next subsection.

• H_1: the species traits are related to the characteristics of the sites.

Test statistic

Compute a χ^2 statistic (*G* here) on the contingency table (matrix **D**) and use it as the reference value for the remainder of the test.

Distribution of the test statistic

Under H_0, the species found at any one site could have been observed at any other site. Where the species have actually been observed is due to chance alone. So, a realization of H_0 is obtained by permuting at random the values in matrix **A**, using one of the methods described in the next subsection. After each permutation of matrix **A**, recompute the χ^2 statistic on **D**.

• Repeat the permutation a large number of times (say, 999 or 9999 times). The different permutations produce a set of values of the χ^2 statistic, obtained under H_0.

• Add to this set the reference value of the statistic, computed for the unpermuted data matrix. Together, the unpermuted and permuted values (for a total of 1000 values, 10000 values, etc.) form an estimate of the sampling distribution of χ^2 under H_0.

Statistical decision

As in any other statistical test, the decision is made by comparing the reference value of the χ^2 statistic to the distribution obtained under H_0. If the reference value of χ^2 is one likely to have been obtained under the null hypothesis, H_0 is not rejected. If it is too extreme (i.e. located out in a tail) to be considered a likely result under H_0, then H_0 is rejected.

Individual values d_{ij} in matrix **D** can also be tested for significance, as shown below in the numerical example and the ecological application.

In addition, a global test of significance can be carried out for the fourth-corner relationship involving all variables in matrices **A** and **C** in the analysis. The global test uses statistic S_{RLQ}, which is the trace of a cross-product matrix computed from the fourth-corner matrix **D**. See Dray & Legendre (2008, eq. 8). This quantity is equal to the total inertia of an RLQ analysis (Dolédec *et al.*, 1996).

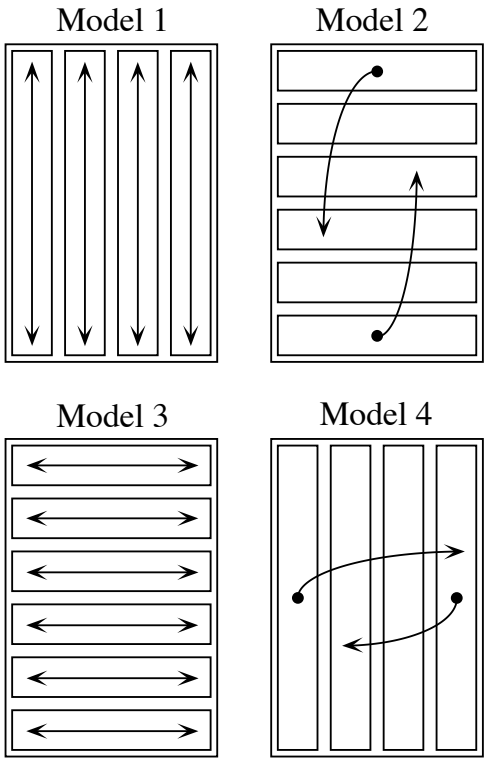


Figure 10.25 Permutations of matrix **A** may be performed in different ways which correspond to different null ecological models.

(1) The occurrence of a species in the study area is constant, but positions are random; permute at random within columns.

(2) Positions of species assemblages are random; permute whole rows (assemblages).

(3) Lottery hypothesis: the species that arrived first occupied a site; permute at random within rows.

(4) Species have random attributes; permute whole columns.

3 — Permutational models

Permutations may be conducted in different ways, depending on the ecological hypotheses to be tested against observations. Technically, the fourth-corner statistical method can accommodate any of the permutation models described below. The random component is clearly the field information about the species found at the sampling sites, i.e. matrix **A**. It is thus matrix **A** that should be permuted (randomized) for the purpose of hypothesis testing. This may be done in various ways (Fig. 10.25). Models 1 to 4 were described by Legendre *et al*. (1997a), model 5 by Dray & Legendre (2008).

Model 1: Environmental control over individual species. — Permute the species presence-absence or abundance data within each column of matrix **A**, independently from column to column. This not only destroys the link between **A** and **C**, but also the relationship between **A** and **B**, as shown by Dray & Legendre (2008, Appendix A). The null hypothesis (H_0) states that individuals of a species are randomly distributed with respect to the site characteristics. The corresponding alternative hypothesis (H_1) states that individuals of a species are distributed according to their preferences for site conditions. Under this permutation model, the number of sites occupied by each

species is kept constant, as in permutation method 2b of the Raup & Crick similarity coefficient (S_{27}, Chapter 7).

Model 2: Environmental control over species assemblages. — Permute entire rows of matrix **A** at random. This method destroys the link between **A** and **C** but keeps **A** linked to **B**; it is equivalent to permuting the rows of **C**. H_0 states that species assemblages are randomly attributed to sites, irrespective of the site characteristics. The corresponding alternative hypothesis (H_1) states that species assemblages are dependent upon the physical characteristics of the locations where they are found. This method preserves the covariances among the species throughout the permutations, as well as the number of sites occupied by each species.

Model 3: Lottery. — Permute the species data within each row of matrix **A**, independently from row to row. This not only destroys the link between **A** and **B**, but also the relationship between **A** and **C**. H_0 states that the distribution of the presences of various species at a site is the result of a random allocation process (the lottery for space model advocated by Sale, 1978); it is not due to the adaptation of the species traits to the sites. The alternative hypothesis (H_1) states that due to their traits, species have some competitive advantages over chance settlers in the habitats where they are found. Under this model, the number of species present in a given site (i.e. species richness) is kept constant.

Model 4: Random species attributes. — Permute entire columns of matrix **A** at random. This destroys the link between **A** and **B** but keeps **A** linked to **C**; it is equivalent to permuting the columns of **B**. H_0 states that species are distributed according to their preferences for site conditions, but irrespective of their traits or other characteristics included in **B**. The alternative hypothesis (H_1) states that the distributions of the species among the sites, which are related to their preferences for site conditions, depend on the adaptations (traits) of the species. Under this model, the number of species present at each site (i.e. the species richness) is kept constant.

Model 5: Permute rows and columns. — Permute entire rows of **A** at random, then (or before) permute entire columns at random. The links between **A** and **B** and between **A** and **C** are destroyed. An alternative, equivalent method is to permute at random the rows of **C** and the columns of **B** while keeping **A** fixed, which was the method used by Dolédec *et al*. (1996). H_0 states that the species distributions among the sites are not related to the site conditions nor to the traits of the species. The alternative hypothesis (H_1) states that the species distributions across the sites are related to species traits, and/or that species assemblages are dependent upon the environmental conditions.

The type I error rate and power of these permutation models were studied by Dray & Legendre (2008) under six data generation scenarios. All permutation models have nearly equal power to detect a relationship between **B** and **C** mediated by **A** when such a relationship is present in the data. In the opposite situation, when there is no relationship between **A** and **B** nor between **A** and **C**, all permutation models have correct rates of type I error and reject H_0 at the α significance level. However, in some

simulation situations, the permutation models differ in type I error rates, which can be very high with some models (Dray & Legendre, 2008). This makes it difficult, with some models, to interpret a rejection of H_0: does it mean that there is a relationship between **B** and **C** via **A**, or is it a type I error?

- When one can assume that a relationship exists between **A** and **B** (i.e. the species have fixed trait values) and the test concerns the relationship between **A** and **C**, (i.e. test the species-environment relationship), permutation model 2 has a correct type I error rate; so this permutation model can be used in that situation.

- Similarly, in the opposite situation, when one can assume that a relationship exists between **A** and **C** and the test concerns the relationship between **A** and **B**, permutation model 4 can be used.

- When no *a priori* assumption can be made about existing relationships (the data in **B** and **C** are considered random instead of fixed), the best strategy is to carry out two tests in sequence using permutation models 2 and 4 and take the maximum of the two p-values as the probability of the data under the combined null hypothesis (ter Braak *et al*., 2012).

Numerical example. Let us examine how the fourth-corner method behaves when applied to the data sets introduced in the numerical example of Subsection 10.6.1. The first test case (Table 10.8, left) was constructed to suggest that herbivores are found on turf while carnivores are more ubiquitously distributed. Globally, the G statistic indicates a significant relationship ($\alpha = 0.05$) between behavioural states and types of habitat ($p = 0.0207$ after 9999 random permutations under model 1 above). The expected values in the various cells of matrix **D** determine the tail in which each frequency d_{ij} of the contingency table is to be tested for significance; this value is taken to be the mean frequency expected from all possible permutations of matrix **A**, given the permutation model that has been selected. Looking at individual values d_{ij}, herbivores are clearly positively associated with turf and negatively with coral ($p = 0.0287$, computed from the random permutation results), while carnivores are not significantly associated with either live coral or turf ($p = 0.1890$). These probabilities are very close to the exact probabilities calculated for the same data, which are the values obtained from a complete permutation procedure (E in Table 10.8). Values of exact probabilities E are computed as follows: consider all possible permutations that result from independently permuting the rows of matrix **A** (permutation model 1); count how many of these would produce values equal to, or more extreme than the observed value in each given cell of matrix **D**. This value may differ slightly from the random permutational probability. Globally, the testing procedure for the relationship between behaviour and habitat behaved as expected in this example, and the random permutation procedure produced values quite close to the exact probabilities.

The second test case (Table 10.8, right) illustrates a situation where the null hypothesis is true in all cases, matrix **A** indicating all 10 species to be present everywhere. Indeed, the testing procedure finds all permutation statistics to be equal to the unpermuted ones, so that the probability of the data under the null hypothesis is 1 everywhere. The procedure once more behaved correctly.

4 — Other types of comparisons among variables

Variables in matrices **B** and **C** are not always qualitative. Through lines of reasoning similar to that of Subsection 10.6.1, involving inflated data matrices (as in Table 10.9), fourth-corner statistics can be formulated to accommodate other types of variable comparisons.

• To compare a quantitative variable in **B** to a quantitative variable in **C**, a Pearson correlation coefficient may be computed between the columns of the inflated matrix. A correlation coefficient is directly obtained from the fourth-corner equation $\mathbf{D} = \mathbf{BA}'\mathbf{C}$ if the columns of the inflated data matrix are first standardized and the scalar product is divided by the number of rows of the inflated matrix minus 1.

• When comparing a quantitative variable in **B** to a qualitative variable coded into dummy variables (Subsection 1.5.7) in **C**, or the converse, the fourth-corner matrix product (eq. 10.35) is equivalent to computing an overall *F*-statistic for the pair of variables, as explained in Legendre *et al*. (1997a); the cells d_{ij} of matrix **D** contain measures of within-group homogeneity. Correlations may also be computed between the quantitative variable on the one hand, and each of the dummy variables coding for the qualitative variable on the other hand.

Each of these statistics can be tested for significance using the permutational procedure described in Subsection 10.6.2.

The fourth-corner method offers a way of analysing the relationships between *supplementary variables* associated with the rows and columns of a community composition data matrix. Other types of problems could be studied using this method. Here are two examples.

• In biogeography, consider a matrix **A** of presence/absence or abundance of species; a matrix **B** describing the extensiveness of the species' distributions, their migratory behaviour, etc.; and a matrix **C** of habitat characteristics (environmental variables), as above. The question is again to relate habitat to species characteristics.

• In the study of feeding behaviour, consider a matrix **A** with columns that are *individuals* while rows correspond to sites. The prey ingested by each individual are found in matrix **B**. Matrix **C** may contain either microhabitat environmental variables, or prey availability variables. The question is to determine feeding preferences: choice of prey *versus* availability, or choice of prey *versus* microhabitat conditions. Problems of the same type are found in such fields as sociology, marketing, political science, and the like.

• In studies involving spatial data, matrix **C** may contain spatial eigenfunctions (Chapter 14) representing the spatial relationships among the study sites. A global test of significance can be carried our between the characteristics of the species in **B** and the spatial eigenfunctions in **C** using the global statistic S_{RLQ}.

Ecological application 10.6

Development of the fourth-corner method was motivated by the study of a fish assemblage (280 species) surveyed along a one-km transect across the coral reef of Moorea Island, French Polynesia (Legendre *et al*., 1997a). Biological and behavioural characteristics of the species were used as descriptors (supplementary variables) for the rows, and characteristics of the environment for the columns of the fish presence-absence data matrix **A**. Parameters of the relationship between habitat characteristics (distance from the beach, water depth, and substrate variables) and biological and behavioural traits of the species (feeding habits, ecological niche categories, size classes, egg types, activity rhythms) were estimated and tested for significance. Results were compared to predictions made independently by reef fish ecologists, in order to assess the method as well as the pertinence of the variables subjected to the analysis.

Table 10.10 summarizes the comparison of reef bottom materials to feeding habits. This is an interesting case: the eight "reef bottom materials" variables are relative frequencies; each one represents the proportion of the habitat covered by a category of substrate material, so that non-integer pseudo-frequencies are obtained in the contingency table where the variables are crossed (Table 10.10). The permutation testing procedure allows data in matrices **B** and **C** to be relative or absolute frequencies. Probabilities remain the same under any linear transformation of the frequency values, even though the value of the G statistic is changed. This would not be allowed by a standard χ^2 test.

The relationship is globally significant (G = 15.426, p(G) = 0.0001 after 9999 random permutations following model 1 of Subsection 10.6.3 above); 20 of the 56 fourth-corner statistics d_{ij} were significant (*) after applying Holm's correction for multiple testing (Box 1.3). Compared to the null hypothesis, fish are under-represented on sand and large algae, and are unrelated to stone slab. In addition, herbivores are over-represented on live coral and calcareous algae. Grazers of sessile invertebrates and carnivores of types 1 and 2 are over-represented on coral debris, turf and dead coral, live coral, calcareous algae, and other types of substrate (large echinoderms, sponges, anemones, alcyonarians); this includes all areas where herbivores are found. Copepod eaters are over-represented on live coral and calcareous algae. Omnivores and specialist piscivores (fish-only diet) do not exhibit significant relationships with substrate.

Distance from the beach and size of fish species (adult individuals) are quantitative variables. The fourth-corner statistic that crosses these two variables is thus correlation-like; its value is r = 0.0504, with a probability of 0.001 after 999 random permutations. There is thus a weak but significant correlation, indicating that larger fish are found farther away from the beach than smaller ones. Other comparisons between biological-behavioural and habitat variables are presented in the published paper.

10.7 Software

Functions in the R language are available to carry out all analyses described in this chapter.

1. Linear regression. — In package STATS, function ***lm()*** computes simple or multiple linear regression. Function ***step()*** used in conjunction with ***lm()*** offers model selection by *AIC* using a backward, forward, or stepwise strategy.

Table 10.10 Contingency table comparing feeding habits (7 states) to materials covering reef bottom (8 proportions). From Legendre *et al*. (1997a, Table 6). First row in each cell: pseudo-frequency resulting from the matrix operation **D** = **BA**'**C**; lower row, probability adjusted using Holm's procedure; *: $p \leq 0.05$. Probabilities before correction resulted from 9999 random permutations. Sign indicates whether a statistic is above (+) or below (–) the expected value, estimated as the mean of the permutation results.

	Herbiv-orous	Omniv-orous	Sessile invertebrates	Carniv. 1 diurnal	Carniv. 2 nocturnal	Fish only	Copepod eater
Stone slab	6.20–	5.84+	3.72–	8.42–	5.18+	0.96+	2.40–
p	0.429	0.232	1.535	2.650	2.650	2.650	2.650
Sand	81.22–	54.26–	43.34–	94.38–	35.90–	8.94–	26.26–
p	0.039*	0.799	0.006*	0.006*	0.006*	0.799	0.039*
Coral debris	34.96+	20.22–	24.32+	46.74+	25.60+	4.48+	12.08–
p	1.976	1.976	0.006*	0.009*	0.645	2.650	2.650
Turf, dead cor.	45.46+	27.88+	28.28+	57.58+	33.58+	6.20+	15.76+
p	0.207	2.650	0.081	0.013*	0.029*	1.976	2.650
Live coral	49.86+	28.50+	29.20+	58.28+	40.82+	6.22+	21.06+
p	0.006*	1.976	0.006*	0.006*	0.006*	1.976	0.006*
Large algae	44.66–	37.50+	28.12–	59.68–	32.26–	6.34–	19.20–
p	0.006*	2.650	0.105	0.048*	0.140	2.650	2.650
Calcar. algae	29.12+	16.32+	16.08+	31.00+	26.02+	4.50+	11.32+
p	0.006*	1.030	0.079	0.122	0.006*	0.207	0.036*
Other substrate	2.52+	1.48+	1.94+	2.92+	1.64+	0.36+	0.92+
p	0.105	2.650	0.006*	0.795	1.734	1.976	1.976

Functions ***lmodel2()*** of LMODEL2 and ***sma()*** of SMATR compute model II simple linear regressions. Function ***lmorigin()*** in APE computes regression through the origin with permutation test. Variance inflation factors are computed by function ***vif()*** of packages CAR and DAAG, applied to models computed by ***lm()***.

QR decomposition, carried out by function ***qr()*** of BASE, is an efficient method to compute coefficients in univariate or multivariate linear regression. Multivariate linear regression can be computed using either ***lm()***, which takes either a single variable **y** or a whole matrix **Y** as the response data, or ***qr()*** after incrementing the explanatory

matrix $\mathbf{X}$ with a column of 1's to estimate the intercept, producing matrix $\mathbf{X}_{+1}$. For example, the matrix of fitted values in multivariate regression can be computed as follows: fitted(lm(as.matrix($\mathbf{Y}$) ~ ., data=$\mathbf{X}$)), or qr.fitted(qr($\mathbf{X}_{+1}$), as.matrix($\mathbf{Y}$)).

Ridge regression is available in functions ***lm.ridge()*** of MASS, ***ridge()*** of SURVIVAL, and ***penalized()*** of PENALIZED. Generalized linear models are computed by function ***glm()*** of STATS. Among the generalized linear models, only logistic regression is discussed in detail in the present chapter; it is computed by ***glm(****y~x, family=binomial(logit))*. In STATS, function ***nls()*** computes nonlinear weighted least-squares estimates of the parameters of a nonlinear statistical model; ***optim()*** is a general-purpose nonlinear optimization function offering a variety of optimization algorithms.

2. Partial regression and variation partitioning. — Partial linear regression can be computed by function ***rda()*** of VEGAN. ***varpart()*** of VEGAN is used for variation partitioning; ***plot.varpart()*** plots a Venn diagram with fixed circle and intersection sizes. A Venn diagram with proportional circle and intersection sizes can be obtained with function ***venneuler()*** of package VENNEULER*.

3. Path analysis. — Structural equation modelling, which is a generalized form of analysis encompassing path analysis, is available in package SEM.

4. Matrix comparisons. — Simple Mantel tests are found in functions ***mantel.test()*** of APE and ***mantel.rtest()*** of ADE4. For simple and partial Mantel tests, use ***mantel()*** of VEGAN, ***mantel()*** of ECODIST, ***mantel.test()*** and ***partial.mantel.test()*** of NCF. ***protest()*** in VEGAN computes the Procrustes permutation test. ***anosim()*** in VEGAN computes the ANOSIM test. The ***MRM()*** function in ECODIST carries out multiple regression on distance matrices.

5. Fourth-corner problem. — Functions ***fourthcorner()*** and ***fourthcorner2()*** of ADE4 compute fourth-corner analysis; function ***rlq()*** of ADE4 carries out RLQ analysis.

6. Miscellaneous methods. — Function ***poly()*** of STATS computes ordinary or orthogonal polynomials, the latter of the degree specified by the user, from a data vector. The resulting monomial vectors are normalized (i.e. scaled to length 1, eq. 2.7) and made to be orthogonal to one another. Several packages contain functions for spline and LOWESS smoothing, e.g. STATS, SPLINES and DIERCKXSPLINE.

* *Beware*: the fraction names in the *combinations* option of function ***venneuler()*** follow a different convention than in ***varpart()***. For two explanatory matrices for example, the first element mentioned, e.g. A, is fraction [c] of Fig. 10.10; the second element, e.g. B, is fraction [a]; the intersection [b] is called "A&B". See the examples in the documentation file.

Chapter 11 *Canonical analysis*

11.0 Principles of canonical analysis

Canonical analysis is the simultaneous analysis of two, or possibly several data tables. Canonical analyses allow ecologists to perform *direct comparisons* of two data matrices (also called "direct gradient analysis"; Fig. 10.4, Table 10.1). Typically, one may be interested in the relationship between a first table describing species composition and a second table containing environmental descriptors, observed *at the same locations*; or two tables of environmental descriptors, e.g. a table about the chemistry of lakes and another about drainage basin geomorphology.

Indirect comparison

Direct comparison

In *indirect comparison* (also called "indirect gradient analysis"; Fig. 10.4), the matrix of explanatory variables **X** does not intervene in the calculation producing the ordination of **Y**. Correlation or regression of the ordination vectors on **X** are computed *a posteriori*. In *direct comparison analysis* (canonical analysis) on the contrary, matrix **X** intervenes in the calculation, forcing the ordination vectors to be maximally related to combinations of the variables in **X**. This description applies to all forms of canonical analysis and in particular to the asymmetric forms described in Sections 11.1 to 11.3.

There is a parallel in cluster analysis, when clustering results are constrained to be consistent with explanatory variables in multivariate regression trees (MRT, Section 8.11) or with structural relationships among observations, either temporal (Subsection 12.6.4) or spatial (Subsection 13.3.2), which are inherent to the sampling design. In constrained clustering or canonical ordination, the results differ in most instances from those of unconstrained analysis and are, hopefully, more readily interpretable. Furthermore, direct comparison analysis allows one to directly test *a priori* ecological hypotheses by (1) bringing out *all* the variance of **Y** that is related to **X** and (2) allowing formal tests of these hypotheses to be performed, as detailed below. Further examination of the unexplained variability may help generate new hypotheses, to be tested using new field observations (Section 13.5).

Canonical form

In mathematics, a *canonical form* (from the Greek κανων, pronounced "kanôn", rule) is the simplest and most comprehensive form to which certain functions, relations, or expressions can be reduced without loss of generality. For example, the

canonical form of a covariance matrix is its matrix of eigenvalues. In general, methods of canonical analysis use eigenanalysis (i.e. calculation of eigenvalues and eigenvectors), although some extensions of canonical analysis have been described that use multidimensional scaling (nMDS) algorithms (Section 9.4).

There are two main families of canonical ordination methods: asymmetric and symmetric. In the asymmetric forms of analysis, there is a response data set and an explanatory data set, which are represented by $\mathbf{Y}$ and $\mathbf{X}$, respectively, in this chapter. The asymmetric methods are redundancy analysis (RDA), canonical correspondence analysis (CCA), and linear discriminant analysis (LDA). In contrast, symmetric methods are used in cases where the two data sets, called $\mathbf{Y}_1$ by $\mathbf{Y}_2$ to mark the symmetry, play the same role in the study; this means that an analysis of $\mathbf{Y}_1$ by $\mathbf{Y}_2$ produces the same result as an analysis of $\mathbf{Y}_2$ by $\mathbf{Y}_1$. These methods include canonical correlation analysis (CCorA), co-inertia analysis (CoIA), Procrustes analysis (Proc), and some others.

Interrelationships among the variables involved in canonical analysis may be represented by the following partitioned covariance matrix, resulting from the concatenation of the $\mathbf{Y}$ (or $\mathbf{Y}_1$, order $n \times p$) and $\mathbf{X}$ (or $\mathbf{Y}_2$, $n \times m$) data sets. The joint dispersion matrix $\mathbf{S_{Y+X}}$ contains blocks that are identified as follows for convenience:

$$\mathbf{S_{Y+X}} = \left[\begin{array}{ccc|ccc} s_{y_1,y_1} & \cdots & s_{y_1,y_p} & s_{y_1,x_1} & \cdots & s_{y_1,x_m} \\ \vdots & & \vdots & \vdots & & \vdots \\ \vdots & & \vdots & \vdots & & \vdots \\ \vdots & & \vdots & \vdots & & \vdots \\ s_{y_p,y_1} & \cdots & s_{y_p,y_p} & s_{y_p,x_1} & \cdots & s_{y_p,x_m} \\ \hline s_{x_1,y_1} & \cdots & s_{x_1,y_p} & s_{x_1,x_1} & \cdots & s_{x_1,x_m} \\ \vdots & & \vdots & \vdots & & \vdots \\ \vdots & & \vdots & \vdots & & \vdots \\ \vdots & & \vdots & \vdots & & \vdots \\ s_{x_m,y_1} & \cdots & s_{x_m,y_p} & s_{x_m,x_1} & \cdots & s_{x_m,x_m} \end{array}\right] = \left[\begin{array}{cc} \mathbf{S_{YY}} & \mathbf{S_{YX}} \\ \mathbf{S_{XY}} & \mathbf{S_{XX}} \end{array}\right] = \left[\begin{array}{cc} \mathbf{S_{YY}} & \mathbf{S_{YX}} \\ \mathbf{S'_{YX}} & \mathbf{S_{XX}} \end{array}\right] \qquad \textbf{(11.1)}$$

Submatrices $\mathbf{S_{YY}}$ (order $p \times p$) and $\mathbf{S_{XX}}$ ($m \times m$) concern each of the two sets of descriptors, respectively, whereas $\mathbf{S_{YX}}$ ($p \times m$) and its transpose $\mathbf{S'_{YX}} = \mathbf{S_{XY}}$ ($m \times p$) account for the covariances among the descriptors of the two groups, as in eq. 4.27.

Asymmetric, canonical analysis

Asymmetric canonical analysis combines the concepts of ordination and regression. It involves a response matrix $\mathbf{Y}$ and an explanatory matrix $\mathbf{X}$. As it was the case with the simple ordination methods (Chapter 9 and Fig. 11.1a), the asymmetric methods of canonical analysis produce a single ordination of the objects, which may be plotted in a scatter diagram. With the symmetric methods on the contrary, two different ordinations of the objects are produced, one for each data set; see below.

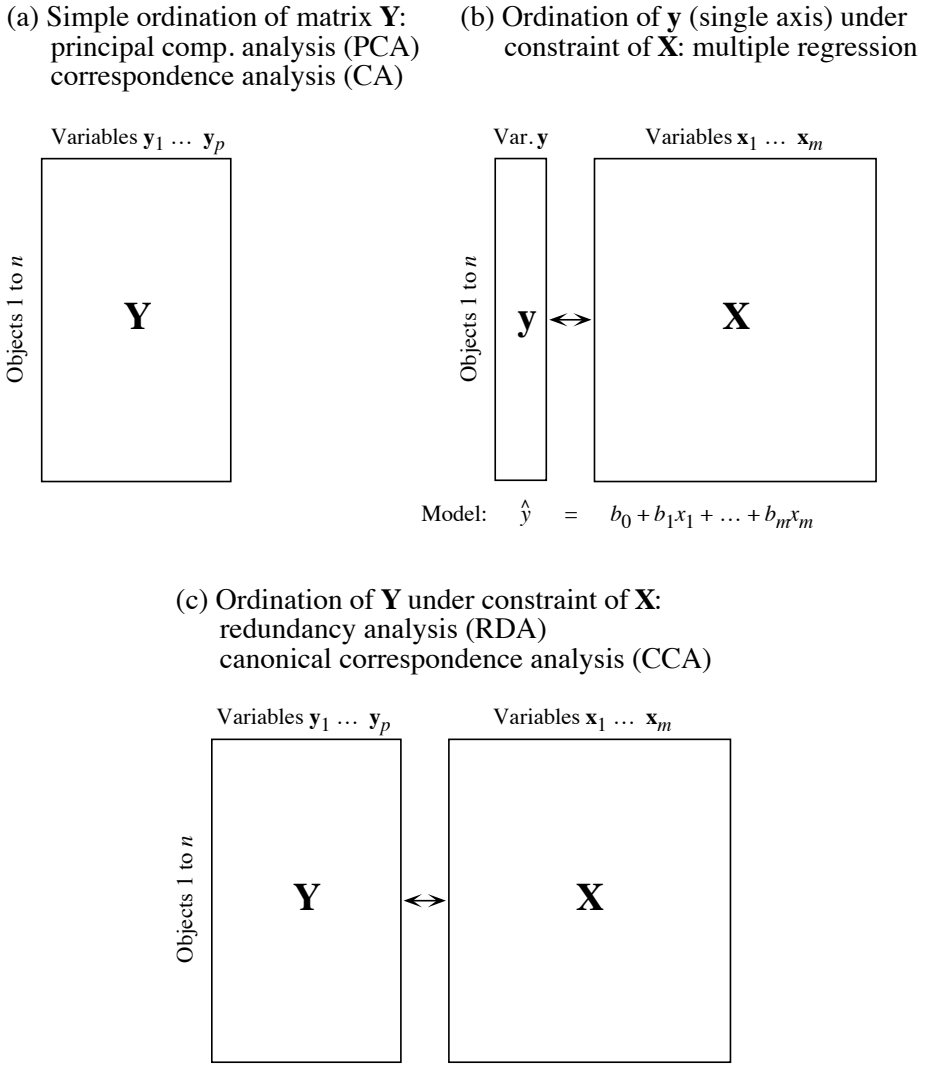


Figure 11.1 Relationships between (a) ordination, (b) regression, and (c) two asymmetric forms of canonical analysis (RDA and CCA). In (c), each canonical axis of **Y** is constrained to be a linear combination of the explanatory variables **X**.

Redundancy analysis (RDA, Section 11.1) and canonical correspondence analysis (CCA, Section 11.2) are related to multiple linear regression. In Subsection 10.3.3, multiple regression was described as a method for modelling a response variable **y** using a set of explanatory variables assembled into a data table **X**. Another aspect of regression analysis must be stressed: while the original response variable **y** provides,

by itself, an ordination of the objects in one dimension, the vector of fitted values (eq. 10.15)

$$\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_p x_{ip}$$

creates a new one-dimensional ordination of the same objects (Fig. 11.1b). The ordinations corresponding to $\mathbf{y}$ and $\hat{\mathbf{y}}$ differ; the square of their correlation is the coefficient of determination (eq. 10.20) of the multiple regression model:

$$R^2_{\mathbf{y}|\mathbf{X}} = [r(\mathbf{y}, \hat{\mathbf{y}})]^2 \qquad \textbf{(11.2)}$$

So, multiple regression creates a correspondence between ordinations $\mathbf{y}$ and $\hat{\mathbf{y}}$, because ordination $\hat{\mathbf{y}}$ is constrained to be optimally (in the least-squares sense) and linearly related to the variables in $\mathbf{X}$. The constraint implemented in multiple regression maximizes R^2. The asymmetric methods of canonical analysis share this property.

Asymmetric canonical analysis combines the properties of two families of methods, i.e. ordination and regression (Fig. 11.1c). It produces ordinations of $\mathbf{Y}$ that are constrained to be linearly related to a second set of variables $\mathbf{X}$, and the results are plotted in reduced space. The way in which the relationship between $\mathbf{Y}$ and $\mathbf{X}$ is established differs among methods of asymmetric canonical analysis.

- In redundancy analysis (RDA, Section 11.1), each canonical ordination axis corresponds to a direction, in the multivariate scatter of objects, that is maximally related to a linear combination of the explanatory variables $\mathbf{X}$. A canonical axis is thus similar to a principal component (Box 9.1). Two ordinations of the objects may be plotted along the canonical axes: (1) linear combinations of the $\mathbf{Y}$ variables (matrix $\mathbf{F}$, eq. 11.17), as in PCA, and (2) linear combinations of the fitted $\hat{\mathbf{Y}}$ variables (matrix $\mathbf{Z}$, eq. 11.18), which are thus also linear combinations of the $\mathbf{X}$ variables. RDA preserves the Euclidean distances among objects in matrix $\hat{\mathbf{Y}}$, which contains values of $\mathbf{Y}$ fitted by regression to the explanatory variables $\mathbf{X}$ (Fig. 11.2); variables in $\hat{\mathbf{Y}}$ are therefore linear combinations of the $\mathbf{X}$ variables.

- Canonical correspondence analysis (CCA, Section 11.2) is similar to RDA. The difference is that CCA preserves the χ^2 distance (as in correspondence analysis), instead of the Euclidean distance among objects in matrix $\hat{\mathbf{Y}}$. Calculations are a bit more complex since matrix $\hat{\mathbf{Y}}$ contains fitted values obtained by weighted linear regression of matrix $\overline{\mathbf{Q}}$ of correspondence analysis (eq. 9.24) on the explanatory variables $\mathbf{X}$. As in RDA, two ordinations of the objects may be plotted.

- In linear discriminant analysis (Section 11.3), the objects are divided into k groups, described by a qualitative descriptor (factor) forming the response matrix $\mathbf{Y}$. The method seeks linear combinations of explanatory variables (matrix $\mathbf{X}$) that explain the classification in $\mathbf{Y}$ by maximizing the dispersion of the centroids of the k groups. This is obtained by maximizing the ratio of the among-object-group dispersion over the pooled within-object-group dispersion (eq. 11.33).

Symmetric, canonical analysis

The *symmetric forms of canonical analysis* described in this book are the following:

- In canonical correlation analysis (CCorA, Section 11.4), the canonical axes maximize the correlation between linear combinations of the two sets of variables $\mathbf{Y}_1$ and $\mathbf{Y}_2$. This is obtained by maximizing the squared among-variable-set correlations (Table 11.10). Two different ordinations of the objects are obtained, one for data set $\mathbf{Y}_1$ and the other for $\mathbf{Y}_2$.

- Co-inertia analysis (CoIA) and Procrustes analysis (Proc) (Section 11.5) search for common structures between two data sets $\mathbf{Y}_1$ and $\mathbf{Y}_2$ describing the same objects. Each object has two representations in the joint plot, one from $\mathbf{Y}_1$ and the other from $\mathbf{Y}_2$.

The application of the various methods of canonical analysis to ecological data was briefly discussed in Section 10.2. In summary, when one of the data sets (**Y**) is to be explained by another (**X**), the asymmetric forms of canonical analysis should be used; the methods are redundancy analysis (RDA) and canonical correspondence analysis (CCA) when **Y** is a full table of response variables, and linear discriminant analysis (LDA) when **Y** contains a classification of the objects. RDA is used when the **X** variables display linear relationships with the **Y** variables, whereas CCA can be used in the cases where correspondence analysis (CA, Section 9.2) would be appropriate for an ordination of **Y** alone. Linear discriminant analysis is applicable when the response data set contains a classification of the objects or an ANOVA factor; in ecology, LDA is used mostly to discriminate among groups of sites using descriptors of the physical environment (Section 11.3). In contrast, canonical correlation analysis (CCorA), co-inertia analysis (CoIA) and Procrustes analysis (Proc) are used to relate two data sets describing the same objects in a correlative framework (Sections 11.4 and 11.5).

Canonical analysis has become an instrument of choice for ecological analysis. A bibliography on the applications of canonical analysis to ecology, covering the period 1986 to 1996, contains a total of 804 entries (Birks *et al.*, 1998). CCorA and discriminant analysis are available in most commercial statistical packages. For RDA, CCA, CoIA and Proc, one must rely on specialized ordination packages and R functions. CANOCO (ter Braak, 1988b) was the first ordination package that made RDA and CCA available to users. These methods are also available in PC-ORD and SYN-TAX 2000. See Section 11.7.

11.1 Redundancy analysis (RDA)

Redundancy analysis (RDA) is the direct extension of multiple regression to the modelling of multivariate response data. The analysis is asymmetric: **Y** ($n \times p$) is a table of response variables and **X** ($n \times m$) is a table of explanatory variables. In RDA, the ordination of **Y** is constrained in such a way that the resulting ordination axes (matrix **Z** below) are linear combinations of the variables in **X**. The difference between

RDA and canonical correlation analysis (CCorA, Section 11.4) is the same as that between simple linear regression (asymmetric analysis) and linear correlation analysis (symmetric); see Box 10.1.

In RDA, the ordination axes are obtained by principal component analysis (PCA, Section 9.1) of a matrix $\hat{\mathbf{Y}}$, computed by fitting the **Y** variables to **X** by multivariate linear regression (details in Subsection 11.1.1). So, in scaling type 1 plots (Subsection 11.1.3), RDA preserves the Euclidean distance among objects (D_1, Chapter 7): the ordination of the points in matrix **Z** is a PCA rotation of the points in $\hat{\mathbf{Y}}$. The ordination axes in **Z** differ, of course, from the principal components that could be computed directly from the **Y** data table because they are constrained to be linear combinations of the variables in **X**. Prior to RDA, the data in **Y** must be at least centred, or transformed following the same principles as in PCA.

1 — Simple RDA

Canonical redundancy analysis was first described by Rao (1964). In his 1973 book (p. 594-595), he proposed the topic to readers as an exercise at the end of his Chapter 8 on multivariate analysis. Rao called the method *Principal components of instrumental variables*. RDA was later rediscovered by Wollenberg (1977) who called the method *Redundancy analysis* by reference to the *redundancy index* of Stewart & Love (1968), which is the proportion of the variance of the response data matrix **Y** that is accounted for by the explanatory matrix **X**. Redundancy is synonymous with explained variance (Gittins, 1985). In his paper, Wollenberg did not refer to Rao's paper (1964) and book (1973). Wollenberg's equation, which only applied to correlation matrices, was less general than that of Rao which involved covariance matrices in general.

Redundancy

Redundancy analysis (RDA) of a response matrix **Y** (with n objects and p variables) by an explanatory matrix **X** (with n objects and m variables) is called simple RDA in Subsections 11.1.1 to 11.1.5, by opposition to partial RDA, described in Subsections 11.1.6 to 11.1.10, which involves a matrix of covariables **W**. Simple RDA involves two computational steps (Fig. 11.2). In the algebraic development that follows, the columns of matrices **Y** and **X** are centred to have means of 0. In computer software, the columns of **X** may be standardized for programming convenience, but this has no effect on the results of the analysis since the matrix of fitted values $\hat{\mathbf{Y}}$ is identical when computed from centred or standardized **X** variables. As in PCA, the variables in **Y** should be standardized if they are not dimensionally homogeneous (e.g. if they are a mixture of temperatures, concentrations, and pH values). Transformations applicable to community composition data (presence-absence or abundance) are described in Section 7.7. As in multiple regression analysis, matrix **X** can contain explanatory variables of different mathematical types: quantitative, multi-state qualitative (e.g. ANOVA factors), or binary variables; see the last five paragraphs of Subsection 10.3.3. If present, collinearity among the **X** variables should be reduced prior to RDA using the methods described for multiple regression in Subsection 10.3.3. Chapters 13 and 14 will show how different expressions of spatial relationships can be used as the explanatory matrix **X** in RDA.

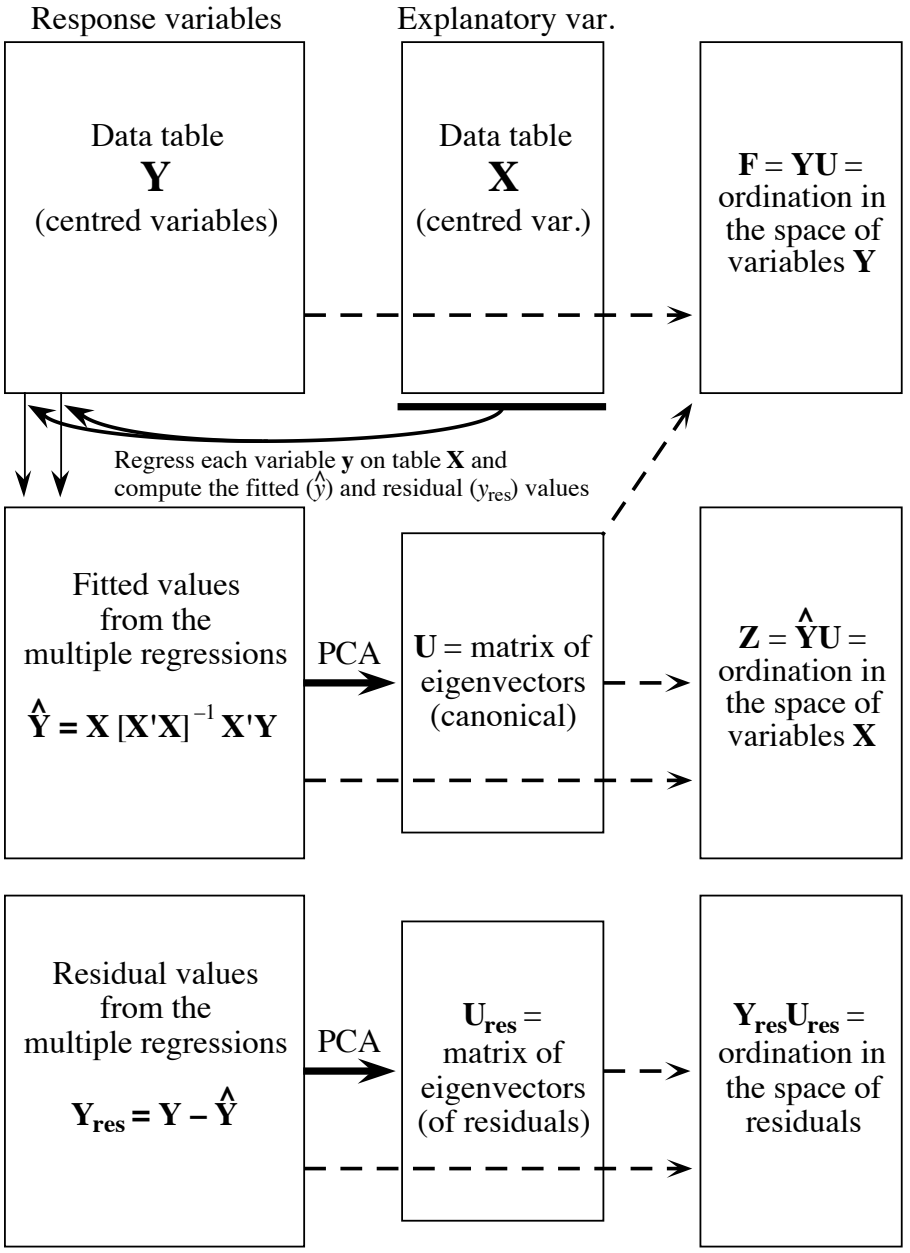


Figure 11.2 Redundancy analysis may be understood as a two-step process: (1) regress each variable in **Y** on all variables in **X** and compute the fitted values; (2) carry out a PCA of the matrix of fitted values to obtain the eigenvalues and eigenvectors. Two ordinations are obtained, one (**F** = **YU**) in the space of the response variables **Y**, the other (**Z** = $\hat{\mathbf{Y}}$**U**) in the space of the explanatory variables **X**. Another PCA ordination can be computed for the matrix of residuals.

The variable distributions should be examined for normality at this stage, as well as bivariate plots within and between the sets **Y** and **X**. Because RDA is a linear model based on multiple linear regression, data transformations (Section 1.5) should be applied as needed to linearize the relationships and make the frequency distributions as symmetric as possible, thus reducing the effect of outliers.

- Step 1 is a *multivariate linear regression* of **Y** on **X** (eq. 10.16), which produces a matrix of fitted values $\hat{\mathbf{Y}}$ through the linear equation:

$$\hat{\mathbf{Y}} = \mathbf{X}\,[\mathbf{X}'\mathbf{X}]^{-1}\,\mathbf{X}'\mathbf{Y} \qquad \textbf{(11.3)}$$

This is equivalent to a series of multiple linear regressions of the individual variables of **Y** on **X** to calculate vectors of fitted values followed by stacking these column vectors side by side into matrix $\hat{\mathbf{Y}}$. In principle, model II regression should be used when the explanatory variables **X** are *random*, by opposition to *controlled* (Subsection 10.3.2). Ordinary least squares (OLS) are used in eq. 11.3 because, among the model II regression methods, OLS produces fitted values with the smallest error for given values of the predictors (Table 10.4). For efficiency reasons in computer software, matrix $\hat{\mathbf{Y}}$ may be computed through QR decomposition instead of eq. 11.3.

- Step 2 is a principal component analysis of $\hat{\mathbf{Y}}$. This PCA produces the canonical eigenvalues and eigenvectors, as well as matrix **Z** containing the canonical axes (object ordination scores, like matrix **F** in PCA). That step is performed to obtain reduced-space ordination diagrams displaying the objects, response variables, and explanatory variables for the most important axes of the canonical relationship. The PCA step is pertinent only if a significant canonical relationship has been found between **Y** and **X** through an appropriate test of significance (Subsection 11.1.2).

Like the fitted values of a multiple linear regression, which are linear combinations of the explanatory variables, the canonical axes (object ordination scores) are also linear combinations of the explanatory variables in **X**. That RDA axes are linear combinations of the explanatory variables is the fundamental property of RDA (ter Braak, 1987c; ter Braak and Prentice, 1988). Individual canonical axes can be tested for significance to determine which ones are important enough to warrant consideration, plotting, and detailed analysis.

2 — Statistics in simple RDA

After step 1 of RDA, one can compute the following informative statistics.

Redundancy statistic (R^2)

1. From matrices **Y** and $\hat{\mathbf{Y}}$, one can calculate the canonical R^2, which Miller and Farr (1971) called the *bimultivariate redundancy statistic*. This statistic measures the strength of the linear relationship between **Y** and **X**:

$$R^2_{\mathbf{Y}|\mathbf{X}} = \frac{\mathrm{SS}\,(\hat{\mathbf{Y}})}{\mathrm{SS}\,(\mathbf{Y})} \qquad \textbf{(11.4)}$$

where SS ($\hat{\mathbf{Y}}$) is the total sum of squares (or sum of squared deviations from the means) of $\hat{\mathbf{Y}}$ and SS($\mathbf{Y}$) is the total sum of squares of $\mathbf{Y}$. The canonical R^2 is constructed in the same way and has the same meaning as the R^2 statistic in multiple regression (eq. 10.20): it is the proportion of the variation of $\mathbf{Y}$ explained by a linear model of the variables in $\mathbf{X}$.

Note: in the absence of relationship between $\mathbf{Y}$ and $\mathbf{X}$, the expected value of R^2 in multiple regression and in RDA is not 0 but $m/(n-1)$. This is because a matrix $\mathbf{X}$ containing $m = (n-1)$ columns of random numbers produces an R^2 of 1; this surprising fact can easily verify numerically by computing a multiple regression or a RDA with a matrix $\mathbf{X}$ containing $(n-1)$ columns of random numbers. Hence, the expected value (E) of the R^2 produced by a single explanatory variable made of random numbers is $E(R^2) = 1/(n-1)$, and $E(R^2) = m/(n-1)$ for m explanatory variables. This is illustrated in the numerical simulation results presented by Peres-Neto *et al*. (2006).

Adjusted R^2

2. The adjusted R^2 (R^2_a) is computed as in eq. 10.21 (Ezekiel, 1930):

$$R^2_a = 1 - (1 - R^2_{\mathbf{Y}|\mathbf{X}}) \frac{(n-1)}{(n-m-1)} \qquad \textbf{(11.5)}$$

where m is the number of explanatory variables in $\mathbf{X}$ or, more precisely, the rank of the variance-covariance matrix of $\mathbf{X}$.

F-statistic
Overall test

3. The F-statistic for the overall test of significance is constructed as follows (Miller, 1975):

$$F = \frac{R^2_{\mathbf{Y}_{stand}|\mathbf{X}} / mp}{(1 - R^2_{\mathbf{Y}_{stand}|\mathbf{X}}) / (n-m-1)\, p} \qquad \textbf{(11.6)}$$

This statistic is used to perform the overall test of significance of the canonical relationship. The null hypothesis of the test is H_0: the strength of the linear relationship, measured by the canonical R^2, is not larger than the value that would be obtained for unrelated $\mathbf{Y}$ and $\mathbf{X}$ matrices of the same sizes.

When the variables of $\mathbf{Y}$ are standardized ($\mathbf{Y}_{stand}$) and the error distribution is normal, the F-statistic (eq. 11.6) can be tested for significance using the Fisher-Snedecor F-distribution with degrees of freedom $\nu_1 = mp$ and $\nu_2 = p(n-m-1)$. p is the number of response variables in $\mathbf{Y}$. Because m parameters were estimated for each of the p multiple regressions used to compute the vectors of fitted values forming the p columns of $\hat{\mathbf{Y}}$, a total of mp parameters were estimated. This is why there are $\nu_1 = mp$ degrees of freedom attached to the numerator of F. Each multiple regression equation has residual degrees of freedom equal to $(n-m-1)$, so the total number of degrees of freedom of the denominator, ν_2, is p times $(n-m-1)$. Miller (1975) conducted numerical simulations in the multivariate normal case, with combinations of m and p

from 2 to 15 and sample sizes of $n = 30$ to 160. He showed that eq. 11.6 produced distributions of F values that were very close to theoretical F-distributions with the same numbers of degrees of freedom. Additional simulations conducted by Legendre *et al*. (2011, Appendix A) confirmed that the parametric test of significance had correct levels of type I error when **Y** was standardized. This was not the case, however, for non-standardized matrices of response variables **Y** generated with equal or unequal population variances, especially when the error was not normal. Permutation tests always had correct levels of type I error in these simulations. The effect of correlations among the standardized response variables in **Y** on the validity of the parametric test remains to be investigated.

In many instances, the response variables should not be standardized prior to RDA. With community composition data (species abundances), for example, the variances of the species should be preserved in most analyses since abundant and rare species do not play the same roles in ecosystems. A permutation test should always be used in that case. For permutation tests, one can simplify eq. 11.6 of the F-statistic by eliminating the constant p from the numerator and denominator:

Permutation test

$$F = \frac{R^2_{\mathbf{Y}|\mathbf{X}} / m}{(1 - R^2_{\mathbf{Y}|\mathbf{X}}) / (n - m - 1)} \qquad \textbf{(11.7)}$$

While the numerator and denominator of eq. 11.6 indicate the numbers of degrees of freedom for a correct parametric test of F, eliminating p from both does not change the computed value of F. Equation 11.7 is the one used for permutation tests in programs of canonical analysis such as CANOCO and VEGAN's ***rda()***. Actually, the degrees of freedom can be entirely eliminated from statistic equations used in permutation tests since they are invariant across all permutations of the data. However, most computer programs and functions that carry out permutation tests display them to allow comparison with the F-statistic used in parametric tests.

Test of individual axes

4. Individual canonical axes can be tested for significance. Since one deals with complex, multivariate data influenced by many factors, several independent structures may coexist in the response data. If these structures are linearly independent, they should appear on different canonical axes. The results of the tests of individual axes allow researchers to determine which of the canonical axes represent variation that is more structured than random. Canonical axes that do not explain more variation than random should be identified since they do not need to be further considered in the interpretation of the results.

Two methods, called the *forward* and *marginal* testing procedures, can be used for testing individual axes. The forward method was developed by Cajo J. F. ter Braak and implemented in the CANOCO package since version 3.10 (ter Braak, 1990). The marginal method was developed by Jari Oksanen for the ***permutest.cca()*** function of the VEGAN R package; that function carries out tests of significance of the canonical axes when users call the ***anova.cca()*** function with parameter *by="axis"*, after

canonical analysis by functions ***rda()*** or ***cca()***. In a simulation study, Legendre *et al*. (2011) showed that these two methods had correct levels of type I error and comparable powers. This latter study also investigated a third method, the simultaneous test of all canonical axes, which was shown to be invalid.

The null hypothesis for the test of significance of the j^{th} canonical axis is H_0: the linear dependence of the response variables **Y** on the explanatory variables **X** is less than *j*-dimensional. More informally, the null hypothesis is that the j^{th} axis under test explains no more variation than a random axis of the same order (*j*), given the variation explained by the previously tested axes. The test of individual canonical axes can also be carried out in partial RDA (Subsection 11.1.6), a form of RDA that incorporates a matrix of covariables **W**.

3 — The algebra of simple RDA

The eigenanalysis equation for redundancy analysis, which is an asymmetric form of analysis, can be obtained from eq. 11.48 of canonical correlation analysis (CCorA, Section 11.4), which is a symmetric form of analysis, by changing the $\mathbf{S}_{\mathbf{YY}}^{-1}$ matrix (called $\mathbf{S}_{11}^{-1}$ in eq. 11.48) into an identity matrix **I**. The latter does not have to be written after matrix $\mathbf{S}_{\mathbf{YX}}'$ and thus disappears from the equation (Rao, 1973; ter Braak, 1987c):

$$(\mathbf{S}_{\mathbf{YX}}\mathbf{S}_{\mathbf{XX}}^{-1}\mathbf{S}_{\mathbf{YX}}' - \lambda_k\mathbf{I})\,\mathbf{u}_k = \mathbf{0} \qquad \textbf{(11.8)}$$

The covariance relationships among the explanatory variables, $\mathbf{S}_{\mathbf{XX}}^{-1}$, remains included in the equation. Equation 11.8 differs from the original formulations by Rao (1964, 1973) and Wollenberg (1977), but it produces the same canonical eigenvalues.

Equation 11.8 is the end result of carrying out the two steps described in the previous subsection, which characterize RDA: (1) a multivariate regression of **Y** on **X** to obtain a matrix of fitted values $\hat{\mathbf{Y}}$, followed by (2) a PCA of that matrix of fitted values. The asymmetric nature of RDA comes from the fact that multivariate regression (eqs. 10.16 and 11.3) is an asymmetric analysis, just as its univariate counterpart, multiple linear regression, where **y** is the response vector and **X** is the explanatory matrix. The developments that follow show that these two computational steps produce eq. 11.8.

1) For *each* response variable in matrix **Y**, compute a multiple linear regression on all variables in matrix **X**. For each regression, the coefficients are computed as follows (eq. 2.19):

$$\mathbf{b} = [\mathbf{X}'\mathbf{X}]^{-1}\,\mathbf{X}'\mathbf{y}$$

The matrix containing all regression coefficients can be obtained by a single matrix operation (equation without number above eq. 10.16 in Subsection 10.3.3):

$$\mathbf{B} = [\mathbf{X}'\mathbf{X}]^{-1}\,\mathbf{X}'\mathbf{Y} \tag{11.9}$$

where $\mathbf{B}$ ($m \times p$) is the matrix of regression coefficients of all p response variables $\mathbf{Y}$ on the m explanatory variables $\mathbf{X}$.

As in multiple regression, the fitted values $[\hat{y}]$ can be computed by a single matrix operation:

$$\hat{\mathbf{Y}} = \mathbf{X}\,\mathbf{B} \tag{11.10}$$

This is the multivariate extension of eq. 10.1. Replacing $\mathbf{B}$ by the expression from eq. 11.9, eq. 11.10 becomes:

$$\hat{\mathbf{Y}} = \mathbf{X}\,[\mathbf{X}'\,\mathbf{X}]^{-1}\,\mathbf{X}'\,\mathbf{Y} \tag{11.11}$$

which is the multivariate linear regression equation (eq. 10.16). Because the variables in $\mathbf{X}$ and $\mathbf{Y}$ were centred on their respective means, there are no intercept parameters in the column vectors of regression coefficients forming $\mathbf{B}$, and the column vectors in $\hat{\mathbf{Y}}$ are also centred.

2) The covariance matrix corresponding to the table of fitted values $\hat{\mathbf{Y}}$ is computed using eq. 4.6:

$$\mathbf{S}_{\hat{\mathbf{Y}}'\hat{\mathbf{Y}}} = [1/(n-1)]\;\hat{\mathbf{Y}}'\hat{\mathbf{Y}} \tag{11.12}$$

Replacing $\hat{\mathbf{Y}}$ by the expression from eq. 11.11, eq. 11.12 becomes:

$$\mathbf{S}_{\hat{\mathbf{Y}}'\hat{\mathbf{Y}}} = [1/(n-1)]\,\mathbf{Y}'\,\mathbf{X}\,[\mathbf{X}'\,\mathbf{X}]^{-1}\,\mathbf{X}'\,\mathbf{X}\,[\mathbf{X}'\,\mathbf{X}]^{-1}\,\mathbf{X}'\,\mathbf{Y} \tag{11.13}$$

This equation reduces to:

$$\mathbf{S}_{\hat{\mathbf{Y}}'\hat{\mathbf{Y}}} = \mathbf{S}_{\mathbf{YX}}\mathbf{S}_{\mathbf{XX}}^{-1}\mathbf{S}'_{\mathbf{YX}} \tag{11.14}$$

where $\mathbf{S}_{\mathbf{YY}}$ is the ($p \times p$) covariance matrix among the response variables, $\mathbf{S}_{\mathbf{XX}}$ the ($m \times m$) covariance matrix among the explanatory variables (it is actually a matrix $\mathbf{R}_{\mathbf{XX}}$ when the $\mathbf{X}$ variables have been standardized), and $\mathbf{S}_{\mathbf{YX}}$ is the ($p \times m$) covariance matrix among the variables of the two sets; the order of its transpose $\mathbf{S}'_{\mathbf{YX}} = \mathbf{S}_{\mathbf{XY}}$ is ($m \times p$). If the $\mathbf{Y}$ variables had also been standardized, this equation would read $\mathbf{R}_{\mathbf{YX}}\mathbf{R}_{\mathbf{XX}}^{-1}\mathbf{R}'_{\mathbf{YX}}$, which is the multivariate form of the equation for the coefficient of multiple determination (eq. 4.31).

3) The matrix of fitted values $\hat{\mathbf{Y}}$ is subjected to principal component analysis to reduce the dimensionality of the solution. This corresponds to solving the eigenvalue problem:

$$(\mathbf{S}_{\hat{\mathbf{Y}}'\hat{\mathbf{Y}}} - \lambda_k \mathbf{I})\, \mathbf{u}_k = \mathbf{0} \qquad \textbf{(11.15)}$$

which, using eq. 11.14, translates into:

$$(\mathbf{S}_{\mathbf{YX}} \mathbf{S}_{\mathbf{XX}}^{-1} \mathbf{S}_{\mathbf{YX}}' - \lambda_k \mathbf{I})\, \mathbf{u}_k = \mathbf{0} \qquad \textbf{(11.16)}$$

This is the equation for redundancy analysis (eq. 11.8). Different programs may express the eigenvalues in different ways: raw eigenvalues, fractions of the total variance in **Y**, or percentages; see Tables 11.2 and 11.4 for examples.

The matrix containing the normalized canonical eigenvectors $\mathbf{u}_k$ is called **U**. The eigenvectors give the contributions of the descriptors in matrix $\hat{\mathbf{Y}}$ to the various canonical axes. Matrix **U**, of size $(p \times p)$, contains only $\min[p, m, n-1]$ eigenvectors with non-zero eigenvalues, since the number of canonical eigenvectors cannot exceed the minimum of p, m and $(n-1)$:

- It cannot exceed p, which is the dimension of the reference space of matrix **Y**. This is obvious in multiple regression where matrix **Y** contains a single variable; the ordination given by the fitted values $\hat{y}$ is one-dimensional.

- It cannot exceed m, which is the number of variables in **X**. Consider an extreme example: if **X** contains a single explanatory variable ($m = 1$), regressing all p variables in **Y** on this single explanatory variable produces p fitted vectors $\hat{y}$ which all point in the same direction of the p-dimensional space; a principal component analysis of matrix $\hat{\mathbf{Y}}$ of these fitted vectors can only produce one common (canonical) axis.

- It cannot exceed $(n-1)$, which is the maximum number of dimensions required to represent n points in Euclidean space.

The canonical coefficients in the normalized matrix **U** give the contributions of the variables of $\hat{\mathbf{Y}}$ to the canonical axes. They should be interpreted as in PCA. Matrix **U** is used to produce scaling 1 biplot or triplot diagrams, described below. For scaling 2 plots, **U** is rescaled in such a way that the length of each eigenvector is $\sqrt{\lambda_k}$.

If **X** and **Y** are made to contain the same data (i.e. $\mathbf{X} = \mathbf{Y}$), eq. 11.16 becomes $(\mathbf{S}_{\mathbf{YY}} - \lambda_k \mathbf{I})\, \mathbf{u}_k = \mathbf{0}$, which is the equation for principal component analysis (eq. 9.1). The result of RDA is then a principal component analysis of data table **Y**, a fact that was pointed out by Rao (1964, 1973) and by Wollenberg (1977). Another way to look at this point is to say that a RDA of **Y** by **Y** is a PCA of **Y** because $\hat{\mathbf{Y}} = \mathbf{Y}$ in that case.

Additional computations must be done to produce the RDA triplot diagram (below), which contains three types of elements: response variables (e.g. species), objects (e.g. sites), and explanatory variables.

4) The ordination of objects in the space of the response variables **Y** is obtained directly from the centred matrix $\mathbf{Y}_c$, using the standard equation for principal components (matrix **F**, eq. 9.4) and matrix **U** of the eigenvectors $\mathbf{u}_k$ found in eq. 11.16:

$$\mathbf{F} = \mathbf{Y}_c\mathbf{U} \qquad \textbf{(11.17)}$$

Site scores

The ordination vectors (columns of **F**) defined in eq. 11.17 are called the vectors of "site scores". They have variances that are close, but not equal to the corresponding eigenvalues. How to represent matrix **F** in biplots is discussed in point 8 (below).

5) Likewise, the ordination of objects in space **X** is obtained as follows:

$$\mathbf{Z} = \hat{\mathbf{Y}}\,\mathbf{U} = \mathbf{X}\mathbf{B}\mathbf{U} \qquad \textbf{(11.18)}$$

Fitted site scores

As stated above, the vectors in matrix $\hat{\mathbf{Y}}$ are centred on their respective means. The right-hand part of eq. 11.18, obtained by replacing $\hat{\mathbf{Y}}$ by it value in eq. 11.10, shows that this ordination is a linear combinations of the **X** variables. For that reason, these ordination vectors (columns of matrix **Z**) are also called "fitted site scores", or "sample scores that are linear combinations of environmental variables" in program CANOCO. The ordination vectors, defined in eq. 11.18, have variances equal to the corresponding eigenvalues. The representation of matrix **Z** in biplots is discussed in point 8 (below).

The "site scores" of eq. 11.17 are obtained by projecting the original data (matrix **Y**) onto axis *k*; they approximate the observed data, which contain residuals ($\mathbf{Y} = \hat{\mathbf{Y}} + \mathbf{Y}_{res}$, Fig. 11.2). In contrast, the "fitted site scores" of eq. 11.18 are obtained by projecting the fitted values of the multiple regressions (matrix $\hat{\mathbf{Y}}$) onto axis *k*; they approximate the fitted data. Either set may be used in biplots; different programs offer one or the other as the default option. These plots may look very different, so users must decide which one they want to obtain and report in published papers. The practical difference between "site scores" and "fitted site scores" is further discussed in the second example below.

6) The correlation r_k between the ordination vectors in spaces **Y** (from eq. 11.17) and **X** (from eq. 11.18) for dimension *k* is called the "species-environment correlation". It measures the strength of the relationship between the two data sets as expressed by each canonical axis *k*. It should be interpreted with caution because a canonical axis with high species-environment correlation may explain but a small fraction of the variation in **Y**, which is given by the amount (or proportion) of variance of matrix **Y** explained by each canonical axis; see example in Table 11.2.

7) The last important information needed for interpretation is the contribution of the explanatory variables **X** to the canonical ordination axes. Either the regression or the correlation coefficients may be considered:

- Matrix **C** of the canonical coefficients,

$$\mathbf{C} = \mathbf{B}\,\mathbf{U} \qquad \textbf{(11.19)}$$

gives directly the weights of the explanatory variables **X** in the formation of the matrix of fitted site scores. The ordination of objects in the space of the explanatory variables can be found directly by computing **XC**; these vectors of site scores are the same as in eq. 11.18. The coefficients in the columns of matrix **C** are identical to the regression coefficients of the ordination scores from eq. 11.18 on the matrix of standardized explanatory variables **X**; they may thus be interpreted in the same way.

• Correlations may also be computed between the variables in **X**, on the one hand, and the ordination vectors, in either space **Y** (from eq. 11.17) or space **X** (from eq. 11.18), on the other hand. The correlations between **X** and the ordination vectors in space **X**, $\mathbf{R_{XZ}} = \text{cor}(\mathbf{X}, \mathbf{Z})$, are used to represent the explanatory variables in biplots.

Biplot
Triplot

8) In RDA, one can draw *biplot diagrams*, called *biplots*, which contain two sets of points as in PCA (Subsection 9.1.4), or *triplot diagrams* (*triplots*) which contain three sets: the site scores (matrices **F** or **Z**, from eqs. 11.17 and 11.18), the response variables from **Y,** and the explanatory variables from **X**. Each pair of sets of points may be drawn in a biplot. Biplots help interpret the ordination of objects in terms of **Y** and **X**. When there are too many objects, or too many variables in **Y** or **X**, separate ordination diagrams for the response and explanatory variables may be drawn and presented side by side. The construction of RDA biplot diagrams is explained in detail in ter Braak (1994); his conclusions are summarized here. As in PCA, two main types of scalings may be used (Table 9.2):

Scalings in RDA

RDA scaling type 1. — The eigenvectors in matrix **U**, representing the scores of the response variables along the canonical axes, are scaled to lengths 1. The site scores in space **X** are obtained from equation $\mathbf{Z} = \hat{\mathbf{Y}}\mathbf{U}$ (eq. 11.18); these vectors have variances equal to λ_k. The site scores in space **Y** are obtained from equation $\mathbf{F} = \mathbf{YU}$; the variances of these vectors are usually slightly larger than λ_k because **Y** contains both the fitted and residual components and has thus more total variance than $\hat{\mathbf{Y}}$. Matrices **Z** and **U**, or **F** and **U**, can be used together in biplots because the products of the eigenvectors with the site score matrices reconstruct the original matrices perfectly: $\mathbf{ZU}' = \hat{\mathbf{Y}}$ and $\mathbf{FU}' = \mathbf{Y}$, as in PCA (Subsection 9.1.4).

Matrix of biplot scores

In scaling type 1, a quantitative explanatory variable **x** is represented in the biplot or triplot using the vector of correlations of **x** with the fitted site scores, $\mathbf{r_{xZ}} = \text{cor}(\mathbf{x}, \mathbf{Z})$, modified by multiplying each correlation by $\sqrt{\lambda_k / \text{Total variance in } \mathbf{Y}}$ where λ_k is the eigenvalue of the corresponding axis *k*. The whole *matrix of biplot scores* in scaling type 1 ($\mathbf{BS}_1$) for the explanatory variables is computed as follows:

$$\mathbf{BS}_1 = (\text{Total variance in } \mathbf{Y})^{-1/2} \mathbf{R_{XZ}} \boldsymbol{\Lambda}^{1/2} \qquad \textbf{(11.20)}$$

This correction accounts for the fact that, in this scaling, the variances of the site scores differ among axes. The correlation matrix $\mathbf{R_{XZ}}$ was obtained in calculation step 7.

The consequences of this scaling, for PCA, are summarized in the central column of Table 9.2. The graphs resulting from this scaling, called *distance biplots* or *triplots*,

focus the interpretation on the ordination of objects because the distances among objects approximate their Euclidean distances in the spaces corresponding to matrices $\mathbf{Y}$ or $\hat{\mathbf{Y}}$.

Distance triplot

The main features of a distance biplot or triplot are the following: (1) Distances among objects in a biplot are approximations of their fitted Euclidean distances. (2) Projecting an object at right angle on a response variable **y** approximates the fitted value (e.g. abundance) of the object along that variable, as in Fig. 9.3a. (3) The angles among variables **y** are meaningless. (4) The angle between two variables **x** and **y** in the biplot reflect their correlation. (5) Binary explanatory variables **x** may be represented as the centroids of the objects possessing state "present" or "1" for that variable. Examples are given in Subsection 11.1.4. Since a centroid represents a "mean object", its relationship to a variable **y** is found by projecting it at right angle on the variable, as for an object. Distances among centroids, and between centroids and individual objects, approximate Euclidean distances.

RDA scaling type 2. — Alternatively, one obtains response variable scores by rescaling the eigenvectors in matrix $\mathbf{U}$ to lengths $\sqrt{\lambda_k}$, using the transformation $\mathbf{U\Lambda}^{1/2}$ as in PCA (eq. 9.10). The site scores in space $\mathbf{X}$ obtained for scaling 1 (eq. 11.18) are rescaled to unit variances using the transformation $\mathbf{Z\Lambda}^{-1/2}$; this is the same transformation as used in PCA (eq. 9.14) to obtain matrix $\mathbf{G}$ of site scores in scaling 2. Likewise, the site scores in space $\mathbf{Y}$ obtained for scaling 1 are rescaled using the transformation $\mathbf{F\Lambda}^{-1/2}$; the variances of these vectors are usually slightly larger than 1 for the reason explained in the case of scaling 1. Matrices $\mathbf{Z\Lambda}^{-1/2}$ and $\mathbf{U\Lambda}^{1/2}$, or $\mathbf{F\Lambda}^{-1/2}$ and $\mathbf{U\Lambda}^{1/2}$, can be used together in biplots because the products of the eigenvectors with the site score matrices reconstruct the original matrices perfectly: $\mathbf{Z\Lambda}^{-1/2}\mathbf{\Lambda}^{1/2}\mathbf{U}' = \hat{\mathbf{Y}}$ and $\mathbf{F\Lambda}^{-1/2}\mathbf{\Lambda}^{1/2}\mathbf{U}' = \mathbf{Y}$, as in PCA (Subsection 9.1.4).

In scaling type 2, a quantitative explanatory variable **x** is represented in the biplot using the vector of correlations of **x** with the fitted site scores, $\mathbf{r_{xZ}} = \text{cor}(\mathbf{x}, \mathbf{Z})$, obtained in calculation step 7, without further transformation. The matrix of biplot scores ($\mathbf{BS}_2$) for the explanatory variables is then:

$$\mathbf{BS}_2 = \mathbf{R_{XZ}} = \text{cor}(\mathbf{X}, \mathbf{Z}) \qquad \textbf{(11.21)}$$

Note that $\text{cor}(\mathbf{X}, \mathbf{Z\Lambda}^{-1/2})$ produces the same correlations as $\text{cor}(\mathbf{X}, \mathbf{Z})$.

The consequences of this scaling, for PCA, are summarized in the right-hand column of Table 9.2. The graphs resulting from this scaling, called *correlation biplots* or *triplots*, focus on the relationships among the response variables (matrix $\mathbf{Y}$ or $\hat{\mathbf{Y}}$).

Correlation triplot

The main features of a correlation biplot or triplot are the following: (1) Distances among objects in the biplot *are not* approximations of their fitted Euclidean distances. (2) Projecting an object at right angle on a response variable **y** approximates the fitted value (e.g. abundance) of the object along that variable. (3) The angle between two variables **x** and **y** in the biplot reflects their correlation. (4) Projecting an object at right angle on a variable **x** approximates the value of that object along the variable.

Table 11.1 Maximum number of non-zero eigenvalues and corresponding eigenvectors that may be obtained from canonical analysis of a matrix of response variables $\mathbf{Y}(n \times p)$ and a matrix of explanatory variables $\mathbf{X}(n \times m)$ using redundancy analysis (RDA) or canonical correspondence analysis (CCA).

	Canonical eigenvalues and eigenvectors	Non-canonical eigenvalues and eigenvectors
RDA	$\min[p, m, n-1]$	$\min[p, n-1]$
CCA	$\min[(p-1), m, n-1]$	$\min[(p-1), n-1]$

(5) Binary explanatory variables may be represented as described above. Their interpretation is done in the same way as in scaling type 1, except for the fact that the distances in the biplot among centroids, and between centroids and individual objects, do not approximate Euclidean distances.

The type of scaling depends on the purpose of the plot: displaying the distances among objects or the correlations among variables. When most explanatory variables are binary, scaling type 1 is probably the most interesting; when most of the variables in set **X** are quantitative, one may prefer scaling type 2. When the first two eigenvalues are nearly equal, the two scalings lead to very similar plots.

9) Redundancy analysis usually does not completely explain the variation in the response variables (matrix **Y**). During the regression step (Fig. 11.2), regression residuals may be computed for each variable **y**; the residuals are the differences between the observed values y_{ij} in matrix **Y** and the corresponding fitted values $\hat{y}_{ij}$ in matrix $\hat{\mathbf{Y}}$. The matrix of residuals ($\mathbf{Y}_{\text{res}}$ in Fig. 11.2) is also a matrix of size $(n \times p)$. Residuals may be analysed by principal component analysis, leading to $\min[p, n-1]$ non-canonical eigenvalues and eigenvectors (Fig. 11.2, bottom). So, the full analysis of matrix **Y** (i.e. the analysis of fitted values and residuals) may lead to more eigenvectors than a principal component analysis of matrix **Y**: there is a maximum of $\min[p, m, n-1]$ non-zero canonical eigenvalues and corresponding eigenvectors, plus a maximum of $\min[p, n-1]$ non-canonical eigenvalues and eigenvectors, the latter being computed from the matrix of residuals (Table 11.1). When the variables in **X** are good predictors of the variables in **Y**, the canonical eigenvalues may be larger than the first non-canonical eigenvalues, but this is not always the case. If the variables in **X** are not good predictors of **Y**, the first non-canonical eigenvalues, computed on the residuals, may be larger than their canonical counterparts.

In the case where **Y** contains a single response variable, redundancy analysis is simply a multiple linear regression analysis. This is why variation partitioning

(Subsection 11.1.11) can be obtained for a single response variable using an R function, ***varpart()***, which was designed for the analysis of multivariate response data.

Different algorithms can be used in computer programs to compute RDA. One may go through the multiple regression and principal component analysis steps described in Fig. 11.2, or calculate the matrix corresponding to $\mathbf{S_{YX}S_{XX}^{-1}S_{YX}'}$ in eq. 11.8 and decompose it into eigenvalues and eigenvectors using standard eigen-analysis (Section 2.9). Computation of the matrix of fitted values $\hat{\mathbf{Y}}$ can be done by QR decomposition, as explained in Subsection 11.1.1, and eigen-decomposition can be replaced by singular value decomposition (SVD, Section 2.11) as shown for PCA (Subsection 9.1.9). Instead of eigen-decomposition or SVD, an iterative algorithm is used in the program CANOCO to calculate the first four canonical eigenvalues and eigenvectors (ter Braak, 1987c).

4 — Numerical examples, simple RDA

As a first example, consider again the data presented in Table 10.6. For RDA, the first five variables were assembled into matrix **Y** whereas the three spatial variables made up matrix **X**. The **Y** variables were standardized at the beginning of the calculations because they were dimensionally heterogeneous. The results of RDA are presented in Table 11.2. There are min[5, 3, 19] = 3 canonical eigenvectors in this example, and 5 non-canonical PCA axes computed from the residuals. This is a case where the canonical analysis is not very successful: the three canonical eigenvalues account together for only 28% (R^2 = 0.2807) of the variation present in the standardized response data **Y**. The first non-canonical eigenvalues are larger than any of the canonical eigenvalues. The correlations shown in Table 11.2 between the two sets of ordination axes (matrices **F** and **Z**) are rather weak. The ordination of objects along the canonical axes (calculation steps 4 and 5 of the previous subsection) as well as the contributions of the explanatory variables to the canonical ordination axes (calculation step 6) are not reported in the table.

A second example was constructed to illustrate the calculation and interpretation of redundancy analysis. In this artificial example, fish have been observed at 10 sites along a transect perpendicular to the beach of a tropical island, with water depths going from 1 to 10 m (Table 11.3). The first three sites are on sand while the other sites alternate between coral and "other substrate". The first six species avoid the sandy area, possibly because there is little food for them there, whereas the last three are ubiquitous. The sums of abundances for the 9 species are in the last row of the table. Species 1 to 6 come in three successive pairs, with distributions forming opposite gradients of abundance between sites 4 and 10. Species 1 and 2 are not associated with a single type of substrate. Species 3 and 4 are found in the coral areas only while species 5 and 6 are found on other substrates only (coral debris, turf, calcareous algae, etc.). The distributions of abundances of the ubiquitous species (7 to 9) have been produced using a random number generator, fitting the frequencies to a predetermined sum; these species will only be used to illustrate CCA in Section 11.2.

Table 11.2 Results of redundancy analysis (selected output). Matrix **Y** contained the first five variables of Table 10.6 and matrix **X**, the last three.

	Canonical axes			Non-canonical axes				
	I	II	III	IV	V	VI	VII	VIII
Eigenvalues (with respect to total variance of the standardized variables in **Y** = 5)								
	0.8044	0.5864	0.0124	1.4517	1.1165	0.5469	0.3715	0.1101
Fraction of total variance in **Y**								
	0.1609	0.1173	0.0025	0.2903	0.2233	0.1094	0.0743	0.0220
Correlations between the ordination vectors in spaces **Y** and **X**								
	0.7996	0.5936	0.1301					
Normalized eigenvectors (the rows correspond to the five standardized variables in matrix **Y**)								
1	0.2977	0.6173	–0.3441	–0.3345	0.5904	–0.1631	–0.5570	–0.4502
2	–0.6286	0.3455	0.0471	0.1753	0.5936	–0.4738	0.1769	0.6010
3	0.1664	0.4049	0.8922	–0.7254	–0.0735	0.3017	–0.2000	0.5808
4	0.6414	–0.2740	0.0928	0.4459	–0.2856	–0.1018	–0.7857	0.3031
5	0.2778	0.5105	–0.2735	0.3638	0.4605	0.8047	–0.0331	0.0832

RDA was computed using the first six species as matrix **Y**. Had the data been real, they would have been subjected to a Hellinger, chord, or chi-square transformation (Section 7.7) prior to RDA, because of the large proportion of zeros in the data. This is not done here in order to simplify the task of readers who would like to replicate the results. These same data, augmented with species 7 to 9, will be analysed using CCA in Section 11.2. Comparison of the RDA results about species 1 to 6 (Tables 11.4 and Fig. 11.3), on the one hand, to the CCA results about species 1 to 9 (Table 11.7 and Fig. 11.9), on the other hand, allows some comparison of the two methods.

The **Y** variables were not standardized: species abundances do not require standardization since they are all in the same physical dimensions. In most ecological studies, it is important to preserve the variances of the individual species in the analyses because abundant and rare species play different roles in ecosystems. Among the **X** variables, the three binary variables coding for substrate types form a collinear group. Including all three in the cross-product matrix [**X**'**X**] would prevent its inversion because the matrix would be singular (Section 2.8); this would jeopardize

Table 11.3 Artificial data set representing observations (fish abundances) at 10 sites along a tropical reef transect. The variables are further described in the text.

Site No.	Sp. 1	Sp. 2	Sp. 3	Sp. 4	Sp. 5	Sp. 6	Sp. 7	Sp. 8	Sp. 9	Depth (m)	Substrate type Coral	Sand	Other
1	1	0	0	0	0	0	2	4	4	1	0	1	0
2	0	0	0	0	0	0	5	6	1	2	0	1	0
3	0	1	0	0	0	0	0	2	3	3	0	1	0
4	11	4	0	0	8	1	6	2	0	4	0	0	1
5	11	5	17	7	0	0	6	6	2	5	1	0	0
6	9	6	0	0	6	2	10	1	4	6	0	0	1
7	9	7	13	10	0	0	4	5	4	7	1	0	0
8	7	8	0	0	4	3	6	6	4	8	0	0	1
9	7	9	10	13	0	0	6	2	0	9	1	0	0
10	5	10	0	0	2	4	0	1	3	10	0	0	1
Sum	60	50	40	30	20	10	45	35	25				

the calculation of the regression coefficients (eq. 11.9) and of the matrix of fitted values $\hat{\mathbf{Y}}$ (eq. 11.11). It is not necessary, however, to eliminate one of the dummy variables: in well-designed programs for canonical analysis, the last dummy variable is automatically eliminated from the calculations leading to $\hat{\mathbf{Y}}$, but its position in the ordination diagram is estimated in the final calculations. A group of dummy variables coding for a qualitative variable, like the substrate types here, can be replaced by a single factor-type variable in R functions such as VEGAN's ***rda()***.

Results of the analysis are presented in Table 11.4. Scaling type 1 was selected for the biplot in order to illustrate the extra calculation step required to transform the correlations into biplot scores for scaling type 1. The data could have produced 3 canonical axes and up to 6 non-canonical eigenvectors. In this example, only 4 of the 6 non-canonical axes had variances larger than 0. An overall test of significance (Subsection 11.1.2) showed that the canonical relationship between matrices **X** and **Y** was very highly significant ($p = 0.001$ after 999 permutations). The canonical axes explained 66%, 22% and 8% of the variance of the response data, respectively, for a total R^2 of 0.9597 and $R^2_a = 0.9396$. The three canonical axes were all significant ($p < 0.05$) and displayed strong species-environment correlations ($r = 0.999$, 0.997, and 0.980, respectively).

In Table 11.4, the eigenvalues are first shown with respect to the total variance of matrix **Y**, as is customary in principal component analysis. They are also presented as proportions of the total variance of **Y**; these are the eigenvalues provided by CANOCO for PCA and RDA. The species and sites are scaled for a distance triplot (RDA scaling type 1). The eigenvectors, normalized to length 1, provide the "species scores". The

Table 11.4 Results of redundancy analysis of the data in Table 11.3 (selected output). Matrix **Y**: species 1 to 6. Matrix **X**: depth and substrate classes.

	Canonical axes			Non-canonical axes			
	I	II	III	IV	V	VI	VII
Eigenvalues (with respect to total variance of **Y** = 112.88889)							
	74.52267	24.94196	8.87611	4.18878	0.31386	0.03704	0.00846
Fraction of total variance of **Y**							
	0.66014	0.22094	0.07863	0.03711	0.00278	0.00033	0.00007
Cumulative fraction of total variance of **Y** accounted for by axes 1 to *k*							
	0.66014	0.88108	0.95971	0.99682	0.99960	0.99993	1.00000
Normalized eigenvectors ("species scores"): mat. **U** for canonical, $\mathbf{U}_{res}$ for non-canonical portions (Fig. 11.2)							
Species 1	0.30127	–0.64624	0.39939	–0.00656	–0.40482	0.70711	–0.16691
Species 2	0.20038	–0.47265	–0.74458	0.00656	0.40482	0.70711	0.16691
Species 3	0.74098	0.16813	0.25690	–0.68903	–0.26668	0.00000	0.67389
Species 4	0.55013	0.16841	–0.26114	0.58798	0.21510	0.00000	0.68631
Species 5	–0.11588	–0.50594	0.29319	0.37888	–0.66624	0.00000	0.12373
Species 6	–0.06292	–0.21535	–0.25679	–0.18944	0.33312	0.00000	–0.06187
Matrix **Z** for the canonical part ("fitted site scores", eq. 11.18) and **F** for the non-canonical part (eq. 9.4)							
Site 1	–6.79498	5.49498	2.24897	0.24712	1.14353	0.23570	0.01271
Site 2	–6.96197	5.91719	0.63774	0.00000	0.00000	–0.47140	0.00000
Site 3	–7.12895	6.33941	–0.97349	–0.24712	–1.14353	0.23570	–0.01271
Site 4	–3.55205	–6.52301	4.39356	2.14250	–0.28230	0.00000	0.00141
Site 5	12.69996	0.24686	3.17159	–3.80923	–0.14571	0.00000	0.10360
Site 6	–3.88603	–5.67858	1.17109	0.71417	–0.09410	0.00000	0.00047
Site 7	12.36599	1.09129	–0.05088	0.22968	0.08889	0.00000	–0.22463
Site 8	–4.22000	–4.83415	–2.05138	–0.71417	0.09410	0.00000	–0.00047
Site 9	12.03201	1.93572	–3.27335	3.57956	0.05682	0.00000	0.12103
Site 10	–4.55398	–3.98972	–5.27384	–2.14250	0.28230	0.00000	–0.00141
Correlations of environmental variables with the **Z** site scores							
Depth	0.42265	–0.55914	–0.71325				
Coral	0.98850	0.15079	–0.01178				
Sand	–0.55652	0.81760	0.14771				
Other subs.	–0.40408	–0.90584	–0.12715				
Biplot scores of environmental variables							
Depth	0.34340	–0.26282	–0.20000				
Coral	0.80314	0.07088	–0.00330				
Sand	–0.45216	0.38431	0.04142				
Other subs.	–0.32831	–0.42579	–0.03565				
Centroids, in the triplot, of the sites with code "1" for the BINARY environmental variables							
Coral	12.36599	1.09129	–0.05088				
Sand	–6.96197	5.91719	0.63774				
Other subs.	–4.05301	–5.25636	–0.44014				

"fitted site scores" (matrix **Z**) are obtained from eq. 11.18. They provide the ordination of the objects, computed from $\hat{\mathbf{Y}}$, in the space of the explanatory variables **X**. These axes are orthogonal to one another because they directly result from the PCA of $\hat{\mathbf{Y}}$. The "site scores" (matrix **F**, not shown) in the space of **Y** would be obtained by eq. 11.17. The columns of matrix **F** are, however, not orthogonal to one another because **Y** contains the "residual" components of the multiple regressions (Fig. 11.2). Both the "site scores" (matrix **F**) and "fitted site scores" (matrix **Z**) may be used in RDA triplots.

Correlations of the environmental variables with the ordination vectors can be obtained in two forms: with respect to either the "site scores" (eq. 11.17) or the "fitted site scores" (eq. 11.18). The latter set of correlations is used to draw triplots containing the sites as well as the variables from **Y** and **X**, as done in Fig. 11.3. There were three binary variables in Table 11.3. Each such variable may be represented by the centroid of the sites possessing state "1" for that variable (or else, the centroid of the sites possessing state "0"). These three variables are represented by both arrows (correlations) and symbols (centroids) in Fig. 11.3 to show the difference between these representations. In real-case triplots, only one of the two representations is used.

The fitted site scores in Table 11.4 have much larger ranges of values than the species scores and the biplot scores of environmental variables. Drawing triplots from these tables of values would produce graphs in which the arrows representing the species and environmental variables would be minute and clustered in the centre of the graph. Two strategies are used in computer software: either the tables of output results are modified to make the three sets of values to be drawn (species, sites, environmental variables) commensurable in the graph (this is the case in CANOCO and in VEGAN's function ***rda()***), or the output tables are those produced by the equations of Subsection 11.1.3 but the species and environmental variable arrows are drawn using a different scale than for the site scores (as done in Fig. 11.3).

5 — RDA and CCA of community composition data

Different approaches are available for the canonical analysis of community composition data (Fig. 11.4): the classical approaches (RDA and CCA), transformation-based RDA (tb-RDA), and distance-based RDA (db-RDA). The three approaches are discussed here in turn.

In the classical approach (Fig. 11.4a), the species-environment relationship is analysed by RDA (this section) or by CCA (Section 11.2). In the early applications of canonical analysis to community ecology, the latter was considered preferable for species data tables sampled in highly diversified regions ("long gradients"), which contain many zeros. This is the case, for example, when sampling communities along extensive spatial or temporal gradients, where the species composition may differ greatly between the two ends of the gradient. For groups of sites that were fairly homogeneous in species composition ("short gradients"), RDA was considered appropriate. A wider array of options is now available.

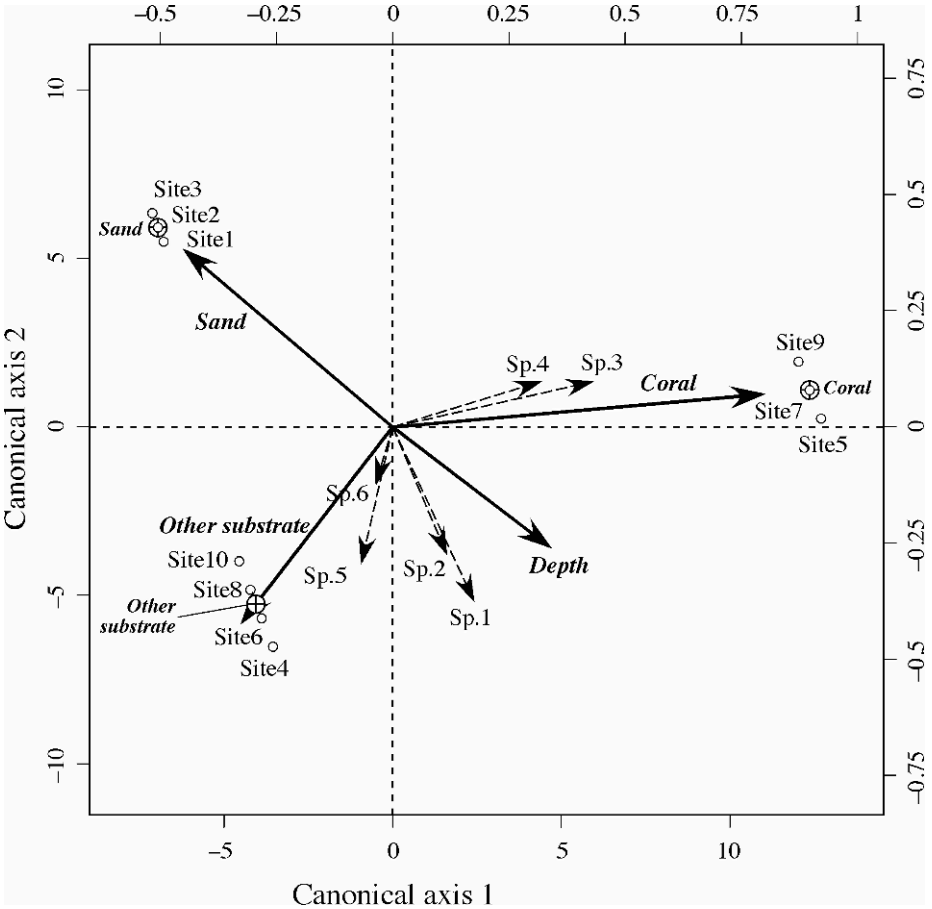


Figure 11.3 RDA triplot of the data in Table 11.3, scaling 1; the numerical results are in Table 11.4. Open circles represent the sites; the site numbers correspond to the site water depths (in m). Dashed arrows are the species. Full-line arrows represent the environmental variables. The sites are positioned in the diagram using the lower and left-hand scales, whereas the species and environmental variables are positioned using the top and right-hand scales. The "centroids of the sites with code 1 for the [three] binary environmental variables" are represented by crossed circles. Binary environmental variables are usually represented by *either* arrows *or* symbols, not both as in this triplot.

Like PCA (Fig. 9.8), RDA can be made to preserve some distance that is appropriate to study composition data along gradients, instead of the Euclidean distance. Figure 11.4b shows that composition data can be transformed using the transformations described in Section 7.7. This is the transformation-based RDA

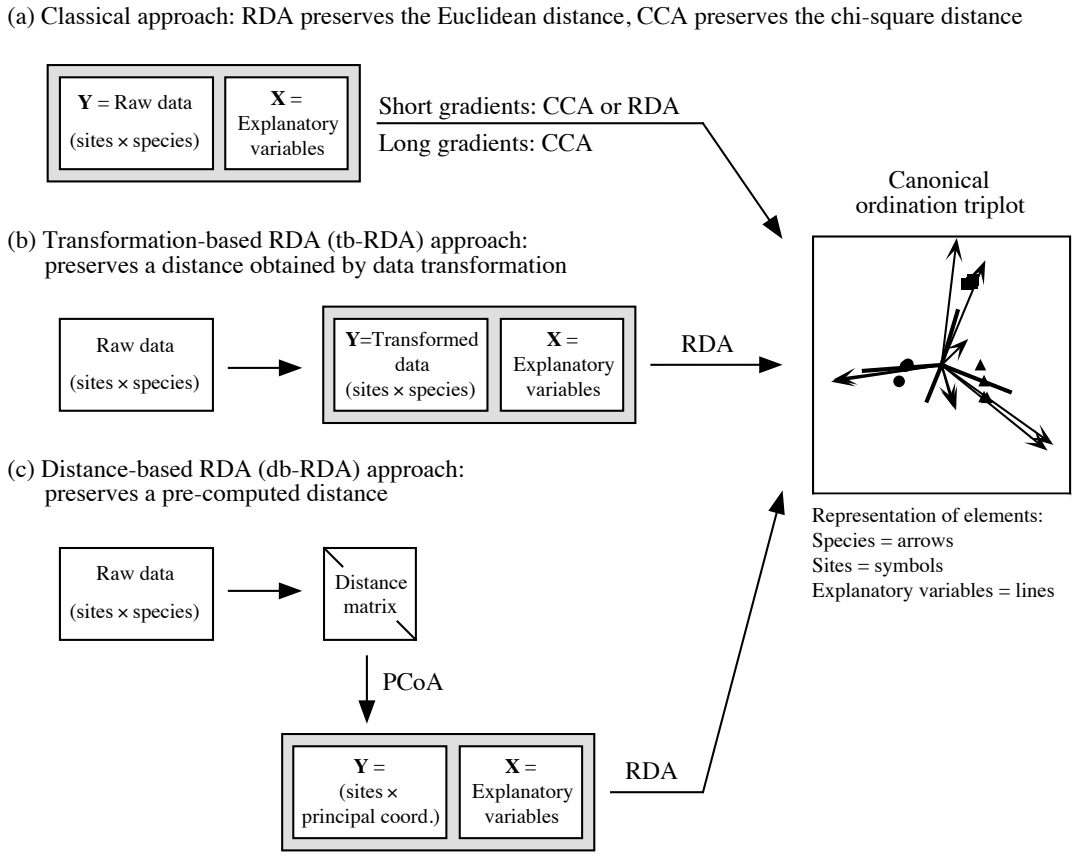


Figure 11.4 Comparison of (a) classical RDA and CCA, and (b and c) alternative approaches forcing RDA to preserve other distances adapted to community composition data. Modified from Legendre & Gallagher (2001).

tb-RDA

(Legendre & Gallagher, 2001), or tb-RDA, approach. RDA computed on data transformed by these equations will actually preserve the chord, profile, Hellinger, chi-square distance or chi-square metric among sites, depending on the transformation used.

One can also (Fig. 11.4c) compute one of the distance functions appropriate for community composition data (Table 7.4), carry out a principal coordinate analysis (PCoA) of the distance matrix, and use *all* the PCoA eigenvectors as input into a RDA.

db-RDA

This is the distance-based RDA, or db-RDA, approach advocated by Legendre & Anderson (1999).

The db-RDA approach must be used in analyses involving distance functions that cannot be obtained by a data transformation followed by RDA (tb-RDA). Among these are most of the coefficients designed for binary data, e.g. Jaccard ($\sqrt{1 - S_7}$) and Sørensen (D_{13} or $\sqrt{1 - S_8}$), as well as quantitative distance measures like the asymmetric Gower coefficient ($\sqrt{1 - S_{19}}$), the geodesic metric (D_4), Whittaker (D_9), Canberra (D_{10}), Clark (D_{11}), percentage difference (D_{14}), and mean character difference modified for species data D_{19}. Distance coefficients intended for other types of data, e.g. symmetric Gower ($\sqrt{1 - S_{15}}$), Estabrook-Rogers ($\sqrt{1 - S_{16}}$), and the generalized Mahalanobis distance for groups of observations, can also be used in canonical ordination through db-RDA. Published studies involving db-RDA include Anderson (1999), Geffen *et al*. (2004) and Lear *et al*. (2008).

6 — Partial RDA

Partial RDA is the analysis of response variables **Y** by explanatory variables **X** in the presence of additional explanatory variables, **W**, called covariables. In partial RDA, the linear effects of the explanatory variables **X** on the response variables **Y** are adjusted for the effects of the covariables **W**, as was done in partial linear regression (Subsection 10.3.5). Partial RDA was first proposed by Davies & Tso (1982, their Section 10.3).

In multiple regression, the partial regression of **y** on **X** in the presence of covariables **W** can be computed in two different ways that were described in Subsection 10.3.5. After computing the residuals of **y** on **W** (noted $\mathbf{y}_{res|\mathbf{W}}$) and the residuals of **X** on **W** (noted $\mathbf{X}_{res|\mathbf{W}}$), one could either (1) regress $\mathbf{y}_{res|\mathbf{W}}$ on $\mathbf{X}_{res|\mathbf{W}}$ or (2) regress **y** on $\mathbf{X}_{res|\mathbf{W}}$. The same partial regression coefficients were obtained in both cases. Between calculation methods, the vectors of fitted values only differed by the value of the intercept of the regression of **y** on $\mathbf{X}_{res|\mathbf{W}}$, which was also the mean of **y**. The R^2 of the first analysis was the partial R^2, whereas that of the second analysis was the semipartial R^2; their square roots were the partial and semipartial correlation coefficients described in Box 4.1.

The same two approaches can be used for partial RDA, which is the extension of partial linear regression to a multivariate response matrix **Y**. First, one computes the residuals of **Y** on **W** (noted $\mathbf{Y}_{res|\mathbf{W}}$) and the residuals of **X** on **W** ($\mathbf{X}_{res|\mathbf{W}}$). Then, one can compute either (1) the RDA of $\mathbf{Y}_{res|\mathbf{W}}$ by $\mathbf{X}_{res|\mathbf{W}}$ or (2) the RDA of **Y** by $\mathbf{X}_{res|\mathbf{W}}$. The two approaches produce the same canonical eigenvalues, eigenvectors and axes. In both approaches, the significance of the canonical axes can be tested using the forward and marginal methods described in Subsection 11.1.2 (paragraph 4). In partial RDA, the canonical axes (matrix **Z**) are linear combinations of the residuals of the explanatory variables **X**, $\mathbf{X}_{res|\mathbf{W}}$, and are orthogonal to the covariables in **W**. The R^2 obtained in the first approach is the partial canonical R^2, whereas that of the second analysis is the semipartial canonical R^2; these two statistics are described in the next subsection. In computer programs, it is customary to use as matrix **F** (eq. 11.17) the matrix obtained from the RDA of $\mathbf{Y}_{res|\mathbf{W}}$ by $\mathbf{X}_{res|\mathbf{W}}$, not the matrix computed in the RDA of **Y** by $\mathbf{X}_{res|\mathbf{W}}$.

Table 11.5 Algorithm for partial RDA in the R language. This is the skeleton of the algorithm used in the ***rda()*** function of the VEGAN package. (Jari Oksanen, personal communication.)

```
pRDA <- function(Y, X = NULL, W = NULL, scale.Y = FALSE)
{
    Y <- scale(as.matrix(Y), center = TRUE, scale = scale.Y)

    if (!is.null(W)) {                  # If covariables W are present
        W <- scale(as.matrix(W), center = TRUE, scale = FALSE)
        Y <- qr.resid(qr(W), Y)
    }
    if (!is.null(X)) {                  # If there are explanatory variables X
        X <- scale(as.matrix(X), center = TRUE, scale = FALSE)
        X <- cbind(X, W)
        Q <- qr(X)
        RDA <- svd(qr.fitted(Q, Y))
        RDA$w <- Y %*% RDA$v %*% diag(1/RDA$d)
        Y <- qr.resid(Q, Y)
    } else {                            # No explanatory variables X nor covariables W
        RDA <- NULL
    }
    RES <- svd(Y)                       # PCA of the residuals
    list(RDA = RDA, RES = RES)
}
```

Table 11.5 presents a very short algorithm for partial RDA, designed by Prof. Jari Oksanen (University of Oulu, Finland). This algorithm handles different cases. (1) If there are covariables (**W**) in the analysis, **Y** is regressed on **W** and residuals $\mathbf{Y}_{\text{res}|\mathbf{W}}$ are computed using QR decomposition (function ***qr()*** in R), which is faster than multivariate regression by matrix inversion (eqs. 10.16 and 11.11). (2) If there are explanatory variables (**X**), RDA is the eigen-decomposition (by SVD through function ***svd()*** in R, Section 2.11) of the fitted values of the multivariate regression of **Y** on **X**. If **X** and **W** are both present, regressing $\mathbf{Y}_{\text{res}|\mathbf{W}}$ on the column concatenation of **X** and **W** produces the same result as a partial regression of $\mathbf{Y}_{\text{res}|\mathbf{W}}$ on $\mathbf{X}_{\text{res}|\mathbf{W}}$ because $\mathbf{Y}_{\text{res}|\mathbf{W}}$ is orthogonal to **W**. (3) A PCA of the residuals is computed. (4) If there are neither explanatory variables **X** nor covariables **W** in the analysis, the result only contains a PCA of **Y** and no RDA is computed.

7 — Statistics in partial RDA

Partial *F*-statistic

For analysis in the presence of **W** containing q covariables (partial RDA), the partial F-statistic is constructed as follows (ter Braak & Smilauer, 2002):

$$F = \frac{\mathrm{SS}(\mathbf{Y}_{\mathrm{fit}}) / m}{\mathrm{SS}(\mathbf{Y}_{\mathrm{res}}) / (n - m - q - 1)} \qquad \textbf{(11.22)}$$

There are several ways of computing the sum of squares of the fitted values $\mathrm{SS}(\mathbf{Y}_{\mathrm{fit}})$ and residuals $\mathrm{SS}(\mathbf{Y}_{\mathrm{res}})$ in the partial RDA case. The most convenient are the following:

$$\mathrm{SS}(\mathbf{Y}_{\mathrm{fit}}) = \mathrm{SS}(\mathbf{Y}_{\mathrm{fit}|(\mathbf{X}+\mathbf{W})}) - \mathrm{SS}(\mathbf{Y}_{\mathrm{fit}|\mathbf{W}})$$

and

$$\mathrm{SS}(\mathbf{Y}_{\mathrm{res}}) = \mathrm{SS}(\mathbf{Y}) - \mathrm{SS}(\mathbf{Y}_{\mathrm{fit}|(\mathbf{X}+\mathbf{W})})$$

where (**X**+**W**) designates the concatenation of **X** and **W** in a single matrix; this is obtained by the operation ***cbind(*****X**,**W***)* in the R language. $\mathbf{Y}_{\mathrm{fit}}$ was noted $\hat{\mathbf{Y}}$ in eq. 11.3 which did not involve covariables **W**.

Semipartial R^2

The semipartial R^2, $R^2_{\mathbf{Y}|\mathbf{X}_{\mathrm{res}|\mathbf{W}}}$, is the proportion of explained variation with respect to the total variation in **Y**. This is the most widely used R^2 statistic in partial RDA because the denominator, which is the total variation in **Y**, forms a common basis for comparisons among analyses using different explanatory matrices **X** and different matrices of covariables **W**. It is the R^2 of the simple RDA of **Y** by $\mathbf{X}_{\mathrm{res}|\mathbf{W}}$:

$$R^2_{\mathbf{Y}|\mathbf{X}_{\mathrm{res}|\mathbf{W}}} = \frac{\mathrm{SS}(\mathbf{Y}_{\mathrm{fit}})}{\mathrm{SS}(\mathbf{Y})} \qquad \textbf{(11.23)}$$

Partial R^2

The partial R^2, $R^2_{\mathbf{Y}_{\mathrm{res}|\mathbf{W}}|\mathbf{X}_{\mathrm{res}|\mathbf{W}}}$, is the proportion of explained variation with respect to the total variation in $\mathbf{Y}$ residualized on the matrix of covariables **W**. Although more rarely used than the semipartial R^2, it is computed as the R^2 of the simple RDA of $\mathbf{Y}_{\mathrm{res}|\mathbf{W}}$ by $\mathbf{X}_{\mathrm{res}|\mathbf{W}}$:

$$R^2_{\mathbf{Y}_{\mathrm{res}|\mathbf{W}}|\mathbf{X}_{\mathrm{res}|\mathbf{W}}} = \frac{\mathrm{SS}(\mathbf{Y}_{\mathrm{fit}})}{\mathrm{SS}(\mathbf{Y}_{\mathrm{res}|\mathbf{W}})} \qquad \textbf{(11.24)}$$

8 — Tests of significance in partial RDA

Permutation test

Tests of significance in partial RDA, using the F-statistic described in eq. 11.22, involve either permutation of the raw data, unrestricted permutation of the residuals of the reduced model (a method proposed by Freedman & Lane, 1983), or unrestricted permutation of the residuals of the full model (a method proposed by ter Braak, 1990, 1992). These methods are described in Anderson & Legendre (1999) for multiple linear regression, which is RDA with a single response variable.

Permute raw data

- In permutation of the raw data (method = "direct" in VEGAN's ***permutest.cca()***), the rows of **Y** are permuted at random to produce the matrix of permuted response data **Y***. This permutation method is used in simple RDA. It can also be used in partial RDA when the covariables do not contain outlying values, e.g. when they represent experimental factors (Subsection 11.1.10, point 4).

Permute residuals of reduced model

- In permutation of the residuals of the reduced model (method = "reduced" in VEGAN's ***permutest.cca()***), one computes the matrix of fitted values $\mathbf{Fit_{Y|W}}$ and the matrix of residuals $\mathbf{Res_{Y|W}}$ of the multivariate regression of **Y** on the matrix of covariables **W**. The rows of $\mathbf{Res_{Y|W}}$ are permuted, producing matrix $\mathbf{Res^*_{Y|W}}$. The matrix of permuted response data, **Y***, is obtained by adding $\mathbf{Fit_{Y|W}}$ (unpermuted) to $\mathbf{Res^*_{Y|W}}$.

Permute residuals of full model

- In permutation of the residuals of the full model (method = "full" in VEGAN's ***permutest.cca()***), one computes the matrix of fitted values $\mathbf{Fit_{Y|XW}}$ and the matrix of residuals $\mathbf{Res_{Y|XW}}$ of the multivariate regression of **Y** on the matrix obtained by concatenation of **X** and **W** by columns into a single matrix. The rows of $\mathbf{Res_{Y|XW}}$ are permuted, producing matrix $\mathbf{Res^*_{Y|XW}}$. The matrix of permuted response data, **Y***, is obtained by adding $\mathbf{Fit_{Y|XW}}$ (unpermuted) to $\mathbf{Res^*_{Y|XW}}$.

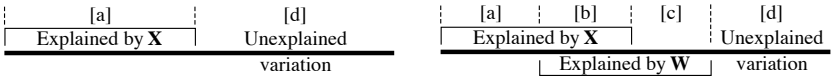
Permutation of the residuals of the reduced and full models were found by Anderson and Legendre (1999) to produce equivalent results. Permutation of the raw data should not be used in partial RDA when the covariables contain outliers. It can, however, be used when partial RDA is used as a form of 2-way MANOVA (Subsection 11.1.10, point 4): in tests of individual factors or the interaction, matrix **W** contains variables coding for the factors or the interaction, and these variables do not have outlier values.

Restricted permutation

Besides these methods, one can also permute the rows of **Y** in a way imposed by the logic of the problem at hand. The most important methods of restricted permutation are: permutation within the levels of a factor or block which is used as a covariable in the study, loop permutation along a time series, and toroidal permutation of the points on a geographic surface (Lotwick & Silverman, 1982).

Methods of permutation of raw data or residuals are compared in Table 11.6 in terms of the permuted portions of variation, in the presence or absence of covariables **W**. *Without covariables*, permutation of raw data involves fraction [a + d] of variation partitioning (Subsection 10.3.5) whereas permutation of residuals of the full model involves [d]. No residual can be computed under a reduced model in the absence of covariables; the method becomes a permutation of raw data. *With covariables*, permutation of residuals may only involve the residuals of the reduced model of the covariables (fraction [a + d]), or the residuals of the full model of the explanatory variables and covariables (fraction [d]). Permutation of the raw data may result in unstable (often inflated) type I error when the covariable contains outliers. This does not occur, however, when using restricted permutations of raw data within groups of a qualitative covariable, which produces an exact test.

Table 11.6 Tests of statistical significance in canonical analysis. Comparison of the methods of permutation of raw data or residuals in terms of the permuted fractions of variation, in the presence or absence of a matrix of covariables **W**. Fractions of variation are noted as in Fig. 10.10: [a] is the variation of matrix **Y** explained by **X** alone, [c] the variation explained by **W** alone, [b] the variation explained jointly by **X** and **W**, and [d] the residual variation.



	Without covariables	With matrix **W** of covariables
Permute raw data	Permute [a + d]	Permute [a + b + c + d]
Permute residuals:		
• reduced model	Equivalent to permuting raw data	Permute [a + d]
• full model	Permute [d]	Permute [d]

9 — Numerical example, partial RDA

Partial RDA provides an answer to the question: what is the partial contribution of one set of explanatory variables when controlling for the effect of another set?

Example 1. — Consider the data in Table 11.3. In that data table, the species can be analysed with respect to *substrate types* while controlling for the effect of *depth*, which is correlated with substrate types. The semipartial R^2 of the analysis is 0.73271; the partial effect of substrate types is highly significant (p = 0.001 after 999 random permutations of the residuals of the reduced model). The two canonical axes produce the triplot shown in Fig. 11.5a.

Example 2. — The converse analysis of the partial effect of *depth* on the distributions of species across the sites while controlling for *substrate types* is also interesting. The semipartial R^2 of this analysis is 0.08274. This is a much weaker effect than that of substrate types, but the partial effect of depth remains significant (p = 0.002 after 999 random permutations of the residuals of the reduced model). A single canonical axis (abscissa of the triplot, Fig. 11.5b) is produced, with the explanatory variable *depth* pointing to the right. Since there is no second canonical axis available, the first axis of the PCA of the residual variation is used as the ordinate of the diagram. This axis separates the coral sites 5, 7 and 9 from the other sites.

These two effects will be considered jointly within the framework of variation partitioning in Subsection 11.1.11 below.

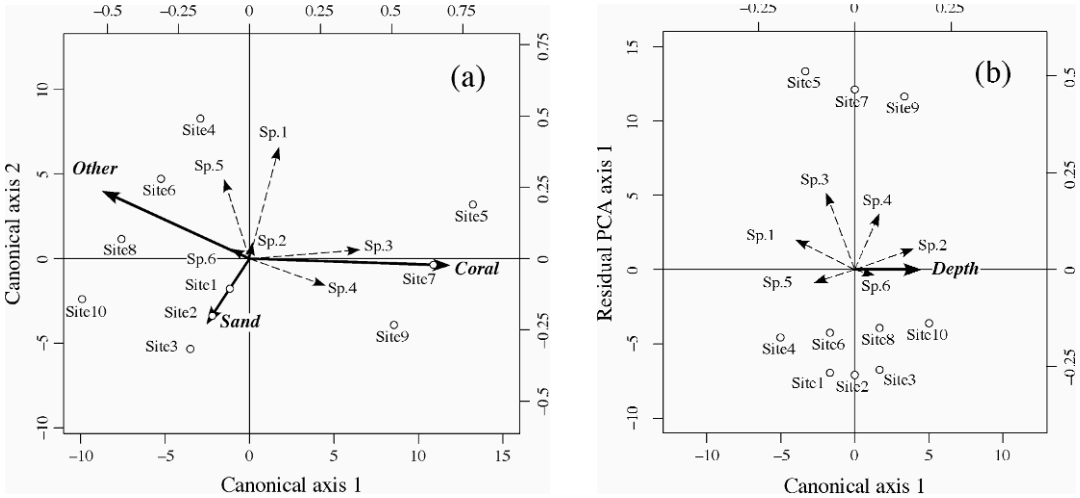


Figure 11.5 Partial RDA triplots of the data in Table 11.3. (a) The explanatory matrix **X** is substrate types, the covariable **W** is depth; (b) the explanatory matrix **X** is depth, the covariable **W** is substrate types. The sites are represented by open circles, the species by dashed arrows, and the explanatory variables in **X** by bold arrows.

10 — Some applications of partial RDA

Partial canonical analysis can be used to investigate a variety of problems. Here are some examples. In most of these applications, CCA (Section 11.2) can be used instead of RDA when **Y** contains frequency data and one wants the analysis to preserve the chi-square distance instead of the Euclidean distance.

Control for effect of **W**

1. *Control for well-known linear effects*. — Consider the case where **W** contains variables whose effects on **Y** are well understood. One wants to control for these well-known effects when analysing the effect of a set of variables **X** on **Y**. For example, one may want to control for the well-known co-variation between phytoplankton assemblages and salinity in a river estuary when analysing the linear effect of nutrient concentrations on phytoplankton. Partial RDA should be used in that case.

Partial effect of a variable

2. *Isolate the effect of a single explanatory variable or factor.* — After conducting a standard RDA as in Subsections 11.1.4, one may want to isolate the partial effect of a single explanatory variable, as in the two examples presented in Subsection 11.1.9 (example 1: a factor with 3 levels; example 2: a quantitative variable). Using all the other explanatory variables as covariates produces a single canonical axis that represents the partial effect of a single quantitative explanatory variable on **Y**. The

corresponding canonical eigenvalue divided by the total variance of **Y** quantifies the partial fraction of the variation of **Y** that is accounted for by that variable (semipartial R^2). The effect of a factor with more than two levels can be isolated by the same method, but then more than one canonical axis are produced because a factor with k levels produces $(k - 1)$ canonical axes.

Related samples

3. *Analysis of related samples*. — Ecological sampling often results in related samples (Box 1.1), where each observation at a site shares some properties with observations at other sites. This is the case, for instance, when sampling different lakes at several depths, the same in all lakes, to study the variation in zooplankton composition. A large portion of the variation in community composition may be associated with the different depths, possibly more than among lakes. Partial RDA offers a way to take this source of variation into account in the analysis of the species-environment relationships. Depth can be coded as a factor or a series of dummy variables, or else as Helmert or polynomial contrasts (Subsection 1.5.7). (Ecologists usually do not hypothesize that zooplankton composition is linearly related to depth, so the covariable depth structuring the sample should be treated as a multi-level factor instead of a quantitative variable.) Including the coding variables in the analysis as a matrix of covariables **W** will effectively control for the effect of the structuring variable. The semipartial R^2 will correctly estimate the partial effect of the environmental variables included in the analysis while controlling for the effect of the structuring variables. Carrying out the analysis for one factor (lakes in this example) while controlling for the effect of the other (depths), and then the opposite, is a form of two-way analysis-of-variance without replication.

Related samples are also obtained when sampling a single lake at different dates and at several depths, or a set of lakes at different dates, the same for all lakes. One may wish to control for the effect of the sampling dates in an analysis of the effect of depths, or lakes, on species composition, bacterial production, or other response variables of interest. As in the previous paragraph, this can be done by using the variable(s) describing the sampling dates as covariable(s) in the analysis. Dates may be represented by dummy variables, or by a quantitative variable whose effect on **Y** is assumed to be linear, or by a sine transformation of the "day of year" (also called "ordinal date", and often "Julian date"[*]), etc. The analysis will effectively control for the effect of dates (days, weeks, years, …) if they only affect the means of the response variables and nothing else. If there is an interaction between sampling dates and the other environmental or spatial variables included in the analysis, the effect of dates cannot be controlled through this simple approach. In the presence of an interaction, the interaction terms must remain in the analysis for the model to be valid (see

[*] The "day of year", also called "ordinal date", is a calendar date starting on 1st January and ranging between 001 and 366. The "Julian day" is used in the "Julian date" system of time measurement, mostly by the astronomy community, where the interval of time is stated in days and fractions of a day since 1st January 4713 BC Greenwich noon. The use of "Julian date" to refer to the day of year, although technically incorrect, is widespread in ecology and other natural sciences. Readers may check the entries "Julian day" and "ordinal date" on Wikipedia.

paragraph 4 below). How to test the space-time interaction in the absence of replication is described in Subsection 14.5.1.

In the same way, one can control for the effect of the sampling locations. Sampling locations may be represented by dummy variables, or by a trend-surface polynomial (Chapter 13) or a set of spatial eigenfunctions (Chapter 14) derived from the geographic coordinates of the sites. The caveat of the previous paragraph concerning interactions applies here as well.

MANOVA by RDA

4. *MANOVA by RDA*. — Partial canonical analysis may be used, instead of MANOVA, to analyse a multivariate response data matrix **Y** in cross-factor experimental designs, including tests of significance for the main effects and the interaction term. For a single experimental factor, the analysis can be conducted using simple RDA or CCA. For two or more factors and their interactions, partial RDA or CCA must be used.

In MANOVA by RDA involving two or more crossed factors, the factors and their interactions are coded by Helmert contrasts (Subsection 1.5.7). The interaction between factors A and B, for example, is represented by a series of variables obtained by the Hadamard product of each Helmert variables coding for factor A by each Helmert variable coding for factor B. Three partial RDAs are necessary to conduct an analysis involving two crossed factors:

- Test the interaction through a RDA of **Y** with the interaction variables in the explanatory matrix **X** and the Helmert variables coding for factors A and B together in the matrix of covariables **W**. If the interaction is significant, analyse separately the effect of factor A in each class of factor B, and conversely the effect of factor B in each class of factor A, because a significant interaction indicates that the effects of factor A on **Y** depend on the levels of factor B, and conversely. If the interaction is not significant, proceed with the next two steps.

- Test the effect of factor A on **Y** through a RDA of **Y** with the variables coding for A in **X** in the presence of a matrix of covariables **W** containing all variables coding for B and the interaction.

- Test the effect of factor B on **Y** through a RDA of **Y** with the variables coding for B in **X** in the presence of a matrix of covariables **W** containing all variables coding for A and the interaction.

The results of the three tests of significance can be assembled in a MANOVA table.

The condition of homogeneity of the variance-covariance matrices applies to this form of MANOVA, as it does to regular MANOVA. It can be tested by the function ***betadisper()*** of the VEGAN package, which implements the testing method described by Anderson (2006). A fully worked out example of MANOVA by RDA, including a test of homogeneity of the multivariate dispersion, is given in Section 6.3.2.8 of Borcard *et al*. (2011).

As stated in Subsection 11.1.5, when **Y** is a matrix of species presence-absence or abundance data, one can either transform **Y** prior to MANOVA by RDA using the transformations described in Section 7.7 (transformation-based RDA, tb-RDA) to force the partial RDA to preserve the distance that is implicit in the transformation, or use partial CCA to preserve the chi-square distance among sites. Else, one can use the distance-based RDA method (db-RDA, Subsection 11.1.5) to preserve some other distance function appropriate for community data.

Principal response curves

5. *Principal response curves (PRC).* — *Principal response curves* is another form of MANOVA; it was developed by van den Brink and co-authors (1998, 1999, 2003, 2009) to analyse the results of experiments conducted over time, that involved multivariate response data (e.g. community composition data). PRC is a special case of RDA with a single factor for treatments and a single factor representing the time series of repeated observations. The method studies the changes in the multivariate (e.g. species) response variables associated with the treatments over time. In this type of analysis, one is interested in displaying the values of the coefficients (contrasts against the control level) computed for the first RDA axis representing the effects of treatment along time. Significance of the canonical relationship and of the first axis can be tested when there is replication in the experimental design. This is an omnibus test: H_0 corresponds to 'no treatment effect'. H_1 includes all functional forms that the treatment effects can take, i.e. main effect and/or interaction. No effect at all produces coinciding treatment lines in the plot. One can also test separately the effect of the main factor (treatment) and, when there is replication, the treatment-time interaction. A significant interaction indicates that the treatment levels had different effects on the response data at different times; it is displayed as non-parallel or crossing treatment lines in the plot.

Partial PCA

6. *Partial PCA.* — Partial principal component analysis is the PCA of a response data table **Y** residualized on a set of explanatory variables. This method allows researchers to examine the multivariate structure of the data after removing the effect of the **X** variables on **Y**, which may already be well understood, by computing residuals. Note that the results of a partial PCA differ from those of a partial RDA.

The three tables represented at the bottom of Fig. 11.2 illustrate how partial PCA is carried out: the residuals of **Y** by **X** are computed, followed by a PCA of the matrix of residuals. Alternatively, since RDA of **Y** by **Y** is a PCA of **Y**, as shown in Subsection 11.1.3, partial PCA can be obtained by computing a partial RDA of **Y** by **Y** with **X** as covariables. In R, the regression function ***lm()*** can be used to easily obtain a matrix of residuals: res = ***residuals(lm(***$\mathbf{Y} \sim \mathbf{X}$***))***. With the data of Table 11.3 for instance, one could examine the residual structure, after controlling for depth and substrate, by plotting a PCA biplot of the non-canonical axes shown in Table 11.4. In spatial analysis, one could detrend the data by computing the regression residuals of **Y** on the geographic coordinates of the sites before computing a PCA.

Selection of explanatory variables

7. *Selection of explanatory variables.* — Different selection methods are available in canonical analysis, as well as in multiple regression (Subsection 10.3.3): backward,

forward, and stepwise. Function ***ordistep()*** in VEGAN offers all three methods of selection. In *forward selection*, the significance of the partial *F*-statistics associated with all candidate variables is tested using permutations, and the explanatory variable that has the most significant partial effect is selected if its p-value satisfies the "p-to-enter" significance level; in case of equality, the variable that has the lowest value of the Akaike Information Criterion (*AIC*, eq. 10.22)* is selected for inclusion in the model. The *backward* option sequentially drops variables from the model using the same criteria of significance (the highest p-value is compared to a "p-to-exclude" significance level) and *AIC* in case of equality (the variable whose removal produces the model with the lowest *AIC* value is excluded). The *stepwise* option tries to eliminate variables from the model (*backward*) after each *forward* step. In this function, "best" refers to the most significant variable.

Functions ***ordiR2step()*** of VEGAN and ***forward.sel()*** of PACKFOR offer the forward method. In these functions, the basic algorithm, developed by ter Braak (1990), is the same as in CANOCO: considering the variables already selected, the explanatory variable with the highest partial R^2 is selected if the additional contribution of that variable is significant (permutation test) at a pre-selected significance level. In these functions, "best" refers to the variable that explains the largest portion of the remaining unexplained variance of **Y**. These two functions offer the option of applying a second stopping criterion proposed by Blanchet *et al*. (2008b): the selection stops either when the tested variable has a p-value higher than the pre-selected significance level or when the adjusted R^2 of the full model, before any selection, is exceeded.

Before applying these variable selection methods, one should look at the collinearity among the variables in **X** by computing variance inflation factors (VIF, eq. 10.17), and remove variables as needed to reduce collinearity. Borcard *et al*. (2011) present examples of forward selection prior to RDA.

11 — Variation partitioning by RDA

Variation partitioning, described in Subsection 10.3.5 for univariate response data, was originally developed for the analysis of multivariate response tables (Borcard *et al*., 1992; Borcard & Legendre, 1994). It is especially useful for partitioning the variation of community composition data with respect to two or more sets of explanatory variables.

Ecological application 11.1a

The method is illustrated here using fish assemblage data (27 species) from 29 sampling sites along the Doubs River in eastern France. The calculations reported in this application were done

* The *AIC* criterion is not meant to identify the "true" model (which is only known in simulation studies) among several alternative models, but to find the best predictive model for new observations.

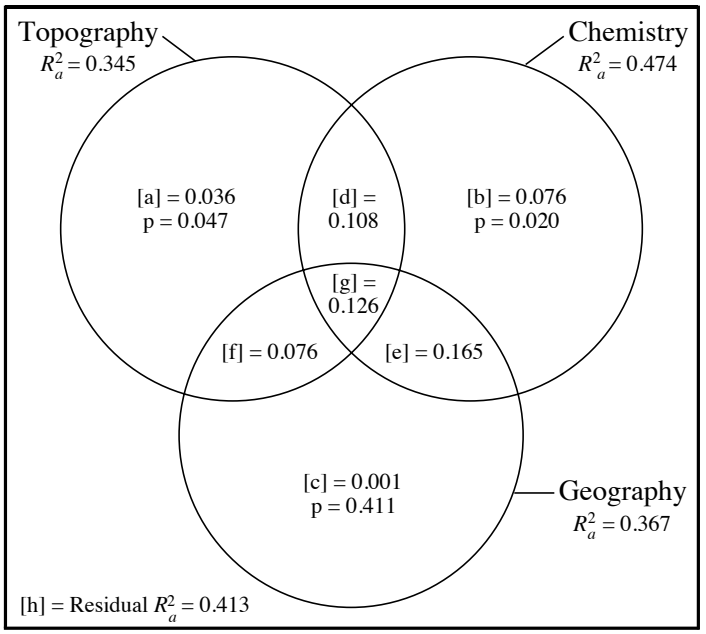


Figure 11.6 Venn diagram illustrating the results of variation partitioning of the Doubs River fish assemblage data (29 sites) among three sets of explanatory variables: Topography, Chemistry and Geography. The fractions of variation are identified by letters [a] to [h]. The value next to each identifier is the adjusted R^2 (R^2_a). The circles, drawn by the plotting function ***plot.varpart()***, are of equal sizes despite differences in the corresponding R^2_a. Circle sizes and shapes can be modified using a graphics editor prior to the publication of the partitioning results.

by RDA, whereas they involved multiple regression in Subsection 10.3.5. The partitioning example reported here uses the species and environmental data collected by Verneaux (1973), which are available in the R package ADE4. The data were reanalysed for variation partitioning by Borcard *et al*. (2011, Section 6.3.2.7); here as in that book, site 8, where no fish were caught, was removed from the original 30-site species and environment data tables.

Partitioning involved three data sets: Topography (variables: altitude, slope, water flow), Chemistry (variables: pH, hardness, concentrations of phosphate, nitrate, ammonia, dissolved O_2, and biological oxygen demand), and Geography (variable: linear distance from the source along the course of the river). The partitioning results, obtained from function ***varpart()*** of the VEGAN package, are illustrated by a Venn diagram (Fig. 11.6); the decomposition into fractions [a] to [h] was done from the adjusted R^2 values (R^2_a) calculated by RDAs involving 1, 2, and all 3 explanatory data tables, as in Subsection 10.3.5. The first finding was that each of the three data sets explained approximately the same fraction of the spatial variation in the fish assemblage along the river (R^2_a = 0.345, 0.474, and 0.367, respectively). A great deal of the variation was shared among two or all three sets of explanatory variables. There was a small but significant portion of the fish variation explained by Topography that was not shared with the

two other data sets (fraction [a]: R^2_a = 0.036, p = 0.047), and likewise for Chemistry (fraction [b]: R^2_a = 0.076, p = 0.020). However, all the fish variation among sites explained by Geography was also explained by one of the other two explanatory data tables, or by both, leaving no significant portion of variation explained solely by Geography (R^2_a = 0.001, p = 0.411). Whereas the three explanatory data sets explained jointly 58.7% of the species variation, it was the Topography and Chemistry data that were the most informative and complementary, adding to the model portions of variation that were not explained by Geography alone. Results of a partitioning involving soil mite assemblages by four explanatory data sets are presented in Borcard *et al*. (2011, Section 7.4.2.5).

In Chapter 14, which describes multiscale spatial analysis, variation partitioning is used to partition the variation of data **Y** between two components, environmental (**X**) and spatial (**W**). Two ecological applications (14.1a and 14.1b) involving variation partitioning by partial canonical analysis are presented.

Ecological application 11.1b

In a classical study of spider community ecology, Aart & Smeenk-Enserink (1975) used canonical correlation analysis (CCorA, Section 11.4) to analyse a portion of the hunting spider data collected in pitfall traps at 100 sites in the Bierlap dune valley of the Netherlands. The paper related the species to environmental descriptors obtained at 28 of the 100 sites. The authors used canonical correlation analysis, a symmetric method of canonical analysis that was available in computer packages in the 1970s, to describe the influence of environmental conditions on the spider assemblages; their objective was to test the hypothesis of an asymmetric relationship between species and environmental conditions. The present example will show original results that we obtained by RDA, which is a more appropriate method to study and test asymmetric relationships. An additional advantage is that RDA can be carried out on unstandardized species data, thus preserving the original variances of the individual species in the analysis (Subsection 11.1.5), whereas the species data are always standardized in CCorA (Subsection 11.4.1). The Aart & Smeenk-Enserink spider data have been reanalysed, after recoding, by ter Braak (1986)* using CCA. The same data (recoded by ter Braak, 1986) were also analysed by De'ath (2002) using multivariate regression tree analysis (MRT, Ecological application 8.11).

At the 28 sites included in the canonical correlation analysis of Aart & Smeenk-Enserink (1975), the community composition data were the abundances of 12 hunting spider species normalized by logarithmic transformation, log(y + 1). Among the 27 environmental descriptors characterizing the light, vegetation, and soil that had been observed, only the 15 that were linearly correlated with the species variables were used by these authors for their canonical correlation analysis in order to ensure linearity of the relationships between the two sets of descriptors.†

* Warning to users: the 28 sites for which environmental data are provided in the Aart & Smeenk-Enserink (1975) paper are presented in a different order in Table 3 of ter Braak (1986).

† The spider (28 sites, 12 species) and environmental data (28 sites, 15 variables) used in this Application are available on the Web page http://numericalecology.com/data.

For the present application, the 15 environmental variables selected by Aart & Smeenk-Enserink (1975) were used as matrix **X** to insure comparability of the present results with theirs. The adjusted R^2 (R^2_a) of the RDA provided a criterion to select the best transformation for the species data: after computing RDA of the spider data, transformed in different ways, with the 15 environmental variables, R^2_a was higher for the log-transformed species data than for any of the other transformations of Section 7.7; so the log-transformed data were used in the RDA. Forward selection (with the stopping criterion $p \leq 0.05$) was carried out among the 15 environmental variables (Subsection 11.1.10, point 7). A parsimonious model containing six environmental variables was selected, which provided the same explanation as the full set of explanatory variables: $R^2_a = 0.761$ for the full set of 15 environmental variables, $R^2_a = 0.768$ for the subset of six variables, i.e. water content of the soil, illuminance under cloudless sky, ground cover by leaves and twigs, cover by the herb layer, cover by *Calamagrostis epigejos* (a grass, family Poaceae), and cover by the tree layer.

A search for species associations was carried out using concordance analysis, described in Subsection 8.9.2. The first statistically significant association comprised three species: *Alopecosa accentuata*, *Alopecosa fabrilis* and *Arctosa perita*; a fourth species, *Pardosa monticola*, was weakly associated with this group. The second significant association contained seven species: *Alopecosa cuneata*, *Arctosa lutetiana*, *Aulonia albimana*, *Pardosa nigriceps*, *Pardosa pullata*, *Trochosa terricola* and *Zora spinimana*. The species *Pardosa lugubris* formed a single-species group.

The RDA triplot (Fig. 11.7) shows the relationships between the species and the environmental variables. The species belonging to association 1 (upper ellipse) were found in greater abundances at very dry and more intensely lit sites. Those belonging to association 2 (right ellipse) were found at sites with higher soil humidity and higher cover by herbs and by *Calamagrostis epigejos*. The single-species association *Pardosa lugubris* exhibited preference for shaded sites with higher soil humidity and higher cover by trees and by leaves and twigs.

The total species variation, which is a measure of beta diversity (Section 6.5), was partitioned between the physical (soil, light) and vegetation influences using variation partitioning (Fig. 11.8). Fractions [a] and [c] were both statistically significant (tested by partial RDA), but fraction [c], which depicted the fraction of beta diversity explained exclusively by vegetation ($R^2_a = 0.43$), was much larger than fraction [a], which corresponded to the variation explained only by the physical environment ($R^2_a = 0.07$). Most of the explanation ([b] = 0.27) provided by the physical variables was shared with the vegetation variables.

11.2 Canonical correspondence analysis (CCA)

Canonical correspondence analysis is a canonical asymmetric ordination method developed by ter Braak (1986, 1987a, 1987c). First implemented in the program CANOCO (ter Braak, 1988b, 1988c, 1990; ter Braak & Smilauer, 1998), it is now available in several computer packages and R functions for community ecology. It is the canonical form of correspondence analysis. Any data table that could be subjected to correspondence analysis (CA, Section 9.2) is a suitable *response matrix* **Y** for CCA; this is the case, in particular, for species presence-absence or abundance data (Subsection 9.2.4).

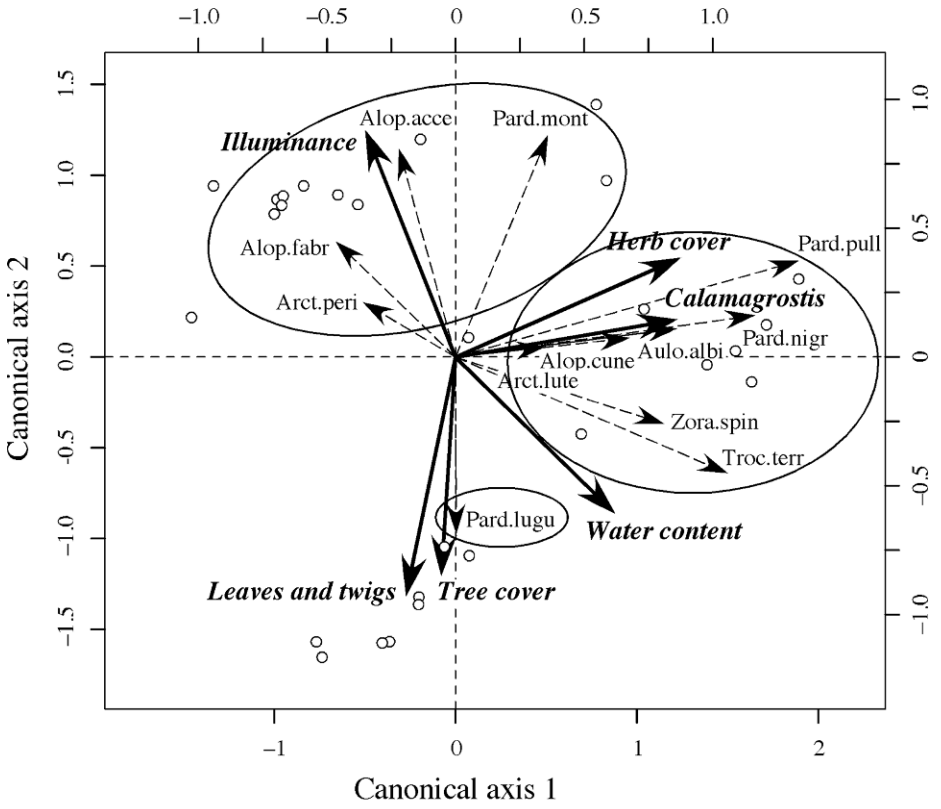


Figure 11.7 RDA triplot relating the spider species (dashed arrows) to the selected environmental variables (full-line arrows). Scaling type 2 was used to emphasize the covariances among the species. Small open circles represent the 28 sites; site names were not printed to keep the diagram simple. The species associations are indicated by ellipses. Association 1: *Alopecosa accentuata* (abbreviation: Alop.acce), *Alopecosa fabrilis* (Alop.fabr), *Arctosa perita* (Arct.peri) and *Pardosa monticola* (Pard.mont, weakly associated with this group). Association 2: *Alopecosa cuneata* (Alop.cune), *Arctosa lutetiana* (Arct.lute), *Aulonia albimana* (Aulo.albi), *Pardosa nigriceps* (Pard.nigr), *Pardosa pullata* (Pard.pull), *Trochosa terricola* (Troc.terr) and *Zora spinimana* (Zora.spin). Single-species group: *Pardosa lugubris* (Pard.lugu).

1 — The algebra of canonical correspondence analysis

The mathematics of CCA is derived from that of RDA. The first difference is that matrix $\overline{\mathbf{Q}}$ is used instead of $\mathbf{Y}$ as the response matrix in the calculations, as it was the case in correspondence analysis (Section 9.2). The second difference is that a diagonal matrix of row weights, $\mathbf{D}(p_{i+})$, is used in the regression portion of the calculation. For

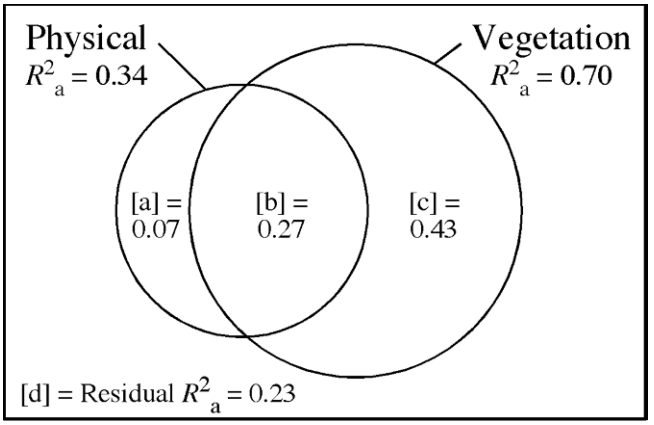


Figure 11.8 Venn diagram partitioning the total spider species variation (rectangle) between physical (water content of the soil, illuminance under cloudless sky) and vegetation influences (ground cover by leaves and twigs, cover by the herb layer, cover by *Calamagrostis epigejos*, and cover by the tree layer). The fraction identifiers [a], [b] and [c], are as in Fig. 10.10. The fractions are expressed as R^2_a, as in Fig. 11.6. Circle sizes are approximate.

each row of **Y**, f_{i+} is the sum of the values in row i, and p_{i+} is f_{i+} divided by the grand total, f_{++}, of the frequencies in **Y**.

Inflated data matrix

To obtain a CCA, the regression portion of the calculation is modified, in eq. 11.25 (below), in such a way as to emulate a RDA carried out on *inflated data matrices* $\mathbf{Y}_{infl}$ and $\mathbf{X}_{infl}$ constructed as follows. **Y** ($n \times p$) contains frequency data, such as species presences or abundances of p species observed at n sites, and **X** ($n \times m$) contains explanatory, e.g. environmental, variables. The presence of *a single individual* in **Y** produces a new row in $\mathbf{Y}_{infl}$, so that there are as many rows in $\mathbf{Y}_{infl}$ as there are individual organisms, or presences, in **Y**. The number of rows of $\mathbf{Y}_{infl}$ is thus f_{++}. $\mathbf{Y}_{infl}$ still has p columns for the p species, but a single individual is present in each row. In $\mathbf{X}_{infl}$, the row vectors of explanatory data are duplicated as many times as needed to make every individual organism (i.e. every species presence) in $\mathbf{Y}_{infl}$ face, in $\mathbf{X}_{infl}$, a copy of the appropriate vector of explanatory data. Compute $\overline{\mathbf{Q}}_{infl}$ from $\mathbf{Y}_{infl}$ using eq. 9.24. CCA is the RDA of $\overline{\mathbf{Q}}_{infl}$ by $\mathbf{X}_{infl}$: the eigenvalues* and matrix of eigenvectors are the same.

* The eigenvalues of RDA of $\overline{\mathbf{Q}}_{infl}$ by $\mathbf{X}_{infl}$ computed on the covariance matrix of $\hat{\mathbf{Y}}$, instead of the cross-product matrix, are smaller than those of CCA by a multiplicative factor ($f_{++} - 1$).

In computer programs, it is possible to use matrices $\overline{\mathbf{Q}}$ and $\mathbf{X}$ for the calculations instead of $\overline{\mathbf{Q}}_{infl}$ and $\mathbf{X}_{infl}$, which would be cumbersome to compute when $\mathbf{Y}$ has a large sum f_{++}. The modified algorithm is the following:

• The dependent data matrix is not $\mathbf{Y}$ centred by columns as in RDA. In this algorithm, CCA uses matrix $\overline{\mathbf{Q}}$ of the contributions to chi-square, also used in correspondence analysis, as the response matrix. $\overline{\mathbf{Q}}$ is derived from matrix $\mathbf{Y}$ through eq. 9.24.

• Matrix $\mathbf{X}$ is standardized to $\mathbf{X}_{stand}$ using weights $\mathbf{D}(f_{i+})$. To achieve this, the inflated matrix $\mathbf{X}_{infl}$ is constructed as described above; it contains f_{++} rows. Then the mean and standard deviation of each column of $\mathbf{X}_{infl}$ are computed and used to standardize the explanatory variables in $\mathbf{X}$. For the standard deviations (eq. 4.5), the maximum likelihood estimator of the variance is used instead of the usual unbiased estimator (eq. 4.3); in other words, the sum of squared deviations from the mean of the variables in $\mathbf{X}_{infl}$ is divided by the number of rows of that matrix (which is equal to f_{++}), instead of the number of rows minus 1.

• To obtain the regression coefficients, weighted multiple regression is used instead of conventional multiple regression. The row weights, written in diagonal matrix $\mathbf{D}(p_{i+})^{1/2}$ (Subsection 9.2.1), are applied to matrix $\mathbf{X}$ everywhere it occurs in the multivariate regression equation, which becomes:

$$\mathbf{B} = [\mathbf{X}_{stand}' \, \mathbf{D}(p_{i+}) \, \mathbf{X}_{stand}]^{-1} \, \mathbf{X}_{stand}' \, \mathbf{D}(p_{i+})^{1/2} \, \overline{\mathbf{Q}}$$

and

$$\hat{\mathbf{Y}} = \mathbf{D}(p_{i+})^{1/2} \, \mathbf{X}_{stand} \, \mathbf{B}$$

The equation for computing $\hat{\mathbf{Y}}$ is then:

$$\hat{\mathbf{Y}} = \mathbf{D}(p_{i+})^{1/2} \, \mathbf{X}_{stand} \, [\mathbf{X}_{stand}' \, \mathbf{D}(p_{i+}) \, \mathbf{X}_{stand}]^{-1} \, \mathbf{X}_{stand}' \, \mathbf{D}(p_{i+})^{1/2} \, \overline{\mathbf{Q}} \qquad \textbf{(11.25)}$$

The matrix of residuals is computed as $\overline{\mathbf{Q}}_{\mathbf{res}} = \overline{\mathbf{Q}} - \hat{\mathbf{Y}}$. This is the equivalent, for CCA, of equation $\mathbf{Y}_{\mathbf{res}} = \mathbf{Y} - \hat{\mathbf{Y}}$ found in Fig. 11.2 for RDA.

• Eigenvalue decomposition (eqs. 11.15 and 11.16) is carried out on matrix $\mathbf{S}_{\hat{\mathbf{Y}}'\hat{\mathbf{Y}}}$ which, in this case, is simply the matrix of sums of squares and cross products, without division by the number of degrees of freedom — as in correspondence analysis:

$$\mathbf{S}_{\hat{\mathbf{Y}}'\hat{\mathbf{Y}}} = \hat{\mathbf{Y}}'\hat{\mathbf{Y}} \qquad \textbf{(11.26)}$$

One can show that $\mathbf{S}_{\hat{\mathbf{Y}}'\hat{\mathbf{Y}}}$ (eq. 11.26) is equal to $\mathbf{S}_{\overline{\mathbf{Q}}\mathbf{X}}\mathbf{S}_{\mathbf{XX}}^{-1}\mathbf{S}'_{\overline{\mathbf{Q}}\mathbf{X}}$ if the covariance matrices $\mathbf{S}_{\overline{\mathbf{Q}}\mathbf{X}}$ and $\mathbf{S}_{\mathbf{XX}}$ are computed as follows, with weights on $\mathbf{X}$ given by matrix $\mathbf{D}(p_{i+})^{1/2}$:

$$\mathbf{S}_{\overline{\mathbf{Q}}\mathbf{X}} = \overline{\mathbf{Q}}' \, \mathbf{D}(p_{i+})^{1/2} \, \mathbf{X} \quad \text{and} \quad \mathbf{S}_{\mathbf{XX}} = \mathbf{X}' \, \mathbf{D}(p_{i+}) \, \mathbf{X}$$

without division by degrees of freedom.

In the modified algorithm, CCA is the eigen-decomposition of $\mathbf{S}_{\hat{\mathbf{Y}}'\hat{\mathbf{Y}}}$ (eq. 11.26). It produces matrices $\mathbf{\Lambda}$ of eigenvalues and $\mathbf{U}$ of eigenvectors. Canonical correspondence analysis is thus a weighted form of redundancy analysis, applied to response matrix $\overline{\mathbf{Q}}$. The solution approximates the chi-square distances among the rows (objects) of the dependent data matrix, subject to the constraint that the canonical ordination vectors be maximally related to weighted linear combinations of the explanatory variables. The method is well suited to analyse the relationships between species presence/absence or abundance data matrices and tables of environmental variables. The number of canonical and non-canonical axes expected from the analysis are shown in Table 11.1. Tests of significance for the total canonical variation and for individual canonical axes are carried out in the same way in CCA as described for RDA in Subsections 11.1.2 and 11.1.8.

- The normalized matrix $\hat{\mathbf{U}}$ is obtained using eq. 9.30:

$$\hat{\mathbf{U}} = \overline{\mathbf{Q}}\,\mathbf{U}\mathbf{\Lambda}^{-1/2}$$

In CCA, matrix $\hat{\mathbf{U}}$ defined here does not contain the loadings of the rows of $\hat{\mathbf{Y}}$ on the canonical axes. It contains instead the loadings of the rows of $\overline{\mathbf{Q}}$ on the ordination axes, as in CA. It will be used to find the site scores (matrices $\mathbf{F}$ and $\hat{\mathbf{V}}$) in the space of the original variables $\mathbf{Y}$. The site scores in the space of the fitted values $\hat{\mathbf{Y}}$ will be found using $\mathbf{U}$ instead of $\hat{\mathbf{U}}$.

Scalings in CCA

- Matrix $\mathbf{V}$ of species scores (for scaling type 1) and matrix $\hat{\mathbf{V}}$ of site scores (for scaling type 2) are obtained from $\mathbf{U}$ and $\hat{\mathbf{U}}$ using the transformations described for correspondence analysis (Subsection 9.2.1):

eq. 9.33 (species scores, scaling 1): $\mathbf{V} = \mathbf{D}(\mathrm{p}_{+j})^{-1/2}\mathbf{U}$

and eq. 9.34 (site scores, scaling 2): $\hat{\mathbf{V}} = \mathbf{D}(\mathrm{p}_{i+})^{-1/2}\,\hat{\mathbf{U}}$

or combining eqs. 9.30 and 9.34: $\hat{\mathbf{V}} = \mathbf{D}(\mathrm{p}_{i+})^{-1/2}\,\overline{\mathbf{Q}}\,\mathbf{U}\mathbf{\Lambda}^{-1/2}$

Scalings 1 and 2 are the same as in correspondence analysis (Subsection 9.2.1). Matrices $\mathbf{F}$ (site scores for scaling type 1) and $\hat{\mathbf{F}}$ (species scores for scaling type 2) are found using eqs. 9.35a and 9.36a:

$$\mathbf{F} = \hat{\mathbf{V}}\,\mathbf{\Lambda}^{1/2} \quad \text{and} \quad \hat{\mathbf{F}} = \mathbf{V}\,\mathbf{\Lambda}^{1/2}$$

Equations 9.35b and 9.36b cannot be used here to find $\mathbf{F}$ and $\hat{\mathbf{F}}$ because the eigenanalysis has been conducted on a covariance matrix (eq. 11.26) computed from the matrix of fitted values $\hat{\mathbf{Y}}$ (eq. 11.25) and not from $\mathbf{Q}$ defined in Subsection 9.2.1.

As mentioned in Subsection 9.2.1 about correspondence analysis, scaling type 3, which is called "symmetric scaling" in program CANOCO, is a compromise between

scalings 1 and 2. This scaling does not preserve the chi-square distances among the species or among the site scores. It is obtained by drawing together matrices $\hat{\mathbf{V}}\,\mathbf{\Lambda}^{1/4}$ (or $\mathbf{F}\mathbf{\Lambda}^{-1/4}$) for sites and $\mathbf{V}\mathbf{\Lambda}^{1/4}$ (or $\hat{\mathbf{F}}\,\mathbf{\Lambda}^{-1/4}$) for species.

The site scores that are linear combinations of the environmental variables, corresponding to eq. 11.18 of RDA, are found from $\hat{\mathbf{Y}}$ using the following equations:

For scaling type 1: $$\mathbf{Z}_1 = \mathbf{D}(p_{i+})^{-1/2}\,\hat{\mathbf{Y}}\,\mathbf{U} \qquad \textbf{(11.27)}$$

For scaling type 2: $$\mathbf{Z}_2 = \mathbf{D}(p_{i+})^{-1/2}\,\hat{\mathbf{Y}}\,\mathbf{U}\mathbf{\Lambda}^{-1/2} \qquad \textbf{(11.28)}$$

For scaling type 3: $$\mathbf{Z}_3 = \mathbf{D}(p_{i+})^{-1/2}\,\hat{\mathbf{Y}}\,\mathbf{U}\mathbf{\Lambda}^{-1/4} \qquad \textbf{(11.29)}$$

Before computing the biplot scores, matrix $\mathbf{Z}_1$ (or $\mathbf{Z}_2$ or $\mathbf{Z}_3$: identical results) must be standardized to $\mathbf{Z}_{stand}$ using the procedure described for the standardization of $\mathbf{X}$: generate the inflated matrix $\mathbf{Z}_{1\,infl}$, compute the vectors of column means and standard deviations (in the computation of the variances, divide the sums of squares by f_{++} instead of $(f_{++} - 1)$) for $\mathbf{Z}_{1\,infl}$, and use these vectors to standardize $\mathbf{Z}_1$. Applying this concept, computational shortcuts can be used to obtain matrix $\mathbf{Z}_{stand}$ without actually generating matrix $\mathbf{Z}_{1\,infl}$. The matrices of biplot scores (**BS**) for the explanatory variables can now be computed using $\mathbf{X}_{stand}$, $\mathbf{Z}_{stand}$, and the diagonal matrix of row weights $\mathbf{D}(p_{i+})$:

For scaling type 1: $$\mathbf{BS}_1 = \mathbf{X}_{stand}'\,\mathbf{D}(p_{i+})\,\mathbf{Z}_{stand}\,\mathbf{\Lambda}^{1/2} \qquad \textbf{(11.30)}$$

For scaling type 2: $$\mathbf{BS}_2 = \mathbf{X}_{stand}'\,\mathbf{D}(p_{i+})\,\mathbf{Z}_{stand} \qquad \textbf{(11.31)}$$

For scaling type 3: $$\mathbf{BS}_3 = \mathbf{X}_{stand}'\,\mathbf{D}(p_{i+})\,\mathbf{Z}_{stand}\,\mathbf{\Lambda}^{1/4} \qquad \textbf{(11.32)}$$

For scaling type 1, triplots are drawn using matrix $\mathbf{V}$ for the species, either $\mathbf{Z}_1$ or $\mathbf{F}$ for the sites, and $\mathbf{BS}_1$ for the explanatory variables. For scaling type 2, matrix $\hat{\mathbf{F}}$ is used for the species, either $\mathbf{Z}_2$ or $\hat{\mathbf{V}}$ for the sites, and $\mathbf{BS}_2$ for the explanatory variables. For scaling type 3, matrix $\mathbf{V}\mathbf{\Lambda}^{1/4}$ is used for the species, either $\mathbf{Z}_3$ or $\hat{\mathbf{V}}\,\mathbf{\Lambda}^{1/4}$ for the sites, and $\mathbf{BS}_3$ for the explanatory variables. The construction and interpretation of CCA triplots is discussed in more detail in ter Braak & Verdonschot (1995).

• Residuals can be analysed by applying eigenvalue decomposition (eq. 11.15) to matrix $\bar{\mathbf{Q}}_{\mathbf{res}}$, producing a matrix of eigenvalues $\mathbf{\Lambda}$ and a matrix of eigenvectors $\mathbf{U}$. Matrix $\hat{\mathbf{U}}$ is obtained using eq. 9.30: $\hat{\mathbf{U}} = \bar{\mathbf{Q}}\,\mathbf{U}\mathbf{\Lambda}^{-1/2}$. Species and site scores are obtained for scaling types 1 and 2 (eqs. 9.33, 9.34, 9.35a and 9.36a) using the matrices of row and column sums $\mathbf{D}(p_{i+})^{-1/2}$ and $\mathbf{D}(p_{+j})^{-1/2}$ of the original matrix $\mathbf{Y}$.

CCA can be computed following the algorithm described in the present subsection*. One may also use the iterative algorithm proposed by ter Braak (1986, 1987a) and implemented in the program CANOCO. The latter algorithm, which has historical significance, is described in Table 11.6 of Legendre & Legendre (1998).

Partial CCA Developed by ter Braak (1988a), partial CCA is computed essentially like partial RDA (Subsection 11.1.6), after residualizing $\mathbf{Q}$ and $\mathbf{X}$ on the covariables $\mathbf{W}$. The weights $\mathbf{D}(p_{i+})^{1/2}$ are used in the computation of these residuals.

CCA can be used for variation partitioning (Subsections 10.3.5 and 11.1.11). The difficulty with CCA resides in the calculation of the adjusted R^2, which is necessary to obtain unbiased estimates of the fractions of explained variation. A method to compute the adjusted R^2 in CCA, involving a permutation procedure, was described by Peres-Neto *et al*. (2006). In Supplements to their paper, Peres-Neto *et al*. (2006) provided a MATLAB package and an executable program to conduct variation partitioning in CCA. At the time this paragraph is written, however, that method has not been incorporated into any major package for community ecology, with the consequence that variation partitioning is not yet generally available for CCA.

2 — Numerical example

Table 11.3 will now be used to illustrate the computation and interpretation of CCA. The 9 species were used in matrix **Y**. Matrix **X** comprised the four columns shown in the right-hand portion of Table 11.3. CCA results are presented in Table 11.7 and Fig. 11.9; the CANOCO program and the CCA functions in R* provide more output tables than presented here. There was a possibility of 3 canonical and 8 non-canonical axes. It turned out that the last 2 non-canonical axes had zero variance; they are consequently not displayed. An overall test of significance showed that the canonical relationship between matrices **X** and **Y** was very highly significant ($p = 0.001$ after 999 permutations of residuals under a full model; Subsection 11.1.8). The canonical axes explained 47%, 24% and 10% of the response table's inertia, respectively. They were all significant ($p < 0.05$) and displayed strong row-weighted species-environment correlations ($r = 0.998, 0.940$, and 0.883, respectively).

Scaling type 2 (Subsection 11.2.1) was used, in this example, to emphasize the relationships among species. As a result, the species (matrix $\hat{\mathbf{F}}$) are at the centroids of the sites (matrix $\hat{\mathbf{V}}$) in Fig. 11.9a, and distances among species approximate their chi-square distances. Species 3 and 4 characterize the sites with coral substrate, whereas species 5 and 6 indicate the sites with "other substrate". Species 1 and 2, which occupy an intermediate position between the sites with coral and other substrate, are not well represented in the biplot of canonical axes I and II; axis III is needed to adequately represent the variance of these species. Among the ubiquitous species 7 to 9, two are well represented in the subspace of canonical axes I and II; they fall near the middle of the area encompassing the three types of substrate. The sites are not perfectly ordered along the depth vector; the site ordering along this variable mainly reflects differences in species composition between the shallow sandy sites (1, 2 and 3) and the other sites.

* Function CCA.R was written to demonstrate the CCA algorithm described in this subsection. It produces results identical to those of CANOCO 4.x. The function is available on the Web page http://numericalecology.com/rcode.

Table 11.7 Results of canonical correspondence analysis of the data in Table 11.3 (selected output). Matrix **Y**: species 1 to 9; **X**: depth and 3 substrate classes. Non-canonical axes VIII and IX not shown.

	Canonical axes			Non-canonical axes			
	I	II	III	IV	V	VI	VII
Eigenvalues (their sum is equal to the total inertia in matrix $\bar{\mathbf{Q}}$ of species data = 0.78417)							
	0.36614	0.18689	0.07885	0.08229	0.03513	0.02333	0.00990
Fraction of the total variance in $\bar{\mathbf{Q}}$							
	0.46691	0.23833	0.10055	0.10494	0.04481	0.02975	0.01263
Cumulative fraction of total inertia in $\bar{\mathbf{Q}}$ accounted for by axes 1 to k							
	0.46691	0.70524	0.80579	0.91072	0.95553	0.98527	0.99791
Eigenvectors ("species scores", scaling 2): matrices $\hat{\mathbf{F}}$ for the canonical and non-canonical portions (eq. 9.36a)							
Species 1	–0.11035	–0.28240	–0.20303	0.00192	0.08223	0.08573	–0.01220
Species 2	–0.14136	–0.30350	0.39544	0.14127	0.02689	0.14325	0.04303
Species 3	1.01552	–0.09583	–0.19826	0.10480	–0.13003	0.02441	0.04647
Species 4	1.03621	–0.10962	0.22098	–0.22364	0.24375	–0.02591	–0.05341
Species 5	–1.05372	–0.53718	–0.43808	–0.22348	0.32395	0.12464	–0.11928
Species 6	–0.99856	–0.57396	0.67992	0.38996	–0.29908	0.32845	0.21216
Species 7	–0.25525	0.17817	–0.20413	–0.43340	–0.07071	–0.18817	0.12691
Species 8	–0.14656	0.85736	–0.01525	–0.05276	–0.35448	–0.04168	–0.19901
Species 9	–0.41371	0.70795	0.21570	0.69031	0.14843	–0.33425	–0.00629
Site scores ("sample scores", scaling 2): matrices $\hat{\mathbf{V}}$ for the canonical and the non-canonical portions (eq. 9.34)							
Site 1	–0.71059	3.08167	0.21965	1.24529	1.07293	–0.50625	0.24413
Site 2	–0.58477	3.00669	–0.94745	–2.69965	–2.13682	0.81353	0.47153
Site 3	–0.76274	3.15258	2.13925	3.11628	2.30660	–0.69894	–1.39063
Site 4	–1.11231	–1.07151	–1.87528	–0.66637	1.10154	1.43517	–1.10620
Site 5	0.97912	0.06032	–0.69628	0.61265	–0.98301	0.31567	0.57411
Site 6	–1.04323	–0.45943	–0.63980	–0.28716	0.57393	–1.44981	1.70167
Site 7	0.95449	0.08470	0.13251	0.42143	0.11155	–0.39424	–0.67396
Site 8	–0.94727	0.10837	0.52611	0.00565	–1.26273	–1.06565	–1.46326
Site 9	1.14808	–0.49045	0.47835	–1.17016	1.00599	0.07350	0.08605
Site 10	–1.03291	–1.03505	2.74692	1.28084	–0.36299	1.98648	1.05356
Correlations of environmental variables with site scores							
Depth	0.18608	–0.60189	0.65814				
Coral	0.99233	–0.09189	–0.04614				
Sand	–0.21281	0.91759	0.03765				
Other subs.	–0.87958	–0.44413	0.02466				
Correlations of environmental variables with fitted site scores (for biplot, scaling 2)							
Depth	0.18636	–0.64026	0.74521				
Coral	0.99384	–0.09775	–0.05225				
Sand	–0.21313	0.97609	0.04263				
Other subs.	–0.88092	–0.47245	0.02792				
Centroids of sites with code "1" for the BINARY environmental variables, scaling 2							
Coral	1.02265	–0.10059	–0.05376				
Sand	–0.66932	3.06532	0.13387				
Other subs.	–1.03049	–0.55267	0.03266				

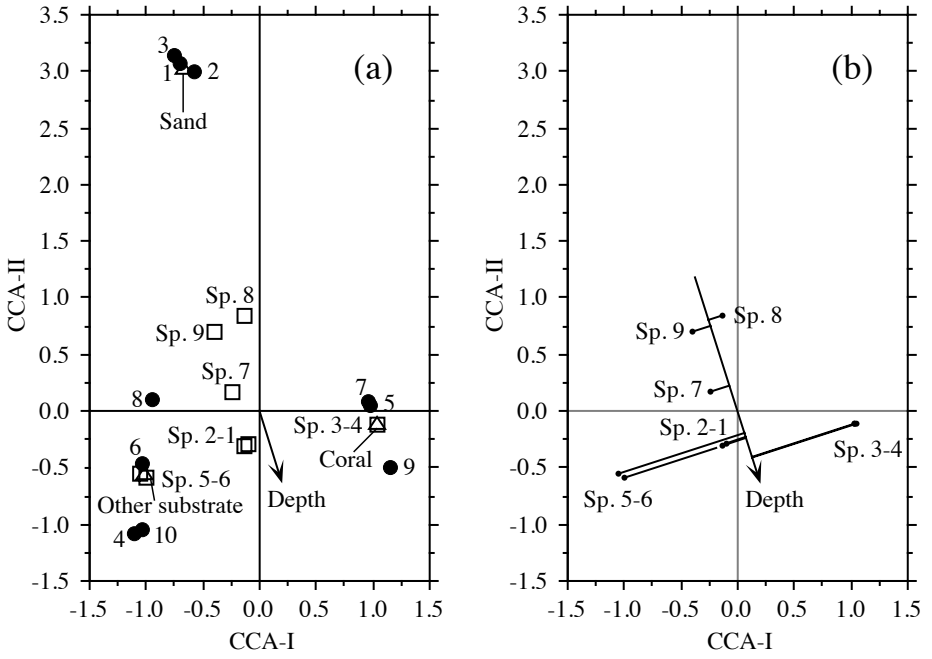


Figure 11.9 CCA ordination triplot (scaling type 2) of the artificial data in Table 11.3; the numerical results of the analysis are in Table 11.7. (a) Triplot representing the species (squares), sites (dots, with site identifiers that also correspond to water depths in m), and environmental variables (full arrow for depth, triangles for the three binary substrate variables). (b) Ranking of the species along the quantitative environmental variable (depth) is inferred by projecting the species at right angle onto the arrow representing that variable.

Figure 11.9b shows how to infer the ranking of species along a quantitative environmental variable. Depth is used in this example. The graphical method simply consists in projecting (at right angle) the species onto the arrow representing that variable. This gives an approximation of the weighted averages of the species with respect to environmental variables. Ecologists like to interpret this ranking as representing the niche optima for the species under consideration. It is important to realize that three rather strong assumptions are made when attempting such an interpretation:

- that the various species have unimodal distributions along the environmental variable of interest (Subsection 9.2.4);

- that the species distributions are under environmental control (Whittaker, 1956; Bray & Curtis, 1957), so that the mode of each species is at its optimum along each environmental variable; and

- that the environmental gradient under study is long enough to allow each species to go from some less-than-optimum low frequency to its high-frequency optimum, and back to some past-optimum low frequency.

In the data of the present example (Table 11.3), only species 1, 3 and 5 were constructed to approximately correspond to these criteria. Species 7, which may also look like it has a unimodal distribution, has actually been constructed using a pseudo-random number generator; so its optimum along depth is fortuitous.

To investigate the similarities among sites or the relationships among species after controlling for the linear effects of depth and type of substrate, one could draw ordination biplots of the *non-canonical axes* in Table 11.7. These axes correspond to a correspondence analysis of the table of regression residuals, as shown in Fig. 11.2.

Ecological application 11.2a

Ecological application 9.2b described the spatial distribution of chaetodontid fish assemblages (butterflyfishes) around a tropical island, using correspondence analysis. This application is continued here. Cadoret *et al*. (1995) next described the relationships between the fish species (quantitative relevés) and some environmental variables, using canonical correspondence analysis. The environmental variables were: the type of environment (qualitative descriptor: bay, lagoon, or outer slope of the reef on the ocean side), geomorphology (qualitative: reef flat, crest, and reef wall of the fringing reefs of bays; fringing reef, shallow, barrier reef, and outer slope for transect sites), depth (quantitative: from 0.5 to 35 m), and exposure to swell (qualitative: low, high, or sites located in bays).

The ordination of sampling sites by CCA is virtually identical to that in Fig. 9.14; this indicates that the first two CA axes are closely related to the environmental variables. The canonical axes account together for 35% of the variation in the species data ($p = 0.001$ after 999 permutations). The description of the ordination of sites presented in Ecological application 9.2b may be compared to Fig. 11.10. This figure shows which types of environment are similar in their chaetodontid species composition and which species are associated with the various types of environment. It indicates that the reef flats of the fringing reefs in bays are similar in species composition to the fringing reefs in the lagoon; likewise, the crests of the fringing reefs in bays are similar to the barrier reefs in the lagoon. The species composition along the reef walls in bays and that on the outer slopes differ, however, from all the other types of environment. The authors discuss the ecology of the most important chaetodontid species in their paper.

Ecological application 11.2b

Canonical correspondence analysis is widely used in palaeoecology, together with regression and calibration, to infer past ecological conditions (climatic, limnological, etc.) from palaeo-assemblages of species. The first 10 years of that literature (1986-1996) was summarized in a bibliography assembled by Birks *et al*. (1998), under the headings *limnology*, *palaeoecology*, *palaeolimnology*, etc. Several applications of CCA are described in a chapter by Legendre & Birks (2012) in a book about numerical methods in palaeoecology edited by Birks *et al*. (2012).

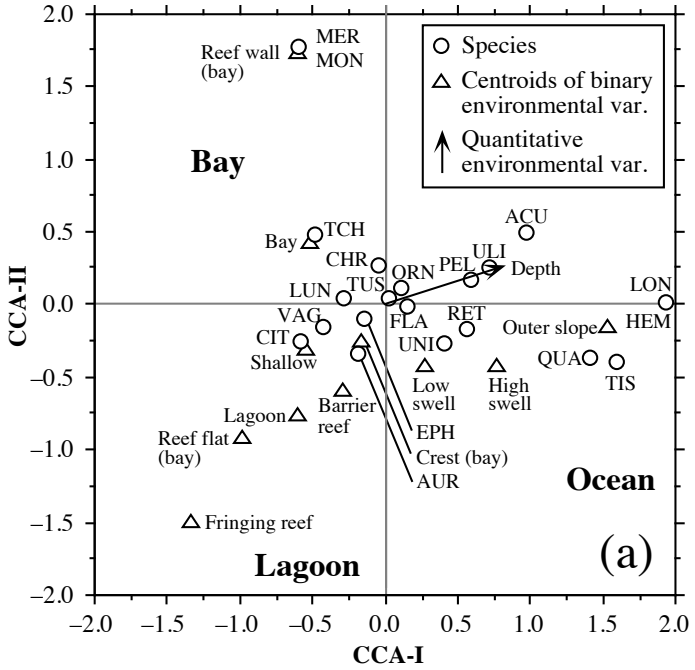


Figure 11.10 (a) CCA ordination diagram: presence/absence of 21 Chaetodontid fish species at 42 sampling sites around Moorea Island, French Polynesia, related to environmental variables. The species (names abbreviated to 3 letters) are represented by circles for readability of the diagram. Axis I: 14.6% of the variation ($p = 0.001$ after 999 permutations); axis II: 7.4% ($p = 0.010$). Redrawn from the original data of Cadoret *et al*. (1995).

One of the classical papers on the subject was written by Birks *et al*. (1990a). Palaeolimnological reconstruction involves two main steps: modelling from a *training data set*, followed by the construction of forecasting models that are then applied to the palaeo-data. In this paper, diatoms were used to reconstruct past water chemistry. The training data set consisted of diatom assemblages comprising 287 species, from present-day surface samples from 138 lakes in England, Norway, Scotland, Sweden, and Wales. Data were also available on pH, conductivity, Ca, Mg, K, SO_4, Cl, alkalinity, total Al, and DOC. Data from more lakes were available for subsets of these variables. CCA was used to relate species composition to water chemistry. The first two canonical eigenvalues were significant and displayed strong species-environment correlations ($r = 0.95$ and 0.84, respectively). The first axis expressed a significant diatom gradient which was strongly and positively correlated with alkalinity and its close correlates, Ca and pH, and negatively but less strongly correlated with total Al; the second axis corresponded to a significant gradient strongly correlated with DOC. This result indicated that pH (or alkalinity), Al, and DOC were potentially reconstructible from the fossil diatom assemblages.

The fossil data set contained 101 slices of a sediment core from a small lake, the Round Loch of Glenhead, in Galloway, southwestern Scotland. The data series covered the past 10000 years. The fossil data (292 diatom taxa) were included in the CCA as passive objects (called *supplementary objects* in Subsection 9.1.9) and positioned in the ordination provided by canonical axes I and II. All fossil objects were well-fitted in that space (they had low squared residual distances), indicating that the pattern of variation in diatom composition can be linked to the modern chemical variables.

Reconstruction of past surface-water chemistry involved two steps. First, the training set was used to model, by regression, the responses of modern diatoms to the chemical variables of interest (one variable at a time). Secondly, the modelled responses were used to infer past chemistry from the composition of fossil diatom assemblages; this phase is called *calibration* (ter Braak, 1987b; ter Braak & Prentice, 1988). Extensive simulations led Birks *et al*. (1990b) to prefer weighted averaging (WA) over maximum likelihood (ML) regression and calibration. Consider pH in lakes, for example. *WA regression* simply consists in applying eq. 9.39 to estimate the pH optimum of each taxon of the training set as the weighted average of all the pH values for lakes in which this taxon occurs, weighted by the taxon's relative abundance. *WA calibration* consists in applying eq. 9.38 to estimate the pH of each lake as the weighted average of the pH optima of all the taxa present. Taxa with a narrow pH tolerance or amplitude may, if required, be given greater weight in WA regression and calibration than taxa with a wide pH tolerance (Birks *et al*., 1990b).

Application of eqs. 9.39 and 9.38 to the data resulted in shrinkage of the range of pH scores. Shrinkage occurred for the same reason as in the TWWA algorithm for correspondence analysis; in step 6.4 of that algorithm (Table 9.8), the eigenvalue was actually estimated from the amount of shrinkage incurred by the site scores after each iteration through eqs. 9.39 and 9.38 (steps 3 and 4). Deshrinking may be done in at least two ways; the relative merits of the two methods are discussed by Birks *et al*. (1990b).

• Deshrinking by classical regression proceeds in two steps. (1) The pH values inferred by WA regression and calibration ($\hat{x}_i$) are regressed on the observed values x_i for the training set, using the linear regression model $\hat{x}_i = b_0 + b_1 x_i + \varepsilon_i$. (2) The parameters of that model are then used to deshrink the $\hat{x}_i$ values, using the equation: final $\hat{x}_i = (\hat{x}_i - b_0)/b_1$. This method was used to deshrink the inferred pH values.

• Another way of deshrinking, advocated by ter Braak & van Dam (1989) for palaeolimnological data, is to use "inverse regression" of x_i on $\hat{x}_i$ (ter Braak, 1987b). Inverse regression was used to deshrink the inferred Al and DOC values.

Training sets containing different numbers of lakes were used to infer pH, total Al, and DOC. Past values of these variables were then reconstructed from the palaeo-assemblages of diatoms, using the pH optima estimated above (eq. 9.39) for the various diatom species, followed by deshrinking. Reconstructed values were plotted against depth and time, together with error estimates obtained by bootstrapping. The past history of the Round Loch of Glenhead over the past 10000 years is discussed in the paper.

This approach involving CCA, WA regression, and WA calibration, is now widely used in palaeolimnology to reconstruct, for example, surface-water temperatures from fossil chironomid assemblages, as well as lake salinity, lake water phosphorus concentrations, or surface water chlorophyll *a* concentrations from fossil diatom assemblages. The WA regression and WA calibration method was further improved by ter Braak & Juggins (1993). ter Braak (1995) made a theoretical comparison of reconstruction methods. For a recent presentation, see the chapter by Birks (2010) in a book edited by Smol & Stoermer (2010). How to carry out the calculations was

described by ter Braak & Juggins (1993) and Line *et al*. (1994). R functions for palaeoenvironmental reconstruction are available in package RIOJA (Juggins, 2009).

A little-known application of CCA is worth mentioning here. Consider a qualitative environmental variable and a table of species presence-absence or abundance data. How can one "quantify" the qualitative states, i.e. give them values along a quantitative scale that would be related in some optimal way to the species data? CCA provides an easy answer to this problem. The species data form matrix **Y**; the qualitative variable, which may be coded as a factor or recoded as a set of dummy variables, is placed in matrix **X**. Compute CCA and take the fitted site scores (or "site scores that are linear combinations of environmental variables"): they provide a quantitative rescaling of the qualitative variable, maximizing the weighted linear correlation between the dummy variables and matrix $\bar{\mathbf{Q}}$. In the same way, RDA may be used to rescale a qualitative variable (factor) with respect to a table of quantitative variables of the objects if linear relationships can be assumed.

McCune (1997) warns users of CCA against inclusion of noisy or irrelevant explanatory variables in the analysis: they may lead to misleading interpretations.

11.3 Linear discriminant analysis (LDA)

A situation that often occurs is to start with an already known grouping of the objects, considered to form a qualitative response variable **y** in this type of analysis, and try to determine to what extent a set of quantitative descriptors, which are the explanatory variables **X**, can actually explain this grouping. In this type of analysis, the grouping is known at the start of the analysis. It may be the result of a cluster analysis computed from a *different* data set, or reflect an ecological hypothesis to be tested. The problem no longer consists in delineating groups, as in cluster analysis, but in interpreting them.

Linear discriminant analysis is a method of linear modelling, like the analysis of variance, multiple linear regression, redundancy analysis, and canonical correlation analysis. It proceeds in two steps. (1) First, one tests for differences in the predictor variables (**X**) among the predefined groups using Wilks' lambda (eq. 11.42). This part of the analysis is identical to the overall test performed in MANOVA. (2) If the test supports the alternative hypothesis of significant differences among groups in the **X** variables, the analysis proceeds to find the linear combinations (called *discriminant functions* or *identification functions*) of the **X** variables that best discriminate among the groups.

Like one-way analysis of variance, discriminant analysis considers a single classification criterion (i.e. division of the objects into groups) and allows one to test whether the explanatory variables can discriminate among the groups. Testing for differences among group means, in discriminant analysis, is identical to ANOVA for a single explanatory variable and to MANOVA for multiple variables (**X**).

When it comes to modelling, i.e. finding the linear combinations of the predictors (**X**) that best discriminate among the groups, discriminant analysis is a form of "inverse analysis" (ter Braak, 1987b), where the classification criterion is considered to be the response variable (**y**) whereas the quantitative variables (matrix **X**) are predictors of the classification. In ANOVA, on the contrary, the objective is to estimate if the variation in a quantitative response descriptor **y** is significantly explained by one or several classification criteria (explanatory variables **X**). As in multiple regression, the discriminatory power of **X** is the same in LDA for **X** standardized or not.

As in multiple regression, discriminant analysis estimates the parameters of a linear model of the explanatory variables that may be used to forecast the response variable (states of the classification criterion). While inverse multiple regression would be limited to two groups (expressed by a single binary variable **y**), discriminant analysis can handle several groups. Discriminant analysis is a canonical method of analysis; its link to canonical correlation analysis (CCorA) is explained at the end of Subsection 11.3.1, after some necessary concepts have been introduced.

After the overall test of significance, the search for discriminant functions may be conducted with two different purposes in mind. One may be interested in obtaining a linear equation to allocate new objects to one of the states of the classification criterion (identification), or simply in determining the relative contributions of various explanatory descriptors to the distinction among these states (discrimination).

Discriminant analysis is also called *canonical variate analysis* (CVA). The method was originally proposed by Fisher (1936) for the two-group case ($g = 2$). Fisher's results were extended to $g > 2$ by Rao (1948, 1952). Fisher (1936) illustrated the method using a famous data set describing the morphology (lengths and widths of sepals and petals) of 150 specimens of irises (Iridaceae) belonging to three species. The data had originally been collected in the Gaspé Peninsula, eastern Québec (Canada), by the botanist Edgar Anderson of the Missouri Botanical Garden who allowed Fisher to publish and use the raw data in his 1936 paper. Fisher showed how to use these morphological measurements to discriminate among the species. The data set is sometimes — erroneously — referred to as "Fisher's irises".

The analysis is based upon an explanatory data matrix **X** of size ($n \times m$), where n objects are described by m descriptors. **X** is meant to discriminate among the groups defined by a separate classification criterion vector (**y**). As in regression analysis, the explanatory descriptors must in principle be quantitative, although qualitative descriptors coded as dummy variables may also be used (Subsection 1.5.7). Other methods are available for discrimination using non-quantitative descriptors (Table 10.1). The objects, whose membership in the various groups of **y** is known before the analysis is undertaken, may be sites, specimens, quadrats, etc.

One possible approach would be to examine the descriptors one by one, either by hand or using analyses of variance, and to note those which have states that characterize one or several groups. This information could be transformed into an

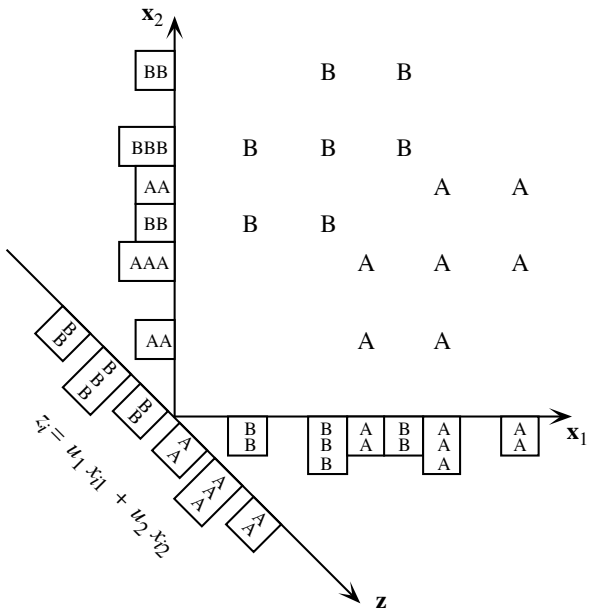


Figure 11.11 Two groups, A and B, with 6 objects each, are overlapping on both axes, $\mathbf{x}_1$ and $\mathbf{x}_2$, as shown by the histograms on the axes. They are perfectly separated, however, along a discriminant axis **z**. The position of each object i is calculated along **z** using the equation $z_i = (\cos 45°)\, x_{i1} - (\cos 45°)\, x_{i2}$. Adapted from Jolicoeur (1959).

identification key, for example. It often occurs, however, that no single descriptor succeeds in separating the groups completely. The next best approach is then to search for a linear combination of descriptors that provides the most efficient discrimination among groups. Figure 11.11 shows an idealized example of two groups (A and B) described by two descriptors only. The groups cannot be separated on either of the two axes taken alone. The solution is a new discriminant descriptor **z**, drawn on the figure, which is a linear combination of the two original descriptors. Along **z**, the two groups of objects are perfectly separated. Note that discriminant axis **z** is parallel to the direction of greatest variability *between groups*. This suggests that the weights u_j used in the discriminant function could be the elements of the eigenvectors of a between-group dispersion matrix. The method can be generalized to more than two groups and several descriptors.

Discriminant function

- *Discriminant functions* (also called standardized discriminant functions) are computed from *standardized descriptors*. The coefficients of these functions are used to assess the relative contributions of the descriptors to the final discrimination.

Identification function

• *Identification functions* (also called unstandardized discriminant functions) are computed from the *original descriptors* (not standardized). They may be used to compute the group to which a new object is most likely to belong. Discriminant analysis is seldom used for this purpose in ecology, whereas it is widely used in that way in taxonomy.

When there are only two groups of objects, the method is called *Fisher's*, or *simple discriminant analysis* (a single function is needed to discriminate between two groups), whereas the case with several groups is called *multiple discriminant analysis* or *canonical variate analysis*. Because the simple discriminant analysis model (two groups) is a particular case of multiple discriminant analysis, it will not be developed here. The solution can be entirely derived from the output of a multiple regression using a dummy variable defining the two groups (used as the dependent variable **y**) against the table of predictor variables **X**.

Analysis of variance is often used for screening variables prior to discriminant analysis: each variable in matrix **X** is tested for its capacity to discriminate among the groups of the classification **y**. Figure 11.11 shows however that there is a danger in this approach; any single variable may not discriminate groups well, although it may have high discriminating power in combination with other variables. One should be careful when using univariate analysis to eliminate variables. If the analysis requires that poorly discriminating variables be eliminated, one should use stepwise discriminant analysis instead, which allows users to identify a subset of good discriminators. Bear in mind, though, that stepwise selection of explanatory variables does not guarantee that the "best" set of explanatory variables will necessarily be found. This is equally true in discriminant analysis and regression analysis (Subsection 10.3.3).

1 — The algebra of discriminant analysis

The problem consists in finding linear combinations of the predictors in matrix **X** ($n \times m$) that maximize the differences among groups while minimizing the variation within the groups. As in regression analysis, the descriptors must be quantitative or binary since they are combined into linear functions. Each descriptor may have already been transformed to meet the condition of multinormality or at least to reduce the asymmetry of its distribution. The discriminant analysis model is robust to departures from this condition, but the parametric statistical tests assume *within-group normality* of the descriptors.

Computations are carried out using either dispersion matrices (**V**, **A**) or matrices of sums of squares and cross-products of centred descriptors (**W**, **B**) (Table 11.8). These matrices are constructed in the same way as in analysis of variance, except that here the predictors form a multivariate data matrix. Matrix **T** is the matrix of scalar products of the centred descriptors, $[x - \bar{x}]$, for all objects irrespective of the groups: $\mathbf{T} = [x - \bar{x}]'[x - \bar{x}]$ (total sums of squares and cross-products). When divided by the total number of degrees of freedom $n - 1$, it becomes the total dispersion matrix **S** used in principal component analysis.

Table 11.8 Discriminant analysis is computed on either dispersion matrices (right-hand column) or matrices of sums of squares and cross-products (centre). Matrices in the right-hand column are simply those in the central column divided by their respective numbers of degrees of freedom. The dimension of all matrices is ($m \times m$).

	Matrices of sums of squares and cross-products	Dispersion matrices
Total dispersion	$\mathbf{T}$	$\mathbf{S} = \mathbf{T}/n - 1$
Pooled within-group dispersion	$\mathbf{W} = \mathbf{W}_1 + \ldots + \mathbf{W}_g$	$\mathbf{V} = \mathbf{W}/n - g$
Among-group dispersion	$\mathbf{B} = \mathbf{T} - \mathbf{W}$	$\mathbf{A} = \mathbf{B}/g - 1$

Matrix **W**, which pools the sums of squares within all groups, is computed by adding up matrices $\mathbf{W}_1$ to $\mathbf{W}_g$ of the sums of squares and cross-products for each of the g groups. Each matrix $\mathbf{W}_j$ is computed from descriptors that have been *centred for the objects of that group only*, just as in ANOVA. In other words, matrix $\mathbf{W}_j$ is the product $[x - \bar{x}]'[x - \bar{x}]$ for the objects that belong to group j only. Dividing the pooled within-group matrix **W** by the within-group number of degrees of freedom, $n - g$, produces the pooled within-group dispersion matrix **V**. In ANOVA involving a single explanatory variable, **W** contains a single value, the residual sum of squares, which is the sum of the within-group sums of squares ($SS_{\text{within groups}}$).

Matrix **B** of the sums of squares among groups is computed by subtracting the pooled within-group matrix **W** from the total matrix of sums of squares **T**. Since $\mathbf{B} = \mathbf{T} - \mathbf{W}$, the number of degrees of freedom by which **B** must be divided to obtain the among-group dispersion matrix **A** is: $(n - 1) - (n - g) = g - 1$. In ANOVA involving a single explanatory variable, **B** contains a single value, the among-group sum of squares ($SS_{\text{among groups}}$).

In analysis of variance involving a single explanatory variable, sums of squares (SS) among and within groups are used to construct the F-statistic to test the hypothesis of no difference between the means of the groups:

$$F = \frac{\text{MS}_{\text{Among groups}}}{\text{MS}_{\text{Within groups}}} = \frac{SS_{\text{Among groups}} / (g - 1)}{SS_{\text{Within groups}} / (n - g)}$$

where MS stands for "mean square".

The matrix of predictor variables, **X**, is multivariate in discriminant analysis. The numerator of F is matrix **A** and its denominator is matrix **V**. One cannot compute **A/V**

but one can compute $\mathbf{V}^{-1}\mathbf{A}$. The eigenvalues of $\mathbf{V}^{-1}\mathbf{A}$ are the canonical *F*-statistics, which can be tested for significance. The eigenvalues are found by eigen-decomposition (eq. 2.22) using the following equation:

$$(\mathbf{V}^{-1}\mathbf{A} - \lambda_k\mathbf{I})\,\mathbf{u}_k = \mathbf{0} \qquad \textbf{(11.33)}$$

This equation maximizes the variation among groups while minimizing the variation within groups. It also produces the matrix of normalized eigenvectors $\mathbf{U} = [\mathbf{u}_k]$. Matrix $\mathbf{V}^{-1}\mathbf{A}$ is asymmetric, as in eqs. 11.48 and 11.50 of CCorA, so its eigenvectors are not orthogonal. The maximum number of discriminant axes needed for the ordination of g groups is $(g - 1)$, so the number of canonical eigenvalues is at most *min*(number of predictors in $\mathbf{X}$, $(g–1)$).

Eigen-decomposition could have been carried out using the matrices of sums of squares and cross-products $\mathbf{W}$ and $\mathbf{B}$ instead of the dispersion matrices $\mathbf{V}$ and $\mathbf{A}$. The eigen-decomposition of

$$(\mathbf{W}^{-1}\mathbf{B} - l_k\mathbf{I})\,\mathbf{u}_k = \mathbf{0} \qquad \textbf{(11.34)}$$

produces the same matrix of eigenvectors $\mathbf{U}$ as eq. 11.33. The eigenvalues l_k are modified by a factor corresponding to the degrees of freedom shown in the equation of the *F*-statistic and in Table 11.8:

$$l_k = \frac{g-1}{n-g}\,\lambda_k \qquad \textbf{(11.35)}$$

which leaves unchanged the percentage of the variance of $\mathbf{V}^{-1}\,\mathbf{A}$ or $\mathbf{W}^{-1}\,\mathbf{B}$ explained by each canonical eigenvalue.

When the non-orthogonal eigenvectors are plotted at right angles, they straighten the reference space and, with it, the ellipsoids of the within-group scatters of objects. If the eigenvectors are now rescaled in an appropriate manner, the within-group scatters of objects can be made circular (Fig. 11.12), insofar as the within-group cross-product matrices $\mathbf{W}_j$ are homogeneous (same dispersion in all groups). This is done by rescaling the eigenvectors (matrix $\mathbf{U}$) using the following formula. The result is matrix $\mathbf{C}$ containing the discriminant function coefficients:

$$\mathbf{C} = \mathbf{U}\left(\mathbf{U}'\frac{\mathbf{W}}{n-g}\mathbf{U}\right)^{-1/2} = \mathbf{U}\,(\mathbf{U}'\mathbf{V}\mathbf{U})^{-1/2} \qquad \textbf{(11.36)}$$

where $(\mathbf{U}'\mathbf{V}\mathbf{U})^{-1/2}$ is a diagonal matrix.

Matrix $\mathbf{C}$ contains the rescaled eigenvectors defining the *canonical space* of the discriminant analysis. After this transformation, the variance among group centroids is maximized even if the group dispersion matrices are not homogeneous. This leads to the conclusion that the principal axes describe the dispersion *among groups*. The first

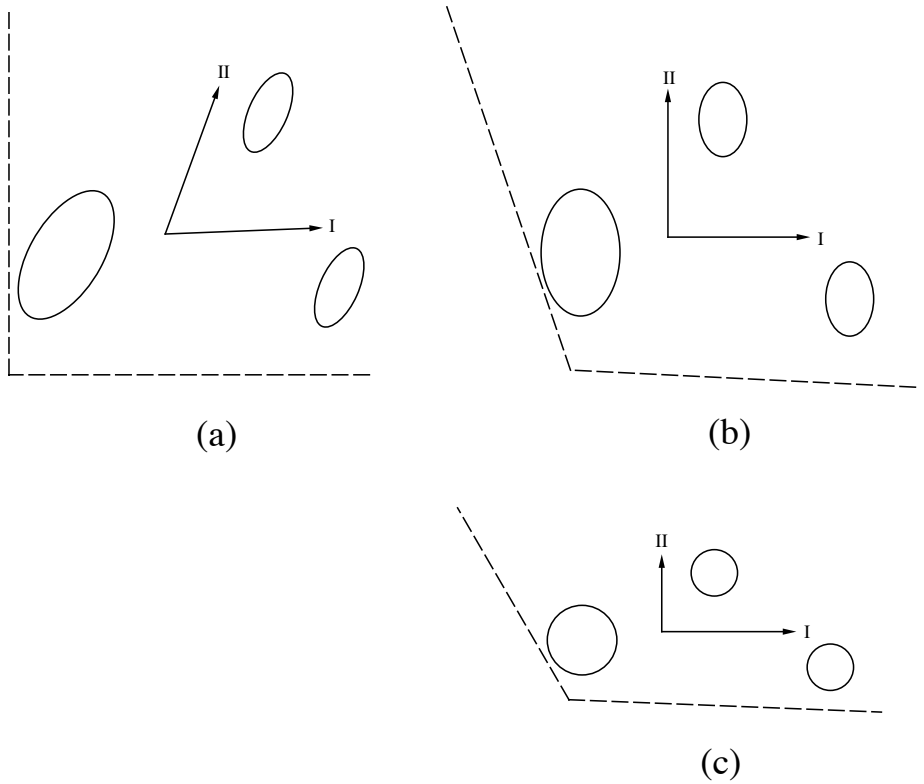


Figure 11.12 Basic principles of multiple discriminant analysis. Dashed: two original descriptors. Full lines: canonical axes. The within-group dispersion matrices are homogeneous in this example. (a) The canonical axes are not orthogonal in the reference space of the original descriptors. (b) When they are used as the orthogonal reference space, the ellipsoids of the within-group scatters of objects are straightened up. (c) Rescaling the eigenvectors to produce **C** (eq. 11.36) makes the within-group dispersions circular if they are homogeneous.

principal axis indicates the direction of largest variation among group centroids, and so on for the successive canonical axes, after the reference space has been straightened up to make each group spherical. The SAS and STATISTICA packages, among others, as well as function ***lda()*** of the MASS package in R, offer the normalization of eq. 11.36.

Other methods for normalizing the eigenvectors are found in the literature, i.e. to lengths 1 or $\sqrt{\lambda}$. Some statistical packages unfortunately compute the positions of the objects along the canonical axes (matrix **F**, eq. 11.37) directly from matrix **U** of the eigenvectors normalized to length 1. In that case, the group dispersions remain nonspherical; it is then difficult to compare the eigenvectors because they describe a combination of within-group and among-group dispersion. It is not always easy to

understand, from the documentation, what a specific statistical program does. A simple way to find out what kind of normalization is used by a program or R function is to run the small example presented in Subsection 11.3.3.

The last step of the computation is to find the positions of the objects in the space of the canonical axes. The matrix of discriminant scores **F** is obtained by multiplying the matrix of centred data with the matrix of normalized eigenvectors **C**:

$$\mathbf{F} = [x - \bar{x}]\,\mathbf{C} \qquad \textbf{(11.37)}$$

Since the matrix of centred data $[x - \bar{x}]$ is used in eq. 11.37, the origin of the discriminant axes is located at the centroid of all objects, as in Fig. 11.12. It is common practice to also compute the positions of the centroids of the g groups of objects in canonical space, by multiplying the matrix of the original group centroids (computed from data centred over all objects in the analysis) with matrix **C**. The centroid of a group is a point whose coordinates are the mean values of the objects of that group for all descriptors. The matrix of group centroids therefore has g rows and m columns.

As in principal component analysis, equation $\mathbf{F} = [x - \bar{x}]\,\mathbf{C}$ contains the scores of the objects, i, on each canonical axis k:

$$f_{ik} = (x_{i1} - \overline{x_1})\, c_{1k} + \ldots + (x_{ip} - \overline{x_p})\, c_{pk} \qquad \textbf{(11.38)}$$

The columns of matrix **F** are called *canonical variates* in discriminant analysis. The distances among objects in discriminant space are Mahalanobis distances (eq. 7.38). The positions of the group centroids in discriminant space can be found by computing these same functions for the mean values of the groups along the **X** variables.

If the analysis is carried out on the *non-standardized* descriptors, the columns of matrix **C** are called *identification functions*. Identification functions are used to place new objects in the canonical space. To do so, values of the various descriptors of a new object are centred using the same descriptor means as in eq. 11.38, and the centred values are multiplied by the weights c_{jk}. This provides the position of this object on the canonical axes. By plotting the point representing this object in the canonical ordination space together with the original set of objects, it is possible to identify the group to which the new object most likely belongs.

There are other ways of assigning objects to groups. *Classification functions*[*] are linear equations that can be used for that purpose. A separate classification function is computed as follows for each group j:

$$\text{Classification function for group } j = -0.5\ \bar{\mathbf{x}}'_j \mathbf{V}^{-1} \bar{\mathbf{x}}_j + \mathbf{V}^{-1} \bar{\mathbf{x}}_j \qquad \textbf{(11.39)}$$

[*] This terminology is unfortunate. In biology, classification consists in forming groups, using clustering methods for instance (Chapter 8), whereas identification is to assign objects to preestablished groups.

where $\mathbf{V}$ is the pooled within-group dispersion matrix (Table 11.8) and $\bar{\mathbf{x}}_j$ is the vector describing the centroid of group j for all m variables of matrix $\mathbf{X}$. Each classification function looks like a multiple regression equation; eq. 11.39 provides the weights ($\mathbf{V}^{-1}\bar{\mathbf{x}}_j$) to apply to the various descriptors of matrix $\mathbf{X}$ combined in the linear equation, as well as a constant ($-0.5\ \bar{\mathbf{x}}'_j\mathbf{V}^{-1}\bar{\mathbf{x}}_j$). The classification score of each object is calculated for each of the g classification functions; an object is assigned to the group for which it receives the highest classification score. Another way is to compute Mahalanobis distances (eq. 7.38) of the objects from each of the group centroids. An object is then assigned to the group to which it is the closest.

Confusion or classification table

A *confusion* or *classification table* (also called *confusion or classification matrix*) can be constructed; this is a contingency table comparing the original assignment of objects to groups (usually in rows) to the group assignments made by the classification functions (in columns). From that table, users can determine the number and percentage of the objects correctly classified by the discriminant functions.

An alternative way to obtain the eigenvalues and eigenvectors in discriminant analysis is through the canonical correlation equation (eq. 11.50). The method is described by Tatsuoka & Lohnes (1988, Section 7.8). The vector of classification levels is first transformed into a matrix $\mathbf{Y}$ containing ($g - 1$) binary or Helmert-coded variables (Subsection 1.5.7). One then computes the eigenvalues and eigenvectors of matrix $\mathbf{S}_{22}^{-1}\mathbf{S}'_{12}\,\mathbf{S}_{11}^{-1}\mathbf{S}_{12}$. The notation is the same as in eq. 11.46: $\mathbf{S}_{11}$ is the covariance matrix of $\mathbf{Y}$, $\mathbf{S}_{22}$ is the covariance matrix of $\mathbf{X}$, and $\mathbf{S}_{12}$ is the covariance matrix crossing the two groups of variables:

$$(\mathbf{S}_{22}^{-1}\mathbf{S}'_{12}\,\mathbf{S}_{11}^{-1}\mathbf{S}_{12} - h_k\mathbf{I})\,\mathbf{u}_k = 0 \qquad \textbf{(11.40)}$$

$\mathbf{S}_{22}^{-1}\mathbf{S}'_{12}\,\mathbf{S}_{11}^{-1}\mathbf{S}_{12}$ is the transpose of the matrix in eq. 11.50 (Tatsuoka & Lohnes, 1988, eq. 7.26). Because this matrix is of order ($m \times m$), the computed matrix of eigenvectors $\mathbf{U}$ has m rows and contains the weights associated with the predictors $\mathbf{X}$.

The discriminant eigenvalues l_k are found by transforming the eigenvalues h_k as follows:

$$l_k = h_k / (1 - h_k)$$

It is the eigenvalues l_k of eq. 11.34 that are found here because eq. 11.40 does not involve the degrees of freedom of Table 11.8. The discriminant eigenvalues λ_k are obtained by applying eq. 11.35 in reverse. The matrix of eigenvectors $\mathbf{U}$ is the same as obtained from eqs. 11.33 and 11.34. Matrix $\mathbf{C}$ of discriminant function coefficients can then be computed using eq. 11.36.

This approach is further supported by the demonstration made by Gittins (1985) that $\mathbf{S}'_{12}\,\mathbf{S}_{11}^{-1}\mathbf{S}_{12}$ is matrix $\mathbf{B}$ of Table 11.8, while $\mathbf{S}_{22}$ is matrix $\mathbf{T}$. Hence eq. 11.40, which is the form of the CCorA equation applied to discriminant analysis, can be written $(\mathbf{T}^{-1}\mathbf{B} - \lambda_k\mathbf{I})\,\mathbf{u}_k = 0$.

2 — Statistics in linear discriminant analysis

Spherical within-group dispersions are obtained only if the condition of homogeneity of the within-group dispersion matrices is fulfilled. Even if discriminant analysis is moderately robust to departures from this condition, it remains advisable to examine whether this condition is met prior to LDA. Several statistics have been developed to test the hypothesis of homogeneity of the within-group dispersion matrices. One of them is Kullback's statistic (1959) which is approximately distributed as χ^2:

$$\chi^2 = \sum_{j=1}^{g} \frac{(n_j - 1)}{2} \log_e \frac{|\mathbf{V}|}{|\mathbf{V}_j|} \qquad \textbf{(11.41)}$$

with $(g-1)m(m+1)/2$ degrees of freedom, where n_j is the number of objects in group j, $|\mathbf{V}|$ is the determinant of the pooled within-group dispersion matrix $\mathbf{V}$, and $|\mathbf{V}_j|$ is the determinant of the within-group dispersion matrix of group j. When the test value is larger than or equal to the critical χ^2 value, the hypothesis of homogeneity is rejected. Another method, which is robust to departures from normality, is the test of homogeneity of multivariate dispersions developed by Anderson (2006); this is the multivariate analogue of Levene's (1960) univariate test for homogeneity of variances. Anderson's test can be computed for any dissimilarity measure of choice. It is available in VEGAN's function ***betadisper()***.

Several important tests in discriminant analysis are based on Wilks' Λ (lambda) statistic (1932). This statistic can be used in an overall test to assess if the groups significantly differ in the positions of their centroids, given the within-group dispersions. Λ is computed as the ratio of the determinants of the matrices of sums of squares and cross-products $\mathbf{W}$ and $\mathbf{T}$:

$$\Lambda = \frac{|\mathbf{W}|}{|\mathbf{T}|} \qquad \textbf{(11.42)}$$

This ratio produces values in the range from near 0 (maximum dispersion of the centroids) to 1 (no dispersion among groups). It can be transformed to a X^2 (chi-square) statistic with $m(g-1)$ degrees of freedom (Bartlett, 1938):

$$X^2 = -\left[(n-1) - \frac{1}{2}(m+g) \right] \log_e \Lambda \qquad \textbf{(11.43)}$$

Alternatively, Wilks' Λ can be transformed into an F-statistic following Rao (1951). It is a generalization of Student's t-test to several groups and several explanatory variables. Another multidimensional generalization of t, for two groups, is Hotelling's T^2 (eq. 7.41), which has been discussed with reference to the Mahalanobis generalized distance (eq. 7.39).

As explained above, discrimination among g groups requires a maximum of $(g-1)$ discriminant functions. To test the significance of the $(g-k-1)$ eigenvalues that remain after examining the first k, Wilks' ratio is computed as the product

$$L = \prod_{j=k+1}^{g-1} \frac{1}{1+l_j} \qquad \textbf{(11.44)}$$

where the l_j are the eigenvalues of eq. 11.34. The value L computed for all eigenvalues produces the value Λ of eq. 11.42. Transformation of this statistic to X^2, as above (eq. 11.43), allows one to estimate the significance of the discriminant power of the axes remaining after accepting the first k eigenvalues as significant (Bartlett, 1948):

$$X^2 = \left[(n-1) - \frac{1}{2}(m+g)\right] \log_e \left[\prod_{j=k+1}^{g-1} (1+l_j)\right] \qquad \textbf{(11.45)}$$

with $(m-k)(g-k-1)$ degrees of freedom. (The logarithm of L from eq. 11.44 is equal to minus the logarithm of the product of the terms $(1+l_j)$ in the denominator.) When the last $(g-k-1)$ canonical eigenvalues, taken together, do not reach the chosen critical χ^2 value, the null hypothesis that the centroids of the groups do not differ on the remaining $(g-k-1)$ discriminant functions cannot be rejected. This indicates that the detectable discriminant power is limited to the first k discriminant functions.

3 — Numerical example

Discriminant analysis is illustrated by means of a numerical example in which seven objects, allocated to three groups, are described by two descriptors. The calculation of *identification functions* is shown first (raw data), followed by *discriminant functions* (standardized data). Normally, these data should not be submitted to discriminant analysis since the variances of the group matrices are not homogeneous; they are used here to illustrate the steps involved in the computation. The data set is the following:

$$\begin{array}{lcccccccc} \text{Groups} = & & 1 & & & 2 & & 3 & \text{Means} \\ \mathbf{X}' = & \left[\begin{array}{ccccccc} 1 & 2 & 2 & 8 & 8 & 8 & 9 \\ 2 & 2 & 1 & 7 & 6 & 3 & 3 \end{array}\right] & & & & & & & \begin{bmatrix} 5.42857 \\ 3.42857 \end{bmatrix} \end{array}$$

The centred data for the objects and the group centroids are the following:

$$\text{Groups} = \quad \underline{\quad 1 \quad} \quad \underline{\quad 2 \quad} \quad \underline{\quad 3 \quad}$$

$$[\mathbf{X}\text{ centred}]' = [x - \bar{x}]' = \begin{bmatrix} -4.429 & -3.429 & -3.429 & 2.571 & 2.571 & 2.571 & 3.571 \\ -1.429 & -1.429 & -2.429 & 3.571 & 2.571 & -0.429 & -0.429 \end{bmatrix}$$

$$[\text{Centroids}]' = \begin{bmatrix} -3.762 & 2.571 & 3.071 \\ -1.762 & 3.071 & -0.429 \end{bmatrix}$$

The matrix of sums of squares and cross-products is:

$$\mathbf{T} = [x - \bar{x}]'[x - \bar{x}] = \begin{bmatrix} 75.71429 & 32.71429 \\ 32.71429 & 29.71429 \end{bmatrix}$$

The pooled within-groups matrix **W** is computed by adding up the three group matrices of sums of squares and cross-products $\mathbf{W}_1$, $\mathbf{W}_2$ and $\mathbf{W}_3$:

$$\mathbf{W} = \begin{bmatrix} 0.66667 & -0.33333 \\ -0.33333 & 0.66667 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 0.5 \end{bmatrix} + \begin{bmatrix} 0.5 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1.16667 & -0.33333 \\ -0.33333 & 1.16667 \end{bmatrix}$$

The determinants of matrices **W** and **T** are 1.25000 and 1179.57, respectively. The ratio of these two values is Wilks' Λ (eq. 11.42: $\Lambda = 0.00106$; eq. 11.43: $X^2 = 23.97$, $p < 0.001$). This indicates that there are significant differences among the groups in the **X** variables. Hence, the analysis can proceed with the calculation of identification and discriminant functions.

To obtain the *identification functions*, the matrix of sums of squares among groups is computed as:

$$\mathbf{B} = \mathbf{T} - \mathbf{W} = \begin{bmatrix} 74.54762 & 33.04762 \\ 33.04762 & 28.54762 \end{bmatrix}$$

The characteristic equation $|\mathbf{B} - l\mathbf{W}| = 0$ is used to calculate the two eigenvalues:

$$l_1 = 106.03086 \quad \Rightarrow \quad \lambda_1 = \frac{(7-3)}{(3-1)} \times 106.03086 = 212.06171\,(93.13\%)$$

$$l_2 = 7.81668 \quad \Rightarrow \quad \lambda_2 = 2 \times 7.81668 = 15.63336\,(6.87\%)$$

In this example, canonical axes 1 and 2 explain 93.13 and 6.87% of the among-group variation, respectively. The two eigenvalues are used to compute the eigenvectors, by

means of matrix equation $(\mathbf{B} - l_k\mathbf{W})\,\mathbf{u}_k = \mathbf{0}$. These eigenvectors, normalized to length 1, are the columns of matrix **U**:

$$\mathbf{U} = \begin{bmatrix} 0.81202 & -0.47849 \\ 0.58363 & 0.87809 \end{bmatrix}$$

The vectors are not orthogonal since $\mathbf{u}_1' \mathbf{u}_2 = 0.12394$. In order to bring the eigenvectors to their final lengths, the following scaling matrix is computed:

$$\left(\mathbf{U}'\frac{\mathbf{W}}{n-g}\mathbf{U}\right)^{1/2} = \begin{bmatrix} 0.46117 & 0 \\ 0 & 0.60141 \end{bmatrix}$$

The component terms of each eigenvector $\mathbf{u}_j$ are *divided* by the corresponding diagonal term from this matrix, to obtain the final vectors (identification functions):

$$\mathbf{C} = \begin{bmatrix} 1.76077 & -0.79562 \\ 1.26553 & 1.46006 \end{bmatrix}$$

Multiplication of the centred matrices of the raw data and centroids by **C** gives the positions of the objects (matrix **F**) and centroids in canonical space (Fig. 11.13):

$$\text{Groups} = \quad \underline{\quad 1 \quad} \quad \underline{\quad 2 \quad} \quad \underline{\quad 3 \quad}$$

$$\mathbf{F} = [\mathbf{X}\ \text{centred}]\,\mathbf{C} = [x - \bar{x}]\,\mathbf{C} = \begin{bmatrix} -9.606 & -7.845 & -9.111 & 9.047 & 7.783 & 3.985 & 5.747 \\ 1.438 & 0.642 & -0.818 & 3.169 & 1.708 & -2.672 & -3.466 \end{bmatrix}'$$

$$[\text{Centroids}]\,\mathbf{C} = \begin{bmatrix} -8.854 & 8.415 & 4.866 \\ 0.420 & 2.438 & -3.069 \end{bmatrix}'$$

One can verify that, in canonical space, the among-group dispersion matrix **A** is equal to the matrix of eigenvalues and that the pooled within-groups dispersion matrix **V** is the identity matrix **I**. Beware: some computer programs calculate the discriminant scores as **XU** instead of [**X** centred] **U** or [**X** centred] **C**.

The *classification functions*, computed from eq. 11.39, are the following for descriptors x_1 and x_2 of the example:

Group 1: $\text{Score}_i = -13.33333 + 8.00000\, x_{i1} + 8.00000\, x_{i2}$

Group 2: $\text{Score}_i = -253.80000 + 36.80000\, x_{i1} + 32.80000\, x_{i2}$

Group 3: $\text{Score}_i = -178.86667 + 34.93333\, x_{i1} + 20.26667\, x_{i2}$

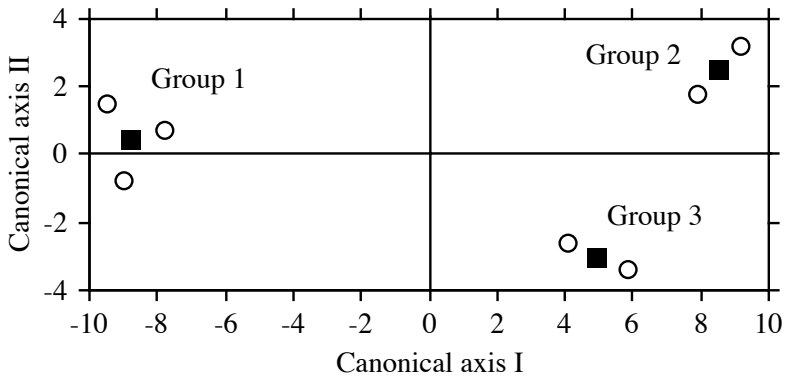


Figure 11.13 Ordination diagram of the seven objects (circles) and group centroids (squares) of the example in the canonical discriminant space.

The scores of the 7 objects i, computed from these functions, are the following:

Object number	Observed group	Function 1	Function 2	Function 3	Assigned to group
1	1	10.66667	–151.40000	–103.40000	1
2	1	18.66667	–114.60000	–68.46667	1
3	1	10.66667	–147.40000	–88.73334	1
4	2	106.66667	270.20000	242.46666	2
5	2	98.66667	237.40000	222.19999	2
6	3	74.66667	139.00000	161.39998	3
7	3	82.66667	175.80000	196.33331	3

Each object is assigned (right-hand column) to the group corresponding to the function giving it the highest score. The *classification table* can now be constructed; this is a contingency table comparing the original group assignment of the objects (from the second column in table above) to the assignment made from the classification functions (last column in table above):

Observed group	Assigned to group			Total and % correct
	1	2	3	
1	3	0	0	3 (100%)
2	0	2	0	2 (100%)
3	0	0	2	2 (100%)
Total	3	2	2	7 (100%)

In order to compute *discriminant functions*, the descriptors are standardized at the start of the analysis:

$$\text{Groups} = \quad \underline{\quad 1 \quad} \quad \underline{\quad 2 \quad} \quad \underline{\quad 3 \quad}$$

$$[\mathbf{X}\ \text{standardized}]' = \left[\frac{x-\bar{x}}{s_x}\right]' = \begin{bmatrix} -1.247 & -0.965 & -0.965 & 0.724 & 0.724 & 0.724 & 1.005 \\ -0.642 & -0.642 & -1.091 & 1.605 & 1.155 & -0.193 & -0.193 \end{bmatrix}$$

$$[\text{Centroids}]' = \begin{bmatrix} -1.059 & 0.724 & 0.865 \\ -0.792 & 1.380 & -0.193 \end{bmatrix}$$

The remaining calculations are the same as for the identification functions (above):

$$\mathbf{T} = \left[\frac{x-\bar{x}}{s_x}\right]'\left[\frac{x-\bar{x}}{s_x}\right] = \begin{bmatrix} 6.00000 & 4.13825 \\ 4.13825 & 6.00000 \end{bmatrix}$$

$$\mathbf{W} = \begin{bmatrix} 0.05283 & -0.04217 \\ -0.04217 & 0.13462 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 0.10096 \end{bmatrix} + \begin{bmatrix} 0.03962 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0.09246 & -0.04217 \\ -0.04217 & 0.23558 \end{bmatrix}$$

$$\mathbf{B} = \mathbf{T} - \mathbf{W} = \begin{bmatrix} 5.90755 & 4.18042 \\ 4.18042 & 5.76441 \end{bmatrix}$$

$$l_1 = 106.03086 \quad \Rightarrow \quad \lambda_1 = \frac{(7-3)}{(3-1)} \times 106.03086 = 212.06171\ (93.13\%)$$

$$l_2 = 7.81668 \quad \Rightarrow \quad \lambda_2 = 2 \times 7.81668 = 15.63336\ (6.87\%)$$

The amounts of among-group variation explained by the canonical axes (93.13 and 6.87%) are the same as those obtained above with the unstandardized data.

$$\mathbf{U} = \begin{bmatrix} 0.91183 & -0.65630 \\ 0.41057 & 0.75450 \end{bmatrix}$$

$$\left(\mathbf{U}'\frac{\mathbf{W}}{n-g}\mathbf{U}\right)^{1/2} = \begin{bmatrix} 0.14578 & 0 \\ 0 & 0.23221 \end{bmatrix} \quad \Rightarrow \quad \mathbf{C} = \begin{bmatrix} 6.25473 & -2.82627 \\ 2.81631 & 3.24918 \end{bmatrix}$$

$$\text{Groups} = \quad \underline{\quad 1 \quad} \quad \underline{\quad 2 \quad} \quad \underline{\quad 3 \quad}$$

$$\mathbf{F} = [\mathbf{X} \text{ standardized}]\,\mathbf{C} = \begin{bmatrix} -9.606 & -7.845 & -9.111 & 9.047 & 7.783 & 3.985 & 5.747 \\ 1.438 & 0.642 & -0.818 & 3.169 & 1.708 & -2.672 & -3.466 \end{bmatrix}'$$

$$[\text{Centroids}]\,\mathbf{C} = \begin{bmatrix} -8.854 & 8.415 & 4.866 \\ 0.420 & 2.438 & -3.069 \end{bmatrix}'$$

The raw and standardized data produce exactly the same ordination of the objects and group centroids.

The classification functions computed using the standardized descriptors differ from those reported above for raw data, but the classification table is the same in both cases.

Computer packages usually have an option for variable selection using forward entry, backward elimination, or stepwise selection, as in multiple regression (Subsection 10.3.3). These procedures are useful for selecting only the descriptors that significantly contribute to discrimination, leaving the others out of the analysis. This option must be used with caution. As it is the case with any stepwise computation method, the step-by-step selection of s successively most discriminant descriptors does not guarantee that they form the most discriminant set of s descriptors.

The following ecological application is an example of multiple discriminant analysis among groups of observations, using physical and chemical descriptors as discriminant variables. Steiner *et al*. (1969) applied discriminant analysis to the agronomic interpretation of aerial photographs, based upon a densimetric analysis of different colours.

Ecological application 11.3

Sea ice is an environment with a rich and diversified biota. This is because ice contains a network of brine cells and channels in which unicellular algae, heterotrophic bacteria, protozoa, and small metazoa can develop and often reach very high concentrations. Legendre *et al*. (1991) investigated the environmental factors controlling the growth of microscopic algae in the sea ice of southeastern Hudson Bay, Canadian Arctic.

Ice cores were taken at eight sites along a transect that extended from the mouth of the Great Whale River to saline waters 25 km offshore. Ice thickness ranged from 98 to 125 cm. The cores were used to determine the crystallographic structure of the ice, at 2 cm intervals from the top to the bottom of each core, together with several chemical and biological variables (nutrients, algal pigments, and taxonomic composition of algal assemblages) along the cores. The chemical and biological variables were determined on melted 10-cm thick sections of the cores; using crystallographic information, the chemical and biological data were transformed into values per unit of brine volume. The rate of ice growth for each 10-cm interval of each core was calculated

Table 11.9 Standardized canonical coefficients for the first two canonical variates.

Discriminant variable	Canonical variate 1	Canonical variate 2
Nitrate	–0.63	0.69
Phosphate	0.55	–0.08
Silicate	0.29	0.44
Rate of ice growth	0.89	0.54

by combining the mean daily air temperatures since the start of ice formation with the ice thickness at the date of sampling. Data on taxonomic composition of the algal assemblages in the brine cells were analysed as follows: (1) Similarities (χ^2 similarity; S_{21}, eq. 7.28) were computed among all pairs of core sections, on the basis of their taxonomic composition. (2) The similarity matrix was subjected to flexible clustering (Subsection 8.5.10) to identify groups of core sections that were taxonomically similar. (3) Discriminant analysis was used to determine which environmental variables best accounted for differences among the groups of core sections. Chlorophyll *a* is not a descriptor of the environment but of the ice algae, so that it was not used as discriminant variable; it is, however, the response variable in the path analysis mentioned below. Another approach to this question would have been to look directly at the relationships between the physical and chemical data and the species, using RDA or CCA.

Cluster analysis evidenced five groups among the 10-cm ice sections. The groups were distributed at various depths in the cores, sometimes forming clusters of up to 5 adjacent ice sections from within the same core. Discriminant analysis was conducted on standardized descriptors. The first canonical variate accounted for 62% of the variation among groups, and the second one 29%.

The standardized canonical coefficients for the first two canonical variates (Table 11.9) indicate that the environmental descriptors that best accounted for the among-group variation were the rate of ice growth (first variate) and nitrate (second variate). Figure 11.14 shows the position of the centroids of the 5 groups of core sections, plotted in the space of the first two canonical axes, with an indication of the role played by the environmental variables in discriminating among the groups of core sections. According to the figure, the groups of core sections are distributed along two gradients, one dominated by ice growth rate (with groups 1, 3 and 5 in faster-growing ice) and the other by nitrate (with group 1 in low-nitrate and group 5 in high-nitrate environments). These results are consistent with those of a path analysis (Section 10.4) conducted on the same data, showing that algal biomass (chl *a*) was inversely related to the rate of ice growth. The paper concluded that slower ice growth favoured the colonization of brine cells by microalgae (path analysis) and that the rate of ice growth had a selective effect on taxa, with nutrient limitation playing a secondary role in some brine cells (discriminant analysis).

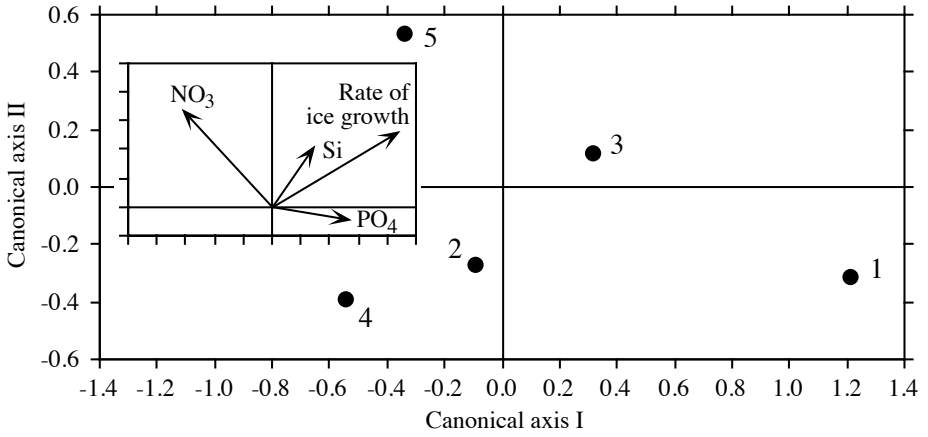


Figure 11.14 Centroids of the five groups of taxonomically similar core sections plotted along the first two canonical axes. Insert: contributions (from Table 11.9) of the four environmental variables (arrows) to the formation of the canonical axes. The groups of core sections are distributed along two gradients, one dominated by ice growth (groups 4, 2 and 3), the other by nitrate (groups 1, 3 and 5). Modified from Legendre *et al*. (1991).

11.4 Canonical correlation analysis (CCorA)

Canonical correlation analysis (CCorA; Hotelling, 1936), differs from redundancy analysis (RDA) in the same way as linear correlation differs from simple linear regression (Box 10.1). In CCorA, the two matrices under consideration are treated in a symmetric way whereas, in RDA, the **Y** matrix is considered to be dependent on an explanatory matrix **X**. The algebraic consequence is that, in CCorA, the matrix whose eigenvalues and eigenvectors are sought (eq. 11.48) is constructed from all four parts of eq. 11.1 whereas, in the asymmetric RDA method, eq. 11.8 does not contain the $\mathbf{S_{YY}}$ matrix.

In CCorA, the objects (sites) under study are described by two sets of quantitative descriptors between which a general form of correlation is sought; for example, a first set $\mathbf{Y}_1$ of p_1 chemical and a second set $\mathbf{Y}_2$ of p_2 geomorphological descriptors of the sampling sites. The dispersion matrix **S** of these $p_1 + p_2$ descriptors contains four sub-matrices, as in eq. 11.1:

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S'}_{12} & \mathbf{S}_{22} \end{bmatrix} \qquad \textbf{(11.46)}$$

The data matrices are designated $\mathbf{Y}_1$ and $\mathbf{Y}_2$, instead of $\mathbf{Y}$ and $\mathbf{X}$, to emphasize the fact that the two matrices play equivalent roles in CCorA. Matrices $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are designates by numbers (1, 2) to simplify the writing in the present section. Submatrices $\mathbf{S}_{11}$ (order $p_1 \times p_1$) and $\mathbf{S}_{22}$ (order $p_2 \times p_2$), represent, respectively, the covariance matrices of $\mathbf{Y}_1$ and $\mathbf{Y}_2$, whereas $\mathbf{S}_{12}$ (order $p_1 \times p_2$) and its transpose $\mathbf{S}'_{12} = \mathbf{S}_{21}$ (order $p_2 \times p_1$) represent the covariance matrix between the two sets of descriptors.

Gittins (1985) presents a comprehensive review of the theory and applications of CCorA in ecology. CCorA has limited applications nowadays because many two-matrix problems encountered in ecology are asymmetric and should be analysed by RDA (Section 11.1) or CCA (Section 11.2), whereas symmetric analyses are often conducted using the more flexible method of co-inertia analysis (Section 11.5).

1 — The algebra of canonical correlation analysis

Consider two response data sets $\mathbf{Y}_1$ ($n \times p_1$) and $\mathbf{Y}_2$ ($n \times p_2$), containing different variables about the same objects. They are to be related and compared in an analysis. CCorA does not invoke the directional hypothesis that $\mathbf{Y}_1$ may influence $\mathbf{Y}_2$ or the opposite.

The correlation coefficient between two variables is computed as $r_{jk} = s_{jk}/(s_j s_k)$ (eq. 4.7). Matrix $\mathbf{K}$ is constructed like a correlation coefficient:

$$\mathbf{K} = \mathbf{S}'^{-0.5}_{11}\mathbf{S}_{12}\mathbf{S}^{-0.5}_{22} \qquad \textbf{(11.47)}$$

$\mathbf{K}$ summarizes the correlation structure between data matrices $\mathbf{Y}_1$ and $\mathbf{Y}_2$. In this equation, $\mathbf{S}^{-0.5}_{11}$ is the inverse of the Cholesky root* of $\mathbf{S}_{11}$, and similarly for $\mathbf{S}^{-0.5}_{22}$. $\mathbf{K}$ would be identical if computed from correlation matrices $\mathbf{R}_{11}$, $\mathbf{R}_{12}$ and $\mathbf{R}_{22}$; this is why the same eigenvalues and eigenvectors are found in CCorA based on either covariance (raw data) or correlation matrices (standardized data). In this symmetric analysis, matrix $\mathbf{K}$ would be transposed if computed after inverting the roles of $\mathbf{Y}_1$ and $\mathbf{Y}_2$ in the equation. The algebra in the present section applies equally well to $\mathbf{S}$ matrices defined as matrices of sums of squares and cross products ($\mathbf{S_{YY}} = \mathbf{Y}'\mathbf{Y}$) or dispersion (variance-covariance) matrices ($\mathbf{S_{YY}} = (1/(n-1))\,\mathbf{Y}'\mathbf{Y}$).

The canonical correlation approach consists in maximizing the between-set dispersion with respect to the two within-set dispersions. The expression to be optimized is $\mathbf{S}_{12}\mathbf{S}^{-1}_{22}\mathbf{S}'_{12}\,\mathbf{S}^{-1}_{11}$ since $\mathbf{S}_{12}\mathbf{S}'_{12} / \mathbf{S}_{11}\mathbf{S}_{22}$ does not exist in matrix algebra.

* The Cholesky root of a matrix $\mathbf{A}$ is an upper triangular matrix $\mathbf{L}$ such that $\mathbf{L}'\mathbf{L} = \mathbf{A}$. Cholesky factorization is easier than computing the true square root of $\mathbf{A}$ using eq. 2.29.

Finding solutions to this optimization problem calls for eigenvalues and eigenvectors. Canonical eigenvalues are obtained by solving the characteristic equation:

$$\left|\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}'_{12}\,\mathbf{S}_{11}^{-1} - \lambda_k\mathbf{I}\right| = 0 \qquad \textbf{(11.48)}$$

In this equation, $\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}'_{12}\,\mathbf{S}_{11}^{-1}$ is an asymmetric matrix, as was that of numerical example 2 in Section 2.9. Its eigenvalues can be used in turn in the following equation, which results from the multiplication of the left and right members of eq. 11.48 by $\mathbf{S}_{11}$:

$$(\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}'_{12} - \lambda_k\mathbf{S}_{11})\,\mathbf{v}_k = 0 \qquad \textbf{(11.49)}$$

This equation is used to estimate the matrix of eigenvectors $\mathbf{V} = [\mathbf{v}_k]$ of the *first data set*. Because the analysis is symmetric, the same non-zero eigenvalues are found in the solution of the following equation, which is the dual of eq. 11.48:

$$\left|\mathbf{S}'_{12}\,\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1} - \lambda_k\mathbf{I}\right| = 0 \qquad \textbf{(11.50)}$$

The eigenvalues are now used in the following equation, which results from the multiplication of both sides of eq. 11.50 by $\mathbf{S}_{22}$:

$$(\mathbf{S}'_{12}\,\mathbf{S}_{11}^{-1}\mathbf{S}_{12} - \lambda_k\mathbf{S}_{22})\,\mathbf{u}_k = 0 \qquad \textbf{(11.51)}$$

This equation is used to estimate the matrix of eigenvectors $\mathbf{U} = [\mathbf{u}_k]$ of the *second data set*. Matrix $\mathbf{U}$ is also the matrix of eigenvectors of $\mathbf{K}'\mathbf{K}$ whereas $\mathbf{V}$ is the matrix of eigenvectors of $\mathbf{K}\mathbf{K}'$.

Equations 11.49 and 11.51 cannot be solved using regular eigenvalue decomposition as described in Section 2.9. So in practice, the solution is found by singular value decomposition of $\mathbf{K}$ (SVD, Section 2.11):

$$\mathbf{K}(p_1\times p_2) = \mathbf{V}(p_1\times c)\,\mathbf{W}(\text{diagonal}, c\times c)\,\mathbf{U}'(c\times p_2)^{*} \qquad \textbf{(11.52)}$$

The canonical correlations (r_k) are the singular values found on the diagonal of $\mathbf{W}$. The eigenvalues are the squared singular values, so the diagonal matrix of eigenvalues is:

$$\mathbf{\Lambda} = \mathbf{W}^2$$

The eigenvalues found here are the same as those of eqs. 11.48 to 11.51. The rank of the solution (i.e. the number of canonical axes) is equal to the number, c, of eigenvalues larger than 0, where $c \leq min(p_1, p_2)$.

---

* As explained in Section 2.11, the symbolism for SVD used in this book differs from that of the R language. Matrix $\mathbf{V}$ is component \$u of the output object of R function ***svd()*** while matrix $\mathbf{U}$ is component \$v.

The *canonical coefficients* give the contributions of the two sets of variables to the canonical axes. They are computed as follows:

$$Coeff_{\mathbf{Y}1} = \mathbf{S}_{11}^{-0.5}\mathbf{V} \quad \textbf{(11.53)}$$

$$Coeff_{\mathbf{Y}2} = \mathbf{S}_{22}^{-0.5}\mathbf{U} \quad \textbf{(11.54)}$$

The scores of the objects on the canonical axes form matrices $\mathbf{C}_{\mathbf{Y}1}$ and $\mathbf{C}_{\mathbf{Y}2}$, called the *canonical variates*. These matrices, with order $(n \times c)$, are computed as follows:

$$\mathbf{C}_{\mathbf{Y}1} = \mathbf{Y}_1 Coeff_{\mathbf{Y}1} = \mathbf{Y}_1 \mathbf{S}_{11}^{-0.5}\mathbf{V} \quad \textbf{(11.55)}$$

$$\mathbf{C}_{\mathbf{Y}2} = \mathbf{Y}_2 Coeff_{\mathbf{Y}2} = \mathbf{Y}_2 \mathbf{S}_{22}^{-0.5}\mathbf{U} \quad \textbf{(11.56)}$$

The canonical correlations found above are the Pearson correlations between the object scores in corresponding columns of $\mathbf{C}_{\mathbf{Y}1}$ and $\mathbf{C}_{\mathbf{Y}2}$. The interpoint distances in the canonical space are Mahalanobis distances. Indeed, CCorA of a data set by itself produces identical matrices $\mathbf{C}_{\mathbf{Y}1}$ and $\mathbf{C}_{\mathbf{Y}2}$; in these matrices, the Euclidean distances among objects are the Mahalanobis distances (D_5, eq. 7.38) among the objects in the original data matrices.

The variables of both sets are drawn in the canonical space using correlation matrices computed as follows:

plot variables $\mathbf{Y}_1$ in space $\mathbf{Y}_1$ using $\text{cor}(\mathbf{Y}_1,\mathbf{C}_{\mathbf{Y}1})$ **(11.57)**

plot variables $\mathbf{Y}_2$ in space $\mathbf{Y}_2$ using $\text{cor}(\mathbf{Y}_2,\mathbf{C}_{\mathbf{Y}2})$ **(11.58)**

One could also plot the variables of one set in the space of the other set, although this is rarely done:

plot variables $\mathbf{Y}_1$ in space $\mathbf{Y}_2$ using $\text{cor}(\mathbf{Y}_1,\mathbf{C}_{\mathbf{Y}2})$

plot variables $\mathbf{Y}_2$ in space $\mathbf{Y}_1$ using $\text{cor}(\mathbf{Y}_2,\mathbf{C}_{\mathbf{Y}1})$

These equations explain why analyses based upon unstandardized or standardized descriptors produce CCorA biplots with the same projections of the objects and variables: the projections of the objects are the same because the covariance matrices, which differ between unstandardized and standardized descriptors, are included in eqs. 11.55 and 11.56, and the projections of the variables are computed using correlations (eqs. 11.57 and 11.58), which are the same for unstandardized and standardized descriptors.

2 — Statistics in canonical correlation analysis

Several statistics derived from multivariate analysis of variance (MANOVA) can be used to test the significance of the canonical correlation between two matrices. The most commonly used is Wilks' Lambda likelihood ratio test, which is also used in discriminant analysis (eq. 11.42). These statistics may provide diverging diagnostics especially when they are tested parametrically. Pillai & Hsu (1979) showed that

Pillai's trace

Pillai's trace (V) is quite robust to non-normality; it also performs well in MANOVA. It is computed as follows:

$$V = \text{trace}(\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}'_{12}\,\mathbf{S}_{11}^{-1}) = \text{sum of the canonical eigenvalues} \qquad \textbf{(11.59)}$$

Permutation test

For normal data, V can be tested parametrically with reference to the F-distribution. For non-normal data, a permutation test of V is available in VEGAN's function ***CCorA()*** in addition to the parametric test. For that test, the rows of either $\mathbf{Y}_1$ or $\mathbf{Y}_2$ are permuted, $\mathbf{S}_{12}$ is recomputed using the permuted data, and $V = \text{trace}(\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}'_{12}\,\mathbf{S}_{11}^{-1})$ is recomputed. This operation is repeated a large number of times to obtain the sampling distribution of V under H_0.

When one of the matrices only contains one descriptor ($p_1 = 1$ and $p_2 > 1$ for example), there is only one positive eigenvalue. The canonical correlation problem reduces to the problem of finding the linear combination of variables in $\mathbf{Y}_2$ that is maximally correlated with the single variable $\mathbf{y}_1$; this is simply a problem of multiple correlation (Subsection 4.5.1). The general equation from which eigenvalues are computed in eq. 11.48 simplifies to:

$$\lambda = r^2 = \mathbf{s}_{12}\mathbf{S}_{22}^{-1}\mathbf{s}'_{12} / s_1^2 \qquad \textbf{(11.60)}$$

where $\mathbf{s}_{12}$ is a vector of covariances. This equation corresponds to that of multiple correlation (eq. 4.31), expressed in Chapter 4 in terms of $\mathbf{r}$ instead of $\mathbf{s}$. Finally, when the two sets contain only one descriptor each ($p_1 = p_2 = 1$), eq. 11.60 becomes:

$$\lambda = r^2 = s_{12}s_{22}^{-1}s_{12}s_{11}^{-1} = \frac{(s_{12})^2}{s_{11}s_{22}} = \frac{(s_{12})^2}{s_1^2 s_2^2} \qquad \textbf{(11.61)}$$

which is the formula for the square of the Pearson linear correlation (eq. 4.7). The parametric F-test of Pillai's trace gives the exact same p-value as the test of the Pearson correlation coefficient in that case.

3 — Applications of CCorA

CCorA cannot handle data sets with p_1 or p_2 greater than $(n - 1)$ because the covariance matrices $\mathbf{S}_{11}$ and $\mathbf{S}_{22}$ must be inverted (eqs. 11.48 to 11.51). This precludes the analysis of community composition data that contain more species than sites.

The interpretation of canonical correlation analyses is more difficult than that of other multidimensional analyses. The main use of this technique is to test the significance of the correlation between two multidimensional data sets, then explore the structure of the data by computing the correlations (which are the square roots of the CCorA eigenvalues) that can be found between linear functions of two groups of descriptors (Kendall & Stuart, 1966). The detailed (graphical) study of pairs of eigenvectors is usually restricted to the first few canonical correlations, although Blackith & Reyment (1971) give an example taken from Blackith & Albrecht (1959) where the lowest canonical correlations were of interest; the corresponding canonical eigenvectors made it possible to isolate a "phase" vector in locusts which was independent of the "size" vector.

When using CCorA, one should remember that strong canonical correlations do not necessarily mean that the corresponding vectors of ordination scores $\mathbf{C}_{\mathbf{Y}1}$ and $\mathbf{C}_{\mathbf{Y}2}$ explain a large fraction of the variation in $\mathbf{Y}_1$ or $\mathbf{Y}_2$; indeed, strong canonical correlations may be produced between members of a pair of canonical variates that may not explain large portions of the variance of the two data sets. *Redundancy coefficients* are used in CCorA to measure the proportion of the variance of $\mathbf{Y}_1$ (or $\mathbf{Y}_2$) that is explained by a linear combination of the variables in $\mathbf{Y}_2$ (or $\mathbf{Y}_1$); they should always be computed together with canonical correlations to help interpret them (Stewart & Love, 1968).

Ecological application 11.4

The Doubs river data of Verneaux (1973) were described and analysed by variation partitioning in Ecological application 11.1a. In the present example, CCorA is used to compare 3 geographic and topographic (linear distance from the source along the course of the river, slope, mean minimum discharge) to 7 water chemistry variables (pH, hardness, concentrations of phosphate (PO_4), nitrate (NO_3), ammonia (NH_4) and dissolved oxygen (O_2), and biological oxygen demand (BOD)) observed at 30 sites along the main course of the river. A similar analysis is presented by Borcard *et al*. (2011), where one topographic variable differs from the present analysis and some of the variables are pre-transformed to make their distributions more symmetrical.

The data sets individually explain fairly well the variation in fish assemblages along the river: a RDA of the Hellinger-transformed fish abundances by geography and topography (site 8 removed, where no fish were captured) produced $R^2_a = 0.45$, whereas a RDA of fish by chemistry produced $R^2_a = 0.47$. The present analysis will try to determine to what extent the water chemistry variables are correlated with the geographic and topographic variables.

Pillai's trace ($V = 1.54387$) is very highly significant, which shows that there is a significant correlation between the two groups of variables. RDA of geography and topography on chemistry produced a very high R^2_a of 0.49, whereas the opposite analysis, RDA of chemistry by geography and topography, produced a fairly high R^2_a of 0.29. The canonical correlations are high on the first two canonical axes: 0.93 for axis 1 and 0.72 for axis 2.

Two biplots (Fig. 11.15) show the canonical relationships. In both biplots, the sites are clearly divided between sites 1-15 on the left (highest portion of the river), associated with high

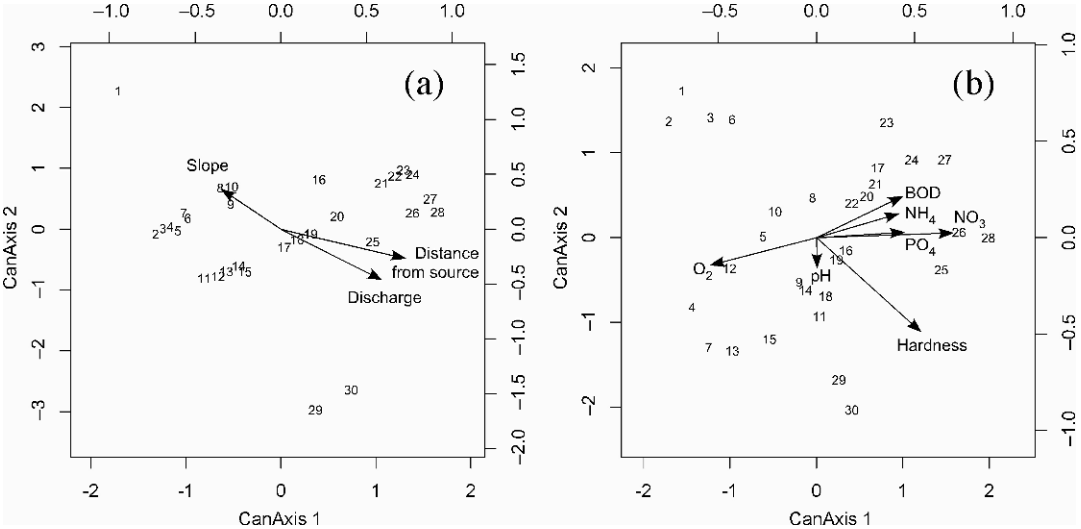


Figure 11.15 (a) CCorA biplot of the sites and the geographic and topographic variables, canonical axes 1 and 2. (b) Biplot of the sites and the water chemistry variables, same canonical axes. Results computed using function ***CCorA()*** of the VEGAN package. The sites are drawn in the graph using the lower and left-hand scales whereas the variables are positioned using the top and right-hand scales.

values of slope in Fig. 11.15a and O_2 in Fig. 11.15b, and sites 16-30 on the right (lowest portion of the river) which are associated with high values of distance from the source and discharge in Fig. 11.15a and BOD, NH_4, PO_4, NO_3 and hardness in Fig. 11.15b.

11.5 Co-inertia (CoIA) and Procrustes (Proc) analyses

Co-inertia analysis (CoIA) is an alternative to canonical correlation analysis,; it was proposed by Dolédec & Chessel (1994) to search for common structures between two data sets describing the same objects.The method is closely related to Procrustes analysis (Proc), described below. CoIA and Proc are symmetric forms of analysis, meaning that they are appropriate when either data set can be equally used as $\mathbf{Y}_1$ or $\mathbf{Y}_2$ in the analysis. This characteristic distinguishes the symmetric canonical ordination methods (CCorA, CoIA, Proc) from the asymmetric methods (RDA, CCA, LDA).

In CoIA, the variables of both data sets are projected onto the axes obtained by eigen-analysis of the cross-set covariance matrix. Various transformations can be used to correctly model the structure in each data set prior to CoIA (Dray *et al*., 2003).

1 — The algebra of co-inertia analysis (CoIA)

Consider two response data sets $\mathbf{Y}_1$ ($n \times p_1$) and $\mathbf{Y}_2$ ($n \times p_2$), containing different variables about the same n objects. The analysis will relate and compare the two sets. Like CCorA, CoIA does not invoke the directional hypothesis that $\mathbf{Y}_1$ may influence $\mathbf{Y}_2$, or the opposite. CoIA is compared to canonical correlation analysis (CCorA) in Subsection 11.5.3.

The analysis for two data sets is conducted as follows:

Total co-inertia

- Compute the covariance matrix crossing the variables of the two data sets. The sum of the squared covariances is the total co-inertia. Compute the eigenvalues and eigenvectors of that matrix. The eigenvalues represent a partitioning of the total co-inertia.

- Project the objects and variables of the two original data sets on the co-inertia axes. Different graphs are produced to compare the projections of the two data sets in the common co-inertia space.

Here is the algebraic development. Co-inertia analysis is implemented in R by the ADE4 function ***coinertia()***, and some of its computational details are mentioned below to clarify what the function does. Compute the matrix of covariances crossing the two data sets $\mathbf{Y}_1$ ($n \times p_1$) and $\mathbf{Y}_2$ ($n \times p_2$):

$$\mathbf{Cov}_{12} = \frac{1}{n-1}\mathbf{Y}'_{1.\text{cent}}\mathbf{Y}_{2.\text{cent}} \qquad \textbf{(11.62)}$$

The notation $\mathbf{Y}_{1.\text{cent}}$ and $\mathbf{Y}_{2.\text{cent}}$ indicates that the two matrices are centred to have column means of 0 before computing $\mathbf{Cov}_{12}$. $\mathbf{Cov}_{12}$ is matrix $\mathbf{S}_{12}$ of Subsection 11.5.1. Compute the singular value decomposition (Section 2.11) of $\mathbf{Cov}_{12}$, with the following result:

$$\mathbf{Cov}_{12}\ (p_1 \times p_2) = \mathbf{V}\ (p_1 \times c)\ \mathbf{W}(\text{diagonal}, c \times c)\ \mathbf{U}'\ (c \times p_2) \qquad \textbf{(11.63)}$$

The value c is defined a few lines down. As mentioned above, the total co-inertia is the sum of the squared covariances in $\mathbf{Cov}_{12}$. It is partitioned among the CoIA eigenvalues, which are the squares of the singular values found on the diagonal of $\mathbf{W}$. So, the diagonal matrix of eigenvalues is:

$$\mathbf{\Lambda} = \mathbf{W}^2 \qquad \textbf{(11.64)}$$

One could have carried out an eigen-decomposition of $\mathbf{Cov}_{12}'\ \mathbf{Cov}_{12} = \mathbf{S}'_{12}\,\mathbf{S}_{12}$ or $\mathbf{Cov}_{12}\mathbf{Cov}_{12}' = \mathbf{S}_{12}\mathbf{S}'_{12}$, instead of a SVD of $\mathbf{Cov}_{12}$: the eigenvalues of these decompositions are the same as those found by squaring the singular values in $\mathbf{W}$; the matrix of eigenvectors of $\mathbf{Cov}_{12}'\ \mathbf{Cov}_{12}$ is matrix $\mathbf{U}$ whereas that of $\mathbf{Cov}_{12}\mathbf{Cov}_{12}'$ is matrix $\mathbf{V}$. The rank of the solution (i.e. the number of co-inertia axes) is equal to the number (c) of eigenvalues larger than 0, where $c \leq \min(p_1, p_2)$.

The objective of co-inertia analysis is to project the objects and variables of the two data sets onto this common multivariate space and compare their positions. This is done as follows.

• To obtain the positions of the objects of $\mathbf{Y}_1$ in the common space, compute $\mathbf{F}_1 = \mathbf{Y}_{1.\text{cent}}\mathbf{V}$, then normalize each column of $\mathbf{F}_1$ to length 1 (eq. 2.7). Multiply the resulting matrix by $\sqrt{n-1}$ and by $\mathbf{\Lambda}^{1/2}$. As a result, column k of $\mathbf{F}_1$ has variance λ_k; this preserves the Euclidean distance among the objects as in PCA scaling type 1. Proceed in the same way for the second data set: compute $\mathbf{F}_2 = \mathbf{Y}_{2.\text{cent}}\mathbf{U}$, then normalize $\mathbf{F}_2$ and multiply by $\sqrt{n-1}$ and by $\mathbf{\Lambda}^{1/2}$. Use the normalized $\mathbf{F}_1$ and $\mathbf{F}_2$ to construct *a single plot* showing the two sets of objects; add arrows going from the representation of each object in $\mathbf{F}_1$ to the representation of the corresponding object in $\mathbf{F}_2$; invert the order of the input data sets to obtain arrows going in the other direction. The objects that have very close representations (i.e. short arrows) in the joint plot contribute more to the co-inertia (overall similarity) between the data sets than objects that are linked by long arrows; see the ecological application below. One may choose to further norm the columns of $\mathbf{F}_1$ and $\mathbf{F}_2$ to variances of 1 and use the resulting matrices for plotting, thus preserving the Mahalanobis distances among objects in the joint plot; this is done in ADE4 function ***coinertia()***, as discussed in the notes below.

• Project the variables of $\mathbf{Y}_1$ and $\mathbf{Y}_2$ onto the canonical axes: draw arrows anchored at the zero-origin of the plot using the coordinates provided by matrices $\mathbf{V}$ for the variables of $\mathbf{Y}_1$ and $\mathbf{U}$ for the variables of $\mathbf{Y}_2$.

Function ***coinertia()*** of ADE4 carries out the computation slightly differently from the description above. The differences are mentioned here to allow comparison between the results obtained with the above equations and those of the ADE4 function.

• In ***coinertia()***, the covariance matrix is computed as $\mathbf{Cov}_{21} = \frac{1}{n}\mathbf{Y}'_{2.\text{cent}}\mathbf{Y}_{1.\text{cent}}$.

• Function ***coinertia()*** was designed to handle data sets where the rows within each set may have different weights, representing for instance different sizes of sampling units or different row sums of abundances in CA. CoIA requires, however, that the weights be the same for $\mathbf{F}_1$ and $\mathbf{F}_2$; the row weights are scaled to sum to 1. In ***coinertia()***, the lengths of the column vectors are computed with these row weights. When the weights are all equal, this amounts to multiplying all values in $\mathbf{F}_1$ and $\mathbf{F}_2$ by $\sqrt{n}$; the column vectors normalized in that way have lengths of $\sqrt{n}$. For equal row weights, this multiplication has no incidence on the joint plot other than changing the numerical scales along the axes of the graph.

• The final normalization of matrices $\mathbf{F}_1$ and $\mathbf{F}_2$ to variances equal to the respective eigenvalues is not done in function ***coinertia()***. With equal row weights, the vectors in $\mathbf{F}_1$ and $\mathbf{F}_2$ are normalized to constant lengths of $\sqrt{n}$ (or variance = 1); this preserves the Mahalanobis distance among objects as in PCA scaling type 2. Compared to scaling type 1, this operation shrinks the objects along axis 1 and stretches them along axis 2. The plot no longer preserves the Euclidean distances among objects, but the two sets of objects are more easily represented into a square plot and the arrows joining the corresponding objects are more easily seen. Co-inertia plots resulting from different pre-treatments of the data sets, e.g. a PCA and a PCoA, are also easier to read.

• Function ***coinertia()*** requires the prior computation of separate ordinations, one for $\mathbf{Y}_1$ and the other for $\mathbf{Y}_2$, using a ***dudi.xxx()*** function. The raw data are included in the ***dudi.xxx()*** output lists. The ***coinertia()*** function retrieves them from these output objects to compute the covariance matrix. For ordinary data sets or community composition data pre-transformed using for instance the Hellinger transformation, use ***dudi.pca()***. To apply a distance function other than the Euclidean distance, use ***dudi.pco()*** which carries out principal coordinate analysis (PCoA).

• Function ***coinertia()*** also produces graphs of the principal axes of the two data sets $\mathbf{Y}_1$ and $\mathbf{Y}_2$. These additional graphs, which are not an essential part of co-inertia analysis, indicate how the principal axes of $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are related to the axes of the common co-inertia solution.

Data may have to be transformed prior to co-inertia analysis: standardize the data (eq. 1.12) if a set contains variables expressed in different physical units, or carry out a Hellinger, chord, or chi-square transformation (Section 7.7) for presence-absence or abundance community composition data with many zeros. No transformation is required if the Euclidean distance among objects is to be preserved.

Else, compute distance matrices $\mathbf{D}_1$ and $\mathbf{D}_2$ using distance coefficients appropriate to each type of data (Chapter 7). Carry out a principal coordinate analysis (PCoA, Section 9.3) of each $\mathbf{D}$ matrix and obtain tables of principal coordinates $\mathbf{PC}_1$ and $\mathbf{PC}_2$. If negative eigenvalues are present, retain only the axes corresponding to the positive eigenvalues, or apply a correction method to make all eigenvalues positive (Subsection 9.3.4). Use matrices $\mathbf{PC}_1$ and $\mathbf{PC}_2$ as input into co-inertia analysis.

Note that if the preliminary analysis incorporates vectors of row weights (row weight can be imposed in functions ***dudi.pca()*** and ***dudi.pco()*** of ADE4), CoIA requires that the weights must be the same for both data sets, a condition that precludes using CoIA with the results of two correspondence analyses (CA). Indeed, CA is a weighted regression method, and the row weights differ between data sets since they depend on the data in each matrix. Applying one or the other vector of weights to both data sets would lead to different CoIA solutions, and there seems to be no logical way of deciding between these two solutions. Co-correspondence analysis (ter Braak & Schaffers, 2004) offers a way to handle that problem; it is available in function ***cocorresp()*** of package COCORRESP. Another way of producing an analysis of two community composition data sets that preserves chi-square distances (eq. 7.55) is to apply a chi-square transformation (eq. 7.70) to each table and use them as input into CoIA; with ADE4, pre-process these tables using ***dudi.pca()***.

RV coefficient

Two overall statistics of co-inertia are available. The first one is the *RV* coefficient (Escoufier, 1973; Robert & Escoufier, 1976), which is a multivariate generalization of the Pearson correlation coefficient; for two vectors $\mathbf{x}_1$ and $\mathbf{x}_2$, $RV(\mathbf{x}_1,\mathbf{x}_2) = \text{cor}(\mathbf{x}_1,\mathbf{x}_2)^2$. The second one is the Procrustes statistic described in the next subsection. Coefficient *RV* is computed as follows for two rectangular data matrices with corresponding objects as rows, centred to column means of 0:

$$RV(\mathbf{Y}_1, \mathbf{Y}_2) = \frac{trace\,(\mathbf{Y}_1\mathbf{Y}'_1\mathbf{Y}_2\mathbf{Y}'_2)}{\sqrt{trace\,(\mathbf{Y}_1\mathbf{Y}'_1\mathbf{Y}_1\mathbf{Y}'_1)\;trace\,(\mathbf{Y}_2\mathbf{Y}'_2\mathbf{Y}_2\mathbf{Y}'_2)}} \qquad \textbf{(11.65)}$$

The *RV* coefficient can also be computed from the matrices of sums of squares and cross-products $\mathbf{SS}_{12} = \mathbf{Y}_1'\mathbf{Y}_2$, $\mathbf{SS}_{11} = \mathbf{Y}_1'\mathbf{Y}_1$, and $\mathbf{SS}_{22} = \mathbf{Y}_2'\mathbf{Y}_2$:

$$RV(\mathbf{Y}_1, \mathbf{Y}_2) = \frac{sum\,(\mathbf{SS}_{12}^{(2)})}{\sqrt{sum\,(\mathbf{SS}_{11}^{(2)})\; sum\,(\mathbf{SS}_{22}^{(2)})}} \qquad \textbf{(11.66)}$$

where the notation $\mathbf{SS}_{12}^{(2)}$ means that each element of $\mathbf{SS}_{12}$ is squared before summation. Significance of the *RV* coefficient can be tested by permutation (Heo & Gabriel, 1998) or parametrically (Josse *et al*., 2008). The null hypothesis of the test is: the two data sets are no more related than random data sets would be; this is the same kind of null hypothesis as in correlation analysis.

Permutation test

Ecological application 11.5, part 1

Oribatid mites (Acari: Oribatida) are a very diversified group of small (0.2 to 1.2 mm) soil-dwelling arthropods. In June 1989, Daniel Borcard collected 70 soil cores from the peat blanket of a bog lake on the territory of *Station de biologie des Laurentides* of Université de Montréal, Québec, Canada. During the following weeks, he extracted, identified, and counted 9800 individuals found therein, which he separated into 35 morphospecies. The mite and environmental data are fully described in a paper by Borcard & Legendre (1994); the data set is available in the VEGAN and ADE4 R packages. For the present ecological application, Dr. Borcard divided the species into a group of 23 panphytophagous species, which eat vegetation debris (most of their regime) as well as algae, fungi, spores, pollen grains, and bacteria, and 12 microphagous species which feed mostly on algae, fungi, spores, and pollen grains. The presence of vegetation debris in the diet of the panphytophagous species differentiates the two groups, which were almost equally represented in the data sets: there were 5667 panphytophagous and 4133 microphagous individuals.

The two data sets, which were considered to represent different communities, were Hellinger-transformed (eq. 7.69) separately, then subjected to co-inertia analysis using the ADE4 function ***coinertia()***. All row weights were equal. The *RV* coefficient (*RV* = 0.36038) was highly significant (p = 0.0001 after 9999 permutations), indicating a strong relationship between the two data sets. The first two canonical axes represented respectively 83% and 10% of the co-inertia. Figure 11.16a shows the 70 sites from the two data sets projected in the co-inertia space and linked by arrows (tail of each arrow = panphytophagous, head = microphagous). The low-numbered sites were physically located near the margin of the forest surrounding the bog lake whereas the high-numbered sites were near the free water portion of the bog. In Fig. 11.16a, most sites are equally distant from the origin of the plot and their arrows have about the same lengths, showing that they contribute fairly equally to the total co-inertia; only a few sites, with high site numbers, have long arrows. Panels b and c show that most of the 12 microphagous species contribute to the dispersion of the sites in the co-inertia plane (long arrows), whereas only 4 or 5 of the 23 panphytophagous species contribute strongly to that dispersion.

Two mite species associations were identified by Legendre (2005). It is interesting to note that the species with fairly long arrows (important contributions) in Fig. 11.16 (panels b and c), found in the same quadrant of the two species projections, also belong to the same species association: species 9, 31, 34 and 35 in quadrant 1, species 1, 10 and 15 in quadrant 2, species 13, 14 and 27 that point left along axis 1, species 7 and 11 in quadrant 3, and species 16 and 23

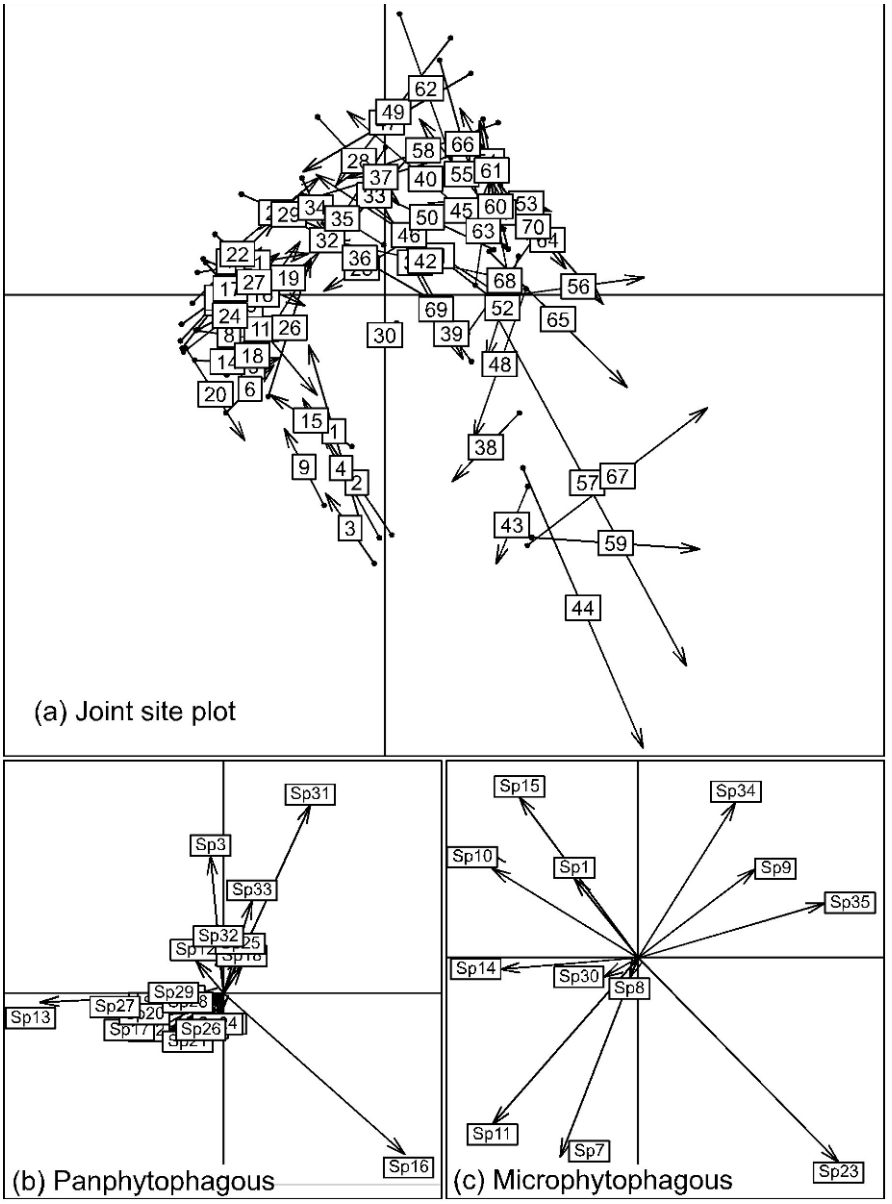


Figure 11.16 Results of co-inertia analysis of the mite data. (a) Joint site plot for the two sets of objects represented in the plane of canonical axes 1 and 2, which accounts for 93% of the total co-inertia. Arrows link objects from set 1 (panphytophagous species at the arrow tails) to the corresponding objects from set 2 (microphagous species at the arrow heads). (b) Projection of the panphytophagous and (c) of the microphagous species onto the co-inertia plane. Figure produced by the ***plot.coinertia()*** function of ADE4, then modified using a graphics editor.

in quadrant 4. In addition, a majority of the panphytophagous species point toward the left-hand site of their plot (panel b), where, in panel a, are found the sites that are closer to the forest and contain coarser vegetation debris.

Interestingly, site-species associations are also identifiable in Fig. 11.16. For example, species 16 (in panel b) and 23 (in panel c) dominate the communities in sites 38, 43, 44, 59 and 67. These site-species associations can be verified in the matrices of Hellinger-transformed mite data.

It is also interesting to look at arrow directions in panel (a): for most of the low-numbered sites (with the exceptions of sites 1 to 4, 9 and 15), the arrows point from left to right, indicating that the panphytophagous species put these sites at larger distances from the other sites than the microphytophagous species do. So for these sites, the panphytophagous species contribute more to beta diversity than the microphytophagous species. Indeed, for sites 5-8, 10-14 and 16-29, the sum of the variances of the Hellinger-transformed data, which is a measure of beta diversity (Legendre *et al*., 2005), is 0.25638 for the panphytophagous and 0.13374 for the microphytophagous species. This method of calculation produces a value of 1 when the sites have completely different species compositions; this corresponds to the situation where beta diversity is maximum (Subsection 6.5.3).

Co-inertia analysis is appropriate to compare pairs of data sets that play equivalent roles in the analysis. The method finds a common space onto which the objects and variables of these data sets can be projected and compared. The analysis may, for example, concern two segments of the species forming an ecological community (as in Ecological application 11.5, part 1). One could also compare data sets representing the physical or biological characteristics of organisms (individuals, species) to their behavioural characteristics. Moretti & Legg (2009) used co-inertia analysis to compare plant and invertebrate animal functional traits across forest sites with different fire and cutting histories. Several other examples are described in Dray *et al*. (2003). Compared to CCorA, co-inertia analysis imposes no constraint regarding the number of variables in the two sets, so that it can be used to compare ecological communities even when they are species-rich; see the comparison of methods in Subsection 11.5.3.

Co-inertia analysis is not well-suited, however, to analyse pairs of data sets that contain the same variables, because the analysis does not establish one-to-one correspondences between variables in the two data sets; the method does not 'know' that the first variable is the same in the first and the second data sets, and likewise for the other variables. Data of that type are found in before-after (BA) or control-impact (CI) studies. When the two data sets contain most or all of the same species, they can be analysed by placing the data 'before' on top of those 'after' in a joint data file, and computing a PCA of the combined data; the before-after pairs can then be linked by arrows in the common PCA ordination graph. Else, the difference between the two sections of the data table can be tested by RDA for the effect of a 'before-after' factor, in the presence of covariables representing the pairing of the sampling sites; see *Analysis of related samples* in Subsection 11.1.10, point 3. For data of that type, the RDA test has greater power to detect a difference than a co-inertia test because it uses the information more efficiently. The null hypothesis of the RDA test is H_0: there is no difference between 'before' and 'after' for data described by the same variables,

whereas the null hypothesis in CoIA is H_0: the two data sets have no more co-inertia structure than random data sets would have, without any reference to the variables being the same in the two data sets.

Multiple factor analysis

Multiple factor analysis (MFA) can be used to compare several data sets describing the same objects (Escofier & Pagès, 1994). MFA consists in projecting objects and variables of two or more data sets on a global PCA, computed from all data sets, in which the sets receive equal weights. For the comparison of two data sets, the algebra of MFA differs from that of CoIA. This method is implemented in functions ***mfa()*** of ADE4 and ***MFA()*** of FACTOMINER; the latter offers more options. A summary of the theory as well as an ecological application are presented in Section 6.10 of Borcard *et al*. (2011).

2 — Symmetric Procrustes analysis (Proc)

Orthogonal Procrustes analysis

Co-inertia analysis is closely related to the orthogonal Procrustes analysis of two data sets. The orthogonal Procrustes problem was first formulated by Hurley & Cattell (1962) who called their computer program PROCRUSTES after the villain of Greek mythology*; later authors referred to the method by that name. The problem consists in finding the best superposition of two sets of corresponding objects (i.e. the *n* objects of the two sets are the same) by rotation and mirror reflection, if necessary, of one of the data sets with respect to the other, in such a way as to minimize the sum of squared distances between the corresponding objects. A general least-squares solution was described by Schönemann (1966) and Schönemann & Carroll (1970) and perfected by Gower (1971b). The Procrustes rotation solution can be asymmetric, meaning that one matrix is rotated to maximum fit while the other is kept fixed, or symmetric in the sense described below.

Asymmetric Procrustes rotation

In symmetric Procrustes rotation, described here, each of the two data sets, $\mathbf{Y}_1$ ($n \times p_1$) and $\mathbf{Y}_2$ ($n \times p_2$), is standardized to have its total variance equal to 1 prior to rotation. This is obtained by Gower's standardization (Gower, 1971b), which consists in dividing each value in a column-centred data matrix by the square root of the total variance of the matrix, which is also the square root of the sum of its eigenvalues. The covariance of the two Gower-standardized matrices, $\mathbf{Y}_{1.\text{Gower}}$ and $\mathbf{Y}_{2.\text{Gower}}$, is then computed using the same covariance formula as in co-inertia analysis (Subsection 11.5.1):

$$\mathbf{Cov}_{12} = \frac{1}{n-1}\mathbf{Y}'_{1.\text{Gower}}\mathbf{Y}_{2.\text{Gower}} \qquad \textbf{(11.67)}$$

Note that co-inertia analysis (Subsection 11.5.1) of two Gower-standardized matrices produces the same relative eigenvalues, *RV* coefficient, and plots, as the CoIA of the

* For a brief description of the story of Procrustes in Greek mythology, see the first paragraph of Subsection 10.5.4.

original data sets. Procrustes analysis differs from co-inertia analysis in that it uses different output matrices for the joint plot of the two sets of objects.

The singular values of $\mathbf{Cov}_{12}$ are computed by singular value decomposition (Section 2.11):

$$\mathbf{Cov}_{12}\ (p_1 \times p_2) = \mathbf{V}\ (p_1 \times c)\ \mathbf{W}(\text{diagonal}, c \times c)\ \mathbf{U}'\ (c \times p_2) \qquad \textbf{(11.68)}$$

and the trace of **W** is computed:

$$\text{Trace}\mathbf{W} = \sum(\text{singular values}) \qquad \textbf{(11.69)}$$

The singular values are the diagonal values of **W**. Because singular values are positive or null, Trace**W** is non-negative. The Procrustes residual sum-of-squares statistic (Gower, 1971b, 1975*; Davis, 1978) is

$$m_{12}{}^2 = 1 - \text{Trace}\mathbf{W}^2 \qquad \textbf{(11.70)}$$

Following that, the rotation matrix that provides the best adjustment of the objects of $\mathbf{Y}_2$ to the objects of $\mathbf{Y}_1$ is computed as:

$$\mathbf{H} = \mathbf{U}\mathbf{V}' \qquad \textbf{(11.71)}$$

The rotated matrix $\mathbf{Y}_{2.\text{rot}}$ is computed as follows:

$$\mathbf{Y}_{2.\text{rot}} = \text{trace}\mathbf{W}\ \mathbf{Y}_{2.\text{Gower}}\ \mathbf{H} \qquad \textbf{(11.72)}$$

Symmetric Procrustes rotation

where trace**W** acts as a scaling factor. The Procrustes analysis is called symmetric when the two data sets are subjected to Gower standardization; it remains asymmetric in the fact that $\mathbf{Y}_{2.\text{Gower}}$ is projected after optimal rotation ($\mathbf{Y}_{2.\text{rot}}$) onto $\mathbf{Y}_{1.\text{Gower}}$. Objects are plotted on a graph using matrices $\mathbf{Y}_{1.\text{Gower}}$ and $\mathbf{Y}_{2.\text{rot}}$. It is thus more interesting (but not compulsory) in most instances to start Procrustes analysis with a matrix $\mathbf{Y}_{1.\text{ord}}$ that represents an ordination of $\mathbf{Y}_1$, e.g. by PCA or PCoA. Differences in positions between corresponding objects of the two data sets can be interpreted as in co-inertia analysis (Subsection 11.5.1).

Permutation test

For permutation testing, one can use either the $m_{12}{}^2$ statistic, or its complement the Procrustes $R^2 = (1 - m_{12}{}^2) = \text{Trace}\mathbf{W}^2$, or else Trace**W**. The latter is a Procrustean form of the correlation coefficient; its value is always positive or null. This permutation test, called PROTEST (Jackson, 1995; Peres-Neto & Jackson, 2001), is available in VEGAN's ***protest()*** function described in Subsection 10.5.4. It tests the same hypothesis as the test of the *RV* coefficient in CoIA. Two-matrix (or *Classical*) Procrustes rotation has been extended to *m* matrices in *Generalized Procrustes analysis* (Gower, 1975).

Generalized Procrustes analysis

* Gower called the residual sum of squares statistic R^2 in 1971b and $m_{12}{}^2$ in his 1975 paper.

Symmetric Procrustes analysis is appropriate for the same types of questions as co-inertia analysis, the difference between the two methods residing in the matrices used to plot the objects. Likewise, the situations where symmetric Procrustes analysis is inappropriate are the same as for CoIA analysis. Procrustes analysis is also appropriate to compare the results of ordinations derived from two sets of distances, for example the distances computed among sets of morphological landmarks (which form the data rows) measured on two organisms, or ordinations obtained by different methods, e.g. PCA and CA of the same data; no test of significance is possible in that case, however, because the original data are the same in the two ordinations.

Ecological application 11.5, part 2

The mite data used to illustrate co-inertia analysis (Subsection 11.5.1) were subjected to a symmetric Procrustes rotation using VEGAN's function ***procrustes()***. Principal components of the Hellinger-transformed panphytophagous species were compared to the rotated (Hellinger-transformed) microphagous species. The value of Trace**W**, which is used as the test statistic in PROTEST (Subsection 10.5.4), was 0.53994; the test was highly significant (p = 0.0001 after 9999 permutations). Figure 11.17 illustrates the symmetric Procrustes rotation results. Note that the axes of the graph are not the canonical axes of the co-inertia analysis, Fig. 11.16. The rotated vectors (species in this example) of the microphagous data set are shown as crossed hairs with numbers in the centre of the plot.

3 — Canonical correlation, Procrustes, or co-inertia analysis?

Which method should be used for symmetric analysis of two data sets? Table 11.10 compares the properties and requirements of canonical correlation analysis, on the one hand, to those of co-inertia and Procrustes analyses, on the other hand. A particularly interesting feature of CoIA and Proc, compared to CCorA, is that they allow the joint analysis of two community composition data sets with more species than there are sites. In addition, as in PCA, collinearity among the variables in one or the other data sets produces no problem in CoIA and Proc. This is not the case in CCorA where collinearity may prevent the computation of the inverses of the covariance matrices of the separate data sets, $\mathbf{S}_{11}^{-1}$ and $\mathbf{S}_{22}^{-1}$.

Note that CoIA carried out between two sets of principal components or principal coordinates is not equivalent to CCorA of these same matrices: the eigenvalues and eigenvectors of the two-table analyses are not the same. The reason is that $\mathbf{S}_{11}$ and $\mathbf{S}_{22}$ are diagonal matrices of eigenvalues in that case, not identity matrices $\mathbf{I}$, so that the matrix $[\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}'_{12}\,\mathbf{S}_{11}^{-1}]$, which is decomposed in CCorA (eq. 11.48), is not equal to matrix $[\mathbf{S}_{12}\mathbf{S}'_{12}]$ which is subjected to eigen-decomposition in CoIA.

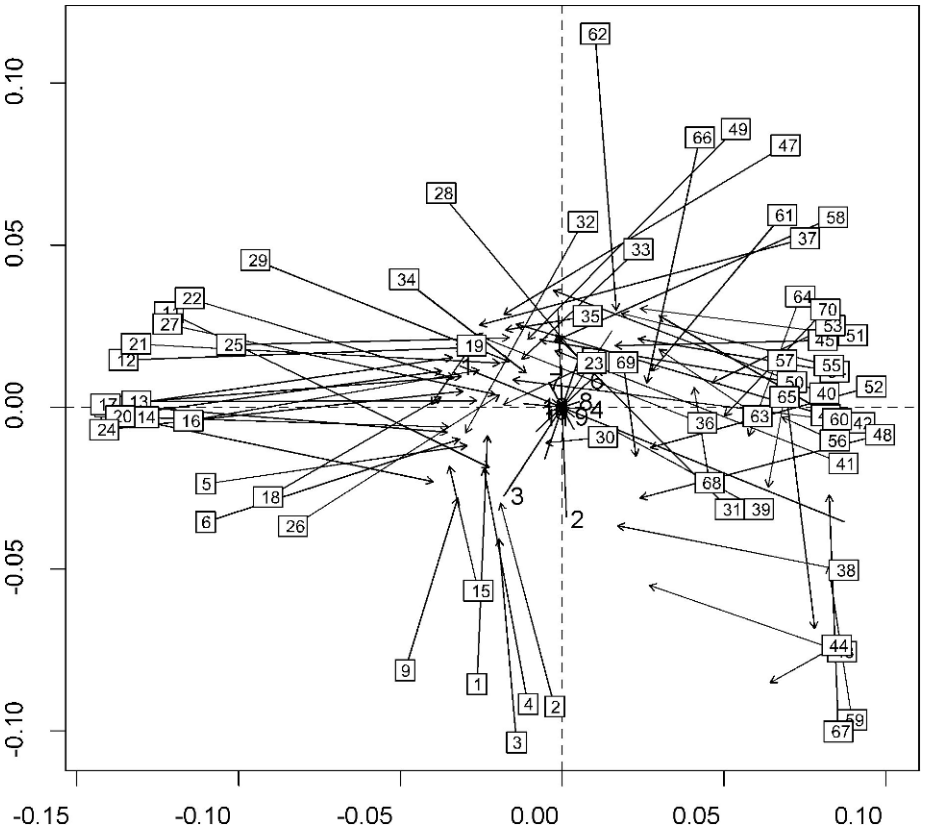


Figure 11.17 Results of symmetric Procrustes rotation of the mite data showing the two sets of objects along axes 1 and 2. Arrows link objects (numbers in boxes) from set 1 (principal components of the panphytophagous species at the arrow tails) to the corresponding objects from set 2 (microphagous species at the arrow heads). The plot was produced by VEGAN's function ***plot.procrustes()***.

11.6 Canonical analysis of community composition data

In early numerical ecology papers, canonical correlation analysis and discriminant analysis were used to analyse tables of species presence/absence or abundance data. In many applications, however, the assumptions of linearity and the algebraic constraints imposed by the models make these methods unsuitable for such data. RDA and CCA provide alternatives that are often more appropriate. Let us consider different types of situations that may involve species data.

Table 11.10 Comparison of canonical correlation analysis (CCorA), on the one hand, to co-inertia analysis (CoIA) and symmetric Procrustes analysis (Proc), on the other.

	CCorA	CoIA, Proc
Matrix sizes	p_1 and $p_2 < n$	No constraint
Physical dimensions	The variables in each set are standardized in CCorA	All variables in each set must have the same physical dimensions*
Eigen-analysis of …	$\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}'_{12}\,\mathbf{S}_{11}^{-1}$	$\mathbf{S}_{12}\mathbf{S}'_{12}$ (or SVD of $\mathbf{SS}_{12}$)
The canonical axes …	Maximize squared correlations among sets	Maximize squared covariances (co-inertia) among sets
Distances among objects preserved	Mahalanobis distances**	Euclidean distances
Test of significance of the relationship	Statistic: Pillai's trace, etc.	Statistics: RV, Trace$\mathbf{W}$

* Standardize (eq. 1.12) the variables of data sets that have heterogeneous physical dimensions. This can be done before the analysis, or by the computer programs performing the analysis.

** Mahalanobis distances among points (eq. 7.38) are preserved in CCorA. They are equal to Euclidean distances computed from matrix $\mathbf{G}$ used to position the objects in PCA scaling 2 (correlation biplot). In this scaling, the ordination axes are stretched to account for the correlations among variables (Subsection 9.1.4).

1. *Species in* $\boldsymbol{Y}$. — The first case involves a matrix $\mathbf{Y}$ of species presence-absence or abundance data and a matrix $\mathbf{X}$ of habitat characteristics. One may wish to find support for the ecological hypothesis of environmental control of the species distributions (Whittaker, 1956; Bray & Curtis, 1957), and/or describe in what way the species are related to the environmental variables. The analysis is not symmetric; the species clearly form the response variables, to be explained by the environmental variables. Hence a symmetric form of analysis such as CCorA or CoIA is not appropriate; one should rely instead on the asymmetric forms of analysis, RDA and CCA.

In some applications, it is interesting to compare two groups of species found together at a series of sampling sites. There are often more species than sites when these analyses involve species-rich communities. CCorA is unable to analyse such data because it cannot handle more variables in any one of the data sets than there are sites minus 1. Rare species would have to be dropped from the analysis to satisfy the requirements of the method. Co-inertia and Procrustes analyses do not have this limitation and can be used for this type of analysis.

When **X** contains dummy variables describing e.g. types of habitat (qualitative variable) recoded as in Subsection 1.5.7, RDA or CCA may be used to test the hypothesis that groups of sites, identified *a priori*, do not differ in species composition. The question is of the same type as in multivariate analysis of variance. Likewise, when **X** codes for factors of an experiment, RDA or CCA may be used for analysis.

2. *Species presence-absence*. — Ecologists may wish to use the environmental variables in **X** to forecast a classification criterion (**y**) representing the presence or absence of a single species, a group of species, or a functional trait at various locations.

Linear discriminant analysis (LDA) is a suitable choice to separate the sites where the species is present from those where it is absent. The classification functions (eq. 11.39) will attribute the observations to the rightful group if the descriptors allow it. Another appropriate statistical model is logistic regression (Subsection 10.3.7) because the forecasted response is binary, 0 (absence) or 1 (presence).

Between linear discriminant analysis (Section 11.3) and logistic regression (Subsection 10.3.7), which method is the most appropriate? Efron (1975) has shown that when the groups are drawn from populations with multinormal distributions in the space of the explanatory variables, discriminant analysis is more effective than logistic regression. On the other hand, logistic regression is more robust than discriminant analysis to departures from multivariate normality. This finding is important for the analysis of species with unimodal distributions along environmental gradients (Subsection 9.2.4): a species may be absent *under both low and high values* of an environmental variable. One can plot scatter diagrams of the presence/absence of the species of interest against each of the environmental variables in matrix **X**. When a unimodal response is detected, a quadratic [orthogonal] polynomial function of that explanatory variable should be used in the logistic model (see Gaussian logistic response, Subsection 10.3.7). The multivariate dispersions of groups of observations representing the presence and absence of a species in the space of an explanatory variable to the powers 1 and 2 (or 1, 2 and 3) cannot be multivariate normal. So in this situation, Gaussian logistic regression should be preferred to discriminant analysis.

3. *Indicator species*. — Species may represent the explanatory variables (matrix **X**). What are the species that characterize different types of habitat? In such cases, the types of habitat form the classification criterion **y**. The method of choice is indicator species analysis (Subsection 8.9.3).

Discriminant analysis is ill-adapted to this type of problem because it requires that there be more objects (n) than descriptors (m) in **X**. Actually, Williams & Titus (1988) recommend that the total number of observations *per group* be at least three times the number of variables in **X**; ter Braak (1987c) recommends that n be much larger than the number of species (m) plus the number of groups (g).

4. *Inverse analysis*. — RDA may be used as a form of inverse analysis to relate species assemblages to types of habitat. The classification criterion (e.g. types of habitat) is represented by a set of dummy variables, written into response matrix **Y**.

The species data are the explanatory variables **X**; they should most likely be transformed using one of the transformations of Section 7.7. The condition of more objects (n) than species (m) must be satisfied in this analysis. For species-rich communities, a solution may be to replace the abundances of individual species by the abundances of species associations identified using an appropriate statistical method; see Section 8.9.

Matrix **X**, in which each species is represented by a vector, may be transformed prior to RDA or discriminant analysis, by replacing the m species vectors by m ordination axes produced by PCA of the transformed species data, or by $(m - 1)$ ordination axes obtained by CA of the raw species data. An alternative is to compute a similarity or distance matrix among sites using the species data and obtain new axes by principal coordinate analysis (PCoA). PCA, CA or PCoA axes might relate to the environmental descriptors better than the original species data. The solution using PCoA is implemented in the CAP method of Anderson & Willis (2003).

One may wish to use the species data in **X** to predict or reconstruct one or more environmental variables in **Y**. This case, which is related to Ecological application 11.2b, is like CCA but with **X** and **Y** interchanged. A solution, which circumvents the too-many-species-problem, is Weighted Averaging Partial Least Squares (WA-PLS), which extends PLS regression in the correspondence analysis framework (ter Braak, 1995).

11.7 Software

Among the methods of canonical analysis, commercial statistical packages usually offer canonical correlation analysis and linear discriminant analysis. RDA and CCA are available in CANOCO* as well as in other packages, in particular PC-ORD and SYN-TAX 2000†.

The R language offers functions for all methods described in this chapter:

1. Redundancy analysis (RDA). — Simple and partial RDA is available in function ***rda()*** in VEGAN, and in package RDATEST found on the Web page http://numericalecology.com/rcode. Selection of explanatory variables in RDA: ***ordistep()*** and ***ordiR2step()*** in VEGAN, ***forward.sel()*** in PACKFOR. Principal response curves are computed by function ***prc()*** in VEGAN. Function ***varpart()*** is available in VEGAN for variation partitioning by RDA. db-RDA: function ***capscale()*** in VEGAN

* CANOCO is available from Plant Research International, Wageningen, The Netherlands. http://www.canoco.com/.

† PC-ORD (http://www.pcord.com) is available from MjM Software, P.O. Box 129, Gleneden Beach, Oregon 97388, USA. SYN-TAX 2000 (http://ramet.elte.hu/~podani) is available from Exter Software, 47 Route 25A, Suite 2, Setauket, New York 11733-2870, USA.

offers db-RDA based on any of the distance functions in ***vegdist()***. Multivariate analysis of variance: function ***manovRDa()*** for two-way crossed-factor MANOVA, for fixed or random factors, is available on the Web page http://www.elaliberte.info/. A similar function ***anova.2way.unbalanced()*** for two fixed factors, balanced or unbalanced designs, is available on the Web page http://numericalecology.com/rcode; type III sums-of-squares are used in the analysis of unbalanced designs. Function ***nested.anova.dbrda()*** for nested MANOVA with two levels (the main factor and one nested factor) is available in the BIODIVERSITYR package.

2. Canonical correspondence analysis (CCA). — Simple and partial CCA: function ***cca()*** in VEGAN, ***cca()*** in ADE4. Function ***CCA()*** is available on the Web page http://numericalecology.com/rcode; this function was written to demonstrate the CCA algorithm described in Subsection 11.2.1; it is fully functional for calculation of CCA and plotting triplots, but tests of significance are not available in that function.

3. Linear discriminant analysis (LDA). — ***lda()*** in MASS, ***discrimin()*** in ADE4. Test of homogeneity of multivariate dispersions: ***betadisper()*** in vegan. Selection of explanatory variables in LDA can be carried out with function ***stepclass()*** of package KLAR (direction = "forward", "backward" or "both").

4. Canonical correlation analysis (CCorA). — Functions ***cancor()*** of STATS, ***CCorA()*** of VEGAN, and ***cc()*** of CCA.

5. Co-inertia (CoIA) and Procrustes (Proc) analyses. — Co-inertia analysis in ***coinertia()*** of ADE4. Test of the *RV* coefficient: permutational test in ***RV.rtest()*** of ADE4, parametric test in ***coeffRV()*** of FACTOMINER. One can also apply function ***randtest.coinertia()*** to the output of function ***coinertia()***. Asymmetric and symmetric Procrustes analysis in ***procrustes()*** of VEGAN and ***procustes()*** of ADE4. Permutation test of the Procrustes statistic: function ***protest()*** of VEGAN. Co-correspondence analysis in ***cocorresp()*** of package COCORRESP. Multiple factor analysis (MFA) in ***mfa()*** of ADE4 and ***MFA()*** of FACTOMINER; the latter offers more options.

6. Miscellaneous methods. — Functions for palaeoenvironmental reconstruction, in particular ***MAT()***, ***MLRC()***, ***WA()*** and ***WAPLS()***, are available in package RIOJA. QR decomposition is computed by ***qr()*** of BASE; this is an efficient computation method for regression coefficients in linear models, e.g. in RDA.

PERMANOVA (permutational ANOVA/MANOVA) is an add-on package for PRIMER 6* that carries out permutational multivariate analysis of variance. This program tests the simultaneous response of one or more variables to one or more factors in an ANOVA experimental design on the basis of any distance measure, using permutation methods. The latest version of the program can handle any balanced ANOVA design up to nine factors.

* Available from PRIMER-E Ltd., 3 Meadow View, Lutton, Ivybridge, PL21 9RH, England.

Chapter

12

Ecological data series

12.0 Ecological series

The use and analysis of *data series* has become increasingly popular in ecology, especially because many terrestrial, aquatic and atmospheric observing stations measure and record environmental variables either automatically or with human intervention. Ecological data series contain continuous or discrete (discontinuous) variables sampled over time or along transects in space.

Stochastic process

A data series is a sequence of observations that are *ordered* along a temporal or spatial axis. As mentioned in Section 1.0, a series is one of the possible realizations of a *stochastic process*. A *process* is a phenomenon (response variable), or a set of phenomena, which is organized along some independent axis. In most cases, the independent axis is time, but it may also be space, or a trajectory through both time and space (e.g. sampling during a cruise). *Stochastic processes* generally exhibit three types of components, i.e. deterministic, systematic, and random. Methods for the numerical analysis of data series are designed to characterize the deterministic and systematic components present in series, given the probabilistic environment resulting from the presence of random components.

The most natural axis along which processes may be studied is *time* because temporal phenomena develop in an irreversible way, and independently of any decision made by the observer. The temporal evolution of populations or communities, for example, provides information that can unambiguously be interpreted by ecologists. Ecological variability is not a characteristic limited to the time domain, however; it may also be studied across space. In that case, the decisions to be made concerning the observation axis and its direction depend on the working hypothesis. In ecology, the distinction between space and time is not always straightforward. At a fixed sampling location, for example, time series analysis may be used to study the spatial organization of a moving system (e.g. migrating populations, plankton in a current), whereas a spatial series is required to assess temporal changes in that same

Eulerian
Lagrangian

system. The first approach (i.e. at a fixed point in space) is called *Eulerian*, and the second (i.e. at a fixed point within a moving system) is known as *Lagrangian*.

Periodic phenomena

Ecologists are often interested in *periodic* changes. This follows in part from the fact that many ecological phenomena are largely determined by geophysical rhythms, but there are also rhythms that are endogenous to organisms or ecosystems. The geophysical cycles of glaciations, for example, or, at shorter time scales, the solar (i.e. seasons, days) or lunar (tides) periods, play major roles in ecosystems. Endogenous rhythms, also called biological clocks (including the well-known circadian, i.e. 24-hour, rhythms), are extensively described in the scientific literature.

The analysis of data series often provides unique information concerning ecological phenomena. However, the quality of the results depends to a large extent on the *sampling design*. As a consequence, data series must be sampled following well-defined rules, in order (1) to preserve the spatio-temporal variability, which is often minimized on purpose in other types of ecological sampling design, and (2) to take into account the various conditions prescribed by the methods of numerical analysis. These conditions will be detailed later in the present chapter. An even more demanding framework prevails for *multidimensional series*, which result from sampling several variables simultaneously. Most numerical methods require that the series be made up of *large numbers of observations* ($n > 100$, or even $n > 1000$) for the analysis to have enough statistical power to provide conclusive results, especially when large random variation is present. Long series require extensive sampling. This is often carried out, nowadays, using equipment that automatically measures and records the variables of ecological interest. There are also a few methods that have been especially designed for the analysis of short time series; they are discussed below.

Observational window
Lag
Period
Frequency
Wavelength
Wavenumber

The most fundamental constraint in periodic analysis is the *observational window*. The width of this window is determined by the number of observations in the data series (n) and the interval (time or distance) between successive observations. This interval is called the *lag*, Δ; for the time being, it is assumed to be uniform over the whole data series. These two characteristics set the time or space domain that can be "observed" when analysing data series (Table 12.1). For temporal data, one refers to either the *period* (T, in time units) or the *frequency* ($f = 1/T$) whereas, for spatial data, the corresponding concepts are the *wavelength* (λ, in spatial distance units) and the *wavenumber* ($1/\lambda$).

The length of the series (Δn) sets, for temporal data, the *fundamental period* ($T_0 = \Delta n$) or *fundamental frequency* ($f_0 = 1/T_0 = 1/\Delta n$) and, for spatial data, the *fundamental wavelength* ($\lambda_0 = \Delta n$) or *fundamental wavenumber* ($1/\lambda_0 = 1/\Delta n$). *Harmonic periods* and *wavelengths* are *integral fractions* of the fundamental period and wavelength, respectively ($T_i = T_0/i$ and $\lambda_i = \lambda_0/i$, where $i = 1, 2, \ldots n$), whereas *harmonic frequencies* and *wavenumbers* are *integral multiples* of the fundamental frequency and wave number, respectively ($f_i = if_0$ and $1/\lambda_i = i/\lambda_0$). Concerning the actual limits of the observational window, the *longest* period or wavelength that can be statistically investigated is, at best, equal to *half the length* of the series ($\Delta n/2$). For

Table 12.1 Characteristics of the observational window in periodic analysis. Strictly speaking, the length of a data series is $(n-1)\Delta$ but, for simplicity, one assumes that the series is long, hence $(n-1) \approx n$.

Harmonic i	Period (T_i) Wavelength (λ_i)	Frequency (f_i) Wavenumber (i)	
1	$n\Delta$	$1/n\Delta$	Fundamental value, i.e. the whole series
2	$n\Delta/2$	$2/n\Delta$	Limit of observational window
.	.	.	
.	.	.	
i	$n\Delta/i$	$i/n\Delta$	ith harmonic
.	.	.	
.	.	.	
$n/2$	2Δ	$1/2\Delta$	Limit of window: Nyquist frequency

example, in a study on circadian (24-h) rhythms, the series must have a *minimum* length of two days (better 4 days or more). Similarly, in an area where spatial structures are of the order of 2 km, a transect must cover *at least* 4 km (better 8 km or more). Similarly, the *shortest* period or wavelength that can be resolved is equal to *twice the interval* between observations (2Δ). In terms of frequencies, the highest possible frequency that can be resolved, $1/2\Delta$, is called the *Nyquist frequency*. For example, if one is interested in hourly variations, observations must be made *at least* every 30 min. In space, in order to resolve changes at the metre scale, observations must be collected along a transect *at least* every 50 cm, or closer.

Nyquist frequency

To summarize the above notions concerning the observational window, let us consider a variable observed every month during one full year. The data series would allow one to study periods ranging between (2×1 month = 2 months) and (12 months/2 = 6 months). Periods shorter than 2 months and longer than 6 months are outside the observational window. In other words, statistical analysis cannot resolve frequencies higher than 1/(2 months) = 0.5 cycle month^{-1} = 6 cycles year^{-1} (Nyquist frequency), or lower than 1/(6 months) = 0.167 cycle month^{-1} = 2 cycles year^{-1}. The longest period (or lowest frequency) of the observational window is easy to understand, by reference to the usual notion of degrees of freedom (Box 1.2). Indeed, in order to have minimum certainty that the observed periodic phenomenon is real, this phenomenon must be observed at least twice, which provides only one degree of freedom. For example, if an annual cycle was observed over a period of one year

only, there would be no indication that it would occur again during a second year (i.e. no degree of freedom). A similar reasoning applies to the shortest period (or highest, Nyquist frequency) detectable in the observational window. For example, if the observed phenomenon exhibits monthly variation (e.g. oscillations between maximum and minimum values over one month), two observations a month would be the absolute minimum required for identifying the presence of that cycle.

Most methods described in the present chapter are limited to the observational window. However, some methods are mathematically capable of going beyond the upper limit (in terms of periods) of the window, because they can fit incomplete cycles of sine and cosine functions to the data series. This is the case of Dutilleul's modified periodogram (Section 12.4) and spectral analysis (Section 12.5). A significant period found in this region (e.g. a 3-month period in a data series 4 months long) should be interpreted with care. It only indicates that a longer time series should be observed and analysed (e.g. > 1 year of data) before drawing ecological conclusions.

Aliasing

There exists another constraint, which is also related to the observational window. This constraint follows from a phenomenon known as *aliasing*. It may happen that the observed variable exhibits fluctuations whose frequency is *higher than the Nyquist frequency*. This occurs when a period T or wavelength λ of the observed variable is smaller than 2Δ. Undersampling an important high-frequency fluctuations may generate an artificial signal in the series, whose frequency is *lower than the Nyquist frequency* (Fig. 12.1). Researchers unaware of the phenomenon could attempt to interpret this artificial low frequency in the series; this would obviously be incorrect. To avoid aliasing, the sampling design must provide at least four data points per cycle of the *shortest* important period or wavelength of the variable under study. The latter period or wavelength may be determined either from theory or from a pilot study.

The sections that follow explore various aspects of series analysis. The methods discussed are those best adapted to ecological data. Additional details may be found in the biologically-oriented textbook of Diggle (1990) and the review paper of Fry *et al.* (1981), or in other textbooks on time series analysis, e.g. Jenkins & Watts (1968), Bloomfield (1976), Box & Jenkins (1976), Brillinger (1981), Priestley (1981a, b), Kendall *et al.* (1983), Chatfield (1989), Kendall & Ord (1990), Venables & Ripley (2002), Dutilleul (2011) and Shumway & Stoffer (2011, with R examples). Methods for analysing time series of ecological and physiological chronobiological data were reviewed by Legendre & Dutilleul (1992).

12.1 Characteristics of data series and research objectives

Signal
Trend
Noise

Observed data series may be decomposed into various components, which can be studied separately since they have different statistical and ecological meanings. Figure 12.2 shows an artificial data series constructed by adding three components: a periodic signal, a trend, and a noise component. Series may be analysed in terms of

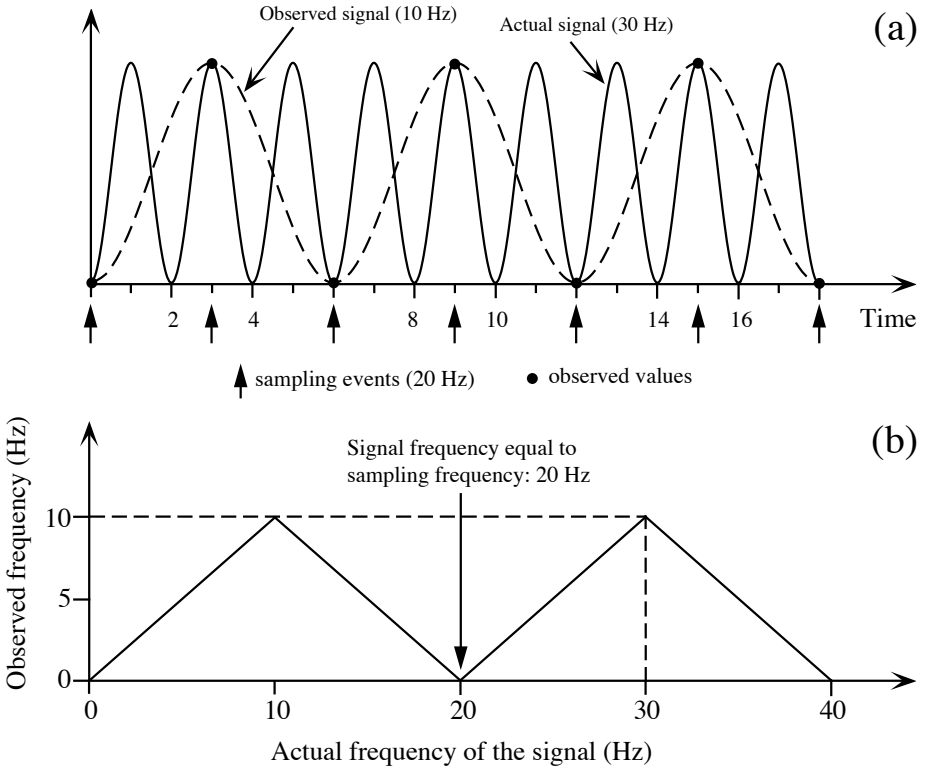


Figure 12.1 Aliasing. (a) The artificial signal detected in the data series (dashed line) is caused by observations (dots) made at a frequency lower than twice that present in the series under study (solid line). Along the abscissa, 1 time unit = 1/60 s. (b) With a sampling frequency of 20 Hz, the observed frequency (ordinate) varies between 0 and 10 Hz, as the actual frequency of the signal increases (abscissa). The observed frequency would be equal to the frequency of the signal (no aliasing) only for a signal ≤ 10 Hz, which is half the 20 Hz sampling frequency. In the example, the frequency of the signal is 30 Hz and the observed (aliased) frequency is 10 Hz (dashed line).

deterministic change (trend), *systematic* (periodic) variability, and *random* fluctuations (noise). Data series may be recorded with different objectives in mind (Table 12.2), to which are associated different methods of time series analysis. The following presentation of objectives is largely drawn from Legendre & Dutilleul (1992).

Trend

Objective 1. — Ecological data series often exhibit a deterministic component, known as the *trend*. The trend may be linear, polynomial, cyclic, etc. This deterministic component underlies the evolution of the series (Fig. 12.2a). It must be extracted as the first step of the analysis.

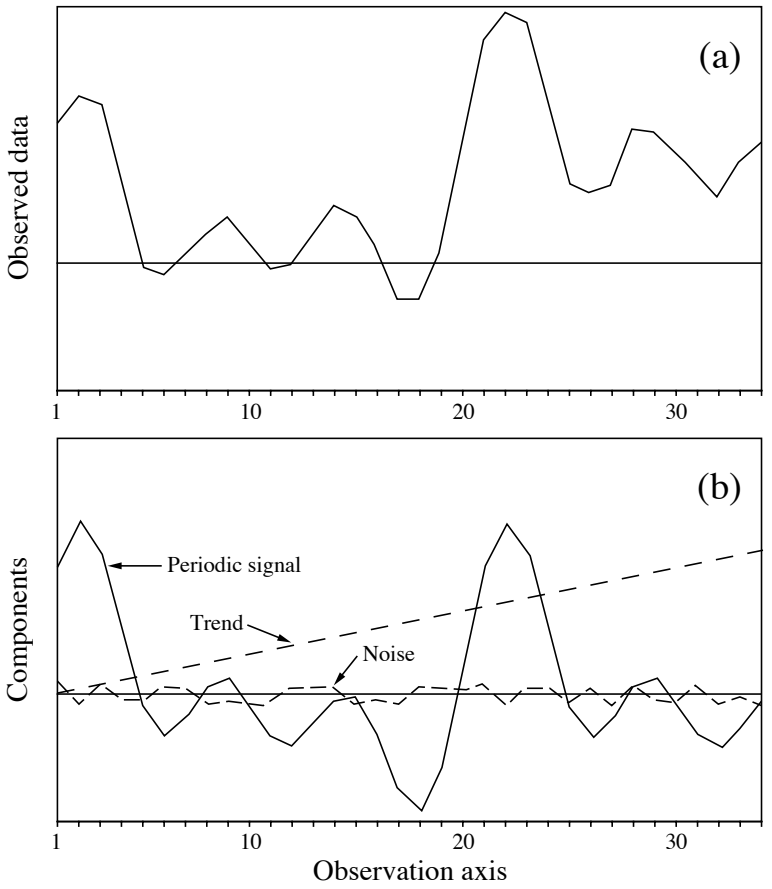


Figure 12.2 Artificial data series (a) constructed by adding the three components shown in (b), i.e. a periodic signal and a noise component, whose combination gives a stationary series (not illustrated), and a linear trend. The periodic signal is the same as in Fig. 12.13. There are $n = 34$ data points sampled at regular intervals. The overall mean of the noise signal is zero, by definition.

In some cases, determining the *trend* is the chief objective of the study. For example, progressive changes in the characteristics of an ecosystem, over several years, may be used to assess whether this system is responding or not to anthropogenic effects. In such a case, the problem would be to characterize the long-term trend, so that the annual cycle as well as the high-frequency noise component would be of no interest. Long-term trends in data series may be modelled by regression (Section 10.3). Linear regression is used when the trend is (or seems to be) linear. In other cases, the ecological hypothesis or a preliminary examination of the series may indicate that the trend is of some other mathematical form (e.g. logistic), in which case the methods of polynomial or nonlinear regression should be used (e.g. Ross, 1990).

Stationarity

Detrending
Trend extraction

In contrast, ecologists primarily interested in the *periodic* component of data series (Objective 2) consider the long-term trend as a nuisance. Even when the trend is not of ecological interest, it must be extracted from the series because most methods of analysis require that the series be *stationary*, i.e. that the mean, variance, and other statistical properties of the distribution be constant over the series. In the numerical example of Fig.12.2, the observed data series (a) is obviously not stationary. It becomes so if the linear trend shown in (b) is removed by subtraction; this operation is called *detrending* (or *trend extraction*). The trend may be estimated, in this case, by linear regression over a reasonably long segment of data; detrending consists in calculating the regression residuals. In practice, the analysis of series only requires *weak*, or *second-order*, or *covariance stationarity*, i.e. the mean and variance are constant along the series and the autocovariance (or autocorrelation) function depends only on the distance between observations along the series; two observations separated by a given interval have the same autocovariance no matter where they occur in the series. Extracting trends may be done in various ways, which are detailed in Section 12.2.

Some *low-frequency periodic* components may also be considered as trends, especially when these are both trivial and known *a priori* (e.g. an annual cycle). A long-term trend as well as broad-scale periodic components may be extracted in order to focus the analysis on finer components of the data series. Again, regression or other statistical methods (Section 12.2) may be used to model the low-frequency components and compute residuals on which the analysis could be carried out.

Objective 2. — Identifying *characteristic periods* is a major objective of series analysis in ecology. It is generally done for one variable at a time, but it is also possible to study multidimensional series (i.e. several variables, often analysed two at a time). Ecological series always exhibit irregular and unpredictable fluctuations, called *noise* (Fig. 12.2b), which are due to non-permanent perturbation factors. The larger the noise, the more difficult it is to identify characteristic periods when analysing stationary series. Table 12.3 summarizes the methods available to do so; several of these are described in Sections 12.3 to 12.5.

Objective 3. — One method for identifying characteristic periods is spectral analysis. In this analysis, the variance of the data series is partitioned among frequencies (or wavenumbers) in order to estimate a *variance spectrum*. Section 12.5 shows that the spectrum is a global characteristic of the series, and presents examples where the spectra are interpreted as reflecting ecological processes.

Objective 4. — There are data series that do not behave in a periodic manner. This may be because only one or even part of a cycle has been sampled or, alternatively, because the variables under study are not under the control of periodic processes. Such series may exhibit structures other than periodic, along time or a spatial direction. In particular, one may wish to identify *discontinuities* along multidimensional data series. Such discontinuities may, for example, characterize *ecological succession*. A commonly-used method for finding discontinuities is cluster analysis. To make sure that the multidimensional series gets divided into blocks, each one containing a set of

Table 12.2 Analysis of data series: research objectives and related numerical methods. Adapted from Legendre & Legendre (1984b) and Legendre & Dutilleul (1992).

Research objective	Numerical methods
1) Characterize the trend	• Regression (linear or polynomial)* • Moving averages • Variate difference method
2) Identify characteristic periods	→ Details in Table 12.3
3) Characterize series by spectrum	• Spectral analysis
4) Detect discontinuities in multivariate series	• Clustering the data series (with or without constraint) • Hawkins & Merriam or Webster segmentation methods
5) Correlate variations in a series with changes in other series	
5.1) Univariate target series	• Regression*: simple / multiple linear, nonlinear, splines • Cross-correlation
5.2) Multivariate target series	• Canonical analysis** • Mantel test*
6) Formulate a forecasting model	• Box-Jenkins modelling

Methods described in * Chapter 10 or ** Chapter 11.

temporally contiguous observations, authors have advocated to constrain clustering algorithms so that they are forced to only group observations that are contiguous. Various methods to do so are discussed in Section 12.6.

Objective 5. — Another objective is to *correlate variations* in the data series of interest (i.e. the *target* or *response variable*) with variations in series of some potentially *explanatory variable(s)*, with a more or less clearly specified model in mind. There are several variants. (1) When the sampling interval between observations is large, the effect of the explanatory variables on the target variable may be considered as instantaneous. In such a case, various forms of regression analysis may be used. When no explicit model is known by hypothesis, spline regression may be used to describe temporal changes in the target variable as a function of another variable (e.g. Press *et al*., 2007). These methods are described in Section 10.3. (2) When the interval between consecutive data is short compared to the periods in the target variable, it is sometimes assumed that the target variable responded to events that occurred at some previous time, although the exact delay (*lag*) may not be known. In such a case, the method of cross-correlation may be used to identify the time lag that maximises the correlation between the explanatory and target variables (Section 12.3). When the optimal lag has been found for each of the explanatory variables in a model, multiple regression can then be used, each explanatory variable being lagged by the

Table 12.3 Analysis of data series: methods for identifying characteristic periods. The approaches best suited to *short* data series are: the contingency periodogram, Dutilleul's modified periodogram, and maximum entropy spectral analysis. Adapted from Legendre & Legendre (1984b) and Legendre & Dutilleul (1992).

Type of series	Methods	
	Quantitative variables only	All precision levels
1) A single variable	• Autocorrelogram	• Spatial correlogram* (quantitative, qualitative)
	• Periodograms (Whittaker & Robinson, Schuster, Dutilleul)	• Contingency periodogram for qualitative data
	• Spectral analysis	• Kedem's spectral analysis for binary data
2) Two variables	• Parametric cross-correlation	• Nonparametric cross-correlation
	• Coherence and phase spectra	• Cross-contingency analysis
3) Multivariate series	• Multivariate spectral analysis	• Multivariate variogram*, Mantel correlogram*

* Methods described in Chapter 13.

appropriate number of sampling intervals. (3) The previous cases apply to situations where there is a single target variable in the series under study. When there are several target variables, the target series is multivariate; the appropriate methods of data analysis are globally called canonical analysis (Chapter 11). Two forms are of special interest here: redundancy analysis and canonical correspondence analysis. (4) Finally, the relationship between two distance matrices based on two multivariate data sets can be analysed using the Mantel test or its derived forms (Section 10.5 and Subsection 13.1.6) when the question strictly concerns distances.

Objective 6. — A last objective is to formulate a model to *forecast* the future behaviour of the target series. Following the tradition in economics, one way of doing that is to model the data series according to its own past behaviour (Section 12.7).

Testing for the presence of trends

The first problem encountered when analysing data series is to decide whether a *trend* is present or not. Visual examination of the series, which may be combined with previous knowledge about the process at work, is often sufficient to detect one or several trends. These may be monotonic (e.g. gradient in latitude, altitude, or water depth) or not (e.g. daily, lunar, or annual cycles). Four methods can be used to test for the presence of trends (extraction of trends: see Section 12.2).

• 1. The most widely used method is to regress the response data series **y** on the time variable. A significant regression coefficient indicates the presence of a linear trend, either positive or negative, in the series. Researchers must beware of a situation where

a trend is sought in a series that contains high variability nested into the series. For example, when looking for a trend among years in a series **y** covering several years, linear regression of **y** on the variable *years* may fail to detect a significant trend if the variation among months is high. To circumvent that problem, one can use a qualitative variable (or *factor*) coding for the months as covariable in the analysis. In practice, one can simply compute a linear model of **y** as a function of the quantitative variable *years* and the factor *months*, and check if the regression coefficient associated with *years* is significant.

• 2. The numbers of positive and negative *differences between successive values* in the series are counted. These are then subjected to a *sign test* (Table 5.2), where the null hypothesis (H_0) is that the plus and minus signs correspond to a population in which the two signs are present in equal proportions. Rejecting H_0 is indication of a trend.

• 3. All values in the series are *ranked* in increasing (or decreasing) order. *Kendall's rank correlation coefficient* (τ) (Subsection 5.3.2) may be used to assess the degree of resemblance between the rank-ordered series and the original one; this is done by computing the Kendall correlation between the original data series and the observation rank labels: $1, 2, 3, \ldots, n$. When τ is significantly different from zero, one can conclude that the series exhibits a *monotonic* trend. These two methods are described in Kendall & Ord (1990, pp. 21-22). The approach based on Kendall's τ is preferable to the sign test because it uses the actual data in the series instead of the differences between neighbouring values.

Up and down runs test

• 4. A nonparametric test, called the *up and down runs* test, is well suited to detect the presence of various types of trends. Consider again n values and, for each one, the sign of the difference from the previous value. The $(n - 1)$ signs would all be the same if the observations were monotonically increasing or decreasing. Cyclical data, on the other hand, would produce more long *runs* of "+" or "–" signs than expected for random data, or more short *runs*, depending on the sampling frequency within each cycle. A *run* is a set of like signs, preceded and followed (except at the end of the series) by opposite signs. Count the *number of runs* in the data series, including those of length 1 (e.g. a single "+" sign, preceded and followed by a "–"). The up and down runs test, described for instance in Sokal & Rohlf (1995), compares this number to the number of runs expected from a same-length sequence of random numbers.

When there is a *trend* in the series, it must be extracted using one of the methods discussed in Section 12.2. If, after detrending, the mean of the series is still not stationary, a second trend must be searched for and removed. When the series does not exhibit any trend, or after detrending, one must decide, before looking for periodic variability (Sections 12.3 to 12.5), whether the stationary series presents some kind of systematic variability or if, on the contrary, it simply displays the kind of variation expected from a random process. In other words, one must test whether the series is simply *random*, or if it exhibits *periodic variability* that could be analysed.

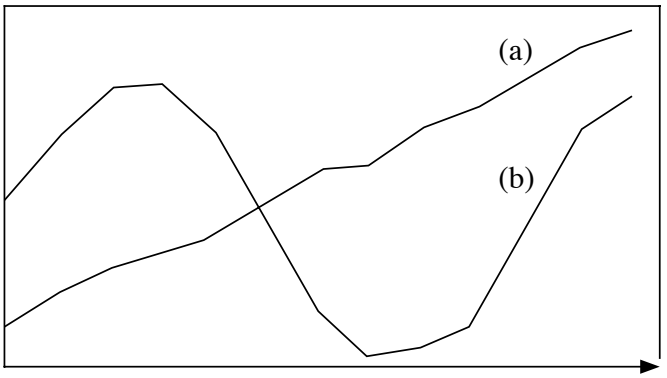


Figure 12.3 Two artificial series; (a) would be random if the linear trend was extracted, whereas (b) displays a cyclic trend.

Test series randomness

In some instances, as in Fig. 12.3, it is useless to conduct sophisticated tests, because the random or systematic character of the series is obvious. Randomness of a series may be tested as follows: identify the *turning points* (i.e. the peaks and troughs) in the series and record the distribution of the *number of intervals* (*phase length*) between successive turning points. It is possible to test whether these values correspond or not to those of a random series (Kendall & Ord, 1990, p. 20). This procedure actually tests the same null hypothesis as the up and down runs test described above. In practice, any ecological series with an average phase longer than two intervals may be considered non-random.

The overall procedure for analysing data series is summarized in Fig. 12.4. The following sections describe the most usual methods for extracting trends, as well as various approaches for analysing stationary series. It must be realized that, in some instances, variations in stationary series may be so small that they cannot be analysed, because they are of the same order of magnitude as the background noise.

If parametric statistical tests are to be conducted during the course of the analysis, *normality* must be checked (Section 4.6) and, if the data are not normally distributed, they must be *transformed* as explained in Subsection 1.5.6. In addition, several of the methods discussed in the following sections require that observations in the series be *equally spaced*. If they are not, data may be eliminated to make them *equispaced*, or else, missing data may be estimated by regression or other interpolation methods (Section 1.6); most methods of series analysis cannot handle missing values. Obviously, it is preferable to consider the requirement of equispaced data when designing a sampling program than to have to modify the data at the stage of analysis.

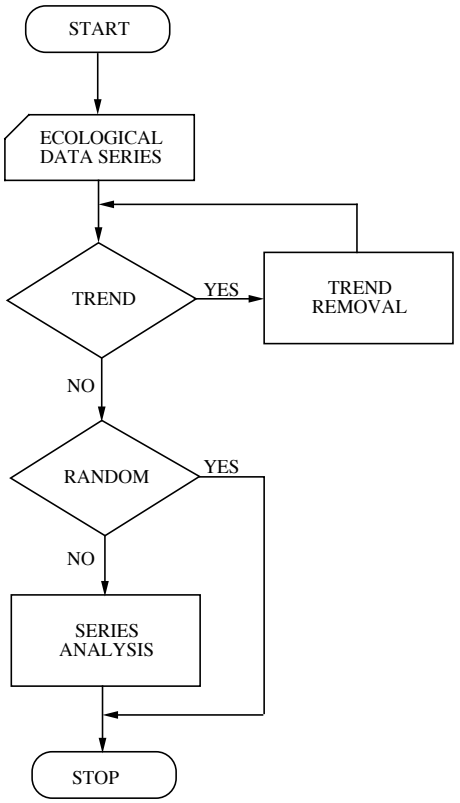


Figure 12.4 Flow diagram summarizing the steps involved in the analysis of data series.

In addition to the numerical methods discussed in the following sections, ecologists may find it useful to have a preliminary look at the data series, using the techniques of exploratory data analysis described by Tukey (1977, his Chapters 7 and 8). These are based on simple arithmetic and easy-to-draw graphs, and they may help decide which numerical treatments would be best suited for analysing the series. Exploratory data analysis for time series is also described in Chapter 14 of Venables & Ripley (2002) and in Chapter 2 of Shumway & Stoffer (2011).

12.2 Trend extraction and numerical filters

When there is a trend in a series (which is not always the case), it must be extracted from the data prior to further numerical analyses. As explained in the previous section, this is because most methods of analysis require that the series be *stationary*.

When the trend itself is of interest, it can be analysed in ecological terms (Objective 1 above). For example, Fortier *et al*. (1978) interpreted a cyclical trend in temporal changes of estuarine phytoplankton in terms of physical oceanographic forcing. In a long-term monitoring study of bacteria of sanitary importance at a lake beach, St-Louis & Legendre (1982) interpreted the significant negative slope of a water quality index computed from bacterial data (363 water samples analysed over 9 years) as an indication of deterioration of the water quality. Borcard *et al*. (2004) identified a linear trend in marine zooplankton size-class data across a coastal reef lagoon in Guadeloupe (spatial series). The trend was related to increasing salinity from the coast to the outer reef, and to decreasing phytoplankton biomass, wind speed and dissolved oxygen. In the analysis of fossil diatom assemblages along a sediment core from south-western Scotland covering the past 10000 years (101 core levels, 139 species), Legendre & Birks (2012) identified a significant temporal trend and related it to changes in the relative abundances of eight diatom species that were highly correlated with their positions along the core.

When the study goes beyond the identification of a trend (Objectives 2 *et seq*.), the analysis is normally conducted on the *residual* (or *detrended*) data series. The residual (i.e. stationary) series is obtained, for each data point i along the series, by subtracting the value estimated by the trend function at position x_i from the observed value y_i:

Residuals

$$y_{\text{res}.i} = \text{residual of } y_i = \text{observed value } (y_i) - \text{value of the trend at } x_i \qquad \textbf{(12.1)}$$

Detrending

There are cases where several trends of different natures must be extracted successively before reaching stationarity. However, because each trend extraction distorts the residuals, one must proceed with caution with detrending. The success of trend extraction may be assessed by plotting and examining the resulting trend (Objective 1) or the detrended series.

Moving averages

Window

The method of *moving averages* is often used to estimate trends, e.g. in climate-change related studies. One calculates successive arithmetic averages over $2m + 1$ contiguous data as one moves along the data series. The interval $(2m + 1)$ over which a moving average is computed is called *window*. For example, with $m = 2$, the first moving average $\bar{y}_3$ is computed over the first 5 values y_1 to y_5, the second moving average $\bar{y}_4$ is calculated over values y_2 to y_6, the third one ($\bar{y}_5$) is the average of values y_3 to y_7, and so forth. Each average value is positioned at the centre of its window. For a series of n observations, there are $(n - 2m)$ moving averages:

x_1	x_2	x_3	x_4	…	x_{n-2}	x_{n-1}	x_n
y_1	y_2	y_3	y_4	…	y_{n-2}	y_{n-1}	y_n

moving averages

$$\bar{y}_3 = \frac{1}{5}\sum_{h=1}^{5} y_h \qquad \bar{y}_4 = \frac{1}{5}\sum_{h=2}^{6} y_h \qquad \cdots \qquad \bar{y}_{n-2} = \frac{1}{5}\sum_{h=n-4}^{n} y_h$$

The general formula for moving averages is thus:

$$\bar{y}_i = \frac{1}{2m+1} \sum_{h=-m}^{m} y_{(i+h)} \qquad \textbf{(12.2)}$$

The h values corresponding to the above example, where $m = 2$, would be: –2, –1, 0, +1, and +2, respectively.

Moving averages may also be *weighted*. In such a case, each of the $2m + 1$ values within the window is multiplied by a weight w_h. Usually, values closer to the centre of the window receive larger weights. The general formula for the weighted moving average corresponding to any position (or object) x_i is:

$$\bar{y}_i = \sum_{h=-m}^{m} y_{(i+h)} w_h \Big/ \sum_{h=-m}^{m} w_h \qquad \textbf{(12.3)}$$

Choosing values for the weights depends on the underlying hypothesis. Kendall & Ord (1990, p. 3) give coefficients to be used under hypotheses of polynomial trend of the second, third, fourth, and fifth degrees. Another, simple method for assigning weights is that of *repeated moving averages*. After calculating a first series of non-weighted moving averages (eq. 12.2), a second series of moving averages is calculated using values from the first series. Thus calculation of three successive series of non-weighted moving averages produces the following results ($\bar{y}_i$) and weights w_h (Table 12.4):

first series ($m = 1$) $\quad \bar{y}_i = \frac{1}{3} \sum_{h=-1}^{1} y_{(i+h)} w_h \qquad w_0 = 1, w_{\pm 1} = 1$

second series ($m = 2$) $\quad \bar{y}_i = \frac{1}{9} \sum_{h=-2}^{2} y_{(i+h)} w_h \qquad w_0 = 3, w_{\pm 1} = 2, w_{\pm 2} = 1$

third series ($m = 3$) $\quad \bar{y}_i = \frac{1}{27} \sum_{h=-3}^{3} y_{(i+h)} w_h \qquad w_0 = 7, w_{\pm 1} = 6, w_{\pm 2} = 3, w_{\pm 3} = 1$

It is easy to check the above values by simple calculations, as shown in Table 12.4.

When using moving averages for estimating the trend of a series, one must choose the *width of the window* (i.e. choose m) as well as the *shape* of the moving average (i.e. the degree of the polynomial or the number of iterations). These choices are not simple. They depend in part on the goal of the study, namely the ecological interpretation of the *trend* itself or the subsequent analysis of *residuals* (i.e. detrended series). To estimate a cyclic trend, for instance, it is recommended to set the window width ($2m + 1$) equal to the period of the cyclic fluctuation.

Cyclic trend

Table 12.4 Calculation of repeated moving averages. Development of the numerator for the first and second series of averages.

x_1	x_2	x_3	x_4	…	x_i	…
y_1	y_2	y_3	y_4	…	y_i	…
	$\bar{y}'_2 = y_1 + y_2 + y_3$	$\bar{y}'_3 = y_2 + y_3 + y_4$	$\bar{y}'_4 = y_3 + y_4 + y_5$	…	$\sum_{h=-1}^{1} y_{(i+h)} w_h$ $w_0 = 1, w_{\pm 1} = 1$	…
		$\bar{y}''_3 = \bar{y}'_2 + \bar{y}'_3 + \bar{y}'_4$	$\bar{y}''_4 = \bar{y}'_3 + \bar{y}'_4 + \bar{y}'_5$	…	$\sum_{h=-2}^{2} y_{(i+h)} w_h$	…
		$\bar{y}''_3 = y_1 + 2y_2 + 3y_3 + 2y_4 + y_5$	…	…	$w_0 = 3, w_{\mp 1} = 2, w_{\mp 2} = 1$	

Slutzky-Yule effect

Trend extraction by moving averages may add to the detrended series an artificial periodic component, which must be identified before analysing the series. This phenomenon is called the *Slutzky-Yule effect*, because these two statisticians independently drew attention to it in 1927. According to Kendall (1976, pp. 40-45) and Kendall *et al*. (1983, pp. 465-466), the average period of this artificial component (T) is calculated using the $(2m + 1)$ weights w_i of the moving average formula (eq. 12.3)*:

$$T = 2\pi/\theta \quad \text{for angle } \theta \text{ in radians, or} \quad T = 360°/\theta \quad \text{for angle } \theta \text{ in degrees,}$$

$$\text{where} \qquad \cos\theta = \left| \sum_{h=1}^{2m+1} (w_{h+1} - w_h)(w_h - w_{h-1}) \right| \Big/ \sum_{h=1}^{2m+2} (w_h - w_{h-1})^2 \qquad \textbf{(12.4)}$$

The values of the weights located outside the window are zero: $w_0 = 0$ and $w_{2m+2} = 0$. For example, using the weights of the second series of repeated moving averages above ($m = 2$):

$$[w_h] = [1 \;\; 2 \;\; 3 \;\; 2 \;\; 1]$$

* In Kendall (1976) and Kendall *et al*. (1983) and previous editions of *The Advanced Theory of Statistics, Vol. 3*, there is a printing error in the formula for the Slutzky-Yule effect. In the first parenthesis of the last term of their numerator, the printed sign for the second weight (w_{2m+1}) is positive; this sign should be negative, as in eq. 12.4, giving (0 – 1) in our numerical example. However, their numerical example is correct, i.e. it is computed with $-w_{2m+1}$, not $+w_{2m+1}$.

gives

$$\cos\theta = \frac{|(2-1)(1-0) + (3-2)(2-1) + (2-3)(3-2) + (1-2)(2-3) + (0-1)(1-2)|}{(1-0)^2 + (2-1)^2 + (3-2)^2 + (2-3)^2 + (1-2)^2 + (0-1)^2} = \frac{3}{6}$$

from which it follows that $\theta = 1.047 \text{ rad} = 60°$

and thus: $T = 2\pi/1.047 = 360°/60° = 6$

If, after detrending by this method of *repeated moving averages*, the analysis of the series resulted in a period $T \approx 6$, this period would probably be a by-product of the moving average procedure. It would not correspond to a component of the original data series, so that one should not attempt to interpret it in ecological terms. If a period $T \approx 6$ was hypothesized to be of ecological interest, one should use different weights for trend extraction by moving average analysis.

Analytical method

The most usual approach for estimating trends is the *analytical method*. It consists in fitting a regression model to the whole series, using the least squares approach or some other method. The matter was fully reviewed in Section 10.3. Smoothing methods such as splines and LOWESS can also be used (Subsection 10.3.8). The model for the trend may be linear, polynomial, exponential, logistic, etc. The main advantages of trend extraction based on regression are: the explicit choice of a model by the investigator, and the ease of calculation using a statistical package. The main problem is that a new regression must be calculated upon addition of one or several observations to the data series, which may generate different values for the regression coefficients. However, as the series gets longer, estimates of the regression coefficients become progressively more stable.

Variate difference

Contrary to the above methods, where the estimated trend was subtracted from the observed data (eq. 12.1), the *variate difference method* directly detrends the series. It consists in replacing each value y_i by the difference $(y_{i+1} - y_i)$. As in the case of repeated moving averages, differences may be calculated not only on the original data, but also on data resulting from previous detrending. If this is repeated on progressively more and more detrended series, the variance of the series usually stabilizes rapidly. The variate difference method, when applied once or a few times to a series, can successfully remove any polynomial trend. Only exponential or cyclic trends may

Cyclic trend

sometimes resist the treatment. The method may be used to remove any cyclic trend whose period T is known, by using differences $(y_{i+T} - y_i)$; however, this is fully successful only in cases where T is an integer multiple of the sampling interval Δ. One must remember that this method does not model the trend that is removed from the data series as a data vector; hence, the trend cannot be studied independently.

Filtration

In some instances, ecologists may also wish to eliminate the random *noise* component from the data series, in order to better evidence the ecological phenomenon under study. This operation, whose aim is to remove high-frequency variability from the series, is called *filtration*. In a sense, filtration is the complement of trend

Filter

extraction, since the former removes high-frequency components of the series and the latter, low-frequency components. Specialists of series analysis often use the term *filter* for any preliminary treatment of the series, whether the extraction of low frequencies (trend) or the removal of high frequencies (noise). Within the context of spectral analysis (Section 12.5), filtration of the series is often called "prewhitening". This refers to the fact that filtration flattens the spectrum of a series and makes it similar to the spectrum of white light. The reciprocal operation (called "recolouring") fits the spectrum (calculated on the filtered series) in such a way as to make it representative of the nonfiltered series. The sequence of operations — prewhitening of the series, followed by computation of the spectrum on the filtered series, and finally recolouring of the resulting spectrum — finds its justification in the fact that spectra that are more flat are also more precisely estimated.

In addition to filters, which aim at extracting low frequencies (trends), computer programs for series analysis offer a variety of numerical filters that allow the removal, or at least the reduction, of any component located outside a given frequency band (passband). It is thus possible, depending on the objective of the study, to select the high or low frequencies, or else a band of intermediate frequencies. It is also possible to eliminate a band of intermediate frequencies, which is the converse of the latter filter. Generally, these numerical filters are found in programs for spectral analysis (Section 12.5), but they may also be used to filter series prior to analyses using the methods described in Sections 12.3 and 12.4. In most cases, filtering data series (including trend extraction) requires solid knowledge of the techniques, because filtration always distorts the original series and thus influences further calculations. It is therefore better to do it under the supervision of an experienced colleague.

12.3 Periodic variability: correlogram

The systematic component of a stationary series is called *periodic variability*. There are several methods available for analysing this type of variability. Those discussed in the present section, namely the autocovariance and autocorrelation (serial correlation) and the cross-covariance and cross-correlation, are all extensions, to the analysis of data series, of statistical methods described in earlier chapters. These methods of analysis have been extensively used in ecology.

At this stage of series analysis, it is assumed that the data series is *stationary*, either because it originally exhibited no trend or as the result of detrending (Section 12.2). It is also assumed that variability is large enough to emerge from random noise.

A general approach for analysing periodic variability is derived from the concepts of covariance and correlation defined in Chapter 4. The methods are called *autocovariance* and *autocorrelation analysis*. The approach is to quantify the relationships between successive terms of the data series. These relationships reflect the pattern of periodic variability.

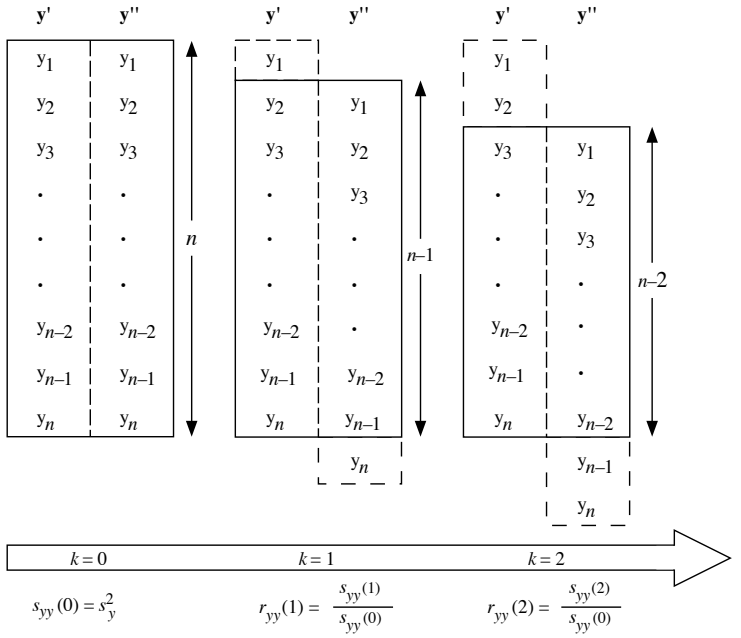


Figure 12.5 Calculation of autocovariance (s_{yy}) and autocorrelation (r_{yy}). Stepwise shift of a data series relative to itself, with successive lags of k units. The number of terms involved in the calculation ($n - k$) decreases as k increases.

1 — Autocovariance and autocorrelation

Autocovariance measures the covariance of the series with itself, computed as the series is progressively shifted with respect to itself (Fig. 12.5). Because second-order stationarity is assumed in the calculation of autocovariance and autocorrelation (Section 12.1), all coefficients will be computed using the same mean and variance, estimated from the whole series, even though individual coefficients involve only part of the data. The overall mean is $\bar{y}$. For the common variance, the sum of squared deviations from $\bar{y}$ is divided by n instead of $(n - 1)$, as in Moran's I coefficient of spatial correlation (eq. 13.1); this is the maximum-likelihood estimator of the variance:

$$s_{yy}(0) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y}) \qquad \textbf{(12.5)}$$

This is the covariance of the series with itself when there is no shift. The notation $s_{yy}(0)$ indicates a lag of zero, or lag $k = 0$.

When the series is shifted relative to itself by one unit (lag $k = 1$), the left-hand copy of the series in Fig. 12.5 loses observation y_1 and the right-hand copy loses observation y_n. The two truncated series, each of length $(n - 1)$, are compared. For a lag of k units, the covariance $s_{yy}(k)$ is computed from the $(n - k)$ terms remaining in the two truncated series, using the mean and n value from the whole series to insure that the covariances remain comparable:

Auto-covariance

$$s_{yy}(k) = \frac{1}{n} \sum_{i=1}^{n-k} (y_{i+k} - \bar{y})\,(y_i - \bar{y}) \qquad \textbf{(12.6)}$$

That equation is similar to that of the covariance (eq. 4.4). In correlograms (below), the autocovariance is estimated for several successive lags k. In specific applications, researchers may decide on biological grounds how long the lag should be to compute the autocovariance of the variable under study.

In eq. 4.7, the Pearson coefficient of linear correlation between variables y_j and y_k is computed by dividing their covariance by the product of their standard deviations:

$$r_{jk} = \frac{s_{jk}}{s_j s_k}$$

In a similar way, the *autocorrelation* of a series $r_{yy}(k)$ is computed as the ratio of its autocovariance $s_{yy}(k)$ (eq. 12.6) to its variance $s_{yy}(0)$ (eq. 12.5):

Auto-correlation

$$r_{yy}(k) = \frac{s_{yy}(k)}{s_{yy}(0)} \qquad \textbf{(12.7)}$$

Equation 12.6 is a good estimator of autocorrelation when $(n - k)$ is reasonably large. The autocorrelation is also called *serial correlation*. It measures the average dependence of the values in the series on values found at a distance of k lags.

One may be tempted to compute $r_{yy}(k)$ using the Pearson linear correlation formula (eq. 4.7) between terms y_i and y_{i+k} of the series, for the $n - k$ pairs of corresponding values in the observed and shifted series (Fig. 12.5). This is not recommended, however, because the mean and variance estimates used for computing eq. 4.7 change with lag k, so that $r_{yy}(k)$ would not produce a set of comparable autocorrelation coefficients (Jenkins & Watts, 1968; Venables & Ripley, 2002).

Since the number of terms $(n - k)$ involved in the calculation of the autocovariance or autocorrelation decreases as k increases, it follows that, as k increases, the precision of the estimate, the number of degrees of freedom available, and consequently the power of the test of significance decrease. The largest interpretable lag is often considered to be about $k_{max} = n/3$; Venables & Ripley (2002) use $10 \log_{10}(n)$ as the default value for the largest lag in function ***acf()*** in R. Table 12.5 gives the values of autocovariance and autocorrelation for the artificial stationary series of Fig. 12.2b.

Table 12.5 Autocovariance and autocorrelation coefficients (eq. 12.7) for the artificial series of Fig. 12.2b, after detrending (i.e. periodic signal + noise components only). For each successive lag, the series is shifted by one sampling interval. Values corresponding to odd lags are not shown. The autocovariance and autocorrelation coefficients are plotted against lag in Fig. 12.6.

Lag	Autocovariance $s_{yy}(k)$	Autocorrelation $r_{yy}(k)$
0	3.07	1.00
2	1.01	0.33
4	–0.90	–0.29
6	–0.28	–0.09
8	–0.29	–0.10
l0	–0.87	–0.28
12	–0.36	–0.12
14	–0.27	–0.09
16	–0.64	–0.21
18	0.33	0.11
20	1.22	0.40
22	0.47	0.15

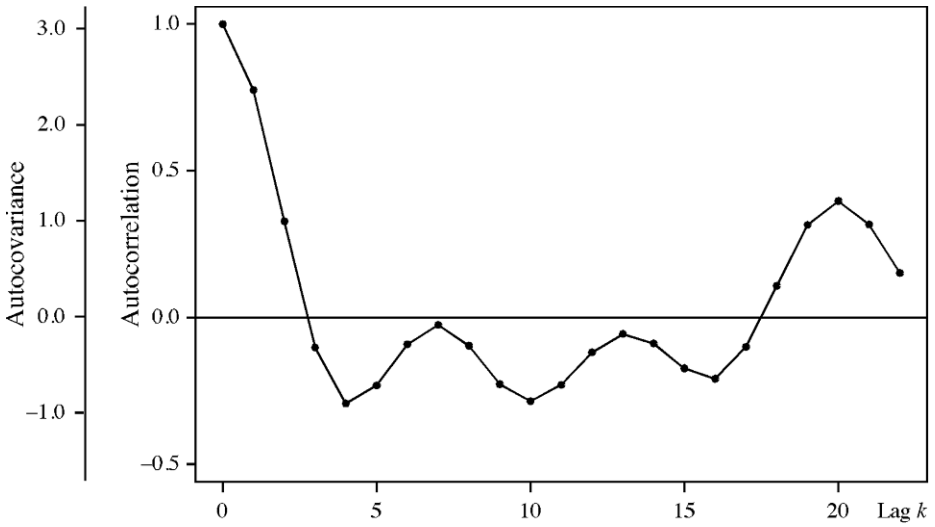


Figure 12.6 Correlogram (autocovariance and autocorrelation; values from Table 12.5) for the artificial series of Fig. 12.2b, after detrending (i.e. periodic signal + noise components only).

Autocorrelogram

The autocorrelation (or autocovariance) coefficients are plotted as a function of lag k (abscissa), in a graph called *autocorrelogram* (or *correlogram*, for simplicity). Autocorrelation coefficients range between +1 and –1. The scale factor between the autocorrelation and autocovariance coefficients is the variance of the series (eq. 12.5). In Fig. 12.6, this factor is $s_{yy}(0) = 3.07$; it is shown in Table 12.5 at lag $k = 0$.

The interpretation of correlograms is based on the following reasoning. At lag $k = 0$, the two copies of the series (**y**' and **y**") have the exact same values facing each other (Fig. 12.5), hence $r_{yy}(0) = +1$. With increasing lag k, corresponding values in the series **y**' and **y**" move farther apart and $r_{yy}(k)$ decreases. This is what is happening, in the numerical example, for lags up to $k = 4$ (Table 12.5 and Fig. 12.6). In series where periodic variability is present (with period T_p), increasing k eventually brings similar values to face each other again (at lag $k = T_p$), with peaks facing peaks and troughs facing troughs, hence a high positive value of $r_{yy}(k)$. The value of $r_{yy}(k = T_p)$ is always smaller than 1, however, because there is always noise in data and because natural periodic phenomena seldom repeat themselves perfectly. Negative autocorrelation often reaches its maximum at $k = T_p/2$ because the signals in **y**' and **y**" are then maximally out of phase.

A practical problem occurs when there are several periodic signals in a series; this may increase the complexity of the correlogram. Nevertheless, high positive values in a correlogram may generally be interpreted as indicative of the presence of periodic variability in the series. For the numerical example, Fig. 12.6 indicates that there is a major periodicity at $k = 20$, corresponding to period $T = 20$; this interpretation is supported by the low value of $r_{yy}(10)$. Period $T = 20$ is indeed the distance between corresponding peaks or troughs in the series of Fig. 12.2b. Other features of the correlogram may be indicative of additional periods (which is the case here, as can be seen by examining Fig. 12.2b) or may simply be the result of random noise.

Confidence intervals can be computed and drawn on a correlogram to identify the values that are significantly different from zero. The confidence interval is usually represented on the correlogram as a two-standard-error band. If the data can be assumed to be normal, independent (in the sense of *not autocorrelated*, Box 1.1) and identically distributed, the confidence interval of r_{yy} can be computed through the usual formula for confidence intervals of correlation coefficients. In most time series analyses, however, there is an assumption that the data are autocorrelated. It is thus more appropriate to compute confidence intervals under a moving average (MA) model (eq. 12.31) (Venables & Ripley, 2002). Both methods of calculation are available in the R function ***plot.acf()*** (Section 12.8).

Harmonic

When the series is long, its correlogram may exhibit significant values for *harmonics* (integer multiples) of the period present in the signal (T_{series}). This is a normal phenomenon, which is generally not indicative of additional periodicity in the data series. However, when a value of the correlogram statistic is noticeably larger for a harmonic period than for the basic period, one can conclude that the harmonic is also a true period of the series.

For short series, autocorrelograms should only be computed when the series include very strong periodic components. This is because the test of significance is not very powerful, i.e. the probability of rejecting the null hypothesis of no autocorrelation is small when a periodic component is present in short series. When there is *more than one periodic component* in a series, correlograms should generally not be used, even with long series, because components of different periods may interfere with one another and prevent the correlogram from showing significance (see also the next paragraph). Periodograms (Section 12.4) should be used instead. Finally, when the data are *not equispaced* and one does not wish to interpolate, methods developed for *spatial correlation* analysis, which do not require equal spacing of the data, may be used (Section 13.1). Special forms of spatial correlation coefficients allow the analysis of series of *qualitative* data (last paragraph of Subsection 13.1.1).

It may happen that periods present in the series do not appear in a correlogram, because they are concealed by other periods accounting for larger fractions of the variance of the series. When one or several periods have been identified using a first correlogram, one may remove these periods from the series using one of the methods recommended in Section 12.2 for cyclic trends and compute a new correlogram for the detrended series. It could bring out previously concealed periods. This is not without risk, however, because successively extracting trends rapidly distorts the residuals. Approaches better adapted to series containing multiple periods are discussed in Sections 12.4 and 12.5.

The following numerical example and ecological applications illustrate the computation and use of correlograms.

Numerical example. Consider the following series of 16 data points (quantitative variable):

$$2, 2, 4, 7, 10, 5, 2, 5, 8, 4, 1, 2, 5, 9, 6, 3$$

Table 12.6 illustrates the computation of the autocorrelation coefficients. These could be plotted as a function of lag (k) to form a correlogram, as in Figs. 12.6 and 12.7b. The coefficients clearly point to a dominant period at $k = 5$, for which autocorrelation is positive and maximum. This approximately corresponds to the average distance separating successive maximum values, as well as successive minima, along the data series.

Ecological application 12.3a

In order to study the spatial variability of coastal marine phytoplankton, Platt *et al*. (1970) measured chlorophyll *a* along a transect 8 nautical miles long, at 10 m depth and intervals of 0.1 naut. mi. (1 naut. mi. = 1852 m). The resulting 80 values are shown in Fig. 12.7a.

The series exhibited a clear linear *trend*, which was extracted at the beginning of the analysis. Autocorrelation coefficients were computed from the residual series, up to lag $k = 10$, because the series was quite short (Fig. 12.7b). The position of the first *zero* in the *correlogram* was taken as indicative of the average apparent *radius* of phytoplankton patches along the transect. The model underlying this interpretation is that of circular patches, separated by average distances equal to their average diameter. In such a case, it is expected that the second

Table 12.6 Computation of the autocorrelation coefficients for the data of the numerical example. Boxes delimit the values included in each calculation. Note how the highest values are facing each other at lag 5, where the autocorrelation coefficient is maximum.

Lag	Data series																	Autocorrelation $r_{yy}(k)$
k=0	2	2	4	7	10	5	2	5	8	4	1	2	5	9	6	3		1.000
	2	2	4	7	10	5	2	5	8	4	1	2	5	9	6	3		
k=1	2	2	4	7	10	5	2	5	8	4	1	2	5	9	6	3		0.313
		2	2	4	7	10	5	2	5	8	4	1	2	5	9	6	3	
k=2	2	2	4	7	10	5	2	5	8	4	1	2	5	9	6	3		–0.544
			2	2	4	7	10	5	2	5	8	4	1	2	5	9	6	3
k=3	2	2	4	7	10	5	2	5	8	4	1	2	5	9	6	3	9	…	–0.472
				2	2	4	7	10	5	2	5	8	4	1	2	5		
k=4	2	2	4	7	10	5	2	5	8	4	1	2	5	9	6	3	5	…	0.105
					2	2	4	7	10	5	2	5	8	4	1	2		
k=5	2	2	4	7	10	5	2	5	8	4	1	2	5	9	6	3	2	…	0.323
						2	2	4	7	10	5	2	5	8	4	1		
k=6	2	2	4	7	10	5	2	5	8	4	1	2	5	9	6	3	1	…	–0.107
							2	2	4	7	10	5	2	5	8	4		
etc.								etc.											etc.

zero would occur at a lag three times that of the first zero, as was indeed observed on the correlogram. In the present case, the average *diameter* of phytoplankton patches and the distance separating them appeared to be ca. 0.5 naut. mi.

Ecological application 12.3b

Steven & Glombitza (1972) sampled tropical phytoplankton and chlorophyll at a site off Barbados during nearly three years. Sampling was approximately fortnightly. The physical environment there is considered to be very stable over the year. The most abundant phytoplankton species, in surface waters, is the filamentous cyanobacterium *Trichodesmium thiebaudii*. Data were concentrations of chlorophyll *a* and of *Trichodesmium* filaments.

The raw data were subjected to two transformations: (1) computation of *equispaced* data at 15-day intervals by interpolation, and (2) *filtration* intended to reduce the importance of non-dominant variations. The filtered data are shown in Fig. 12.8a, where the synchronous variations of the two variables are obvious. Correlograms for the nonfiltered (Fig. 12.8b) and filtered (Fig. 12.8c) series clearly show the same periodic signal, of ca. 8 lags × (15 days lag^{-1}) = 120 days. Nonfiltered data provide the same information as the filtered series, but not quite as clearly. According to the authors, these periodic variations could be an example of free

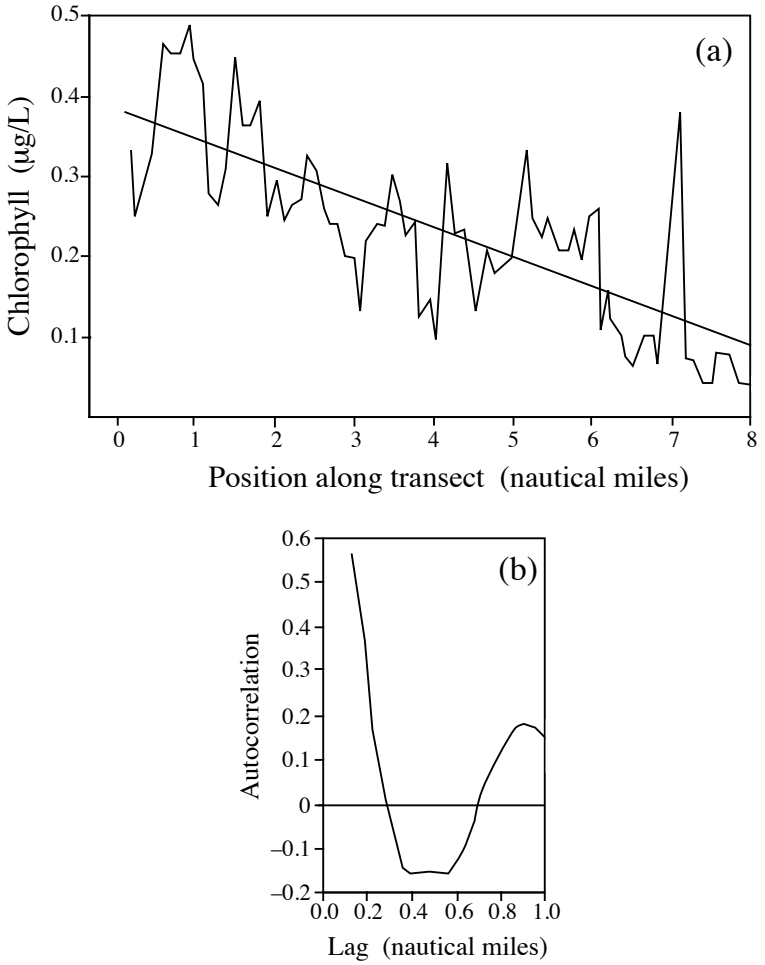


Figure 12.7 Chlorophyll *a* concentrations in a coastal marine environment, along a transect 8 naut. miles long in St. Margaret's Bay (Nova Scotia, Canada). (a) Data series exhibiting a linear trend, and (b) correlogram of the detrended series where lags (abscissa) are given as distances along the transect. After Platt *et al*. (1970).

oscillations, since they seemed independent of any control by the environment, which was stable the year round. The same ecological application will be used again below to illustrate the calculation of cross-correlation (next subsection) and Schuster's periodogram (Section 12.4).

Wilson & Dawe (2006) used autocorrelograms to compare variations in population densities of marine foraminifera with monsoonal rainfall data. Dutilleul (2011, his Sections 6.2.1 and 6.3.2) discussed applications of autocorrelation to several data

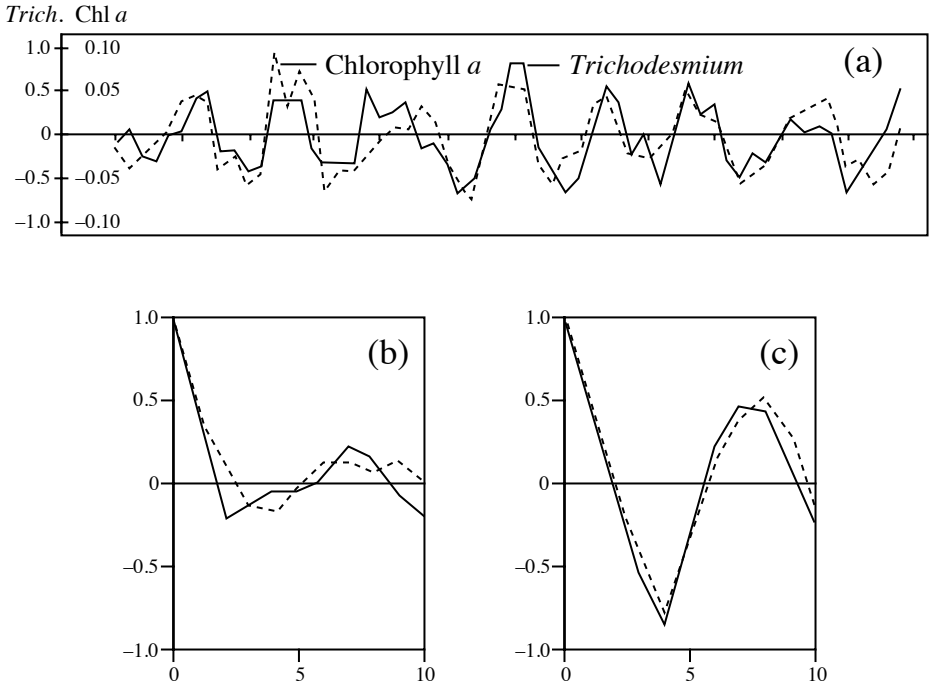


Figure 12.8 (a) Filtered time series of chlorophyll *a* and *Trichodesmium* in tropical surface waters off Barbados. Marks along the abscissa are spaced by 75 days. On the ordinate, units are 10^3 filaments *Trichodesmium* L^{-1} and µg chlorophyll *a* L^{-1}. Correlograms of (b) the nonfiltered series and (c) the filtered series. After Steven & Glombitza (1972).

series: maternal behaviour of the Wistar rat (a strain of albino rats) in the laboratory, observed every two hours during five days (his Fig. 6.7, b1 and b2); yearly mean sunspot numbers for the period 1749-1994 (his Fig. 6.8, b and c); monthly atmospheric CO_2 concentrations at Mona Laua, Hawaii, from 1965 through 2004 (his Fig. 6.9, c and d); daily mean temperatures in air and soil in the Gault Nature Reserve (Québec) over thirty days in June 2004 (his Fig. 6.3, b-c and f-g); and hourly mean temperatures in air and soil in the same nature reserve over eight days in June (his Fig. 6.10, c-d and g-h).

2 — Cross-covariance and cross-correlation

In order to determine the extent to which two data series exhibit concordant periodic variations, a method closely related to autocovariance and autocorrelation can be used. This method has two variants called *cross-covariance* and *cross-correlation* (or *lag correlation*).

Consider two series, $\mathbf{y}_j$ and $\mathbf{y}_l$, of the same length. One is progressively shifted with respect to the other, with lags $k = 1, 2, \ldots$ As the lag increases, the zone of overlap between the two series shortens. *Cross-covariance* of order k is computed in a way analogous to autocovariance. As in eq. 12.6 for autocovariance, the means $\bar{y}_j$ and $\bar{y}_l$ of the full series are used to compute the cross-covariance $s_{jl}(k)$ between the two series for lag k:

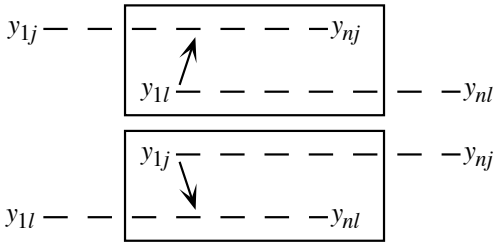
Cross-covariance

$$s_{jl}(k) = \frac{1}{n} \sum_{i=1}^{n-k} [y_{(i+k)j} - \bar{y}_j] [y_{il} - \bar{y}_l] \qquad \textbf{(12.8)}$$

When $k = 0$ (no shift), eq. 12.6 becomes the maximum likelihood estimator of the covariance between the variables:

$$s_{jl}(0) = \frac{1}{n} \sum_{i=1}^{n} (y_{ij} - \bar{y}_j)(y_{il} - \bar{y}_l)$$

Equation 12.8 shows an important difference between *cross-covariance* and *autocovariance*, namely that the relative direction in which a series is shifted with respect to the other must be taken into account. Indeed, shifting series $\mathbf{y}_j$ "to the right" by k units with respect to series $\mathbf{y}_l$ is not equivalent to shifting it "to the left" because the direction of the implied causal relationship (arrows in the figure) is not the same:



The value of cross-covariance for lag k would be different if $\mathbf{y}_j$ and $\mathbf{y}_l$ were interchanged in eq. 12.8; in other words, generally $s_{jl}(k) \neq s_{lj}(k)$. In order to distinguish between the two sets of cross-covariances, one set of shifts is labelled as positive and the other as negative. The choice of the positive and negative directions is arbitrary and without consequence. In eq. 12.8, if the cross-covariance of $\mathbf{y}_j$ relative to $\mathbf{y}_l$ is identified as $s_{jl}(k)$, the converse would be labelled $s_{jl}(-k)$. No distinction was made between the two relative shift directions in autocovariance (eq. 12.6) because $s_{yy}(+k) = s_{yy}(-k)$. When the direction of the causal relationship is known, there is no need to compute cross-covariance for both positive and negative shifts, although computer functions may automatically compute them, e.g. function ***ccf()*** in R.

The cross-covariance is generally plotted as a function of the positive and negative lags k, to the right and to the left of $k = 0$. The alternative is to plot the two sets on the

positive side of the abscissa using two different symbols. Maximum cross-covariance does not necessarily occur at $k = 0$. Sometimes, the dependence between the two series is maximum at a lag $k \neq 0$. In predator-prey interactions for example, cross-covariance may be maximum for a lag corresponding to the response time of the predator population (*target variable*) to changes in the number of prey (*predictor variable*). One then says that the target variable *lags* the causal variable.

Cross-covariance can be transformed into *cross-correlation*. To do so, the cross-covariance $s_{jl}(k)$ is divided by the product of the corresponding standard deviations, which are the square roots of the variance $ss_{jj}(0)$ and $ss_{ll}(0)$ (eq. 12.5):

Cross-correlation

$$r_{jl}(k) = \frac{s_{jl}(k)}{\sqrt{s_{jj}(0)\, s_{ll}(0)}} \qquad \textbf{(12.9)}$$

Cross-correlogram

As for cross-covariance, cross-correlation is defined for $+k$ and $-k$. Values are plotted as a function of k in a *cross-correlogram*. Fortier & Legendre (1979) used Kendall's τ (Section 5.3) instead of Pearson's r for computing cross-correlations between series of quantitative variables which were *not linearly related*. They called this measure *Kendall's cross-correlation*. It may also be applied to series of *semiquantitative* data; Spearman's r (Section 5.3) could be used instead of Kendall's τ. Legendre & Legendre (1982) proposed to extend this approach to *qualitative* data under the name *cross-contingency*. In that case, contingency statistics (X^2 or uncertainty coefficients; Section 6.2) are computed for the two series as one is progressively shifted with respect to the other.

When several ecological variables are observed simultaneously, the resulting *multidimensional series* may be analysed using cross-covariance or cross-correlation. Such methods are obviously of interest in ecology, where variation in one variable is often interpreted in terms of variation in others. However, eq. 12.9 considers only two series at a time; for multidimensional data series, it is sometimes useful to extend the concept of *partial correlation* (Sections 4.5 and 5.3) to the approach of cross-correlation. In Ecological application 12.3d, Fréchette & Legendre (1982) used Kendall's partial (partial τ; Section 5.3) cross-correlation to analyse an ecological situation involving three variables.

Ecological application 12.3c

In their study on temporal variability of tropical phytoplankton (Ecological application 12.3b), Steven & Glombitza (1972) compared the variations in concentrations of chlorophyll *a* and *Trichodesmium*, using cross-correlations (Fig. 12.9). The cross-correlogram shows that changes in the two variables were in phase, with a period of 8 lags × 15 days lag^{-1} = 120 days. Filtration of the data series brought but a small improvement to the cross-correlation. These results confirm the conclusions drawn from the correlograms (Fig. 12.8), and show that variations of chlorophyll *a* concentration, in surface waters, were due to changes in the concentration of *Trichodesmium* filaments. This same application will be further discussed in Section 12.4 (Ecological application 12.4e).

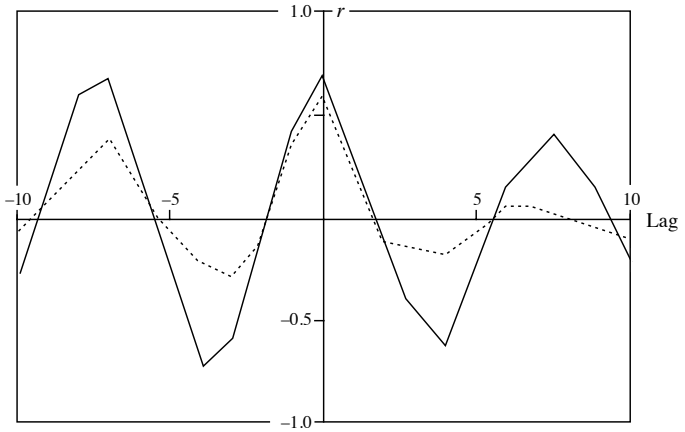


Figure 12.9 Cross-correlations between temporal changes in concentrations of chlorophyll *a* and *Trichodesmium*, in tropical surface waters, computed on nonfiltered (solid line) and filtered (dotted line) data series. After Steven & Glombitza (1972).

Ecological application 12.3d

At an anchor station in the St. Lawrence Estuary (Québec), Fréchette & Legendre (1982) determined the photosynthetic capacity of phytoplankton (P^B_{max}) hourly, during six consecutive days. The sampling area was subjected to internal tides, which drove changes in two important physical variables: (1) vertical oscillations of the water mass (characterized in this study by the depth of isopycnal $\sigma_t = 22$, i.e. the depth where the density of water was 1022 kg m^{-3}), and (2) variations in the vertical stability of the upper water column, estimated as the density gradient between 1 and 25 m. Two hypotheses could explain the observed effects of internal tides on P^B_{max}: (1) upwelling under the effect of incoming internal tides, up to the depths where sampling took place, of deeper water containing phytoplankton with lower P^B_{max}, or (2) adaptation of P^B_{max} to changes in the vertical stability of the upper water column.

Since the two physical variables were controlled by the same mechanism (i.e. internal tides), it was not easy to identify their specific contributions to phytoplankton photosynthesis. This was achieved by computing two Kendall's partial cross-correlations (partial τ): (1) between P^B_{max} and the depth of $\sigma_t = 22$, controlling for the effect of vertical stability, and (2) between P^B_{max} and stratification, controlling for vertical displacement. When calculating the partial cross-correlations, the response variable (P^B_{max}) was shifted relative to the two potentially causal (physical) variables until a maximum value was reached.

The authors concluded that the photosynthetic activity of phytoplankton responded to changes in the vertical stability of the water column, driven by internal tides. This was interpreted as an adaptation of the cells to periodic variations in their light environment.

Another example of cross-correlation applied to an ecological data series can be found in Wilson & Dawe (2006), who used cross-correlograms to compare variations in population densities of marine foraminifera with monsoonal rainfall data.

For series with irregular lag or missing data, the multivariate variogram (Subsection 13.1.4) can be used to detect periodic phenomena in univariate or multivariate quantitative data series. Likewise, the Mantel correlogram (Subsection 13.1.6) can be used to detect periodic phenomena in irregular univariate quantitative, semiquantitative or qualitative data series, and in multivariate series involving variables of any precision level. This type of correlogram is computed from a distance matrix among the observations in the series.

12.4 Periodic variability: periodogram

In addition to the relatively simple methods discussed in the previous section, there is another general approach to the study of periodic variability, called *harmonic analysis*. This approach is mathematically more complex than correlogram analysis, but it is often better adapted to the study of ecological data series. Results of harmonic analysis are generally plotted in a graph called *periodogram*.

1 — Periodogram of Whittaker and Robinson

The simplest way to approach harmonic analysis is to examine a *Buys-Ballot table*. Assume that a series of n quantitative observations is characterized by a period T_{series}. If $T = T_{\text{series}}$ is known, the series can be split into n/T sequences, each containing T observations. A Buys-Ballot table (Table 12.7) is a double-entry table whose rows contain the $r = n/T$ sequences of T observations. The number of columns corresponds to the known or assumed period of the data series. If $T = T_{\text{series}}$, the r successive rows in the table are repetitions of the same oscillation, although the actual values in any column (j) are generally not identical because of noise. Calculating the mean value for each column ($\bar{y}_{T,\,j}$) and comparing these means is a way of characterizing the variation within period T_{series}.

When there exists a hypothesis concerning the value of T_{series} (e.g. a diurnal cycle), Buys-Ballot tables may be constructed for this value and also for neighbouring lower and higher values T_k. As the period of the table (T_k) approaches that of the series (T_{series}), values within each column become more similar, so that all maximum values tend to be located in one column and all minimum values in another. As a result, the difference between the highest and lowest mean values is maximum when period T_k of the table is the same as period T_{series} of the series. The *amplitude* of a Buys-Ballot

Amplitude

Table 12.7 Buys-Ballot table. Allocation of data from a series containing n observations to the rows of the table.

	1	2	3	...	T
1	y_1	y_2	y_3	...	y_T
2	y_{T+1}	y_{T+2}	y_{T+3}	...	y_{2T}
.	.	.	.	...	.
.	.	.	.	...	.
.	.	.	.	...	.
r	$y_{(r-1)T+1}$	$y_{(r-1)T+2}$	$y_{(r-1)T+3}$	...	$y_{rT} = y_n$
$\bar{y}_T$	$\bar{y}_{T,1}$	$\bar{y}_{T,2}$	$\bar{y}_{T,3}$	...	$\bar{y}_{T,T}$

table is some measure of the variation found among the columns of the table. It may be measured by the *range* of the column means (Whittaker & Robinson, 1924):

Range

$$[\bar{y}_{\max} - \bar{y}_{\min}] \qquad \textbf{(12.10)}$$

or by the *standard deviation* of the column means (Enright, 1965):

Standard deviation

$$\sqrt{\frac{1}{T_k}\sum_{j=1}^{T_k}(\bar{y}_{T_k,j} - \bar{y}_{T_k})^2}, \quad \text{where} \quad \bar{y}_{T_k} = \frac{1}{T_k}\sum_{j=1}^{T_k}\bar{y}_{T_k,j} \qquad \textbf{(12.11)}$$

When the period T of interest is not an integer multiple of the interval between two observations, a problem occurs in the construction of the Buys-Ballot table. The solution proposed by Enright (1965) is to construct the table with a number of columns equal to the largest integer that is less than or equal to the period of interest, T. Observations are attributed to the columns in sequence, as usual, leaving out an observation here and there in such a way that the average rate of advance in the series, from row to row of the Buys-Ballot table, is T. This is done, formally, by using the following formula for the mean of each column j:

$$\bar{y}_{T,j} = \frac{1}{r}\sum_{i=1}^{r} y_{[(i-1)T+j]} \qquad \textbf{(12.12)}$$

where r is the number of rows with data in column j of the table. The subscript of y is systematically rounded up to the next integer. Thus, for example, if $T = 24.5$, $\bar{y}_{T,1}$ is estimated from values $\{y_1, y_{26}, y_{50}, y_{75}, y_{99}, y_{124}, \text{etc.}\}$ found in rows $i = \{1, 2, 3, 4, 5, 6, \text{etc.}\}$ of the table;

in other words, intervals of 24 and 25 units are successively used, to give an average period $T = 24.5$. This modified formula is required to understand Ecological application 12.4a, where fractional periods are used.

When studying an empirical data series, the period T_{series} is not known *a priori*. Even when some hypothesis is available concerning the value of T_{series}, one may want to check whether the hypothesized value is the one that best emerges when analysing the data. In both situations, estimating T_{series} becomes the purpose of the analysis. The values of amplitude, computed for different periods T, may be plotted together as a periodogram in order to determine which period best characterizes the data series.

Periodogram

The *periodogram of Whittaker & Robinson* is a graph in which the measures of amplitude (eq. 12.10 or 12.11) are plotted as a function of periods T_k. According to Enright (1965), periodograms based on the statistic of eq. 12.11 are more internally consistent than those based on eq. 12.10. Various ways have been proposed for testing the significance of statistic 12.11 (reviewed by Sokolove & Bushell, 1978); these tests are only asymptotically valid, so that they are not adequate for short time series.

Numerical example. Consider again the series (2, 2, 4, 7, 10, 5, 2, 5, 8, 4, 1, 2, 5, 9, 6, 3) used in Subsection 12.3.1 to compute Table 12.6. In order to examine period $T_k = 4$, for instance, the series is cut into segments of length 4 as follows:

$$2, 2, 4, 7; \quad 10, 5, 2, 5; \quad 8, 4, 1, 2; \quad 5, 9, 6, 3$$

and distributed in the successive rows of the table. Buys-Ballot tables for periods $T_k = 4$ and 5 are constructed as follows:

$T = 4$	1	2	3	4
Row 1	2	2	4	7
Row 2	10	5	2	5
Row 3	8	4	1	2
Row 4	5	9	6	3
Means	6.25	5	3.25	4.25

Range = 3, standard deviation = 1.0951

$T = 5$	1	2	3	4	5
Row 1	2	2	4	7	10
Row 2	5	2	5	8	4
Row 3	1	2	5	9	6
Row 4	3				
Means	2.75	2	4.67	8	6.67

Range = 6, standard deviation = 2.2708

The range is calculated using eq. 12.10 and the standard deviation with eq. 12.11. Repeating the calculations for $k = 2$ to 8 produces the periodogram in Fig. 12.10.

Interpretation of the periodogram may be quite simple. If one and only one oscillation is present in the series, the period with maximum amplitude is taken as the best estimate for the true period of this oscillation. Calculation of the periodogram is made under the assumption that there is *a single stable period* in the series. If several periods are present, the periodogram may be so distorted that its interpretation could lead to erroneous conclusions. Enright (1965) provides examples of such distortions, using artificial series. Other methods, discussed below, are better adapted to series with several periods.

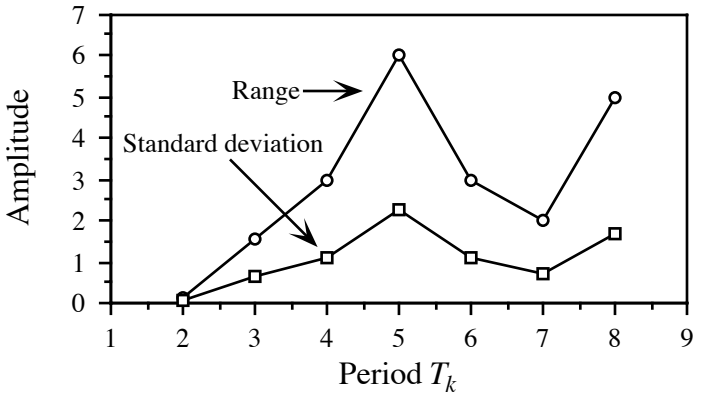


Figure 12.10 Periodogram of Whittaker and Robinson for the artificial data series. The amplitude statistics plotted in the periodogram may be either the range or the standard deviation of the column means in the Buys-Ballot tables.

Ecological application 12.4a

Enright (1965) re-examined 17 time series taken from the literature, which described the activity of animals as diverse as the chaffinch (bird), laboratory rat, crayfish, oyster, quahog (mollusc), and fiddler crab. The purpose of Enrigh's study was to determine, using periodograms, whether the cycles of activity described by the authors of the original papers (solar, i.e. 24 h, or lunar, i.e. 24.8 h) could withstand rigorous numerical analysis.

The approach is exemplified here by a series of 28 days of observations on the perch-hopping activity of a chaffinch, a European songbird, kept under constant light conditions. The periodogram shown in Fig. 12.11a is clearly dominated by a period of 21.8 h. Figures 12.11b-d display the mean values $\bar{y}_{T,\,j}$ of the columns of the Buys-Ballot tables constructed for some of the time periods investigated: $T_k = 21.8$, 24.0 and 24.8 h. (The values $\bar{y}_{T,\,j}$ of Fig. 12.11b-d were used to calculate the amplitudes of the periodogram shown in Fig. 12.11a.) Similar figures could be drawn for each point of the periodogram, since a Buys-Ballot table was constructed for each period considered. Without the array of values in the periodogram, exclusive examination of, say, the Buys-Ballot table for $T_k = 24$ h (Fig. 12.11c) could have led to conclude to the presence of a circadian rhythm. Similarly, examination of the table for $T_k = 24.8$ h (Fig. 12.11d) could have suggested a lunar rhythm. In the present case, the periodogram allowed Enright to (1) reject periods that were intuitively interesting (e.g. $T_k = 24$ h) but whose amplitude was not significantly high, and (2) identify a somewhat unexpected 21.8-h rhythm, which seemed to be of endogenous nature.

The 17 data series re-examined by Enright (1965) had been published with the objective of demonstrating the occurrence of circadian or tidal cycles. Enright's periodogram analyses confirmed the existence of circadian cycles for *only two* of the published series: one for the rat locomotor activity, and one for the quahog shell-opening activity. *None* of the published series exhibited a tidal (lunar) cycle. This stresses the usefulness of periodogram analysis in ecology and the importance of using appropriate numerical methods when studying data series.

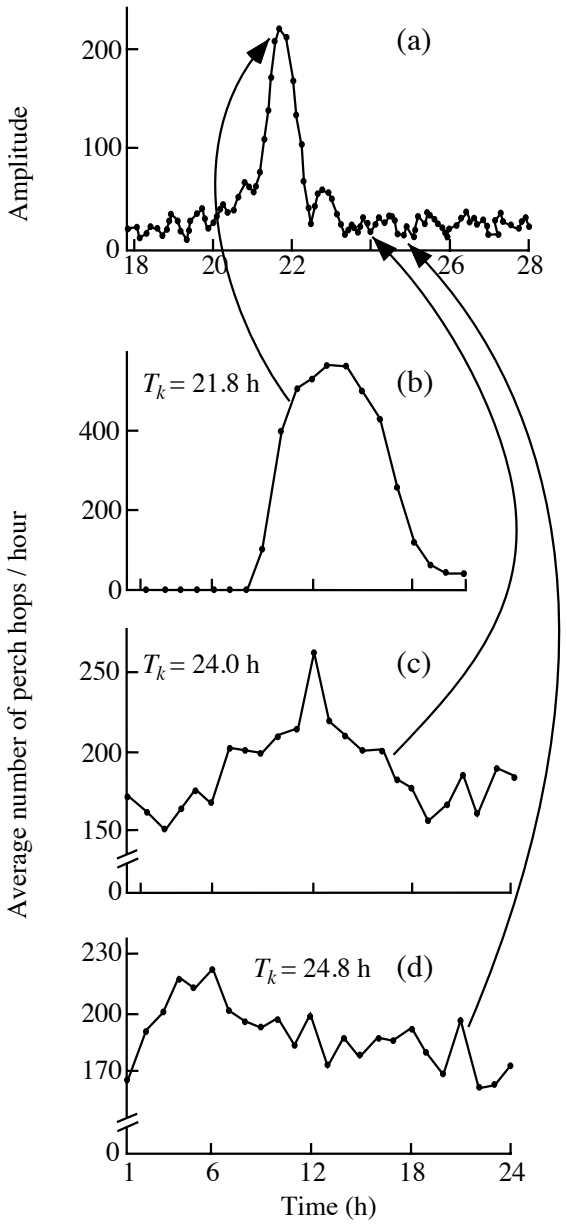


Figure 12.11 (a) Periodogram for the chaffinch perch-hopping activity series (n = 672 data points). The amplitude was calculated using Enright's formula (eq. 12.12). The three lower panels illustrate examples of values from which the amplitudes in (a) were calculated. These graphs show the means $\bar{y}_{T,\,j}$ of the columns in the Buys-Ballot tables, as functions of time, for periods T_k of (b) 21.8 h, (c) 24.0 h, and (d) 24.8 h. After Enright (1965).

Ecological application 12.4b

Nardi *et al*. (2003) used the periodogram of Whittaker & Robinson to study seasonal variations in the free-running period (i.e. circadian rhythm) in two populations of sandhopper (marine amphipods) on Italian beaches that differed in morphodynamics and human disturbance.

2 — Contingency periodogram of Legendre et al.

Another type of periodogram has been proposed by Legendre *et al*. (1981) to identify rhythms in series of *qualitative* ecological data. In this *contingency periodogram*, the Buys-Ballot table is replaced by a *contingency table* (Section 6.2). The columns of the table (Colwell, 1974) are the same as in a Buys-Ballot table, but the rows are the r states of the qualitative descriptor under study. Values in the table are frequencies f_{ij} of the states of the descriptor (rows i), observed at the various times (columns j) of period T_k. As in the periodogram of Whittaker & Robinson (above), a different table is constructed for each period T_k considered in the periodogram.

Information statistic

Information (H) as to the states of the qualitative variable of interest (S), which is accounted for by a given period T_k, is the information in common between S and the sampling axis X (most often, time). This amount of information is computed as the intersection between S and X, for period T_k:

$$H(\mathrm{S} \cap \mathrm{X}) = H(\mathrm{S}) + H(\mathrm{X}) - H(\mathrm{S}, \mathrm{X}) \qquad \textbf{(12.13)}$$

Equation 12.14 is the same as eq. 6.10, used for calculating the information shared by two descriptors (statistic B), so that $H(\mathrm{S} \cap \mathrm{X}) = \mathrm{B}$.

The *contingency periodogram* is a graph of the values $H(\mathrm{S} \cap \mathrm{X}) = \mathrm{B}$ on the ordinate, as a function of periods T_k. Periodograms, as well as spatial correlograms (Section 13.1), are often read from left (shortest periods or lags) to right (larger periods or lags). This is the case when the process that may have generated the periodic or autocorrelated structure of the data, if any, is assumed to be stronger at small lags and to generate short periods before these are combined into long periods.

Section 6.2 has shown that statistic B is related to Wilks' X^2_{W} statistic:

$$X^2_{\mathrm{W}} = 2n\mathrm{B} \quad \text{(when B in nats; eq. 6.13)}$$

or $$X^2_{\mathrm{W}} = 2n\mathrm{B} \log_e 2 = n\mathrm{B} \log_e 4 \quad \text{(when B in bits; eq. 6.14).}$$

Because X^2_{W} can be tested for significance, critical values may be drawn on the periodogram. The critical value of B is found by replacing X^2_{W} in eq. 6.13 by the critical value $\chi^2_{\alpha,\nu}$:

$$\mathrm{B}_{\mathrm{critical}} = \chi^2_{\alpha,\nu}/2n \quad \text{(for B in nats)}$$

where α is the significance level and $\nu = (r - 1)(T_k - 1)$ is the number of degrees of freedom. For the periodogram, an alternative to plotting B is to plot the χ^2_W statistic as a function of periods T_k; the critical value to be used is then $\chi^2_{\alpha,\nu}$ directly. As one proceeds from left (smaller periods) to right (larger periods) in the periodogram, T_k and ν increase; as a consequence, the critical value, $\chi^2_{\alpha,\nu}$ or $B_{critical}$, monotonically increases from left to right in this type of periodogram, as will be shown in the numerical example below.

Since multiple tests are performed in a contingency periodogram, the critical values of B must be corrected (Box 1.3). The simplest approach is the Bonferroni correction, where significance level α is replaced by $\alpha' = \alpha/$(number of simultaneous tests). In a periodogram, the number of simultaneous tests is the number of periods T_k for which the statistic (B or X^2_W) has been computed. Since the maximum number of periods that can be investigated is limited by the observational window (Section 12.0), the maximum number of simultaneous tests is $[(n/2) - 1]$ and the strongest Bonferroni correction that can be made is $\alpha' = \alpha/[(n/2) - 1]$. This is the correction recommended by Oden (1984) to assess the global significance of spatial correlograms (Section 13.1). In practice, when analysing long data series, one usually does not test the significance past some arbitrarily chosen point; if there are h statistics that have been tested for significance, the Bonferroni method would call for a corrected significance level $\alpha' = \alpha/h$.

There are two problems with the Bonferroni approach applied to periodograms and spatial correlograms. The first one is that the correction varies in intensity, depending on the number of periods (in periodograms) or lags (in spatial correlograms) for which statistics have been computed and tested. The second problem is that the interest in the results of the tests of significance decreases as the periods (or lags) get longer, especially in long data series; when a basic period has been identified, its harmonics are of lesser interest. These problems can be resolved by resorting to a *progressive Bonferroni* correction, proposed by P. Legendre in the Hewitt *et al*. (1997) paper. In this method, the first periodogram or spatial correlogram statistic is tested against the α significance level; the second statistic is tested against the Bonferroni-corrected level $\alpha' = \alpha/2$ because, at this point, two tests have been performed; and so forth until the k-th statistic, which is tested against the Bonferroni-corrected level $\alpha' = \alpha/k$. This approach also solves the problem of "where to stop computing a periodogram or spatial correlogram"; one goes on as long as significant values are likely to emerge, considering the fact that the significance level becomes progressively more stringent.

Progressive Bonferroni correction

Numerical example. Consider the following series of qualitative data ($n = 16$), for a qualitative variable with 3 states (from Legendre *et al*., 1981):

1, 1, 2, 3, 3, 2, 1, 2, 3, 2, 1, 1, 2, 3, 3, 1

To analyse period $T_k = 4$, for instance, the series is cut into segments of length 4 as follows:

1, 1, 2, 3; 3, 2, 1, 2; 3, 2, 1, 1; 2, 3, 3, 1

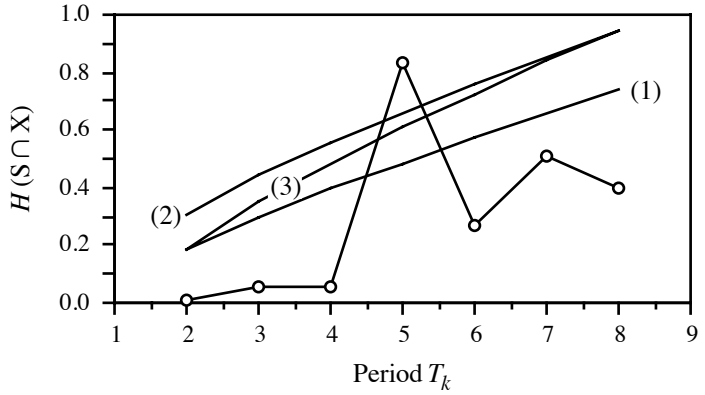


Figure 12.12 Contingency periodogram for the artificial data series (circles). The contingency statistic used here is B = H(S ∩ X). (1) Uncorrected critical values. (2) Bonferroni-corrected critical values, correcting for 7 simultaneous tests in the observational window. (3) Progressive Bonferroni correction; the first value ($T_k = 2$) is without correction, while the last ($T_k = 8$) receives the full Bonferroni correction.

and distributed in the successive rows of the table. The first four data go into columns 1 to 4 of the contingency table, each one in the row corresponding to its code; similarly, observations 5 to 8 are placed into the columns of the table, each in the row corresponding to its code; and so forth. When the operation is completed, the number of occurrences of observations are counted in each cell of the table, so that the resulting table is a contingency table containing frequencies f_{ij}. As an exercise, readers should try to reproduce the contingency tables shown below for $T_i = 4$ and $T_i = 5$. The values of X^2_W and B (in *nats*) are given for these two periods:

$T = 4$	1	2	3	4
State 1	1	1	2	2
State 2	1	2	1	1
State 3	2	1	1	1

B (in *nats*) = 0.055, $X^2_W = 1.76$

$T = 5$	1	2	3	4	5
State 1	3	3	0	0	0
State 2	1	0	3	0	1
State 3	0	0	0	3	2

B (in *nats*) = 0.835, $X^2_W = 26.72$

Repeating the calculations for $k = 2$ to 8 produces the periodogram shown in Fig. 12.12. Only $X^2_W = 26.72$ ($T = 5$) is larger than the corresponding critical value, which may be computed in various ways (as explained above), depending on the need:

• Uncorrected critical value: $\alpha = 0.05$, $\nu = (3 - 1)(5 - 1) = 8$, critical $\chi^2_{\alpha,\nu} = 15.5$. $B_{critical} = 15.5/(2 \times 16) = 0.484$.

• Bonferroni correction for 7 simultaneous tests: $\alpha' = \alpha/(n/2 - 1) = 0.05/7$, $\nu = 8$, critical $\chi^2_{\alpha',\nu} = 21.0$. $B_{critical} = 21.0/32 = 0.656$.

• Progressive Bonferroni correction. Example for the 4th test: $\alpha' = \alpha/4 = 0.05/4$, $\nu = 8$, critical $\chi^2_{\alpha',\nu} = 19.5$. $B_{critical} = 19.5/32 = 0.609$.

Thus, the only significant period in the data series is $T_k = 5$.

The contingency periodogram can be directly applied to qualitative descriptors. Quantitative or semiquantitative descriptors must be divided into states before analysis with the contingency periodogram. A method to do so is described in Legendre *et al.* (1981).

In their paper, Legendre *et al.* (1981) established the robustness of the contingency periodogram in the presence of strong random variations, which often occur in ecological data series, and its ability to identify hidden periods in series of non-quantitative ecological data. Another advantage of the contingency periodogram is its ability to analyse very short data series.

One of the applications of the contingency periodogram is the analysis of multivariate series (e.g. multi-species; Ecological application 12.4c). Such series may be transformed into a single qualitative variable describing a partition of the observations found by clustering (Chapter 8). With the contingency periodogram, it is possible to analyse the data series, now transformed into a single nonordered variable (factor) corresponding to the partition of the observations. An alternative approach would be to carry out the analysis on the multivariate distance matrix among observations using the Mantel correlogram described in Subsection 13.1.6.

Ecological application 12.4c

Phytoplankton was enumerated in a series of 175 water samples collected hourly at an anchor station in the St. Lawrence Estuary (Québec). Using the contingency periodogram, Legendre *et al.* (1981) analysed the first 80 h of that series, which corresponded to neap tides. The original data consisted of six functional taxonomic groups. The *six-dimensional quantitative data* were transformed into a *one-dimensional qualitative descriptor* by clustering the 80 observations using flexible clustering (Subsection 8.5.10). Five clusters of "hours" were obtained; each hour of the series was attributed to one of them. Each cluster thus defined a state of the new qualitative variable resulting from the classification of the hourly data.

When applied to the qualitative series, the contingency periodogram identified a significant period $T = 3$ h, which suggested rapid changes in surface waters at the sampling site. The integer multiples (harmonics) of the basic period (3 h) in the series also appeared in the contingency periodogram. Periods $T = 6$ h, 9 h, and so on, had about the same significance as the basic period, so that they did not indicate the presence of additional periods in the series.

3 — Periodogram of Schuster

Harmonic analysis

For *quantitative* serial variables, there exists another method for calculating a periodogram, which is mathematically more complex than the periodogram of Whittaker and Robinson (Subsection 12.4.1) but is also more powerful. It is sometimes called *harmonic analysis* or *periodic regression*. This method is based on the fact that

the periodic variability present in series of quantitative data can often be represented by a sum of periodic terms, involving combinations of sines and cosines (Fig. 12.13):

Fourier series

$$y(x) = a_0 + \sum_k \left[a_k \cos\left(\frac{2\pi}{T_k} x\right) + b_k \sin\left(\frac{2\pi}{T_k} x\right) \right] \quad \textbf{(12.14)}$$

Equation 12.15 is called a *Fourier series*. Constant a_0 is the mean of the series; parameters a_k and b_k determine the importance of a given period T_k in the resulting signal. Using eq. 12.14, any periodic signal can be partitioned into a sequence of superimposed oscillations (Fig. 12.13). Function $\cos[x(2\pi/T_k)]$ transforms the explanatory variable x into a cyclic variable. Periods T_k are generally chosen in such a way that the sines and cosines, which model the data series, are *harmonics* (Section 12.0) of a fundamental period T_0: $T_k = T_0/k$ (where $k = 1, 2, \ldots, n/2$). Periods T_k become shorter as k increases. Equation 12.15 may be rewritten as:

$$y(x) = a_0 + \sum_{k=1}^{n/2} \left[a_k \cos\left(k\frac{2\pi}{T_0} x\right) + b_k \sin\left(k\frac{2\pi}{T_0} x\right) \right]$$

Generally, T_0 is taken to be equal to the length of the series ($T_0 = n\Delta$, where Δ is the interval between data points), so that:

$$y(x) = a_0 + \sum_{k=1}^{n/2} \left[a_k \cos\left(k\frac{2\pi}{n\Delta} x\right) + b_k \sin\left(k\frac{2\pi}{n\Delta} x\right) \right] \quad \textbf{(12.15)}$$

The purpose of Fourier analysis is not to determine the values of coefficients a_k and b_k, but to find out which periods, among all periods T_k, best explain the variance observed in the response variable $y(x)$. After estimating the values of a_k and b_k, the *amplitude* of the periodogram for each period T_k is computed as the *fraction of the variance* of the series that is explained by the given period. This quantity, which is a sum of coefficients of partial determination, combines the estimates of coefficients a_k and b_k as follows:

$$S^2(T_k) = (a_k^2 + b_k^2)/2 \quad \textbf{(12.16)}$$

Values in the periodogram are thus calculated by fitting to the data series (by least squares) a finite number of sine and cosine functions with different periods. There are $n/2$ such functions in the harmonic case. The shortest period considered is 2Δ ($T_{k\,\max} = T_0/(n/2) = n\Delta/(n/2) = 2\Delta$). It corresponds to the lower limit (expressed in period or wavelength) of the observational window (bottom row of Table 12.1). The amplitude is computed for each period T_k independently.

Schuster periodogram

Plotting the amplitudes from eq. 12.16 as a function of periods T_k produces the *periodogram of Schuster* (1898), which is used to identify significant periods in data series. In usual calculations, frequencies T_k are harmonics of T_0, but it is also possible

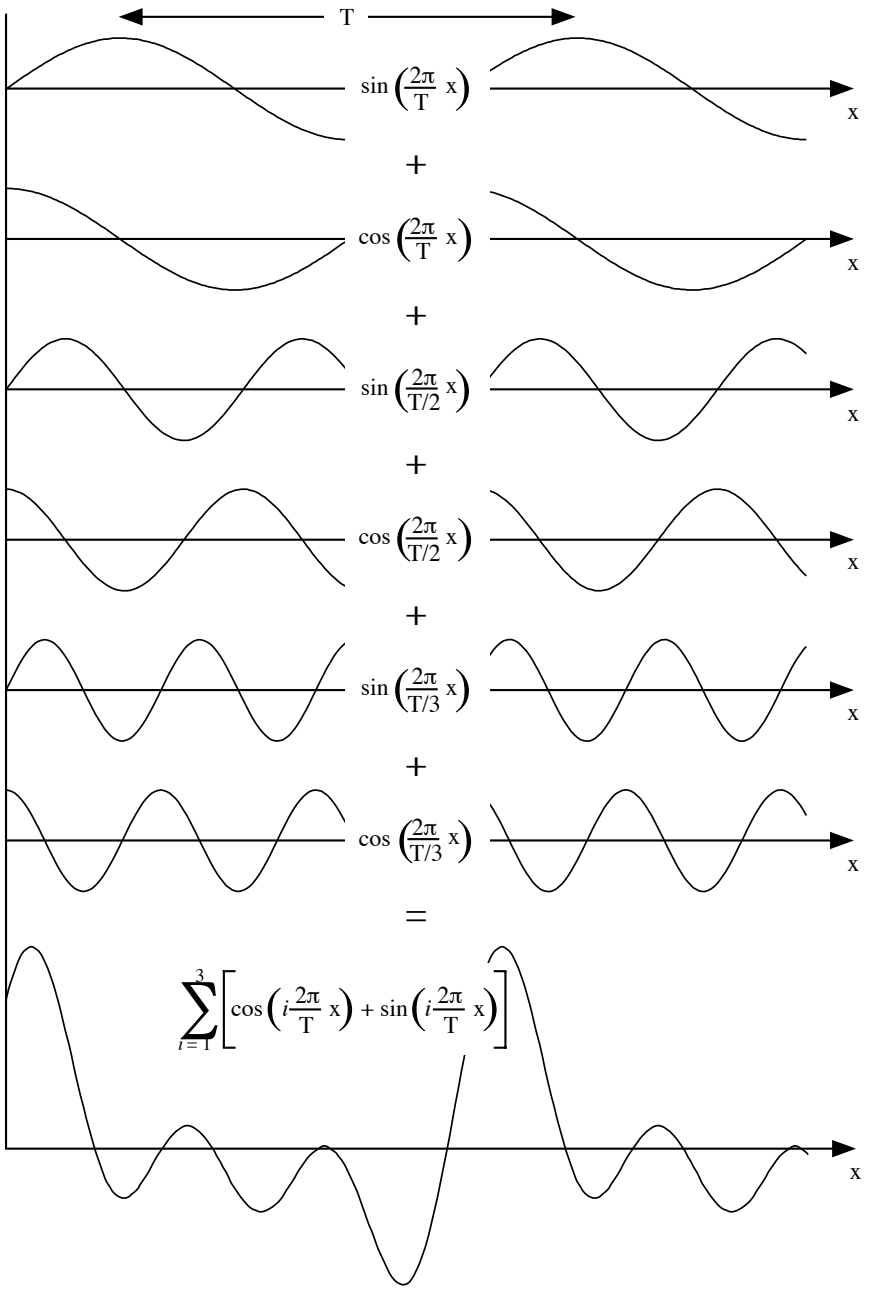


Figure 12.13 Fourier series. The periodic variation in this example (bottom graph, same as the periodic component of Fig. 12.2b) results from the sum of three sines and three cosines, which make up a harmonic sequence ($T_k = T, T/2$ and $T/3$). The mean of the series is 0 ($a_0 = 0$) and the amplitude of each sine and cosine is equal to 1 ($a_k = b_k = 1$).

to choose them to correspond to values of particular interest in the study. Contrary to the periodogram of Whittaker & Robinson, which does not refer to an underlying mathematical model, Schuster's periodogram is based on Fourier series (eqs. 12.14 and 12.15). Indeed, Kendall & Ord (1990, p. 158) have shown that any time series may be decomposed into a set of cycles based on the harmonic frequencies, even if the series does not display periodicity. Spatial eigenfunctions (Sections 14.1 to 14.3), computed along the time series, can be used for the same type of decomposition.

One advantage of Schuster's periodogram is that it can handle series showing several periods, contrary to the periodogram of Whittaker and Robinson which is limited to series with only one stable period (see above). Values in Schuster's periodogram can be tested for significance by reference to a critical value, which is calculated using a formula derived from Anderson (1971, p. 110 *et seq.*):

$$-(2/n)\log_e\left(1-\sqrt[m]{1-\alpha}\right) \qquad \textbf{(12.17)}$$

where n is the number of observations in the series, m is the largest computed harmonic period (usually, $m = n/2$), and α is the significance level.

Ecological application 12.4d

Demers & Legendre (1981) used Schuster's periodogram to analyse a 76-h series of oceanographic data. For a significance level $\alpha = 0.05$, the critical value for the periodogram was:

$$-(2/76)\log_e\left(1-\sqrt[38]{1-0.05}\right) = 0.174 = 17.4\%$$

Hence, any period explaining more than 17.4% of the variance of the series was considered to be significantly different from zero at significance level $\alpha = 0.05$.

Ecological application 12.4e

The time series of chlorophyll *a* and *Trichodesmium* filaments in tropical waters (Steven & Glombitza, 1972), discussed in Ecological applications 12.3b and 12.3c above, were subjected to harmonic analysis. Results are reported in Table 12.8. Each column of results could also be plotted as a periodogram. The period $T = 120$ days, already evidenced by autocorrelation (Fig. 12.8) and cross-correlation (Fig. 12.9), was also clearly identified by harmonic analysis.

Ecological application 12.4f

Crow birds act as a reservoir of the West Nile virus (WNV), which first appeared in North America in 1999. Ludwig *et al*. (2009) used Schuster's periodogram to investigate the population dynamics of crow birds in Québec and evaluate the impact of WNV infection on these dynamics. Their purpose was to develop a predictive algorithm that could be used as a disease surveillance tool and a measure of the impact of WNV on wildlife.

Table 12.8 Harmonic analysis of time series of chlorophyll *a* and *Trichodesmium* filaments, in tropical marine waters. The table reports the amplitudes corresponding to harmonic periods. The dominant period (T_k = 120) is in italics. After Steven & Glombitza (1972).

Harmonic	Period	Nonfiltered series		Filtered series	
k	T_k = 840 days/k	Chl *a*	*Trichodesmium*	Chl *a*	*Trichodesmium*
4	210	0.007	67	0.010	75
5	168	0.007	178	0.006	168
6	140	0.022	113	0.019	129
7	*120*	*0.039*	*318*	*0.038*	*311*
8	105	0.017	147	0.016	162
9	93	0.018	295	0.019	291
10	84	0.020	123	0.020	144

4— Periodogram of Dutilleul

Fractional periods do not correspond to an integer number of cycles in the series. These periods are usually not computed in Schuster's periodogram, although there is nothing that prevents it mathematically except the fact that the test of statistical significance of individual values (eq. 12.17) is only asymptotically valid with fractional periods. As a consequence, Schuster's periodogram is poorly adapted to the analysis of short time series, in which the periods of interest are likely to be fractional. A rule of thumb is to only analyse series that are at least 10 times as long as the longest hypothesized period.

Dutilleul (1990) proposed to modify Schuster's periodogram, in order to compute the portion of total variance associated with periods that do not correspond to integer fractions of the fundamental period T_0 (i.e. *fractional periods*). The method allows a more precise detection of the periods of interest and is especially useful with *short data series*.

Dutilleul periodogram

The statistic in *Dutilleul's modified periodogram* is the exact fraction of the total variance of the time series explained by regressing the series on the sines and cosines corresponding to one or several periodic components. In contrast, Schuster's periodogram is estimated *for a single period at a time*, i.e. each period T_k in eq. 12.14. It follows that, when applied to short series, Schuster's periodogram generally only provides an approximation of the explained fraction of the variance. In general, the number of periodic components actually present in a series is unknown *a priori*, but it may be estimated using a stepwise procedure proposed by Dutilleul (1990; see also

Dutilleul, 1998). The modified periodogram thus offers two major extensions over Schuster's: (1) it may be computed for *several periods at a time* (i.e. it is *multifrequential*) and (2) its maximization over the continuous domain of possible periods provides the maximization of the sum of squares of the corresponding trigonometric model fitted by least squares to the series. Both periodograms lead to the same estimates when computed for a single period over a long data series, or when the period corresponds to an integer fraction of T_0. In all other cases, the modified periodogram has better statistical properties (Dutilleul, 1990; see also Legendre & Dutilleul, 1992; Dutilleul & Till, 1992; Dutilleul, 1998, 2011):

- The explained fraction of the variance tends to be maximum for the true periods present in the time series, even when these are fractional, because the periodogram statistic exactly represents the sum of squares of the trigonometric model fitted by least squares to the series at the frequencies considered, whether these are integers or not (when expressed in number of cycles over the series).

- Assuming normality for the data series, the periodogram statistic is distributed like χ^2 for all periods in small or large samples, which leads to exact tests of significance. With Schuster's periodogram, this is only the case for periods corresponding to integer fractions of T_0 or, outside these periods, only for large samples.

- When the number of frequencies involved in the computation corresponds to the true number of periodic components in the series, the frequencies maximizing the periodogram statistic are unbiased estimates of the true frequencies. The stepwise procedure mentioned above allows the estimation of the number of periodic components present in the series.

In order to compare Dutilleul's periodogram to Schuster's, Legendre & Dutilleul (1992) created a test data series of 30 simulated observations containing two periodic components, which jointly accounted for 70.7% of the total variance in the series, with added noise. The true periods were $T = 12$ and 15 units. Schuster's periodogram brought out only one peak, because the two components were close to each other and Schuster's periodogram statistic was estimated for only one period at a time. When estimated for a single period, Dutilleul's modified periodogram shared this drawback. However, when estimated for the correct number of periods (i.e. two, as found by the stepwise procedure mentioned above), the modified periodogram showed maxima near the two constructed periods, i.e. at $T = 11.3$ and 14.4 units. The authors also compared the results of Dutilleul's method to those obtained with the stepwise procedure of Damsleth & Spjøtvoll (1982), which is based on Schuster's periodogram. Results from the latter (estimated periods $T = 10.3$ and 13.5) were not as good as with Dutilleul's modified periodogram. Dutilleul (1998) also showed the better performance of the modified periodogram over autocorrelograms in the context of scale analysis.

Dutilleul & Till (1992) published an application of the modified periodogram to the analysis of long dendrochronological series. Dutilleul's periodogram clearly detected the annual solar signal in cedar tree-ring series in the Atlas, a sub-tropical region

where, typically, the annual dendrochronological signal is weak. An application to a series of moderate length (river discharge) was published by Tardif *et al*. (1998).

Dutilleul (2011, his Section 6.3.2) discussed applications of Dutilleul's periodogram to several data series: maternal behaviour of the Wistar rat (a strain of albino rats) in the laboratory, observed every two hours during five days (his Fig. 6.7, c1 and c2); yearly mean sunspot numbers for the period 1749-1994 (his Fig. 6.8, d and e); monthly atmospheric CO_2 concentrations at Mona Laua, Hawaii, from 1965 through 2004 (his Fig. 6.9b); and hourly mean temperatures in air and soil in the Gault Nature Reserve (Québec) over eight days in June 2004 (his Fig. 6.10, b and f).

5— Harmonic regression

Legand (1958) proposed to use the first term of the Fourier series (eq. 12.14) to analyse ecological periodic phenomena with known *sinusoidal* periodic variability (e.g. circadian or annual). This method is called *harmonic regression*. As in the case of Fourier series (see above), the explanatory variable x (e.g. time of day) is transformed into a cyclic variable:

$$x' = \cos\left[\frac{2\pi}{T}(x + c)\right] \qquad \textbf{(12.18)}$$

for cosine functions using angles in radians, as in R. In the above expression, which is the first term of a Fourier series, T is the period suggested by hypothesis (e.g. 24 hours); x is the explanatory variable (e.g. local time); and 2π is replaced by 360° when the cosine function uses angles in degrees. Constant c fits the position of the cosine along the abscissa, so that it corresponds to the time of minimum and maximum values in the data set. The regression coefficients are estimated by the least-squares method:

$$\hat{y} = b_0 + b_1 x'$$

The harmonic regression equation can be fitted to data series by nonlinear least squares using function ***nls()*** in R (Section 10.7).

Ecological application 12.4g

Angot (1961) studied the diurnal cycle of marine phytoplankton production near New Caledonia, in the South Pacific. Values of primary production exhibited regular diurnal cyclic variations, which might reflect physiological rhythms. After logarithmic transformation of the primary production values, the author found *significant* harmonic regressions, with $T = 24$ h and $c = 3$ h; the explanatory variable x was the local time. Coefficients of regression b_0 and b_1 were used to compare different sampling sites.

Ecological application 12.4h

Taguchi (1976) used harmonic regression to study the short-term variability of marine phytoplankton production for different irradiance conditions and seasons. Data, which represented a variety of coastal conditions, were first transformed into ratios of production to chlorophyll *a*. The explanatory variable x was local time, $c = 4$ h, and T was generally 24 h. The intercept b_0 represented the mean production and b_1 was the slope of the regression line. The two coefficients decreased with irradiance and varied with seasons. The author interpreted the observed changes of regression coefficients in terms of photosynthetic dynamics.

Periodogram analysis is of interest in ecology because calculations are relatively simple and interpretation is direct. The correlogram and periodogram approaches, however, often give way to *spectral analysis* (next section). Spectral analysis is more powerful than correlogram or periodogram analyses, but it is also a more complex method for studying series. For simple problems where spectral analysis would be an unnecessary luxury, ecologists should rely on correlograms or, better, periodograms.

12.5 Periodic variability: spectral and wavelet analyses

Spectral analysis is the most advanced approach to analyse data series. The general concepts upon which it is founded are described below and illustrated by ecological applications. However, the analysis cannot be conducted without taking into account a number of theoretical and practical considerations, whose discussion exceeds the scope of the present book. Interested readers should refer, for instance, to the review papers by Platt & Denman (1975) and Fry *et al*. (1981). They may also consult the books of Bendat & Piersol (1971) and Muller & Macdonald (2002) as well as the references provided at the end of Section 12.0. Ecologists wishing to use spectral analysis are advised to consult a colleague with *practical experience* of the method.

1 — Series of a single variable

In the previous section, calculation of the Schuster periodogram involved least-squares fitting of a Fourier series to the data (eq. 12.14):

$$y(x) = a_0 + \sum_{k=1}^{n/2} \left[a_k \cos\left(\frac{2\pi}{T_k} x\right) + b_k \sin\left(\frac{2\pi}{T_k} x\right) \right]$$

When calculating the periodogram, the Fourier series was constructed using periods $T_k = T_0/k$. In spectral analysis, frequencies $f_k = 1/T_k$ are used instead of periods T_k. Thus, eq. 12.14 is rewritten as:

$$y(x) = a_0 + \sum_{k=1}^{n/2} \left[a_k \cos(2\pi f_k x) + b_k \sin(2\pi f_k x) \right] \qquad \textbf{(12.19)}$$

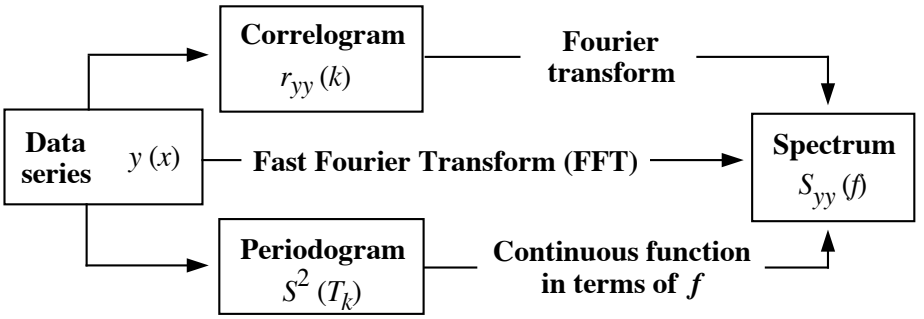


Figure 12.14 Relationships between a data series, its correlogram and periodogram, and its variance spectrum. The figure shows that the correlogram or the periodogram, on the one hand, and the spectrum, on the other hand, form a pair of Fourier transforms.

Using a formula similar to eq. 12.16, the *intensity* of the periodogram, at frequency f_k, is computed using the least-squares estimates of coefficients a_k and b_k:

$$I(f_k) = n(a_k^2 + b_k^2)/2 \qquad \textbf{(12.20)}$$

The intensity of the periodogram is defined only for *harmonic* frequencies $k/n\Delta$. It is possible, however, to turn the intensity of the periodogram into a *continuous* function over all frequencies from zero to the Nyquist frequency (see Table 12.1). This defines the *spectrum* of the series:

Spectrum

$$S_{yy}(f) = n(a_f^2 + b_f^2)/2 \qquad 0 \le f \le f_{n/2} \qquad \textbf{(12.21)}$$

The spectrum is thus a *continuous* function of frequencies, whereas the periodogram is discontinuous. Calculation and interpretation of spectra is the object of *spectral analysis*. Because of its origin in the field of electricity and telecommunications, the spectrum is sometimes called "power spectrum" or "energy spectrum". As shown below, it is also a "variance spectrum", which is the terminology used in ecology.

Power, or variance spectrum

In algebra, there exist mathematically equivalent pairs of equations that are used to go from one independent variable to another. Two mathematically equivalent equations where one is a function of x and the other a function of frequency $f = 1/x$ are called a pair of *Fourier transforms*. It can be shown that the *autocovariance* or *autocorrelation* function (eqs. 12.5-12.7) and the *spectral density* function (eq. 12.21) are a pair of Fourier transforms. In other words, the spectral density function is a Fourier transform of the autocorrelation function, and vice versa. Therefore, both the correlogram (Section 12.3) and periodogram analyses (Section 12.4), when they are generalized, lead to spectral analysis (Fig. 12.14). Classically, the spectrum is

Fast Fourier Transform

computed by Fourier transformation (also called "Fourier transform") of the autocorrelation, followed by smoothing. There is another method, called *Fast Fourier Transform* (FFT), which is faster than the classical approach (shorter computing time) and efficiently computes the pair of Fourier transforms written in discrete form. This last method offers the advantage of computational efficiency, but it involves a number of constraints, which can only be fully mastered after acquiring some *practical* experience of spectral analysis. It is sometimes confusing that, according to the context, the word "transform" is used as a verb (i.e. to transform an equation into another) or as a noun, and in the latter case it either refers to an algebraic operation (e.g. fast Fourier transform) or the result of that operation (e.g. a pairs of Fourier transforms).

Smoothing window

The spectrum computed from a correlogram or autocovariance function is an unbiased estimate of the true spectrum. However, the standard error of this spectral estimate is 100% whatever the length of the series. It follows that the computed spectrum must be *smoothed* in order to reduce its variance. Smoothing is done using a *window*, which is a function by which one multiplies the spectrum itself (spectral window), or the autocovariance estimates (lag window) prior to Fourier transformation. The two types of windows lead to the same results. The main problem of *smoothing* is that reduction of the standard error of the spectral estimates, on the ordinate, always leads to spreading of the variance on the abscissa. As a result, the spectral estimate, at any given frequency, may become contaminated by variance that is "leaking" from neighbouring frequencies. This *leakage* may result in biased smoothed spectral estimates. The various windows found in the literature (e.g. Bartlett, Daniell, de la Valle-Poussin or Parzen, Hamming, von Han, Tukey) provide different compromises between reduction of the standard error of spectral estimates and loss of resolution between adjacent frequencies. As was stressed above, the practical aspects of spectral analysis, including the choice of windows, filters (Section 12.2), and so on, often necessitate the help of an experienced colleague.

The ecological interpretation of spectra is not necessarily the same as that of correlograms or periodograms. First, the spectrum is a true *partition of the variance* of the series among frequencies. Therefore, spectral analysis is a third type of variance decomposition, in addition to the usual partitioning among experimental factors or sampling axes (ANOVA) and the partition among principal axes (Sections 4.4 and 9.1). The *units* of spectral density are [variance × frequency^{-1}], i.e. [(units of the response variable y)^2 × (units of the explanatory variable x)]. Therefore, the *variance* that corresponds to a frequency band is the *area under the curve* between the upper and lower frequencies, i.e. the integration of [variance × frequency^{-1}] over the frequency band. Spectra may be computed to identify harmonics in the data series or they may be regarded as characteristics of whole series, whether they are true sums of harmonics or not (Kendall & Ord, 1990, p. 158). These concepts should become clearer with the following ecological applications.

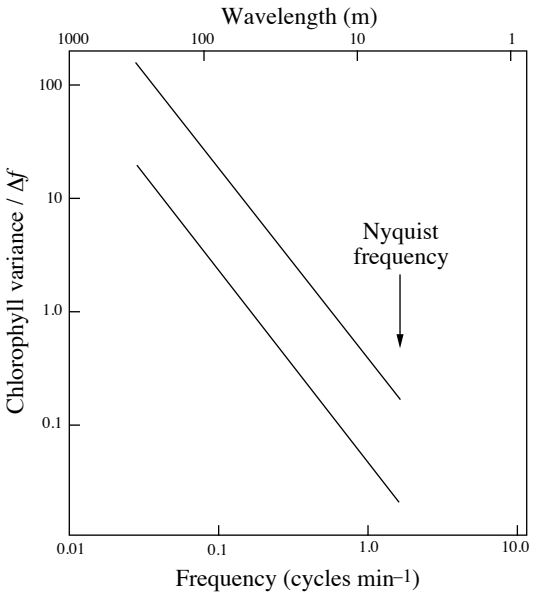


Figure 12.15 Horizontal distribution of chlorophyll *a* (*in vivo* fluorescence; arbitrary units) in surface waters of the Gulf of St. Lawrence. The two parallel lines on the variance spectrum show the envelope of calculated spectral densities. The Nyquist frequency is 1.5 cycle min^{-1}. After Platt (1972).

Ecological application 12.5a

At an anchor station in the Gulf of St. Lawrence, Platt (1972) continuously recorded *in vivo* fluorescence in surface waters as an estimate of phytoplankton chlorophyll *a*. Spectral analysis of the detrended data series (Fourier transform of autocorrelation) resulted in a spectrum characterized by a slope of –5/3, over frequencies ranging between ca. 0.01 and 1 cycle min^{-1}. The average current velocity being ca. 20 cm s^{-1} (ca. 10 m min^{-1}), the time series covered spatial scales ranging between ca. 1000 and 10 m (wavelength = speed × $frequency^{-1}$). This is illustrated in Fig. 12.15.

Interpretation of the spectrum was based on the fact that spectral analysis is a type of variance decomposition in which the total variance of the series is partitioned among the frequencies considered in the analysis (here: 0.03 cycle $min^{-1} < f <$ 1.5 cycle min^{-1}). The slope –5/3 corresponds to that of turbulent processes. This led the author to hypothesize that the local concentration of phytoplankton could be mainly controlled by turbulence. In a subsequent review paper, Platt & Denman (1975) cite various studies, based on spectral analysis, whose results confirm the hypothesis that the mesoscale spatial organization of phytoplankton is controlled by physical processes, in both marine and freshwater environments. This is in fact a modern version of the model proposed in 1953 by Kierstead & Slobodkin, which is discussed in Ecological applications 3.2d and 3.3a. Other references on spectral analysis of *in vivo* fluorescence series include Demers *et al*. (1979), Denman (1976, 1977), Denman & Platt (1975, 1976), Denman *et al*. (1977), Fashman & Pugh (1976), Legendre & Demers (1984), Lekan & Wilson (1978), Platt (1978), Platt & Denman (1975), and Powell *et al*. (1975), among others.

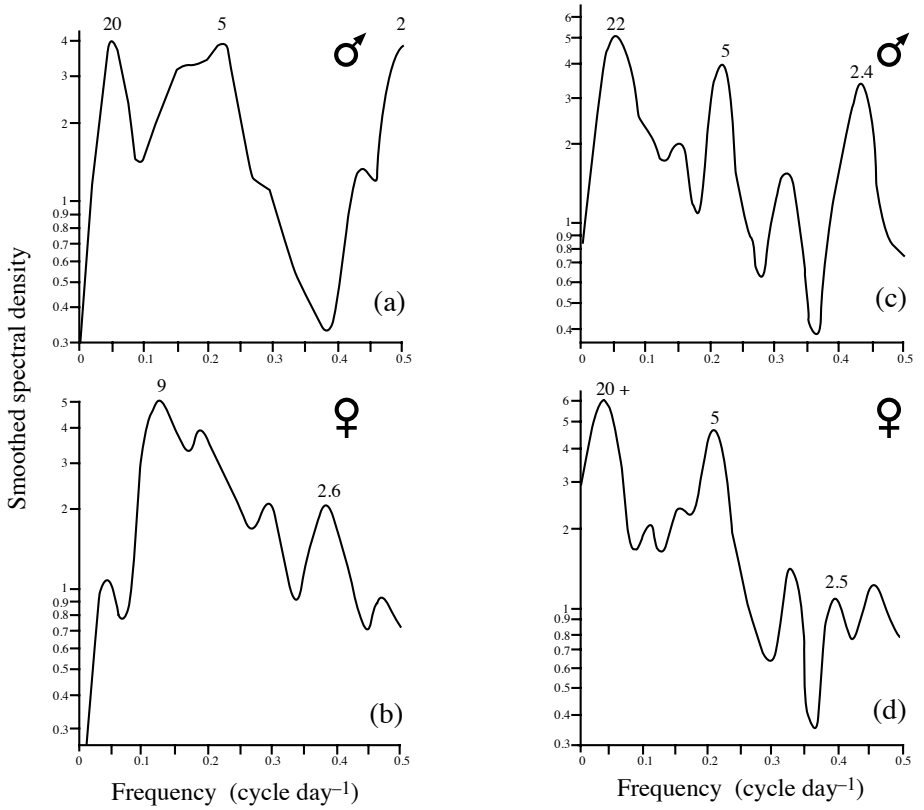


Figure 12.16 Estimates of smoothed spectra for numbers of migrating (a) male and (b) female crickets and for the locomotor activity of (c) male and (d) female crickets in the laboratory. The Nyquist frequency is 0.5 cycle day^{-1}. Periods corresponding to the main peaks are indicated above the curve, in order to facilitate interpretation; periods are the inverse of frequencies (abscissa). After Campbell & Shipp (1974).

Ecological application 12.5b

Campbell & Shipp (1974) tried to explain the migrations of an Australian cricket from observations on rhythms of locomotor activity of the males and females. One summer migration was followed during 100 days, starting in mid-February. In addition, locomotor activity rhythms of the males and females were observed in the laboratory during ca. 100 days. Figure 12.16 shows smoothed spectra for numbers of migrating crickets and locomotor activity, for both sexes.

Peaks corresponding to periods of ca. 2.5, 5, 10, and 20 days were observed in one or several spectra, which suggested a long-term biological rhythm with several harmonics. It followed from spectral analysis that the migratory waves could be explained by synchronization of the

locomotor activity cycles of individuals in the population. Migrations of the males appeared to follow a 20-day cycle, whereas those of females seemed to follow a cycle of ca. 10 days. The authors suggested that, during these periods, males attract females to their burrows and form relatively stable couples.

Ecological application 12.5c

Another ecological example, quite different from those presented above, is provided by the study of Logerwell *et al*. (1998) in the southeastern Bering Sea. There, the authors used spectral analysis to characterise the spatial aggregation patterns of thick-billed murres (*Uria lomvia*) (birds, family *Alcidae*), and their prey (e.g. juvenile fish and krill), whose biomass had been estimated by underwater acoustic surveying.

As a further example, Dutilleul (2011, his Sections 6.2.1 and 6.2.2) applied spectral analysis to time series of daily mean temperatures in air and soil in the Gault Nature Reserve (Québec) sampled over thirty days in June 2004 (his Fig. 6.3, d and h).

2 — Multidimensional series

Spectral analysis can be used not only with univariate but also with *multidimensional* series, when several ecological variables have been recorded simultaneously. This analysis is an extension of *cross-covariance* or *cross-correlation*, in the same way as the variance spectrum is a generalization of autocovariance or autocorrelation (Fig. 12.14).

Two-dimensional series

From two data series, $\mathbf{y}_j$ and $\mathbf{y}_l$, one can compute a pair of smoothed spectra S_{jj} and S_{ll} and a cross-correlation function $r_{jl}(k)$. These are used to define the *co-spectrum* (K_{jl}) and the *quadrature spectrum* (Q_{jl}):

Co-spectrum

$$K_{jl}(f) = \text{Fourier transform of } [r_{jl}(k) + r_{jl}(-k)]/2 \qquad \textbf{(12.22)}$$

Quadrature s.

$$Q_{jl}(f) = \text{Fourier transform of } [r_{jl}(k) - r_{jl}(-k)]/2 \qquad \textbf{(12.23)}$$

The *co-spectrum* (eq. 12.22) measures the distribution, as a function of frequencies, of the covariance between those components of the two series that are in phase, whereas the *quadrature spectrum* (eq. 12.23) provides corresponding information for a phase shift of 90° between the same components. For example, a sine and cosine function are in perfect quadrature. These spectra are used, below, to compute the *coherence*, *phase*, and *gain*.

The *cross-amplitude spectrum* is defined as:

Cross-amplitude spectrum

$$\sqrt{K_{jl}^2(f) + Q_{jl}^2(f)} \qquad \textbf{(12.24)}$$

The spectra for $\mathbf{y}_j$ and $\mathbf{y}_l$ are used to compute the (squared) *coherence spectrum* (C_{jl}) and the *phase spectrum* (Φ_{jl}):

Coherence spectrum

$$C_{jl}^2(f) = \frac{K_{jl}^2(f) + Q_{jl}^2(f)}{S_{jj}(f)\,S_{ll}(f)} \quad \textbf{(12.25)}$$

Phase spectrum

$$\Phi_{jl}(f) = \arctan\left(\frac{-Q_{jl}(f)}{K_{jl}(f)}\right) \quad \textbf{(12.26)}$$

The *squared coherence* (eq. 12.25) is a dimensionless measure of the correlation of the two series in the frequency domain; for frequency f, $C_{jl}^2(f) = 1$ indicates perfect correlation between two series whereas $C_{jl}^2(f) = 0$ implies the opposite. The *phase spectrum* (eq. 12.26) shows the phase shift between the two series. When the phase is a regular function of the frequency, the squared coherence is usually significantly different from zero; when the phase is very irregular, the squared coherence is generally low and not significant.

Gain spectrum

In order to assess the causal relationships between two variables, one can use the *gain spectrum* (R_{jl}^2), which is analogous to a coefficient of simple linear regression. One can determine the response of $\mathbf{y}_j$ to $\mathbf{y}_l$:

$$R_{jl}^2(f) = \frac{S_{jj}(f)\,C_{jl}^2(f)}{S_{ll}(f)} \quad \textbf{(12.27)}$$

or, alternatively, the response of $\mathbf{y}_l$ to $\mathbf{y}_j$:

$$R_{lj}^2(f) = \frac{S_{ll}(f)\,C_{jl}^2(f)}{S_{jj}(f)} \quad \textbf{(12.28)}$$

Ecological application 12.5d

In a study of the spatial variability of coastal marine phytoplankton, Platt *et al*. (1970) repeated, in 1969, the sampling programme of 1968 described in Ecological application 12.3a. This time, data were collected not only on chlorophyll *a* but also on temperature and salinity at 80 sites along a transect. Figure 12.17 shows the coherence spectra for the three pairs of series, recorded on 24 June. Strong coherence between temperature and salinity indicates that these variables well characterized the water masses encountered along the transect. Significant coherence between the series of chlorophyll *a* and those of temperature and salinity, at ca. 3 cycles (naut. mi.)$^{-1}$, were consistent with the hypothesis that the spatial distribution of phytoplankton was controlled to some extent by the physical structure of the environment.

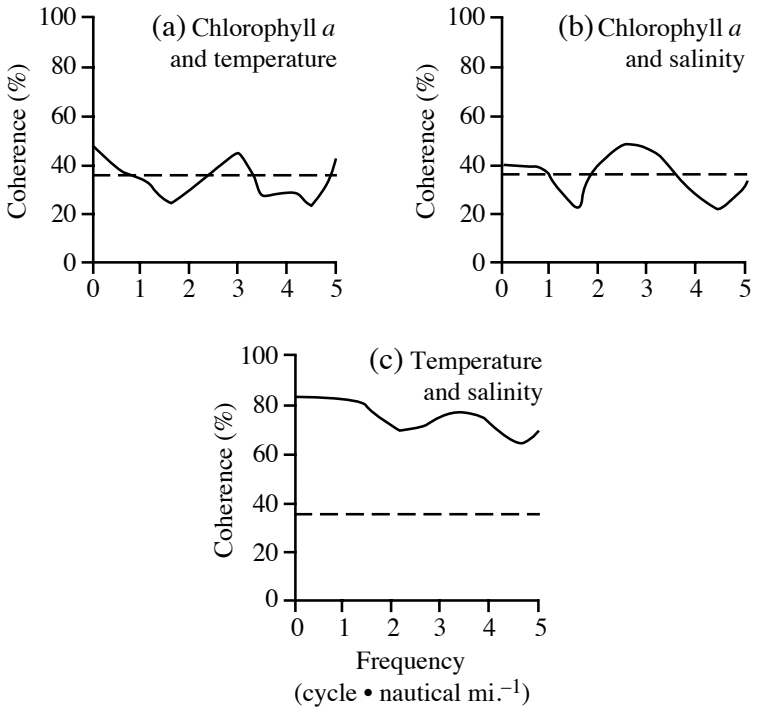


Figure 12.17 Coherence spectra between pairs of variables sampled along a transect 8 nautical miles long in St. Margaret's Bay (Nova Scotia, Canada). Dashed lines: approximate 95% confidence limits. After Platt *et al*. (1970).

Ecological application 12.5e

In order to identify the factors controlling the spatial heterogeneity of marine phytoplankton (patchiness), Denman & Platt (1975) analysed values of chlorophyll *a* and temperature, recorded continuously along a transect in the St. Lawrence Estuary. Two pumping systems were towed, at depths of 5 and 9 m, over a distance of 16.6 km (10 nautical miles). The sampling interval was 1 s, which corresponds to 3.2 m given the speed of the ship. After detrending, computations were carried out using the fast Fourier transform. Four coherence and phase spectra were calculated, as shown in Fig. 12.18.

For a given depth (Fig. 12.18a: 5 m; b: 9 m), the coherence between temperature and chlorophyll *a* was high at low frequencies and the phase was relatively constant. At higher frequencies, the coherence decreased rapidly and the phase varied randomly. The lower panels of Fig. 12.18 indicate the absence of covariation between series from different depths. The authors concluded that physical processes played a major role in the creation and control of phytoplankton heterogeneity at intermediate scales (i.e. from 50 m to several kilometres). Weak coherence between series from the two depths, which were separated by a vertical distance of only 4 m, suggested the presence of a strong vertical gradient in the physical structure. Such

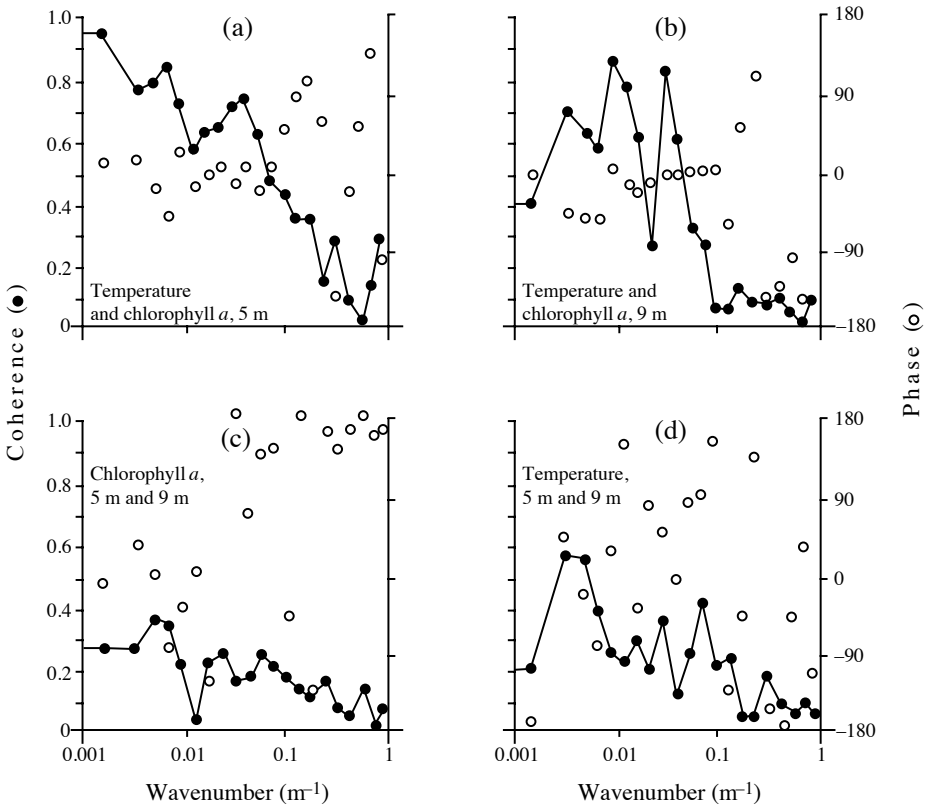


Figure 12.18 Values of coherence (solid lines) and phase (open circles), for pairs of spatial series continuously recorded in the St. Lawrence Estuary. Abscissa: *wavenumber* (= 2π/wavelength = 2π frequency/speed). Adapted from Denman & Platt (1975).

gradients are known to favour the propagation of internal waves (analogous to the propagation of waves at the air-water discontinuity). The authors proposed that the strong coherence between temperature and chlorophyll *a*, at each of the sampled depths, could reflect the presence of internal waves.

Ecological application 12.5f

In the study on the spatial distributions of thick-billed murres (*Uria lomvia*) and their prey (acoustic data) in the southeastern Bering Sea, described in Ecological application 12.5c, Logerwell *et al*. (1998) also used phase and coherence spectra. With these spectra, the authors compared the distribution patterns of birds and prey over a wide range of spatial scales.

Multivariate spectral analysis
Frequency regression

In the last paragraphs, the approach to multidimensional situations was to consider two series at a time. Brillinger (1981) provides the mathematical bases for processing multidimensional series using methods that are fully multivariate. When a stochastic series is a time-invariant function of several other series, the method recommended is *frequency regression*. It is analogous to multiple linear regression (Subsection 10.3.3), computed in the frequency domain. More generally, the method to study relationships among several series is that of *principal components in the frequency domain* (see Ecological application 12.5g). In that case, a spectrum is computed for each of the principal components, which are linear combinations of the serial variables (Section 9.1). The method has been adapted by Laurec (1979), who explained how to use it in ecology.

Another approach to the analysis of multivariate data series is the *Mantel correlogram* (Subsection 13.1.6). Because this type of correlogram is based upon a similarity or distance matrix among observations (Chapter 7), it is suitable to analyse multivariate data. It can also be used to analyse univariate or multivariate series of *semiquantitative*, *qualitative*, or *binary* data, like the species presence-absence data often collected by ecologists. Yet another approach is spatial eigenfunction analysis (Chapter 14). The study of a sediment core representing 10000 years of sedimentation (101 levels, 139 diatom species) by Legendre & Birks (2012), reported near the end of Subsection 14.1.3, is an example of analysis of a multivariate ecological series.

Ecological application 12.5g

Arfi *et al*. (1982, pp. 359-363) reported results from a study on the impact of the main sewage effluent of the city of Marseilles on coastal waters in the Western Mediterranean. During the study, 31 physical, chemical, and biological variables were observed simultaneously, at an anchor station 1 km offshore, every 25 min during 24 h ($n = 58$). Spectra for individual series (detrended) all showed a strong peak at T = ca. 6 h. Comparing the 31 data series two at a time would not have made sense because this would have required $(31 \times 30)/2 = 465$ comparisons. Thus, the 31-dimensional data series was subjected to principal component analysis in the frequency domain. Figure 12.19 shows the 31 variables, plotted in the plane of the first two principal components (as in Fig. 9.5), for $T = 6$ h. The long arrows pointing towards the upper left-hand part of the graph corresponded to variables that were indicative of the effluent (e.g. dissolved nutrients, bacterial concentrations) whereas the long arrows pointing towards the lower right-hand part of the ordination plane corresponded to variables that indicated unperturbed marine waters (e.g. salinity, dissolved O_2, phytoplankton concentrations). The positions of the two groups of variables in the plane show that their variations were out of phase by ca. 180°, for period $T = 6$ h. This was interpreted as a periodic increase in the effluent every 6 h. This periodicity corresponded to the general activity rhythm of the adjacent human population (wake-up, lunch, end of work day, and bedtime).

3 — Maximum entropy spectral analysis

As explained in Subsection 12.5.1, estimating spectra requires the use of spectral or lag *windows*. Each type of window provides a compromise between reduction of the standard error of the spectral estimates and loss of resolution between adjacent

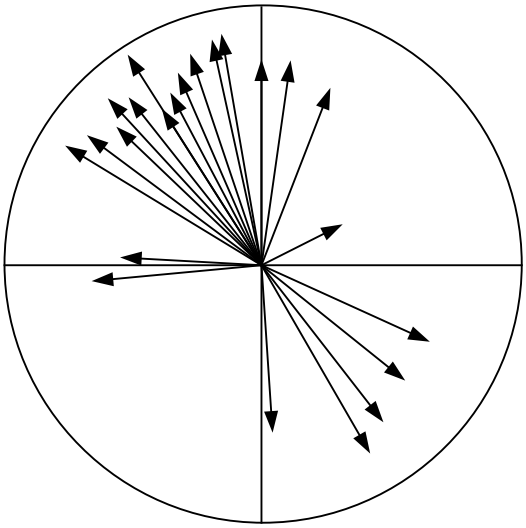


Figure 12.19 Principal component analysis in the frequency domain of 31 simultaneous series of physical, chemical, and biological variables, obtained at an anchor station in the Western Mediterranean. Plot of the 31 variables (arrows), in the plane of the first two principal components, for period $T = 6$ h. Adapted from Arfi *et al*. (1982).

frequencies. As an alternative to windows, Burg (1967) proposed to improve the spectral resolution by *extrapolating* the autocorrelation function beyond the maximum lag ($k_{\max}$), whose value is limited by the length of the series (Subsection 12.3.1). For each extrapolated lag ($k_{\max} + k$), he suggested to calculate an autocorrelation value $r_{yy}(k_{\max} + k)$ that *maximizes the entropy* (Chapter 6) of the probability distribution of the autocorrelation function. Burg's (1967) method will not be further discussed here, because a different algorithm (Bos, 1971; see below) is now used for computing this *maximum entropy spectral analysis* (MESA). Estimation of the spectrum, in MESA, does not require spectral or lag windows. An additional advantage, especially for ecologists, is that it allows the computation of spectra for very short series.

AR model

Data series may be mathematically described as stochastic linear processes. A corresponding mathematical model is the *autoregressive model* (also called *AR model* or *all-pole model*), where each observation in the series $\tilde{y}_t$ (centred on the mean $\bar{y}$ of the series: $\tilde{y}_t = y_t - \bar{y}$) is represented as a function of the q preceding observations:

$$\tilde{y}_t = \phi_1 \tilde{y}_{t-1} + \phi_2 \tilde{y}_{t-2} + \ldots + \phi_q \tilde{y}_{t-q} + a_t \qquad \textbf{(12.29)}$$

q specifies how many steps back one takes into account to forecast value $\tilde{y}_t$. This is called the *order* of the process. The *autoregression coefficients* ϕ are estimated using

White noise

the observations of the data series itself. Residual values a_t must be independent of one another; the series of residual values is called *white noise*. Their overall variance is noted s_a^2. This type of model will be further discussed in Section 12.7.

Concerning maximum entropy spectral analysis, Bos (1971) has shown that the maximum entropy method proposed by Burg (1967) is equivalent to a least-squares fitting of an AR model to the data series. Using the autoregression coefficients ϕ, it is possible to compute the same spectral densities as those resulting from the entropy calculation of Burg (1967). Thus, the spectrum is estimated directly from the autoregression coefficients ϕ of the AR model, which are themselves estimated from the values $\tilde{y}_t$ of the data series. The spectral density for each frequency f is:

$$S(f) = \frac{2s_a^2\Delta}{\left|1 - \sum_{j=1}^{q} \phi_j \exp(-2i\pi fj)\right|^2} \qquad \textbf{(12.30)}$$

where $i = \sqrt{-1}$. Generally, the sampling interval is $\Delta = 1$ time or space unit.

Maximum entropy spectral analysis is not entirely free of problems. Some of these are still the subject of specialized papers. A first practical problem is choosing the *order* q of the AR model for an empirical data series. Various criteria for determining q have been reviewed by Berryman (1978) and Arfi & Dumas (1990). Another problem concerns the estimation of the coefficients of the AR model (see, for instance, Ulrych & Clayton, 1976). A third problem, also discussed by Ulrych & Clayton (1976), is that other processes may fit the data series better than the AR model; for example, an autoregressive-moving average model (ARMA; Section 12.7). Fitting such models may, however, raise other practical problems. The criteria for deciding to use models other than AR are partly intuitive (Section 12.7).

Ulrych & Bishop (1975) briefly reviewed the theoretical bases underlying the algorithms of Burg (1967) and Bos (1971). Barrodale & Erikson (1980) propose another algorithm for estimating the coefficients ϕ of the AR model, based on least squares, which provides a more precise estimation of the spectrum frequencies. The same authors criticize, on an empirical basis, the method of Akaike, and they propose a different approach.

Maximum entropy spectral analysis can handle short series as well as series with data exhibiting measurement errors (Ables, 1974). It may also be used to analyse series with missing data (Ulrych & Clayton, 1976). Arfi & Dumas (1990) compared MESA to the classical Fourier approach, using simulated and real oceanographic data series. For long series ($n = 450$), the two approaches have the same efficiency when noise is low, but MESA is more efficient when noise is high. For short ($n = 49$ to 56) and very short ($n = 30$) series, MESA is systematically more efficient. For long data series with low noise, it may often be simpler to compute the spectrum in the traditional way (Berryman, 1978). However, for many ecological data series, MESA would be the

method of choice. The maximum entropy approach can be generalized to handle multivariate series, since coherence and phase spectra can be computed (Ulrych & Jensen, 1974).

Spectral analysis and, thus, Objective 3 of the analysis of data series (Table 12.2), are presently restricted to *quantitative data*. The only exception is the computation of spectra for long (i.e. $n > 500$ to 1000) series of *binary variables*, using the method of Kedem (1980). Since MESA is not very demanding as to the precision of the data, it could probably be used as well for analysing series of *semiquantitative data* coded using several states.

Ecological application 12.5h

Colebrook & Taylor (1984) analysed the temporal variations of phytoplankton and zooplankton series recorded monthly in the North Atlantic Ocean and in the North Sea during 33 consecutive years (1948 to 1980). Similar series were also available for some environmental variables (e.g. surface water temperature). The series were analysed using MESA. In addition, coherence spectra were computed between series of some physical variables and the series representing the first principal component calculated for the plankton data. For the plankton series, one spectrum was computed for each species in each of 12 regions, after which the spectra were averaged over the species in each region. The resulting 12 species-averaged spectra exhibited a number of characteristic periods, of which some could be related to periods in the physical environment using coherence spectra. For example, a 3 to 4-year periodicity in plankton abundances was associated to heat exchange phenomena at the sea surface. Other periods in the spectra of the physical and biological variables could not easily be explained. Actually, 33-year series are relatively short compared with the long-term meteorological or oceanographic variations, so that some of the identified periods may turn out not to be true cycles.

Ecological application 12.5i

Kim *et al*. (2003) measured the oxygen consumption rates of sublittoral-dwelling Washington clams (*Saxidomus purpuratus*) collected in southern South Korea. Using MESA, they evidenced two endogenous rhythms in clam respiration kept under constant conditions, i.e. during 7-9 days after collection. They found a rhythm that corresponded to the tides in their original environment, followed by a shift to a circadian rhythm.

4 — Wavelet analysis

Subsection 12.5.1 introduced the notion of pairs of mathematically equivalent equations that are called pairs of transforms, and applied it to Fourier transforms. It was then shown that the autocovariance or autocorrelation function (eqs. 12.5-12.7) and the spectral density function (eq. 12.21) are a pair of Fourier transforms. Another type of transform, called wavelet transform, provides a somewhat different approach to the analysis of data series, including ecological series. Although the wavelet transform can be regarded as a generalisation of the Fourier transform, the former may be better adapted to ecological data series than the latter (Cazelles *et al*., 2008). This is because Fourier analysis decomposes the signal into waveforms that have constant

amplitude along the time axis (i.e. the sines and cosines in Fig. 12.13), whereas wavelet analysis uses waveforms (wavelets) that are narrow when the features of the signal are high-frequency and occur over a short period along the time axis, and wide when these features are low-frequency and occur over a long period. In practice, the wavelet transform decomposes the signal over functions (called wavelets) that are narrow in the portions of the data series presenting high-frequency features, and wide where structures in the data series are of low frequency.

Section 12.3 explained that a basic assumption of the correlation-based techniques used in series analysis is stationarity, i.e. the statistical parameters of a stationary time series are constant along the time axis. However, many ecological processes violate the stationarity assumption, including population dynamics (e.g. Cazelles & Hales, 2006). As explained by various authors including Cazelles *et al*. (2008), wavelet analysis overcomes the problems of non-stationarity in time series by performing local time-scale decomposition of the signal, i.e. it estimates different spectral characteristics along the time axis. As in the case of Fourier analysis for multidimensional series (Subsection 12.5.2), it is possible to investigate relationships between two signals using wavelet cross-spectrum and coherence.

In practice, wavelet analysis is only useful to analyse univariate, regular data without gaps. For one-dimensional time series or spatial transects, the data set must be fairly large, the time interval between neighbouring observations (i.e. the lag) must be small, and the series must be long compared to the structures to be extracted. In the context of spatial analysis (Chapter 13), wavelets can be used for the analysis of two-dimensional data on a grid, e.g. remotely sensed data, or forest plots that have been entirely studied; see note in Subsection 6.5.3 about the CTFS permanent forest plots and Ecological application 14.1b where data from one of those plots are analysed.

Basic principles of wavelet analysis, and applications to both artificial data series and real ecological time series, are found in Cazelles *et al*. (2008). In that paper, the authors analyse real ecological time series describing fluctuations in populations of red grouse in Scotland over 100 years, and the association between sunspot numbers and populations of lynx and porcupine over almost 200 years.

Fortin & Dale (2005, their Section 2.6.6) provide a short introduction to wavelet analysis. Readers may refer to Dale & Mah (1998), Percival & Walden (2000), and Keitt & Urban (2005) for more in-depth introductions to this type of analysis. Analysis of ecological time series with the wavelet approach offers a new perspective for the treatment of univariate data series that do not meet the stationarity assumption. For that reason, the number of publications reporting analyses of this kind is rapidly growing.

12.6 Detection of discontinuities in multivariate series

Ecological succession

Detection of discontinuities in *multivariate data series* is a problem familiar to ecologists (Objective 4 of Section 12.1 and Table 12.2). For example, studies on changes in species assemblages over time often refer to the concept of *ecological succession*. According to Margalef (1968), the theory of species succession within ecosystems plays the same role in ecology as evolutionary theory does in general biology.

The simplest way to approach the identification of discontinuities in multivariate series is by *visual inspection* of the curves depicting changes with time (or along a spatial direction) in the abundance of the various taxa or/and in the values of the environmental variables. In most instances, however, simple visual examination of a set of graphs does not allow one to unambiguously identify discontinuities in multivariate series. Numerical techniques must be used.

Methods of series analysis described in Sections 12.3 to 12.5 are not appropriate for detecting discontinuities in multivariate series, because the presence of discontinuities is not the same as periodicity in the data. Four types of methods are summarized here.

Instead of dividing multivariate series into subsets, Orlóci (1981) proposed a multivariate method for identifying successional trends and separating them into monotonic and cyclic components. That method may be viewed as complementary to those described below.

1 — Ordinations in reduced space

Several authors have used *ordinations in reduced space* (Chapter 9) to represent multispecies time series in low-dimensional space. To help identify the discontinuities, successive observations of the time series are connected with lines, as in Figs. 9.20 and 12.24. When several observations corresponding to a bloc of time are found in a small part of the reduced space, they may be thought of as a "step" in the succession. Large jumps in the two-dimensional ordination space are interpreted as discontinuities. This approach has been used, for example, by Williams *et al*. (1969; vegetation, principal coordinates), Levings (1975; benthos, principal coordinates), Legendre *et al*. (1984a; benthos, principal components), Dessier & Laurec (1978; zooplankton, principal components and correspondence analysis), and Sprules (1980; nonmetric multidimensional scaling; zooplankton; Ecological application 9.4a). In studies of annual succession in temperate or polar regions, using ordination in reduced space, one expects the observations to form some kind of a circle in the plane of the first two axes, since successive observations are likely to be close to each other in the multidimensional space, due to climate forcing (temporal correlation, Section 1.1), and the community structure is expected to come back to its original structure after one year; the rationale for this null model of succession is developed in Legendre *et al*.

(1985, Appendix D). Departures from a regular circular pattern are thus interpreted as evidence for the existence of subsets in the data series. In simple situations, such subsets are indeed observed in the plane of the first two ordination axes (e.g. Figs. 9.20 and 12.24). When used *alone*, however, this approach has two major drawbacks.

- Plotting a multivariate data series in two or three dimensions only is not the best way of using the multivariate information. In most studies, the first two principal axes used to represent the data series account together for only 10 to 50% of the multivariate information. In such cases, distances from the main clusters of observations to isolated objects (which are in some particular way different from the major groups) are likely to be expressed by some minor principal axes that are orthogonal (i.e. perpendicular in multidimensional space) to the main projection plane. As a consequence, these objects may be projected, in the reduced-spaced ordination, within a group from which they are actually quite different. Moreover, it has been observed that the "circle" of observations (see previous paragraph) may be deformed in a spoon shape so that groups that are distinct in a third or higher dimension may be packed together in some part of the two-dimensional ordination plane. These problems are common to all ordinations when used alone for the purpose of group recognition. They are not as severe for ordinations obtained by nonmetric multidimensional scaling, however, because that method is, by definition, more efficient than others at flattening multidimensional phenomena into a user-determined small number of dimensions (Section 9.4). The best way to eliminate this first drawback is to associate ordination to clustering results, as explained in Section 10.1. This was the approach of Allen *et al.* (1977) in a study of the phytoplankton succession in Lake Wingra. See also Fig. 12.24.

- The second drawback is the lack of a criterion for assigning observations to groups in an ordination diagram. As a consequence, groups delineated on published ordination diagrams often look rather arbitrary.

2 — Segmenting data series

Hawkins & Merriam (1973, 1974) proposed a method for segmenting a multivariate data series into homogeneous units, by *minimizing the variability* within segments in the same way as in *K*-means partitioning (Section 8.8). Their work followed from the introduction of a contiguity constraint in the grouping of data by Fisher (1958), who called it *restriction* in space or time. The method of Hawkins & Merriam has been advocated by Ibanez (1984) for studying successional steps.

Contiguity constraint

The method has three interesting properties. (a) The multidimensional series is partitioned into homogeneous groups using an *objective clustering criterion*. (b) The partitioning is done with a *constraint of contiguity* along the data series. Within the context of series analysis, contiguity means that only observations that are neighbours along the series may be grouped together. The notion of contiguity has been used by several authors to resolve specific clustering problems: temporal contiguity (Subsection 12.6.5, below) or spatial contiguity (Subsection 13.3.2). (c) The observations do not have to be equispaced.

A first problem with Hawkins & Merriam's method is that users must determine the number of segments that the method is requested to identify. To do so, the increase in explained variation relative to the increase in the number of segments is used as a guide. Any one of the stopping rules used with *K*-means partitioning could also be used here (end of Section 8.8). A solution to this problem is described in Subsection 12.6.4. For community composition data, a second problem is that strings of zeros in multispecies series are likely to result in segments that are determined by the simultaneous absence of species. That problem can be resolved by transforming the species data using one of the transformations described in Section 7.7.

3 — Webster's method

Webster (1973) proposed a rather simple method to detect discontinuities in data series. He was actually working with spatial transects, but his method is equally applicable to time series. Draw the sampling axis as a line and imagine a window that travels along that line, stopping at the mid-points between adjacent observations (if these are equispaced). Divide the window in two equal parts (Fig. 12.20a). There are observations in the left-hand and right-hand halves of the window. Calculate the difference (see below) between the points located in the left-hand and right-hand halves and plot these differences in a graph, as the window is moved from one end of the series to the other (Fig. 12.20c, d). The principle of the method is that the difference should be large at points where the left-hand and right-hand halves of the window contain values that are appreciably different, i.e. where discontinuities occur in the series. The following statistics may be used in the computations:

Window

- For univariate data, calculate the absolute value of the difference between the means of the values in the left-hand and right-hand halves of the window: $\text{Statistic} = \left|\bar{x}_1 - \bar{x}_2\right|$.

- For univariate data again, one may choose to compute the absolute value of a *t*–statistic comparing the two halves of the window: $\text{Statistic} = \left|\bar{x}_1 - \bar{x}_2\right| / s_{\bar{x}_1 - \bar{x}_2}$. If one assumes second-order stationarity of the series and uses the standard deviation of the whole series as the best estimate of the standard deviations in the two halves, this statistic is linearly related to the previous one. Alternatively, one could use the regular *t*-statistic formula for *t*-tests, estimating the variance in each window from the few values that it contains; this is not recommended as it produces values of the *t*-statistic that cannot be compared, and unstable estimates when windows are narrow, which is often the case with this method.

- For multivariate series, compare the two halves of the window using either the Mahalanobis generalized distance (D_5 or D_5^2, eq. 7.39), which is the multivariate equivalent of a *t*-statistic, or the coefficient of racial likeness (D_{12}, eq. 7.52).

The width of the window is an empirical decision made by the investigator. It is recommended to try different window widths and compare the results. The window width is limited, of course, by the spacing of observations, considering the

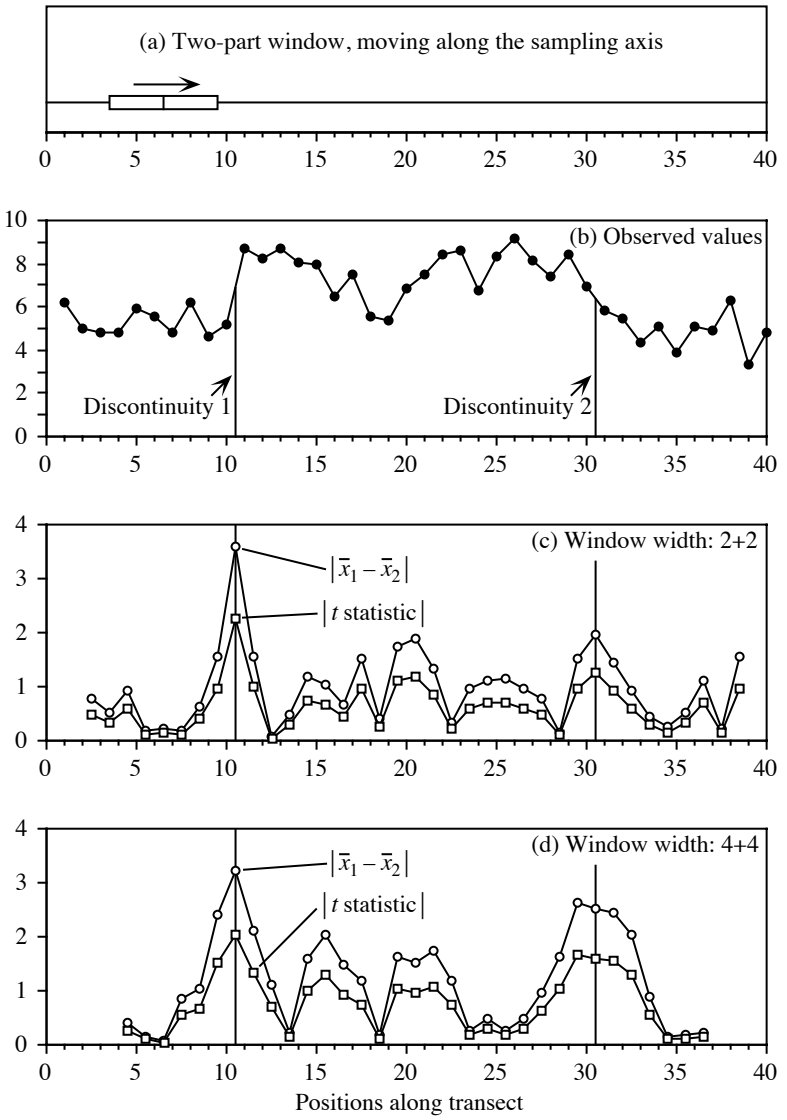


Figure 12.20 Webster's method for detecting discontinuities in data series. (a) Principle of the method. (b) Numerical example (see text). Results using a window that was (c) 4 observations wide, or (d) 8 observations wide.

approximate interval between the expected discontinuities. Webster's method works best with equispaced observations, but some departure from equal spacing, or missing data points, are allowed, because of the empirical nature of the method.

Numerical example. A series of 40 observations was generated using a normal pseudo-random number generator $N(5,1)$. The values of observations 11 to 30 were increased by adding 3 to the generated values in order to artificially create discontinuities between observations 10 and 11, on the one hand, and observations 30 and 31, on the other. It so happened that the first of these discontinuities was sharp whereas the second was rather smooth (Fig. 12.20b).

Webster's method for univariate data series was used with two window widths. The first window had a width of 4 observations, i.e. 2 observations in each half; the second window had a width of 8 observations, i.e. 4 in each half. Both the absolute values of the differences between means and the absolute values of the *t*-statistics were computed. The overall standard deviation of the series was used as the denominator of *t*, so that this statistic was a linear transformation of the difference-between-means statistic. Results (Fig. 12.20c, d) are reported at the positions occupied by the centre of the window.

The sharp discontinuity between observations 10 and 11 was clearly identified by the two statistics and window widths. This was not the case for the second discontinuity, between observations 30 and 31. The narrow window (Fig. 12.20c) estimated its position correctly, but did not allow one to distinguish it from other fluctuations in the series, found between observations 20 and 21 for instance (remember, observations are randomly-generated numbers; so there is no structure in this part of the series). The wider window (Fig. 12.20d) brought out the second discontinuity more clearly (higher values of the statistics), but its exact position was no longer estimated precisely.

D_5^2 to the centroid

Window

Ibanez (1981) proposed a related method to detect discontinuities in multivariate records (e.g. simultaneous records of temperature, salinity, *in vivo* fluorescence, etc. in aquatic environments). He called the method D_5^2 *to the centroid*. For every sampling site, the method computes a generalized distance D_5^2 (eq. 7.39) between the new multivariate observation and the centroid (i.e. multidimensional mean) of the m previously recorded observations, m defining the width of a window. Using simulated and real multivariate data series, Ibanez showed that changes in D_5^2 to the centroid, drawn on a graph like Figs. 12.20c or d, allowed one to detect discontinuities. For multi-species data, however, the method of Ibanez suffers from the same drawback as the segmentation method of Hawkins & Merriam: since the simultaneous absence of species is taken as an indication of similarity, it could prevent changes occurring in the frequencies of other species from producing high, detectable distances. That problem can be resolved by transforming the community composition data, prior to the analysis, using one of the transformations described in Section 7.7.

McCoy *et al*. (1986) proposed a segmentation method somewhat similar to that of Webster, for species occurrence data along a transect. A matrix of Raup & Crick similarities is first computed among sites (S_{27}, eq. 7.31) from the species presence-absence data. A "+" sign is attached to a similarity found to be significant in the upper tail (i.e. when a_{hi} is significantly larger than expected under the random sprinkling hypothesis) and a "–" sign to a similarity that is significant in the lower tail (i.e. when a_{hi} is significantly smaller than expected under that null hypothesis). The number of significant pluses and minuses is analysed graphically, using a rather complex empirical method, to identify the most informative boundaries in the series.

4 — Time-constrained clustering by MRT

Multivariate regression tree analysis (MRT, Section 8.11) can be used as a form of time-constrained clustering. The solution consists in analysing a multivariate response matrix **Y** using a quantitative or rank-ordered variable **x** representing the sampling sequence through time. **Y** may contain community composition data transformed in some appropriate way (Section 7.7). For a weekly time series over a year, for example, the constraining variable **x** may be a vector containing the sampling dates, counted from January 1st, or the numbers 1 to 52; the results will be identical since MRT segments **Y** at cutting points along the explanatory, or constraining, variable **x**. The observations do not have to be equispaced.

MRT is a least-squares algorithm. In the present application, it segments **Y** in such a way that the sum of the within-group multivariate sums of squares is minimum, with the constraint that the sampling dates within each group be adjacent along the sampling sequence. As a consequence, the solution obeys the Hawkins & Merriam criterion described in Subsection 12.6.2. As a bonus, the cross-validation procedure available in MRT helps determine the 'best' number of groups for the data under study; this solves the first problem of the Hawkins & Merriam method mentioned in Subsection 12.6.2. MRT can be used to segment spatial series, e.g. transect data as shown in the following ecological application, as well as time series.

Ecological application 12.6a

Borcard *et al*. (2011, their Section 4.11) used MRT to segment fish assemblage data collected at 29 sites along the Doubs River in eastern France (29 sites, 27 species) by space-constrained clustering along the course of the river. These data were also used in Ecological application 11.1a. In the present application, the data were chord-transformed (eq. 7.67) before MRT analysis. Cross-validation in MRT suggested 5 groups as the best solution; that solution had the smallest CVRE value (eq. 8.23). The five groups are represented on a map of the river in Fig. 12.21. The calculations were done with the R code provided by Borcard *et al*. (2011).

5 — Chronological clustering

Temporal contiguity

Combining some of the best aspects of the methods described above, Gordon & Birks (1972, 1974) and Gordon (1973) introduced a constraint of temporal contiguity in a variety of clustering algorithms to study pollen stratigraphy. Analysing bird surveys repeated at different times during the breeding season, North (1977) also used a constraint of temporal contiguity to cluster bird presence locations on a geographic map and delineate territories. Applications of time-constrained clustering to palaeoecological data (where a spatial arrangement of the observations corresponds to a time sequence) can be found in Bell & Legendre (1987), Hann *et al*. (1994) and Song *et al*. (1996). Algorithmic aspects of constrained clustering are discussed in Subsection 13.3.2.

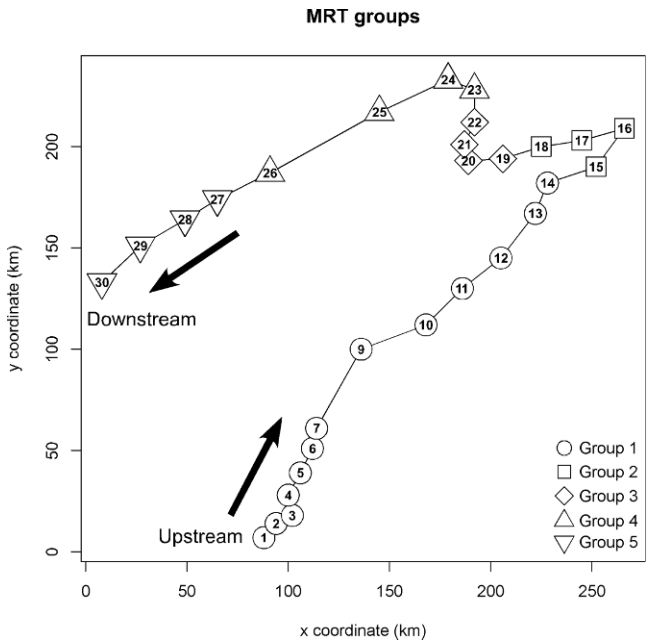


Figure 12.21 Clustering with spatial contiguity constraint of the Doubs River fish assemblage data by multivariate regression tree analysis (MRT). The groups of sites are represented on a map along the course of the river; the arrows indicate flow direction. Sites are numbered 1 to 30; site 8 was removed from the analysis for the reason explained in Ecological application 11.1a.

Succession model

Using the same concept, Legendre *et al*. (1985) developed the method of *chronological clustering*, based on hierarchical clustering (Chapter 8). The algorithm was designed to identify discontinuities in multi-species time series. It has also been successfully used to analyse spatial transects (e.g. Galzin & Legendre, 1987; Ardisson *et al*., 1990; Tuomisto & Ruokolainen, 1994: Ecological application 12.6c; Tuomisto *et al*., 2003). When applied to *ecological succession*, chronological clustering corresponds to a well-defined *model*, in which succession proceeds by steps and the transitions between steps are rapid (see also Allen *et al*., 1977, on this topic). Broad-scale successional steps contain finer-scale steps, which may be identified using a finer analysis if finer-scale data are available. Chronological clustering takes into account the sampling sequence, imposing a constraint of temporal contiguity to the clustering activity.

The method also permits the elimination of *singletons* (in the game of *bridge*, a singleton is a card that is the only one of a suit in the hand of a player). Such singular observations often occur in ecological series. In nature, singletons are the result of

D	Group 1 5	6	Group 2 7	8	9
5	0				
6	0.2	0			(a)
7	0.4	0.7	0		
8	0.6	0.5	0.1	0	
9	0.7	0.8	0.3	0.6	0

D	Group 1 5	6	Group 2 7	8	9
5					
6	0				(b)
7	0	1			
8	1	0	0		
9	1	1	0	1	

Figure 12.22 Numerical example. (a) Distance matrix for two contiguous groups from a multidimensional time series (used also in Fig. 10.23). The lower half of the symmetric matrix is shown. (b) 50% of the distances, i.e. those with the highest values, are coded 1; the others are coded 0.

random fluctuations, migrations, or local changes in external forcing. In an aquatic system studied at a fixed location (Eulerian approach, Section 12.0), such changes may be due to temporary movements of water masses. Singletons may also result from improper sampling or inadequate preservation of specimens.

Hierarchical agglomerative clustering (Section 8.5) proceeds from an association matrix ($n \times n$) among the observations of the data series (length n), computed using an appropriately chosen similarity or distance coefficient (Chapter 7). Any method of agglomerative clustering may be used; Legendre *et al.* (1985) used intermediate linkage clustering (Subsection 8.5.3). The clustering algorithm is modified to include the contiguity constraint; Fig. 13.25 shows how a constraint of spatial or temporal contiguity can be introduced into any agglomerative clustering algorithm. Each clustering step is subjected to a permutation test (Subsection 1.2.2) before the fusion of two objects or groups is authorized.

Permutation test

Consider two adjacent groups of objects pertaining to some data series (Fig. 12.22). The first group ($n_1 = 2$) includes objects 5 and 6 and the second ($n_2 = 3$) contains objects 7, 8 and 9. Assume that an agglomerative clustering algorithm now proposes that these two groups are the next pair to join. Distances among the five objects are given in Fig. 12.22a. Before applying the permutation test of cluster fusion, the distances are divided in two groups: the 50% of the distances (5 in this example) that have the highest values are called "high distances" and are coded 1 (Fig. 12.22b) whereas the other 50% are called "low distances" and are coded 0. The test statistic is the number of high distances (h) in the between-group matrix (shaded area); $h = 4$ in this example. Under the null hypothesis, the objects in the two groups are drawn from the same statistical population and, consequently, it is only an artefact of the agglomerative clustering algorithm that they temporarily form two groups. If the null hypothesis is true, the number of high distances ($h = 4$) presently found in the between-group matrix should be comparable to that found among all possible

permutations of the five objects in two groups with $n_1 = 2$ and $n_2 = 3$ objects. If the null hypothesis is false and the two groups come from different statistical populations (i.e. different steps of the succession), the number of high distances presently found in the between-group matrix should be *higher* than most of the values found after permutation of the objects into two groups with $n_1 = 2$ and $n_2 = 3$ objects. This calls for a one-tailed test. After setting a significance level α, the permutations are performed and results that are higher than or equal to h are counted. The number of distinguishable combinations of the objects in two groups of sizes n_1 and n_2 is $(n_1 + n_2)!/(n_1!\, n_2!)$. If this number is not too large, all possible permutations can be examined; otherwise, permutations may be selected at random to form the reference distribution for significance testing. The number of permutations producing a result as large as or larger than h, divided by the number of permutations performed, gives an estimate of the probability p of observing the data under the null hypothesis.

- If $p > \alpha$, the null hypothesis is not rejected and the two groups are fused.
- If $p \leq \alpha$, the null hypothesis is rejected and fusion of the groups is prevented.

This test may actually be reformulated as a Mantel test (Section 10.5.1) between the matrix of recoded distances (Fig. 12.22b) and another matrix of the same size containing 1's in the among-group rectangle and 0's elsewhere.

Internal validation criterion

The above is not a proper test of significance because the alternative hypothesis (H_1: the two groups actually found by the clustering method differ) is not independent of the data that are used to perform the test; it comes from the data through the agglomerative clustering algorithm. So this is actually an internal validation clustering criterion (Section 8.13). Legendre *et al*. (1985) have shown, however, that this criterion has a correct probability of type I error; when testing on randomly generated data (Monte Carlo simulations) at significance level α, the null hypothesis was rejected in a proportion of the cases approximately equal to α.

Resolution

Significance level α used as the criterion for cluster fusion determines how easy it is to reject the null hypothesis. When α is small (close to 0), the null hypothesis is almost never rejected and only the sharpest discontinuities in the time or space series are identified. Increasing the value of α actually makes it easier to reject the null hypothesis, so that more groups are formed; the resulting groups are thus smaller and bring out more discontinuities in the data series. So, changing the value of α actually changes the resolution of the clustering results.

Singleton

A singleton is defined as a single observation whose fusion has been rejected with the groups located to its right and left in the series. When the test leads to the discovery of a singleton, it is temporarily removed from the series and the clustering procedure is started again from the beginning. This is done because the presence of a singleton can disturb the whole clustering geometry, as a result of the contiguity constraint.

The end result of chronological clustering is a *nonhierarchical partition* of the series into nonoverlapping homogeneous groups. Within the context of ecological succession, these groups correspond to the steps of a succession. *A posteriori* tests are used to assess the relationships between distant groups along the series as well as the origin of singletons. Plotting the clusters of observations onto an ordination diagram in reduced space may help in the overall interpretation of the results.

Legendre (1987b) showed that time-constrained clustering possesses some interesting properties. On the one hand, applying a constraint of spatial or temporal contiguity to an agglomerative clustering procedure forces different clustering methods to produce approximately the same results; without the constraint, the methods may lead to very different clustering results (Chapter 8), except when the spatial or temporal structure of the data (patchiness, gradient: Section 13.0) is very strong. Using autocorrelated simulated data series, he also showed that, if patches do exist in the data, constrained clustering always recovers a larger fraction of the structure than the unconstrained equivalent.

Constrained clustering along a time or spatial sampling axis can also be done by a more general form of constrained hierarchical clustering described in Subsection 13.3.2; see function ***constrained.clust()*** in Section 12.8.

Ecological application 12.6b

In May 1977, the Société d'Énergie de la Baie James impounded a small reservoir (ca. 7 km^2), called Desaulniers, in Northern Québec (77°32' W, 53°36' N). Ecological changes occurring during the operation were carefully monitored in order to use them to forecast the changes that would take place upon impoundment of much larger hydroelectric reservoirs in the same region. Several sampling sites were visited before and after the flooding. Effects of flooding on the zooplankton community of the deepest site (max. depth: 13 m), located ca. 800 m from the dam, were studied by Legendre *et al*. (1985) using chronological clustering. Before flooding, the site was located in a riverbed and only zooplankton drifting from lakes located upstream was found there (i.e. there was no zooplankton community indigenous to the river). Changes observed are thus an example of primary succession.

After logarithmic normalization of the data (eq. 1.14), the Canberra metric (D_{10}, eq. 7.49) was used to compute distances among all pairs of the 47 observations. Homogeneous groups of observations were identified along the data series, using a time-constrained algorithm for intermediate linkage clustering (Subsection 8.5.3) and the permutation test of cluster fusion described above. Results of chronological clustering are shown in Fig. 12.23 for different levels of resolution α. Plotting the groups of observations from Fig. 12.23, for $\alpha = 0.25$, on an ordination diagram obtained by nonmetric multidimensional scaling (Fig. 12.24), led to the following conclusions concerning changes in the zooplankton community. In 1976, as mentioned above, zooplankton was drifting randomly from small lakes located upstream. This was evidenced by low species numbers and highly fluctuating evenness (eq. 6.45), which indicated that no stable community was present. After impoundment of the reservoir, the community departed rapidly from the river status (Fig. 12.24) and formed a fairly well-developed assemblage, with 13 to 20 species in the summer of 1977, despite large chemical and water-level fluctuations. After the autumn overturn and during the 1977-1978 winter period, the

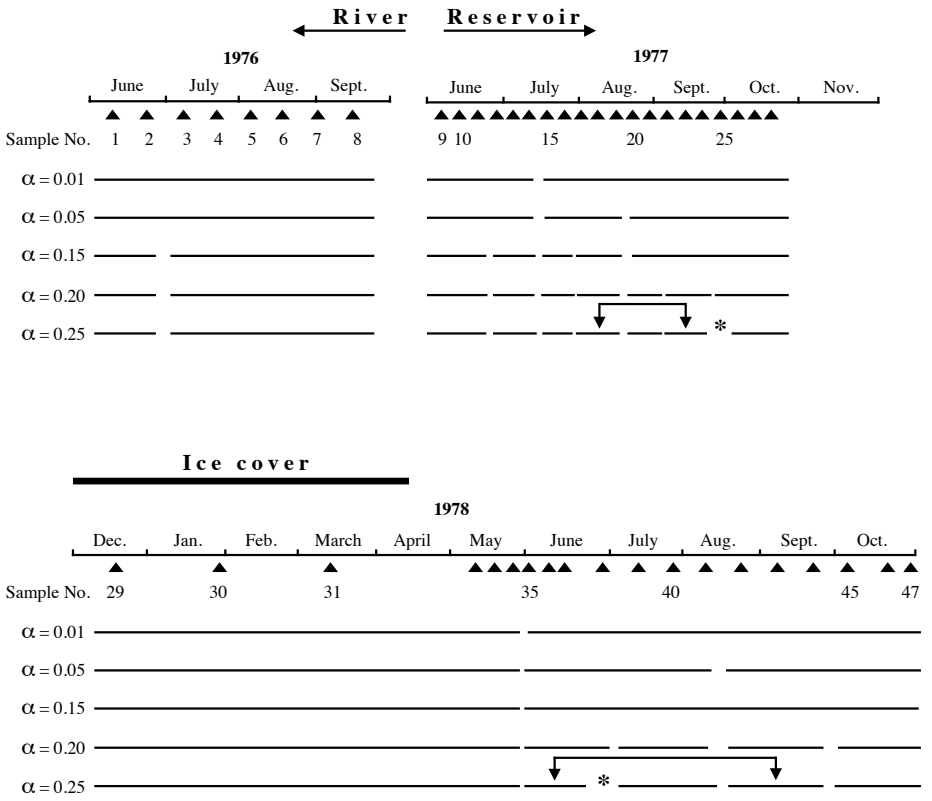


Figure 12.23 Chronological clustering: zooplankton time series. Results for different levels of resolution (α). For α = 0.25, the double arrows identify *a posteriori* tests with probabilities of fusion larger than α. Asterisks (*) identify singletons. Modified from Legendre *et al*. (1985).

community moved away from the previous summer's status. When spring came (observation 35), the community had reached a zone of the multidimensional scaling plane quite distinct from that occupied in summer 1977. Zooplankton was then completely dominated by rotifers, which increased from 70 to 87% in numbers and from 18 to 23% in biomass between 1977 and 1978, with a corresponding decrease in crustaceans, while the physical and chemical conditions had stabilized (Pinel-Alloul *et al*., 1982). When the succession was interrupted by the 1978 autumn overturn, the last group in the series (observations 45-47) was found (Fig. 12.23) near the position of the previous winter's observations (29-34), indicating that the following year's observations might resemble the 1978 succession.

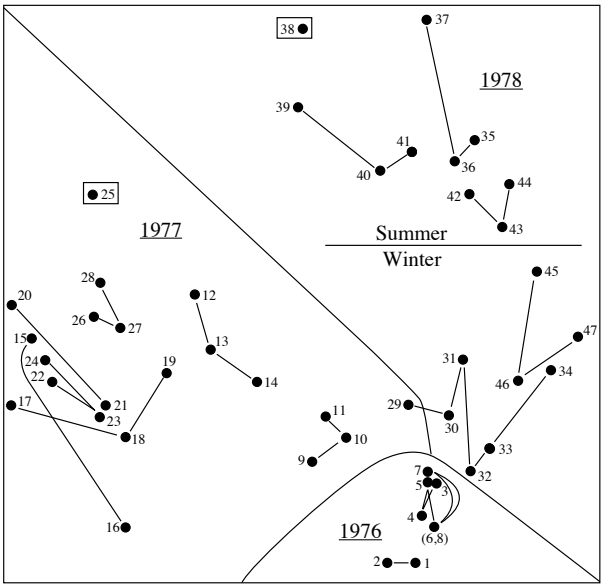


Figure 12.24 Chronological clustering: zooplankton time series. Nonmetric multidimensional scaling plot showing groups of observations from Fig. 12.23, for $\alpha = 0.25$. The groups are the sets of observations that are connected by lines materializing the sampling sequence. Objects in boxes are singletons. From Legendre *et al*. (1985). The regions of the graph delimited by envelopes correspond to sampling years.

Ecological application 12.6c

Tuomisto & Ruokolainen (1994) studied species assemblages of *Pteridophyta* (ferns; 40 species in the study) and *Melastomataceae* (a family of shrubs, vines, and small trees restricted to the American tropics; 22 species in the study) along two spatial transects (replicates) in a non-flooded area of the Amazonian rain forest in Peru, covering an edaphic (i.e. soil-related) and topographic gradient from clay soil on level ground, to quarzitic sand on a hill top. The two 700-m-long and 5-m-wide, parallel transects were 50 m apart. Chronological clustering was applied to the edaphic and floristic variables separately, using different similarity coefficients and three levels of resolution (parameter α). In all cases, the transects could be divided into distinct sections; the results of constrained clustering were more readily interpretable than the unconstrained equivalent. The groups of plants selected proved adequate for the rapid assessment of changes in the floristic composition of the rain forest.

Ecological application 12.6d

Tuomisto *et al*. (2003) studied the community structure of *Pteridophyta* (ferns) and *Melastomataceae*, the same groups as in Ecological application 12.6c, along a 43-km long transect in the Amazonian rain forest in Peru. They segmented the series of pteridophytes and

Melastomataceae data into groups using chronological clustering. They also used chronological clustering to partition a data series of spectral reflectance characteristics of the forest, extracted from a Landsat TM satellite image. The chronological clustering results were fairly consistent; the authors recognize eight groups of sites, which were also related to topography and soil characteristics. *Pteridophyta* and *Melastomataceae* indicator species of these groups of sites were then identified using the *INDVAL* index (Subsection 8.9.3). The results supported the hypothesis that species segregate edaphically at the landscape scale within the rain forest.

12.7 Box-Jenkins models

Forecasting

Objective 6 of time series analysis in ecology (Section 12.1) is to *forecast* future values. The Preface explained that ecological modelling is not, as such, within the scope of numerical ecology. In ecological studies, however, *Box-Jenkins modelling* is often conducted together with other forms of series analysis; this is why it is briefly presented here. This type of technique has already been mentioned in the context of maximum entropy spectral analysis (MESA, Section 12.5.3). The present section summarizes the principles that underlie the approach. Interested readers may refer to Box & Jenkins (1976), Cryer (1986), and Bowerman & O'Connell (1987) for the theory and to user's manuals of computer packages and R functions for actual implementation of the method.

MA model

Stochastic linear models (processes) described here are based on the idea that, in a series where data within a small window are strongly interrelated, the observed values are generated by a number of "shocks" a_t. These shocks are independent of each other and their distribution is purely random (mean zero and variance s_a^2). Such a series $(a_t, a_{t-1}, a_{t-2}, \ldots)$ is called *white noise*. In the *moving average* (*MA*) *model*, each observations in the series ($\tilde{y}_t = y_t - \bar{y}$, i.e. the data are centred on the mean $\bar{y}$ of the series) can be represented as a weighted sum of the values of process a:

$$\tilde{y}_t = a_t - (\theta_1 a_{t-1} + \theta_2 a_{t-2} + \ldots + \theta_q a_{t-q}) \qquad \textbf{(12.31)}$$

where θ are the weights and q is the *order* of the model. The name *moving average* for this process comes from the fact that eq. 12.31 is somewhat similar to that of the moving average (see the right-hand column of Table 12.4). The weights θ are estimated by numerical iteration, using techniques that are described in the above references and available in computer packages and R functions.

When the above model does not fit the series adequately (see below), another possibility is to represent an observation by a weighted sum of the q previous observations plus a random shock:

$$\tilde{y}_t = \phi_1 \tilde{y}_{t-1} + \phi_2 \tilde{y}_{t-2} + \ldots + \phi_q \tilde{y}_{t-q} + a_t$$

AR model

This is the *autoregressive model* (*AR*, or *all-pole* model) already described in eq. 12.29. In this model (of *order* q), q successive terms of the series are used to

forecast term $(q + 1)$, with error a_t. When estimating the autocorrelation coefficients ϕ by least squares, it is easy to compute residual errors $a_t = y_t - \tilde{y}_t$. Residual errors, as specified above for all Box-Jenkins models, must be independent of one another; this implies that a correlogram of the series of residuals a_t should display no significant value. The residuals must also be normally distributed.

ARMA model

Combining the above two models gives the *autoregressive-moving average model* (*ARMA model*), whose general form is:

$$\tilde{y}_t = \phi_1 \tilde{y}_{t-1} + \phi_2 \tilde{y}_{t-2} + \ldots + \phi_q \tilde{y}_{t-q} + a_t - (\theta_1 a_{t-1} + \theta_2 a_{t-2} + \ldots + \theta_q a_{t-q}) \quad \textbf{(12.32)}$$

An important advantage of ARMA models is that they can be fitted to data series using a small number of parameters (i.e. the coefficients ϕ and θ). However, such models may only be estimated for strictly *stationary* series (Sections 12.1 and 12.2).

One approach described in Section 12.2 for extracting the trend from a series is the *variate difference method*. In the computation, each value y_t is replaced by $(y_t - y_{t-T})$ where T is the period of the trend:

$$\tilde{y}_t = y_t - y_{t-T} \quad \textbf{(12.33)}$$

ARIMA model

Since $\tilde{y}_t$ results from a *difference*, y_t is called the *integrated form* of $\tilde{y}_t$. When an ARMA model is applied to a series of values computed with eq. 12.33, it is called an *autoregressive-integrated-moving average model* (*ARIMA model*).

Box-Jenkins analysis normally proceeds in four steps. (1) Identification of the *type of model* to be fitted to the data series (i.e. MA, AR, ARMA, or ARIMA). Even though Box & Jenkins (1976) described some statistical properties of the series (e.g. shape of the autocorrelation) that may guide this choice, identification of the proper model remains a somewhat intuitive step (e.g. Ibanez, 1982). (2) Estimation of the *parameters* of the model. For each case, various methods are generally available, so that one is confronted with a choice. (3) The *residuals* must be independent and normally distributed. If not, either the model is not adequate for the data series, or the parameters were not properly estimated. In such a case, step (2) can be repeated with a different method or, if this does not improve the residuals, a different model must be chosen at step (1). Steps (1) through (3) may be repeated as many times as necessary to obtain a good fit. The procedure of identification of the appropriate model is therefore iterative. (4) Using the model, values can be *forecasted* beyond the last observation.

It may happen that the data series is under external influences, so that the models described above cannot be used as such. For example, in the usual ARIMA model, the state of the series at time t is a function of the previous q observations ($\tilde{y}$) and of the random errors (a). In order to account for the additional effect of external variables, some computer programs allow the inclusion of a *transfer function* into the model (if the external forcing variable is also a random variable) and/or an *intervention component* (if the external variable is binary and not random). It is possible to extend

the forecasting to *multidimensional* data series. References to conduct the analysis are Whittle (1963) and Jones (1964).

It is important to remember that the models discussed here are *forecasting* and not *predictive* models. Indeed, the purpose of Box-Jenkins modelling is to *forecast* values of the series beyond the last observation, using the preceding data. Such forecasting is only valid as long as the environmental conditions that characterize the population under study (demographic rates, migrations, etc.) as well as the anthropogenic effects (exploitation methods, pollution, etc.) remain essentially the same. In order to *predict* with some certainty the fate of the series, causal relationships should be determined and modelled; for example, between the observed numbers of organisms, on the one hand, and the main environmental conditions, population characteristics, or/and anthropogenic factors, on the other. This requires extensive knowledge of the system under study. Forecasting models often prove quite useful in ecology, but one must be careful not to use them beyond their limits.

Ecological application 12.7

Boudreault *et al*. (1977) tried to forecast lobster landings in Îles-de-la-Madeleine (Gulf of St. Lawrence, Québec), using various methods of series analysis. In a first step, they found that an *autoregressive model* (of order 1) accounted for ca. 40% of the variance in the series of landings. This relatively low percentage could be explained by the fact that observations in the series were not very homogeneous. In a second step, external physical variables were added to the model and the data were analysed using *regression on principal components* (Section 10.3). The two external variables were: water temperature in December, 8.5 years before the fishing season, and average winter temperature 3.5 years before. This increased to 90% the variance explained by the model. Lobster landings in a given year would thus depend on: the available stock (autocorrelated to landings during the previous year), the influence of water temperature on larval survival (lobster *Homarus americanus* around Îles-de-la-Madeleine reach commercial size when ca. 8 years old), and the influence of water temperature at the time the animals reached sexual maturity (at the age of ca. 5 years).

12.8 Software

Procedures available in commercial statistical packages are not reviewed here. The R language offers functions for the methods described in Chapter 12.

1. Time series objects. — Function ***ts()*** of STATS creates a time-series object identified to class "ts". ***plot.ts()*** plots a graph for such an object, ***ts.plot()*** plots several time series in a common plot. ***ts.union()*** binds two or more time series into a single R object.

2. Trend extraction. — Function ***lm()*** in STATS is used to detrend data, i.e. extract a linear or polynomial trend and compute residuals.

3. Periodic variability: correlogram. — Function ***acf()*** in STATS computes spatial autocovariance and autocorrelation; ***plot.acf()*** plots confidence intervals under either a white noise or a MA model. ***ccf()*** computes cross-covariance and cross-correlation. For spatial transects or time series with irregular lags, correlograms can be computed using function ***sp.correlogram()*** of package SPDEP; a constant must be written in the second column in the file of geographic coordinates used to create the list of neighbours.

4. Periodic variability: periodogram. — ***buysbal()*** in PASTECS constructs Buys-Ballot tables from time series. ***periodograph()***[*] computes the contingency periodogram (Subsection 12.4.2). ***spec.pgram()*** in STATS estimates the spectral density of a series by a smoothed Schuster periodogram. ***cpgram()*** plots a cumulative periodogram.

5. Periodic variability: spectral analysis. — Function ***spectrum()*** in STATS estimates the spectral density of a time series. ***spec.ar()*** fits an AR model to a time series and computes the spectral density of the fitted model.

6. Wavelet analysis. — Function ***dwt()*** of WAVESLIM is used to compute wavelet analysis for data series, and ***dwt.2d()*** for two-dimensional data[†]. Package WMTSA contains other wavelet methods for time series analysis.

7. Detection of discontinuities in multivariate series. — Function ***chclust()*** of package RIOJA, developed for palaeoecological reconstruction, performs constrained hierarchical clustering from a distance matrix, with clusters constrained by the order of the sampling units in the data file. The method is applicable to temporal or spatial multivariate series, such as sediment core data. Multivariate regression tree analysis (MRT) can also be used for constrained clustering for temporal or spatial multivariate data series, as shown in Subsection 12.6.4. Function ***constrained.clust()*** of package CONST.CLUST[*] carries out constrained hierarchical clustering along a time or spatial series, or on a geographic surface (Section 13.3.2), with cross-validation of the results. Chronological clustering (Section 12.6.5) is implemented in function ***chrono*** of THE R PACKAGE[*] for mainframe computers and Mac OS Classic. This program has not been rewritten yet for the R language.

8. Box-Jenkins models. — Function ***ar()*** in STATS fits an autoregressive model to a univariate or multivariate time series; ***arima()*** fits an ARIMA model to a univariate time series; ***ARMAacf()*** computes the theoretical autocorrelation function for an ARMA process.

9. Miscellaneous methods. — Function ***turnogram()*** in PASTECS computes and plots turnograms; ***turpoints()*** analyses turning points (Section 12.1) and tests the randomness of series.

---

[*] Available on the Web page http://numericalecology.com/rcode.

[†] An introduction and R code for wavelet analysis using WAVESLIM are found on the Web page https://sites.google.com/site/patrickmajames/stuff.

Chapter 13 *Spatial analysis*

13.0 Spatial patterns

Experiment
Gradient
Patch

The analysis of spatial patterns is of prime interest to ecologists because most ecological phenomena investigated by sampling geographic space are structured by forces that have spatial components. Spatial patterns are studied through surveys (called *mensurative experiments* by Hurlbert, 1984), whereas underlying processes can be studied by *manipulative experiments* (Subsection 10.2.3). Ecological processes may give rise to response data displaying recognizable spatial patterns, which may be the subject of spatial analysis. Most ecological patterns may be described as either patches (such as tree groves, phytoplankton patches, and animal herds) or gradients. The latter may be linear or not.

Ecologists investigate the spatial patterns of species assemblages in order to understand the mechanisms that control species distributions. Patchiness is found at all spatial scales — from micrometres to continental and ocean-wide scales. Displaying the spatial variation of an ecological variable in the form of a map shows whether the structure is smoothly continuous or marked by sharp discontinuities. Most field studies cover only a part of any variable's spatial structure. So, gradients or patches displayed by maps may only be interpreted with respect to the scale of the sampling programme, which should be compared to the scale of the phenomenon under study when it is known.

Historical events

It is now understood that species distributions result from the combined action of several forces, some of which are external whereas others are intrinsic to the community. According to the environmental control model (Whittaker, 1956; Bray & Curtis, 1957; Hutchinson, 1957), environmental characteristics are the external forces that control species distributions. The internal forces relate to population dynamics or to top-down or bottom-up biotic interactions within the community (Lindeman, 1942; Southwood, 1987). Both types of forces generate spatial patterns within species or communities. Historical events (Sousa, 1979; Pickett & White, 1985; Reynolds, 1987) are other possible sources of spatial patterns; examples are given in Subsection 14.1.4. The mechanisms that create spatial structures and, hence, spatial correlation in the data, have been discussed in Section 1.1.

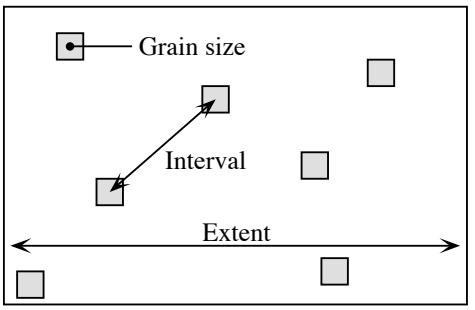


Figure 13.1 Components of a sampling design are the grain size, sampling interval, and extent. In the figure, the sampling units are represented by squares.

The present chapter is not a tutorial discussing all possible questions concerning the analysis of spatial structures. Its scope is more modest; it will describe some methods that allow the investigation of some of the questions of interest in spatial pattern analysis. A fundamental question will be left unanswered in this chapter: that of designing efficient sampling programmes for studying and analysing spatial patterns. The theory of spatial ecological sampling has to be re-written to provide meaningful answers to this question.

Scale

Scale is a key concept in both sampling design and the analysis of spatial (or temporal) patterns. It includes several spatial (or temporal) characteristics of random variables. Definition of these properties, which follows, depends on context.

In sampling theory (Fig. 13.1), spatial scale encompasses three elements of the sampling design (Wiens, 1989; Allen & Hoekstra, 1991; He *et al*., 1994; Dungan *et al*., 2002):

Grain size

- *Grain size* is the size of the elementary sampling units. It may be expressed as the diameter, surface or volume of matter supporting the measurements. In time series analysis, it is the duration over which measurements are integrated. The *resolution* of a study (Schneider, 1994) is equal to the grain size of its sampling design.

Sampling interval

- *Sampling interval* is the average distance between neighbouring sampling units. It is called *lag* in time series analysis (Section 12.0). For fixed extent, the sampling interval is a function of n, the number of sampling units. In turn, n is determined by the total effort that can be allocated to sampling.

Extent

- *Extent* is the total length, area or volume included in the study, or the total duration of the time series. It was called *range* by Schneider (1994) who also defined the *scope* as the ratio of the extent to the grain size. Since extent and grain size are expressed in the same units, scope is a dimensionless variable (Section 3.1).

It may happen that the data consist of contiguous sampling areas that completely cover the extent, instead of small sampling units distant from one another. This may occur in a variety of circumstances where a map is divided into contiguous "picture

Pixel

cells" or *pixels*. These include satellite data, video analysis of a transect, completely inventoried forest plots, and modelling. The linear measurement of grain size is equal to the sampling interval in such a case. Time series may also be entirely studied.

The spatial scale of patterns or processes is described as follows:

Ecological neighbourhood
Unit object
Unit process

- How large is a unit object, or how much space is disturbed by a unit process? This amount of space, which is equivalent to grain size, is called the ecological neighbourhood (Addicott *et al*., 1987) or the area of resolution of individuals (Wiens, 1989). *Unit objects* may be individual plants or animals, bacterial colonies, etc. Examples of measurable structures resulting from *unit processes* are: the neighbourhood occupied by a territorial animal, the width of the wetland zone along a stream or of a tidal sand flat, the size of the patch of soil modified by the root system of a plant, and the size of phytoplankton patches which result from the combined action of primary production and diffusion (see Ecological applications 3.2d and 3.3a).

- What is the average distance between unit objects or processes? This distance is equivalent to the sampling interval.

- Over how much space does this type of object, or this process, occur? This amount of space is equivalent to the extent. For some processes, the extent may be an ocean or the whole planet.

The same notions may be applied to temporally occurring patterns or processes. While they are readily applicable to patterns that concern the distribution of objects, they may sometimes be applied as well to processes.

Sampling design

The scale of the sampling design should follow from what is known (e.g. from a pilot study) about the scale of the pattern or process, and from the ecological question being addressed (Dungan *et al*., 2002). A well-focused question generally reduces the difficulty of choosing the type (simple random, systematic, stratified, etc.) as well as the scale components (grain, interval, extent) of the sampling design.

Sampling grain

- The sampling grain should be larger than a unit object (e.g. an individual organism) and the same as, or preferably smaller than, the structures resulting from a unit process (e.g. a patch), which should be detected by the sampling design.

Sampling interval

- The sampling interval should be smaller than the average distance between the structures resulting from a unit process to be detected by the sampling design.

Sampling extent

- The sampling extent may, in some cases, be the same as the total area covered by the type of objects or by the process under study. In other cases, it is limited to a smaller area, determined by the total allowable effort (n) and the maximum interval that one wishes to maintain between adjacent sampling units. For constant n, the sampling extent can be maximized by turning the sampling area into a transect (see Ecological application 13.1c).

The extent and grain define the observation window in spatial pattern analysis. No structure can be detected that is smaller than the grain or larger than the extent of a study. Wiens (1989) compares them to the overall size and mesh size of a sieve, respectively.

In quantitative ecology, the term "scale" is generally used in a sense opposite to that of cartography. For cartographers, the scale is the ratio between the linear size of an object on a map and its size in nature, so that a small-scale map (e.g. 1:100000) is less detailed than a large-scale map (e.g. 1:25000). For ecologists, scale generally refers to the unit of measurement, e.g. the kilometre sampling scale is bigger than the centimetre scale and weekly observations are broader-scaled than hourly observations.

Broad scale
Fine scale

Confusion is avoided by using "broad scale" for phenomena with large extents and "fine scale" for those with small extents (Wiens, 1989)*. In any case, these terms only have comparative values.

In many instances, not one but several scales may be pertinent for the study of a pattern or process. Different processes are often at work, depending on the scale, to determine spatial patterns. As a consequence, conclusions derived for a spatial scale often cannot be extrapolated to other scales. The scale chosen for any particular study may be considered as a variable-sized window through which one observes nature; see the notion of observational window in Section 12.0. He *et al*. (1994) have shown how species diversity, for example, changes as a function of different components of scale (grain size, sampling interval, and extent). The methods described in Section 13.1, in particular, allow researchers to depict how spatial correlation changes as a function of the sampling interval.

Scale is an important reference to help understand the difference between environmental management problems and the answers that may be found in ecological studies. Most studies are conducted at scales (extents) finer than those of natural or anthropogenic disturbances (Fig. 13.2). As a consequence, environmental problems usually involve scales broader than the information available from field studies — surveys or field experiments. Scaling up from studies to environmental problems is a challenge that ecologists are often facing. New concepts and modelling tools must be developed to do so (Thrush *et al*., 1997). Multiscale spatial analysis of the results of surveys conducted across several spatial scales is one means towards that end.

Heterogeneity

An important concept is that of *heterogeneity* (Kolasa & Rollo, 1991; Dutilleul & Legendre, 1993; Dutilleul, 2011). With reference to spatial patterns, heterogeneity is the opposite of *homogeneity* which means the absence of variation. In everyday's language, heterogeneous means "composed of unlike elements or parts". Pitard (1992) distinguishes *constitution heterogeneity*, which is a property of the objects under study, from *distribution heterogeneity* which can be altered by mixing. In spatial pattern

* Unfortunately these two terms are not antonymic. *Broad* scale refers to the extent; its antonym is *narrow*. *Fine* scale refers to the grain; its antonym is *coarse*.

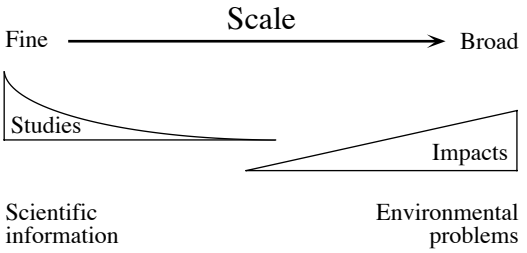


Figure 13.2 Scale differences between environmental management problems and the answers that may be found in ecological studies. From S. F. Thrush (pers. comm.).

analysis, heterogeneous refers to spatial variation in the measurements, in some general sense that applies to quantitative, semiquantitative, or qualitative variables (Subsection 1.4.1). The concept of heterogeneity may also be applied to the time dimension, considering repeated observations made at a single point in space. Heterogeneity can be measured in a univariate (e.g. the variance of a singe variable) or a multivariate way (e.g. the trace of a dispersion matrix). It can be decomposed into orthogonal components (as in PCA, Section 9.1) or with respect to spatial or temporal distance classes (e.g. correlograms for spatial survey data, Section 13.1, or for time series, Section 12.3). Kolasa & Rollo (1991) recognize that "measured heterogeneity", which reflects the observer's perspective, may be inadequate in that it may differ from "functional heterogeneity", which is the heterogeneity that influences the organisms. Functional heterogeneity may not be the same for different groups of organisms because the processes that are important for different groups may act at different temporal or spatial scales. In the sea, for instance, the doubling time of organisms is of the order of 1 day for phytoplankton, 10 to 40 days for zooplankton, 100 to 900 days for fish, and 120 to 500 days for mussels. Spatially, the horizontal scales that characterize patches are of the order of 0.1 to 1 km for phytoplankton and zooplankton, and 1 to 100 km for fish (Legendre *et al*., 1986). Measured heterogeneity converges towards functional heterogeneity as our knowledge of a system increases and, with it, our ability to use our measures to characterize important properties of the system (Kolasa & Rollo, 1991; Dutilleul & Legendre, 1993).

Point pattern

The analysis of spatial ecological patterns comprises two families of methods. *Point pattern analysis* is concerned with the distribution through space of individual objects — for instance individual plants or animals. Its chief purpose is to determine whether the geographic distribution of data points is random or not and to describe the type of pattern, in order to infer what kind of process may have generated it. In this family of methods, the quadrat-density and nearest-neighbour approaches have been widely used in vegetation science (Galiano, 1982; Carpenter & Chaney, 1983). Point pattern analysis will not be discussed further in this chapter. It has been authoritatively reviewed by a number of authors, including Pielou (1977), Cicéri *et al*. (1977), Getis

& Boots (1978), Ripley (1981, 1987), and Upton & Fingleton (1985), and by Dale (1999) and Fortin & Dale (2005) for ecological data.

Regionalized variable
Surface
Surface pattern

Values of a variable observed over a delimited geographic area form a *regionalized variable* (Matheron, 1965), also called a *surface* (Oden *et al*., 1993; Legendre & McArdle, 1997), if the sites where the variable has been observed may be viewed as a sample from an underlying continuous surface. The second family of methods to analyse spatial ecological patterns, called *surface pattern analysis*, deals with the study of spatially continuous phenomena. The spatial distributions of the variables are known, as usual, through sampling (or measurements on aerial photos or satellite maps) at discrete sampling sites. One or several variables are observed or measured at the observation sites, each site representing its surrounding portion of the geographic space. The analysis of continuous surfaces, where pixels cover the whole map (including data obtained by echolocation or remote sensing), is not specifically discussed here.

Surface pattern analysis includes a large number of methods developed to answer a variety of questions (Table 13.1). Several of these methods are discussed in the present chapter. General references are: Cliff & Ord (1981), Ripley (1981), Upton & Fingleton (1985, 1989), Griffith (1987), Legendre & Fortin (1989), and Rossi *et al*. (1992). The geostatistical literature is briefly reviewed in Subsection 13.2.2. The comparison of surfaces, i.e. univariate measures over the same area repeated at two or more sampling times, has been discussed by Legendre & McArdle (1997).

Multiscale modelling of univariate or multivariate ecological data can be done using spatial eigenfunction analysis, described in Chapter 14. Wavelet analysis, described in Subsection 12.5.4 for ecological data series, offers another method of multiscale modelling for *univariate spatial data on a regular grid*, e.g. remotely sensed data or entirely inventoried map areas.

The book of Fortin & Dale (2005) describes point pattern as well as surface pattern methods of spatial analysis. Section 13.6 provides a list of computer programs and functions for surface pattern analysis; most of these methods are not available in the major statistical packages.

Line pattern

Geographers have also developed *line pattern analysis* which is a topological approach to the study of networks of connections among points. Examples are: roads, electric or telephone lines, and river networks.

For a point pattern, heterogeneity refers to the distribution of individuals across space; one often compares the observed density variation of organisms to that expected for randomly distributed objects. For a surface pattern, heterogeneity refers to the variability of quantitative or qualitative descriptors across space. Dutilleul & Legendre (1993) provide a summary of the main statistical tools available to ecologists to quantify spatial heterogeneity in both the point and the surface pattern cases. Dutilleul (1993) describes in more detail how experimental designs can be accommodated to the

Table 13.1 Surface pattern analysis: research objectives and related numerical methods. Modified from Legendre & Fortin (1989).

Research objective	Numerical methods
1) Description of spatial structures and testing for the presence of spatial correlation (Descriptions using structure functions should always be complemented by maps.)	Univariate structure functions: correlogram, variogram, etc. (Section 13.1) Multivariate structure functions: multivariate variogram, Mantel correlogram (Section 13.1) Testing for a gradient in multivariate data: canonical ordination between the multivariate response data and the geographic coordinates of the sites (Section 13.4).
2) Mapping; estimation of values at given locations	Univariate data: local interpolation map, kriging; trend-surface map (global statistical model) (Section 13.2) Multivariate data: space-constrained clustering, search for boundaries (Section 13.3); interpolated map of the 1st (2nd, etc.) ordination axis (Section 13.4); multivariate trend-surface map obtained by canonical analysis (Section 13.4)
3) Modelling species-environment relationships while taking spatial structures into account	Spatial modelling through canonical analysis (Section 13.5); Multiscale analysis: spatial eigenfunction modelling (Chapter 14).
4) Performing valid statistical tests on autocorrelated data	Subsections 1.1.2 and 14.5.3.

spatial heterogeneity found in nature; spatial heterogeneity may be a nuisance for the experimenter, or a characteristic of interest. The analysis of spatial patterns is the study of organized arrangements of [ecological] heterogeneity across space.

These concepts, and more, are discussed in the book of Dutilleul (2011) on spatio-temporal heterogeneity. The book describes the study designs (field surveys and experiments) as well as the methods of analysis to interpret point and surface patterns in data collected to answer questions about the spatial, temporal, and spatio-temporal heterogeneity of ecosystems.

13.1 Structure functions

Ecologists are interested in describing spatial structures in quantitative ways and testing for the presence of spatial correlation in data. The primary objective is to:

- either support the null hypothesis that no significant spatial correlation is present in a data set, or that none remains after detrending (Subsection 13.2.1) or after controlling for the effect of explanatory (e.g. environmental) variables, thus insuring valid use of the standard univariate or multivariate statistical tests of hypotheses,

- or reject the null hypothesis and show that significant spatial correlation is present in the data, in order to use it in conceptual or statistical models.

Tests of spatial correlation coefficients may only support or reject the null hypothesis of the absence of significant spatial structure. When a significant spatial structure is found, it may correspond to induced spatial dependence (Subsection 1.1.1, model 1) or true spatial autocorrelation (model 2).

Map

Spatial structures may be described through *structure functions*, which allow one to quantify the spatial dependence and partition it amongst distance classes. Interpretation of that description is usually supported by maps of the univariate or multivariate data (Sections 13.2 to 13.4). The most commonly used spatial structure functions are correlograms, variograms, and periodograms.

Spatial correlogram

A *correlogram* is a graph in which spatial correlation values are plotted, on the ordinate, against *distance classes* among sites on the abscissa. Correlograms (Cliff & Ord, 1981) can be computed for single variables (Moran's I or Geary's c, Subsection 13.1.1, or the spatial correlation function, Subsection 13.1.5) or for multivariate data (multivariate variogram, Subsection 13.1.4, and Mantel correlogram, Subsection 13.1.6). In all cases, a test of significance is available for each individual spatial correlation coefficient plotted in a correlogram.

Variogram

Similarly, a *variogram* is a graph in which semi-variance is plotted, on the ordinate, against *distance classes* among sites on the abscissa (Subsection 13.1.3). In the geostatistical tradition, semi-variance statistics are not tested for significance, although they could be through the test developed for Geary's c, when the condition of second-order stationarity is satisfied (Subsection 13.1.1). Statistical models may be fitted to variograms (linear, exponential, spherical, Gaussian, etc.); they allow the investigator to relate the observed structure to hypothesized generating processes or to produce interpolated maps by kriging (Subsection 13.2.2).

Because they measure the relationship between pairs of observation points located a certain distance apart, correlograms and variograms may be computed either for preferred geographic directions or, when the phenomenon is assumed to be isotropic in space, in an all-directional way.

2-D periodogram

A *two-dimensional Schuster* (1898) *periodogram* may be computed when the structure under study is assumed to consist of a combination of sine waves propagated through space. The basic idea is to fit sines and cosines of various periods, one period at a time, and to determine the proportion of the series' variance (r^2) explained by each period. In periodograms, the abscissa is either a period or its inverse, a frequency; the ordinate is the proportion of variance explained. Two-dimensional periodograms may be plotted for all combinations of directions and spatial frequencies. The technique is applicable to regular grids of points; it is described Priestley (1964), Ripley (1981), Renshaw and Ford (1984) and Legendre & Fortin (1989). It is not discussed further in this book. Spatial eigenfunction analysis, described in Chapter 14, carries out a similar form of analysis and is more general since it can be used on irregularly-spaced points.

1 — Spatial correlograms

For quantitative variables (univariate data), spatial correlation can be estimated by Moran's *I* (1950) or Geary's *c* (1954) spatial correlation statistics* (Cliff & Ord, 1981):

Moran's *I*:
$$I(d) = \frac{\frac{1}{W}\sum_{h=1}^{n}\sum_{i=1}^{n} w_{hi}(y_h - \bar{y})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad \text{for } h \neq i \qquad \textbf{(13.1)}$$

Geary's *c*:
$$c(d) = \frac{\frac{1}{2W}\sum_{h=1}^{n}\sum_{i=1}^{n} w_{hi}(y_h - y_i)^2}{\frac{1}{(n-1)}\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad \text{for } h \neq i \qquad \textbf{(13.2)}$$

y_h and y_i are the values of the observed variable at sites h and i, and d is the distance class considered in the calculation. Before computing spatial correlation coefficients, a matrix of geographic distances $\mathbf{D} = [D_{hi}]$ among observation sites must be calculated. Statistical details about these coefficients are available in Cliff & Ord (1981) and d'Aubigny (2006).

In the presence of explanatory variables generating spatial structure in the variable of interest, true spatial autocorrelation must be estimated on the *residuals* of a model that takes these explanatory variables into account. This is in agreement with the definition of spatial autocorrelation (Section 1.1), which is the spatial dependence *among the error components* of the observed data (eq. 1.2).

* These statistics are often called spatial *autocorrelation* coefficients. This terminology is misleading since the coefficients measure any type of spatial structure, be it due to induced spatial dependence (eq. 1.1) or true spatial autocorrelation (eq. 1.2).

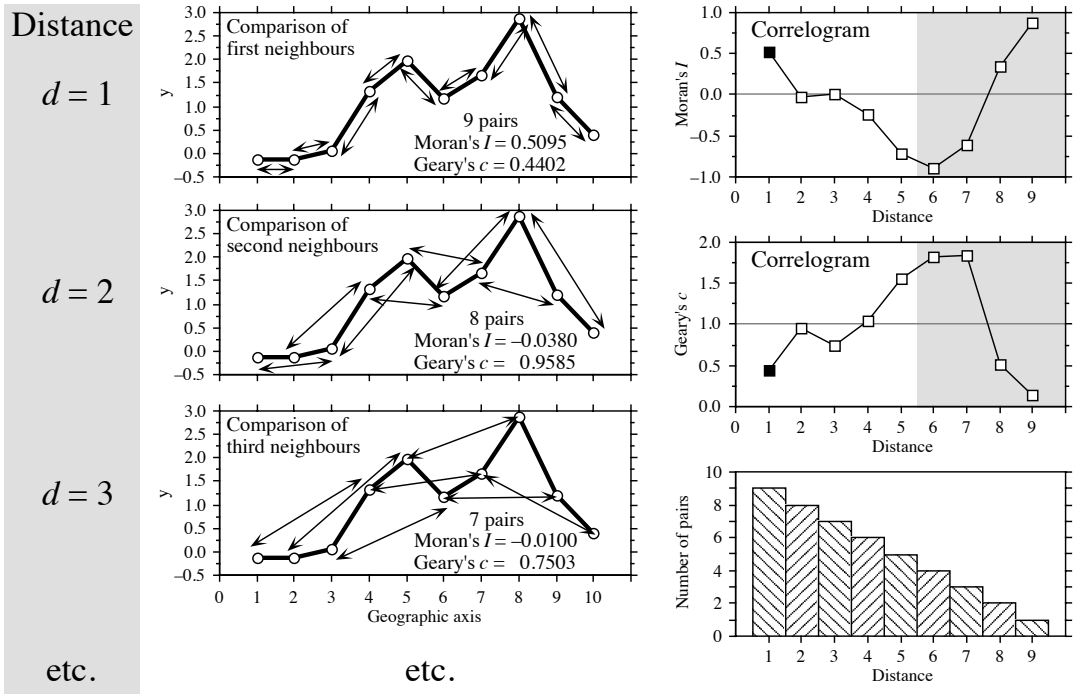


Figure 13.3 Construction of correlograms. Left: data series observed along a single geographic axis (10 equispaced observations). Moran's I and Geary's c statistics are computed from pairs of observations found at preselected distances ($d = 1$, $d = 2$, $d = 3$, etc.). Right: correlograms are graphs of the spatial correlation statistics plotted against distance. Dark squares: significant correlation statistics ($p \leq 0.05$). Lower right: histogram showing the number of pairs in each distance class. Coefficients for the larger distance values (grey zones in correlograms) should not be considered in correlograms, nor interpreted, because they are based on a small number of pairs (test with low power) and exclude some points found in the centre of the series or surface.

In the construction of a correlogram, spatial correlation coefficients are computed, in turn, for the various distance classes d. The weights w_{hi} are Kronecker deltas (as in eq. 7.21); the binary weights take the value $w_{hi} = 1$ when sites h and i are at distance d and $w_{hi} = 0$ otherwise. In this way, only the pairs of sites (h, i) within the stated distance class (d) are taken into account in the calculation of any given coefficient. This approach is illustrated in Fig. 13.3. W is the sum of the weights w_{hi} for the given distance class, i.e. the number of pairs used to calculate the coefficient. For a given distance class, the weights w_{ij} are written in a $(n \times n)$ spatial weighting matrix **W**; an example of a binary spatial weighting matrix is matrix **X**(1) of Fig. 13.14. Jumars *et al.* (1977) present ecological examples where the distance^{-1} or distance^{-2} among adjacent sites is used for weight instead of 1's.

The numerators of eqs. 13.1 and 13.2 are written with summations involving each pair of objects twice; in eq. 13.2 for example, the terms $(y_h - y_i)^2$ and $(y_i - y_h)^2$ are both used in the summation. This allows for cases where the distance matrix **D** or the weight matrix **W** is asymmetric. In studies of the dispersion of pollutants in soil, for instance, drainage may make it more difficult to go from A to B than from B to A; this may be recorded as a larger distance from A to B than from B to A. In spatio-temporal analyses, an observed value may influence a later value at the same or a different site, but not the reverse. An impossible connection may be coded by a very large value of distance or by $w_{hi} = 0$. In most applications, however, the geographic distance matrix among sites is symmetric and the coefficients can be computed from the half-matrix of distances; the formulae remain the same, because W and the sum in the numerator are half the values computed over the whole distance matrix **D**.

One may use distances along a network of connections (Subsection 13.3.1) instead of straight-line geographic distances; this includes the "chess moves" for regularly-spaced points as obtained from systematic sampling designs: rook's, bishop's, or king's connections (see Fig. 13.21). For very broad-scale studies, involving a whole ocean or continent, "great-circle distances", i.e. distances along the earth's curved surface, should be used instead of straight-line distances through the earth crust.

Moran's I formula is related to the Pearson correlation coefficient; its numerator is a covariance, comparing the values found at all pairs of points in turn, while its denominator is the maximum-likelihood estimator of the variance (i.e. division by n instead of $n - 1$); in Pearson r, the denominator is the product of the standard deviations of the two variables (eq. 4.7), whereas in Moran's I there is only one variable involved. Moran's I mainly differs from Pearson r in that the sums in the numerator and denominator of eq. 13.1 do not involve the same number of terms; only the terms corresponding to distances within the given class are considered in the numerator whereas all pairs are taken into account in the denominator. Moran's I usually takes values in the interval [–1, +1] although values lower than –1 or higher than +1 may occasionally be obtained. Positive spatial correlation in the data translates into positive values of I; negative correlation produces negative values.

Readers who are familiar with correlograms in time series analysis (Section 12.3) will be reassured to know that, when a problem involves equispaced observations along a single physical dimension, as in Fig. 13.3, calculating Moran's I (eq. 13.1) for the different distance classes is nearly the same as computing the autocorrelation coefficient of time series analysis (Fig. 12.5, eq. 12.7).

Geary's c coefficient is a distance-type function; it varies from 0 to some unspecified value larger than 1. Its numerator sums the squared differences between values found at the various pairs of sites being compared. A Geary's c correlogram varies as the reverse of a Moran's I correlogram; strong spatial correlation produces high values of I and low values of c (Fig. 13.3). Positive spatial correlation translates in values of c between 0 and 1 whereas negative correlation produces values larger than 1. Hence, the reference 'no correlation' value is $c = 1$ in Geary's correlograms.

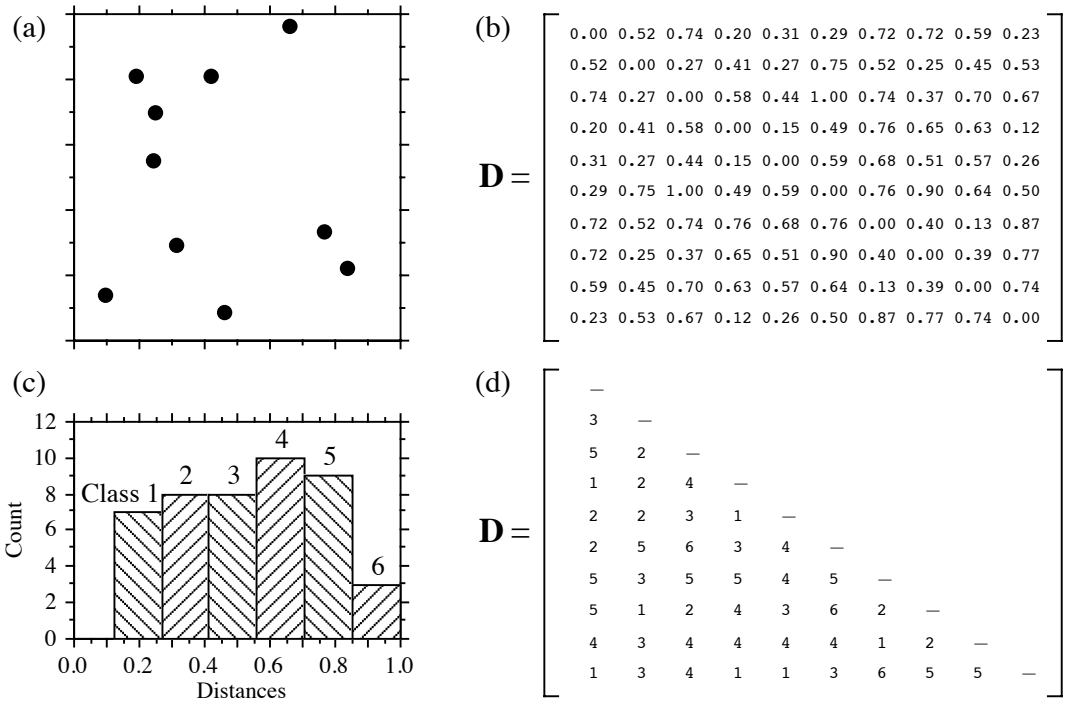


Figure 13.4 Calculation of distance classes, artificial data. (a) Map of 10 sites in a 1-km^2 sampling area. (b) Geographic distance matrix (**D**, in km). (c) Frequency histogram of distances (classes 1 to 6) for the upper (or lower) triangular portion of **D**. (d) Distances recoded into 6 classes.

For sites lying on a surface or in a volume, geographic distances do not naturally fall into a small number of values; this is true for regular grids as well as random or other forms of irregular sampling designs. Distance values must be grouped into distance classes; in this way, each spatial correlation coefficient can be computed using several comparisons of sampling sites.

Numerical example. In Fig. 13.4 (artificial data), 10 sites have been located at random into a 1-km^2 sampling area. Euclidean (geographic) distances were computed among sites. The number of classes is arbitrary and left to the user's decision. A compromise has to be made between resolution of the correlogram (more resolution when there are more, narrower classes) and power of the test (more power when there are more pairs in a distance class). Sturges' (1926) rule is often used to decide about the number of classes in histograms; it was used here and gave:

$$\text{Number of classes} = 1 + 3.322 \log_{10}(m) = 1 + 3.3 \log_{10}(45) = 6.46 \qquad \textbf{(13.3)}$$

where m is the number of distances in the upper triangular matrix and 3.322 is $1/\log_{10}2$; the number was rounded to the nearest integer (i.e. 6). The distance matrix was thus recoded into 6 classes, ascribing class numbers (1 to 6) to all distances within a class of the histogram.

An alternative to distance classes with equal widths would be to create distance classes containing the same number of pairs (notwithstanding tied values); distance classes formed in this way are of unequal widths. The advantage is that the tests of significance have the same power across all distance classes because they are based upon the same number of pairs of observations. The disadvantages are that limits of the distance classes are more difficult to find and correlograms are harder to draw.

Spatial correlation coefficients can be tested for significance and confidence intervals can be computed. With proper correction for multiple testing, one can determine if a significant spatial structure is present in the data and what are the distance classes showing significant positive or negative correlation. Tests of significance require, however, that certain conditions specified below be fulfilled.

Second-order stationarity

The tests require that the condition of *second-order stationarity* be satisfied. Second-order stationarity refers to the vectors separating pairs of values in the study area. This rather strong condition states that the mean of the variable is constant over the study area, and the spatial covariance (numerator of eq. 13.1) depends only on the length and orientation of the vector between any two points, not on its position in the study area (David, 1977). The variance (denominator of eq. 13.1) must be the same for all points in the study area (homogeneity of the variance; Dutilleul, 2011).

Intrinsic stationarity

A relaxed form of stationarity, called *intrinsic stationarity*, states that the differences $(y_h - y_i)$ for any distance d (numerator of eq. 13.2) must have zero mean and constant and finite variance over the study area, independently of the location where the differences are calculated. Here, one considers the *increments* of the values of the regionalized variable instead of the values themselves (David, 1977). As shown below, the variance of the increments is the variogram function. In layman's terms, this means that a single spatial correlation function is adequate to describe the entire surface under study. An example where intrinsic stationarity does not hold is a region which is half plain and half mountains; such a region should be divided in two subregions in which the variable "altitude" could be modelled by separate spatial correlation functions. Second-order stationarity implies intrinsic stationarity, but the reciprocal is not true. Intrinsic stationarity is a weaker form of stationarity compatible with a broader range of models. This condition must always be met when variograms or correlograms (including multivariate Mantel correlograms) are computed, even for descriptive purpose.

Cliff & Ord (1981) describe how to compute confidence intervals and test the significance of spatial correlation coefficients. For any normally distributed statistic *Stat*, a confidence interval at significance level α is obtained as follows:

$$Pr\left(Stat - z_{\alpha/2}\sqrt{\mathrm{Var}\left(Stat\right)} < Stat_{\mathrm{Pop}} < Stat + z_{\alpha/2}\sqrt{\mathrm{Var}\left(Stat\right)}\right) = 1 - \alpha \qquad \textbf{(13.4)}$$

For significance testing with large samples, a one-tailed critical value $Stat_\alpha$ at significance level α is obtained as follows:

$$Stat_\alpha = z_\alpha \sqrt{\mathrm{Var}\,(Stat)} + \text{Expected value of } Stat \text{ under } H_0 \qquad \textbf{(13.5)}$$

It is possible to use this approach because both I and c are asymptotically normally distributed for data sets of moderate to large sizes (Cliff & Ord, 1981). Values $z_{\alpha/2}$ or z_α are found in a table of standard normal deviates. Under the hypothesis (H_0) of random spatial distribution of the observed values y_i, the expected values (E) of Moran's I and Geary's c are:

$$E(I) = -(n-1)^{-1} \quad \text{and} \quad E(c) = 1 \qquad \textbf{(13.6)}$$

Under the null hypothesis, the expected value of Moran's I approaches 0 as n increases. The variances are computed as follows under a randomization assumption, which simply states that, under H_0, the observations y_i are independent of their positions in space (second-order stationarity assumption) and, thus, are exchangeable:

$$\mathrm{Var}\,(I) = \mathrm{E}\,(I^2) - [\mathrm{E}\,(I)]^2 \qquad \textbf{(13.7)}$$

$$\mathrm{Var}\,(I) = \frac{n[(n^2-3n+3)S_1 - nS_2 + 3W^2] - b_2[(n^2-n)S_1 - 2nS_2 + 6W^2]}{(n-1)(n-2)(n-3)W^2} - \frac{1}{(n-1)^2}$$

$$\mathrm{Var}\,(c) = \frac{(n-1)S_1[n^2-3n+3-(n-1)b_2]}{n(n-2)(n-3)W^2} \qquad \textbf{(13.8)}$$

$$+ \frac{-0.25(n-1)S_2[n^2+3n-6-(n^2-n+2)b_2] + W^2[n^2-3+(-(n-1)^2)b_2]}{n(n-2)(n-3)W^2}$$

In these equations,

- $S_1 = \frac{1}{2}\sum_{h=1}^{n}\sum_{i=1}^{n}(w_{hi}+w_{ih})^2$ (there is a term of this sum for *each cell* of matrix **W**);
- $S_2 = \sum_{i=1}^{n}(w_{i+}+w_{+i})^2$ where w_{i+} and w_{+i} are respectively the sums of row i and column i of matrix **W**;
- $b_2 = n\sum_{i=1}^{n}(y_i-\bar{y})^4 \Big/ \left[\sum_{i=1}^{n}(y_i-\bar{y})^2\right]^2$ measures the kurtosis of the distribution;
- W is as defined in eqs. 13.1 and 13.2.

In most cases in ecology, tests of spatial correlation are one-tailed because the sign of the correlation is stated in the ecological hypothesis; for instance, contagious biological processes such as growth, reproduction, and dispersal, all suggest that ecological variables are positively correlated at short distances. To carry out an approximate test of significance, select a value of α (e.g. $\alpha = 0.05$) and find z_α in a table of the standard normal distribution (e.g. $z_{0.05} = +1.6452$). Critical values are found as in eq. 13.5, with a correction factor that becomes important when n is small:

- $I_\alpha = z_\alpha \sqrt{\mathrm{Var}\,(I)} - k_\alpha\,(n-1)^{-1}$ in all cases, using the value in the upper tail of the z distribution when testing for positive spatial correlation (e.g. $z_{0.05} = +1.6452$), and the value in the lower tail in the opposite case (e.g. $z_{0.05} = -1.6452$);

- $c_\alpha = z_\alpha \sqrt{\mathrm{Var}\,(c)} + 1$ when $c < 1$ (positive spatial correlation), using the value in the lower tail of the z distribution (e.g. $z_{0.05} = -1.6452$);

- $c_\alpha = z_\alpha \sqrt{\mathrm{Var}\,(c)} + 1 - k_\alpha\,(n-1)^{-1}$ when $c > 1$ (negative spatial correlation), using the value in the upper tail of the z distribution (e.g. $z_{0.05} = +1.6452$).

The value taken by the correction factor k_α depends on the values of n and W. If $4\,(n - \sqrt{n}) < W \leq 4\,(2n - 3\sqrt{n} + 1)$, then $k_\alpha = \sqrt{10\alpha}$; otherwise, $k_\alpha = 1$. If the test is two-tailed, use $\alpha^* = \alpha/2$ to find z_{α^*} and k_{α^*} before computing critical values. These corrections are based upon simulations reported by Cliff & Ord (1981, Section 2.5).

Other formulas are found in Cliff & Ord (1981) for conducting a test under the assumption of normality, where one assumes that the y_i's result from n independent draws from a normal population. When n is very small, tests of I and c should be conducted by permutation (Subsection 1.2.2).

Moran's I and Geary's c are sensitive to extreme values and, in general, to asymmetry in the data distributions, as are the related Pearson r and Euclidean distance coefficients. Asymmetry increases the variance of the data. It also increases the kurtosis and hence the variance of the I and c coefficients (eqs. 13.7 and 13.8); this makes it more difficult to reach significance in statistical tests. So, practitioners usually attempt to normalize the data before computing correlograms and variograms.

Statistical testing in correlograms implies multiple testing since a test of significance is carried out for each spatial correlation coefficient. Oden (1984) has developed a Q statistic to test the global significance of spatial correlograms; his test is an extension of the Portmanteau Q-test used in time series analysis (Box & Jenkins, 1976). An alternative global test is to check whether the correlogram contains at least one correlation statistic that is significant at the Bonferroni-corrected significance level (Box 1.3). Simulations by Oden (1984) showed that the power of the Q-test is not appreciably greater than the power of the Bonferroni procedure, which is computationally a lot simpler. A practical question remains, though: how many distance classes should be created? This determines the number of simultaneous tests that are carried out. More classes mean more resolution but fewer pairs per class and,

thus, less power for each test; more classes also mean a smaller Bonferroni-corrected α' level, which makes it more difficult for a correlogram to reach global significance.

When the overall test has shown global significance, one may wish to identify the individual spatial correlation statistics that are significant, in order to reach an interpretation (Subsection 13.1.2). One could rely on Bonferroni-corrected tests for all individual correlation statistics, but this approach would be too conservative; a better solution is to use Holm's correction procedure (Box 1.3). Another approach is the *progressive Bonferroni correction* described in Subsection 12.4.2; it is only applicable when the ecological hypothesis indicates that significant spatial correlation is to be expected in the smallest distance classes and the purpose of the analysis is to determine the extent of the spatial correlation (i.e. which distance class it reaches). With the progressive Bonferroni approach, the likelihood of emergence of significant values decreases as one proceeds from left to right, i.e. from the small to the large distance classes of the correlogram. In addition, one does not have to limit the correlogram to a small number of classes to reduce the effect of the correction, as it is the case with Oden's overall test and with the Bonferroni and Holm correction methods. This approach will be used in the examples that follow.

Spatial correlation coefficients and tests of significance also exist for qualitative (nominal) variables (Cliff & Ord, 1981); they have been used, for example, to analyse spatial patterns of sexes in plants (Sakai & Oden, 1983; Sokal & Thomson, 1987). Special types of spatial correlation coefficients have been developed to answer specific problems (e.g. Galiano, 1983; Estabrook & Gates, 1984). The paired-quadrat variance method, developed by Goodall (1974) to analyse spatial patterns of ecological data by random pairing of quadrats, is related to correlograms.

2 — Interpretation of all-directional correlograms

Isotropy
Anisotropy

When the spatial correlation function is the same for all geographic directions considered, the phenomenon is *isotropic*. The opposite of isotropy is *anisotropy*. When a variable is isotropic, a single correlogram can be computed over all directions of the study area. The correlogram is said to be *all-directional* or *omnidirectional*. Directional correlograms, which are computed for a single spatial direction, are discussed together with anisotropy and directional variograms in Subsection 13.1.3.

Correlograms are analysed mostly by looking at their shapes. Examples will help clarify the relationship between spatial structures and all-directional correlograms. The important message is that, although correlograms may give clues as to the underlying spatial structure, the information they provide is not specific; a blind interpretation may be misleading and should be supported by examination of maps (Section 13.2).

Numerical example. Artificial data were generated that correspond to a number of spatial patterns. The data and resulting correlograms are presented in Fig. 13.5.

1. Nine bumps. — The surface in Fig. 13.5a is made of nine bi-normal curves. 225 points were sampled across the surface using a regular 15 × 15 grid (Fig. 13.5f). The "height" was noted at each sampling point. The 25 200 distances among points found in the upper-triangular portion of the distance matrix were divided into 16 distance classes, using Sturges' rule (eq. 13.3), and correlograms were computed. According to Oden's test, the correlograms were globally significant at the $\alpha = 5\%$ level since several individual values were significant at the Bonferroni-corrected level $\alpha' = 0.05/16 = 0.00312$. In each correlogram, the progressive Bonferroni correction method was applied to identify significant spatial correlation coefficients: the coefficient for distance class 1 was tested at the $\alpha = 0.05$ level; the coefficient for distance class 2 was tested at the $\alpha' = 0.05/2$ level; and, more generally, the coefficient for distance class k was tested at the $\alpha' = 0.05/k$ level. Spatial correlation coefficients are not reported for distance classes 15 and 16 (60 and 10 pairs, respectively) because they only include the pairs of points bordering the surface, to the exclusion of all other pairs.

There is a correspondence between individual significant spatial correlation coefficients and the main elements of the spatial structure. The correspondence can clearly be seen in this example, where the data generating process is known. This is not the case when analysing field data, for which the existence and nature of the spatial structures must be confirmed by mapping the data. The presence of several equispaced patches produces an alternation of significant positive and negative values along the correlograms. The first spatial correlation coefficient, which is above 0 in Moran's correlogram and below 1 in Geary's, indicates positive spatial correlation in the first distance class; the first class contains the 420 pairs of points that are at distance 1 of each other on the grid (i.e. the first neighbours in the N-S or E-W directions of the map). Positive and significant spatial correlation in the first distance class confirms that the distance between first neighbours is smaller than the patch size; if the distance between first neighbours in this example were larger than the patch size, the first neighbours would be dissimilar in values and the correlation would be negative for the first distance class. The next peaking positive correlation value (which is smaller than 1 in Geary's correlogram) occurs at distance class 5, which includes distances from 4.95 to 6.19 in grid units; this corresponds to positive spatial correlation between points located at similar positions on neighbouring bumps, or neighbouring troughs; distances between successive peaks are 5 grid units in the E-W or N-S directions. The next peaking positive spatial correlation value occurs at distance class 9 (distances from 9.90 to 11.14 in grid units); it includes value 10, which is the distance between second-neighbour bumps in the N-S and E-W directions. Peaking negative correlation values (which are larger than 1 in Geary's correlogram) are interpreted in a similar way. The first such value occurs at distance class 3 (distances from 2.48 to 3.71 in grid units); it includes value 2.5, which is the distance between peaks and troughs in the N-S and E-W directions on the map. If the bumps were unevenly spaced, the correlograms would be similar for the small distance classes, but there would be no other significant values afterwards.

The main problem with all-directional correlograms is that the diagonal comparisons are included in the same calculations as the N-S and E-W comparisons. As distances become larger, diagonal comparisons between, say, points located near the top of the nine bumps tend to fall in different distance classes than comparable N-S or E-W comparisons. This blurs the signal and makes the spatial correlation coefficients for larger distance classes less significant and interpretable.

2. Wave (Fig. 13.5b). — Each crest was generated as a normal curve. Crests were separated by five grid units; the surface was constructed in this way to make it comparable to Fig. 13.5a. The correlograms are nearly indistinguishable from those of the nine bumps. All-directional

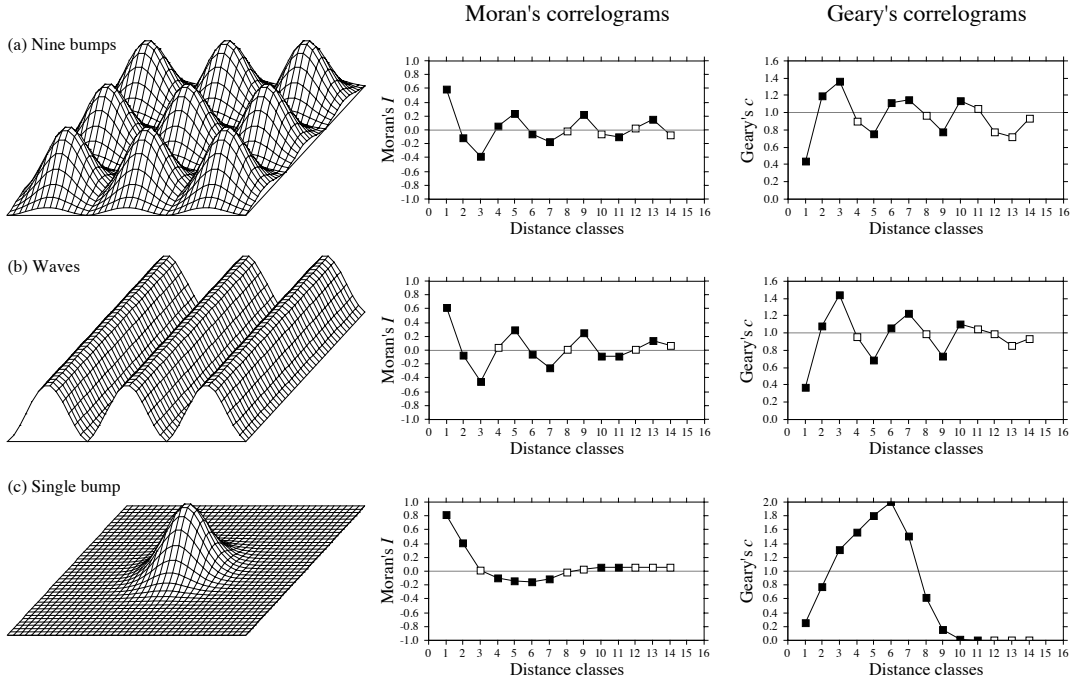


Figure 13.5 Spatial correlation analysis of artificial spatial structures shown on the left: (a) nine bumps; (b) waves; (c) a single bump. Centre and right: all-directional correlograms. Dark squares: correlation statistics that remain significant after progressive Bonferroni correction ($\alpha = 0.05$); white squares: non-significant values. (The figure continues next page.)

correlograms alone cannot tell apart regular bumps from regular waves; directional correlograms or maps are required.

3. Single bump (Fig. 13.5c). — One of the normal curves of Fig. 13.5a was plotted alone at the centre of the study area. Significant negative spatial correlation, which reaches distance classes 6 or 7, delimits the extent of the "range of influence" of this single bump, which covers half the study area. It is not limited here by the rise of adjacent bumps, as this was the case in (a).

4. Linear gradient (Fig. 13.5d). — The correlogram is monotonic decreasing. Nearly all spatial correlation values in the correlograms are significant.

True, false gradient

There are actually two kinds of gradients (Legendre, 1993). *True gradients*, on the one hand, are deterministic structures. Model 1 of Subsection 1.1.1 (induced spatial dependence, eq. 1.1) can generate a true gradient; see Fig. 1.5, case 4. That gradient can be modelled using trend-surface analysis (Subsection 13.2.1). The observed values have independent error terms,

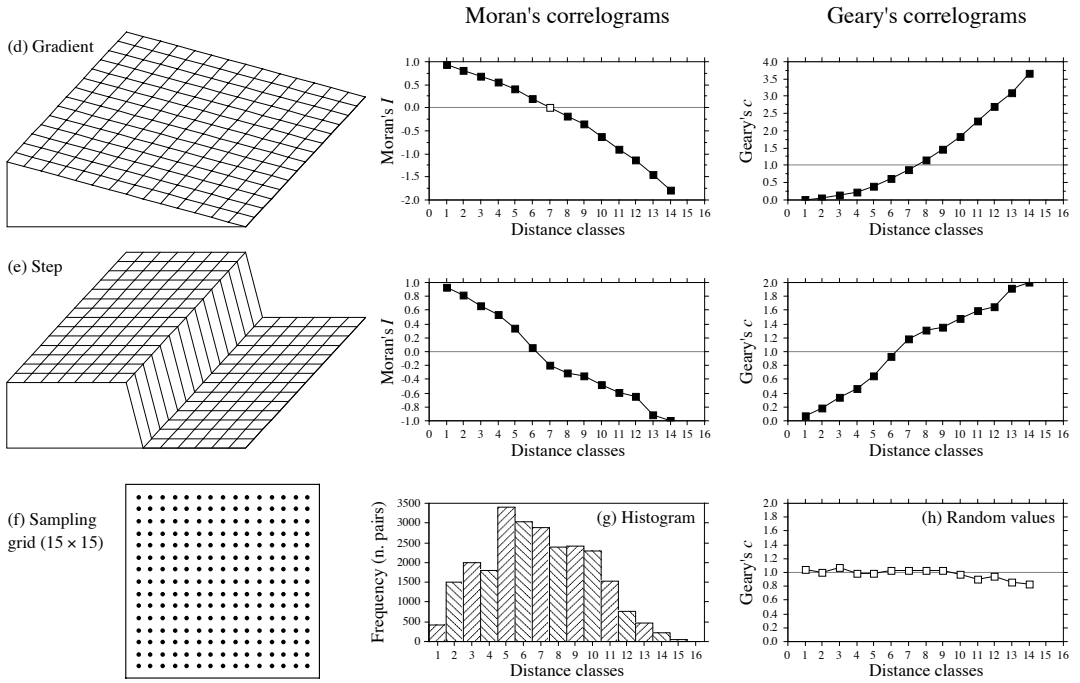


Figure 13.5 **(continued)** Spatial correlation analysis of artificial spatial structures shown on the left: (d) gradient; (e) step. (h) All-directional correlogram of random values. (f) Sampling grid used on each of the artificial spatial structures to obtain 225 "observed values" for spatial correlation analysis. (g) Histogram showing the number of pairs in each distance class. Distances, from 1 to 19.8 in units of the sampling grid, were grouped into 16 distance classes. Spatial correlation statistics (I or c) are not shown for distance classes 15 and 16; see text.

i.e. error terms that are not autocorrelated. *False gradients*, on the other hand, are structures that look like gradients, but actually correspond to spatial correlation generated by some spatial process. Model 2 of Subsection 1.1.1 (spatial autocorrelation, eq. 1.2) can generate a false gradient, especially when the sampling area is small relative to the range of influence of the generating process; see Fig. 1.5, case 3.

In the case of "true gradients", spatial correlation coefficients should not be tested for significance because the condition of second-order stationarity is not satisfied (definition in Subsection 13.1.1); the expected value of the mean is not the same over the whole study area. In the case of "false gradients", however, tests of significance are warranted. For descriptive purposes, correlograms may still be computed for "true gradients" (without tests of significance) because intrinsic stationarity is satisfied. One may also choose to extract a "true gradient" using trend-surface analysis, compute residuals, and look for spatial correlation among the residuals. This is equivalent to trend extraction prior to time series analysis (Section 12.2).

How does one know whether a gradient is "true" or "false"? This is a moot point. When the process generating the observed structure is known, one may decide whether it is likely to have generated spatial correlation in the observed data, or not. Otherwise, one may empirically look at the *target population* of the study. In the case of a spatial study, this is the population of potential sites in the larger area into which the study area is embedded, the study area representing the *statistical population* about which inference can be made. Even from sparse or indirect data, a researcher may form an opinion as to whether the observed gradient is deterministic ("true gradient") or is part of a landscape displaying spatial correlation at broader spatial scale ("false gradient").

5. Step (Fig. 13.5e). — A step between two flat surfaces is enough to produce a correlogram that is indistinguishable, for all practical purposes, from that of a gradient. Correlograms alone cannot tell apart regular gradients from steps; maps are required. As in the case of gradients, there are "true steps" (deterministic) and "false steps" (resulting from an autocorrelated process), although the latter is rare. The presence of a sharp discontinuity in a surface generally indicates that the two parts should be subjected to separate analyses. The methods of boundary detection and constrained clustering (Section 13.3) may help detect such discontinuities and delimit homogeneous areas prior to spatial correlation analysis.

6. Random values (Fig. 13.5h). — Random numbers drawn from a standard normal distribution were generated for each point of the grid and used as the variable to be analysed. Random data are said to represent a "pure nugget effect" in geostatistics. The spatial correlation coefficients were small and non-significant at the 5% level. Only the Geary correlogram is presented.

Sokal (1979) and Cliff & Ord (1981) described, in general terms, where to expect significant values in correlograms, for some spatial structures such as gradients and large or small patches. Their summary tables are in agreement with the test examples above. The absence of significant coefficients in a correlogram must be interpreted with caution, however.

- The absence may indicate that the surface under study is free of spatial correlation at the study scale. This conclusion is subject to *type II error*. Type II error depends on the power of the test, which is a function of (1) the α significance level, (2) the size of effect (i.e. the minimum amount of spatial correlation) one wants to detect, (3) the number of observations (n), and (4) the variance of the sample of data (Cohen, 1988):

$$\text{Power} = (1 - \beta) = f(\alpha, \text{size of effect}, n, s_x^2)$$

Is the test powerful enough to warrant such a conclusion? Are there enough observations to reach significance? The easiest way to increase the power of a test, for a given variable and fixed α, is to increase n.

- The absence may also indicate that the sampling design is inadequate to detect the spatial correlation that may exist in the system. Are the grain size, extent and sampling interval (Section 13.0) adequate to detect the type of spatial correlation one can hypothesize from knowledge about the biological or ecological process under study?

Ecologists can often formulate hypotheses about the mechanism or process that may have generated a spatial phenomenon and deduct the shape that the resulting

surface should have. When the model specifies a value for each geographic position (e.g. a spatial gradient), data and model can be compared by correlation analysis. In other instances, the biological or ecological model only specifies the process generating the spatial correlation, not the exact geographic position of each resulting value. Correlograms may be used to support or reject a biological or ecological hypothesis. As in the examples of Fig. 13.5, one can construct an artificial model-surface corresponding to the hypothesis, compute a correlogram of that surface, and compare the correlograms of the real and model data. For instance, Sokal *et al*. (1997a) generated data corresponding to several gene dispersion mechanisms in populations and showed the kind of spatial correlogram that may be expected from each model. Another application concerning phylogenetic patterns of human evolution in Eurasia and Africa (space-time model) is found in Sokal *et al*. (1997b).

Bjørnstad *et al*. (1999) and Bjørnstad & Falck (2001) proposed a spline correlogram, which provides a continuous and model-free function for the spatial covariance. The spline correlogram may be seen as a modification of the nonparametric covariance function of Hall and co-workers (Hall & Patil, 1994; Hall *et al*., 1994). A bootstrap algorithm estimates a confidence envelope around the entire correlogram. Confidence envelopes allow one to test the similarity between correlograms of real or simulated data. See package NCF in Section 13.6.

Ecological application 13.1a

During a study of the factors potentially responsible for the choice of settling sites of *Balanus crenatus* larvae (Cirripedia) in the St. Lawrence Estuary (Hudon *et al*., 1983), plates of artificial substrate (plastic laminate) were subjected to colonization in the infralittoral zone. Plates were positioned vertically, parallel to one another. Pictures of plates were taken during the course of the study. The present ecological application uses data obtained from a picture of a plate taken after a 3-month immersion at a depth of 5 m below low tide, during the summer 1978. The picture was divided into a (10 × 15) grid, for a total of 150 pixels of 1.7 × 1.7 cm. Barnacles were counted by C. Hudon and P. Legendre for the present application (Fig. 13.6a; not published in *op. cit*.). The hypothesis to be tested was that barnacles had a patchy distribution. Barnacles are gregarious animals; their larvae are chemically attracted to settling sites by arthropodine secreted by settled adults (Gabbott & Larman, 1971).

A gradient in larval concentration was expected in the top-to-bottom direction of the plate because of the known negative phototropism of barnacle larvae at the time of settlement (Visscher, 1928). Some kind of border effect was also expected because access to the centre of the plates located in the middle of the pack was more limited than to the fringe. These large-scale effects create violations to the condition of second-order stationarity. A trend-surface equation (Subsection 13.2.1) was computed to account for it, using only the Y coordinate (top-to-bottom axis). Indeed, a significant trend surface was found, involving Y and Y^2, that accounted for 10% of the variation. It forecasted high barnacle concentration in the bottom part of the plate and near the upper and lower margins. Residuals from this equation were calculated and used in spatial correlation analysis.

Euclidean distances were computed among pixels; following Sturge's rule (eq. 13.3), the distances were divided into 14 classes (Fig. 13.6b). Significant positive spatial correlation was

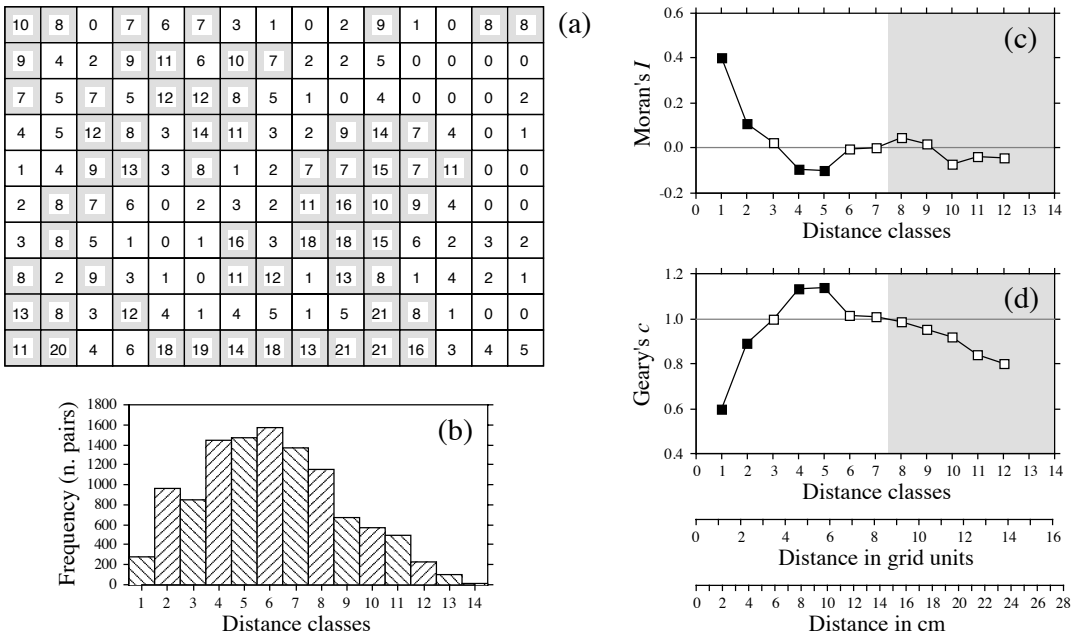


Figure 13.6 (a) Counts of adult barnacles in 150 (1.7 × 1.7 cm) pixels on a plate of artificial substrate (17 × 25.5 cm). The mean concentration is 6.17 animals per pixel; pixels with counts ≥ 7 are shaded to display the aggregates. (b) Histogram of the number of pairs in each distance class. (c) Moran's correlogram. (d) Geary's correlogram. Dark squares: spatial correlation statistics that remain significant after progressive Bonferroni correction ($\alpha = 0.05$); white squares: non-significant values. Grey zones: coefficients that should not be interpreted because they exclude some points in the centre of the study area. Coefficients for distance classes 13 and 14 are not given because they only include the pairs of points bordering the surface. Distances are also given in grid units and in cm.

found in the first distance classes of the correlograms (Fig. 13.6c, d), supporting the hypothesis of patchiness. The size of the patches, or "range of influence" (i.e. the distance between zones of high and low concentrations), is indicated by the distance at which the first maximum negative Moran's *I* correlation value is found. This occurs in classes 4 and 5, which correspond to a distance of about 5 in grid units, or 8 to 10 cm. The patches of high concentration are shaded on the map of the plate of artificial substrate (Fig. 13.6a).

A spatial correlogram is an overall function of spatial correlation across a study area. It is not meant to display details of the structure across the area. Anselin (1995) proposed to decompose the global spatial correlation coefficients into *Local Indicators of Spatial Association* (LISA), producing a local statistic for each sampling unit compared to its surrounding units. LISA can be computed using Moran's *I* or Geary's *c* formulas (eqs. 13.1 and 13.2), and the resulting values can be plotted on maps. Fortin & Dale (2005) give examples of such maps of LISA computed for

LISA

simulated data. Readers can also run the example provided in the documentation file of function ***lisa()*** of package NCF in R.

In anisotropic situations, directional correlograms should be computed in two or several directions. Description of how the pairs of points are chosen is deferred to Subsection 13.1.3 on variograms. One may choose to represent either a single, or several of these correlograms, one for each of the aiming geographic directions, as seems fit for the problem at hand. A procedure for representing in a single figure the directional correlograms computed for several directions of a plane was proposed by Oden & Sokal (1986); Legendre & Fortin (1989) gave an example for vegetation data. Another method is illustrated in Rossi *et al*. (1992).

Another way to approach anisotropic problems is to compute two-dimensional spectral analysis. This method, described by Priestley (1964), Rayner (1971), Ford (1976), Ripley (1981) and Renshaw & Ford (1984), differs from spatial correlation analysis in the structure function it uses. As in time-series spectral analysis (Section 12.5), the method assumes the data to be stationary (second-order stationarity; i.e. no "true gradient" in the data) and made of a combination of sine patterns. A spatial correlation function $r_{dX,dY}$ for all combinations of lags (dX, dY) in the two geographic axes of a plane, as well as a periodogram with intensity I for all combinations of frequencies in the two directions of the plane, are computed. Details of the calculations are also given in Legendre & Fortin (1989), with an example.

3 — Variogram

Like correlograms, semi-variograms (called *variograms* for simplicity) decompose the spatial (or temporal) variability of observed variables among distance classes. The structure function plotted as the ordinate, called *semi-variance*, is the numerator of eq. 13.2:

$$\gamma(d) = \frac{1}{2W(d)} \sum_{h=1}^{n} \sum_{i=1}^{n} w_{hi} (y_h - y_i)^2 \quad \text{for } h \neq i \qquad \textbf{(13.9)}$$

or, for symmetric distance and weight matrices, which is the most common case:

$$\gamma(d) = \frac{1}{2W(d)} \sum_{h=1}^{n-1} \sum_{i=h+1}^{n} w_{hi} (y_h - y_i)^2 \qquad \textbf{(13.10)}$$

$\gamma(d)$ is thus a non-standardized form of Geary's c coefficient. γ may be seen as a measure of the error mean square of the estimate of y_i using a value y_h distant from it by d. The two equation forms produce the same numerical value in the case of symmetric distance and weight matrices. The calculation is repeated for different values of d. This provides the *sample variogram*, which is a plot of the empirical values of variance $\gamma(d)$ as a function of distance d.

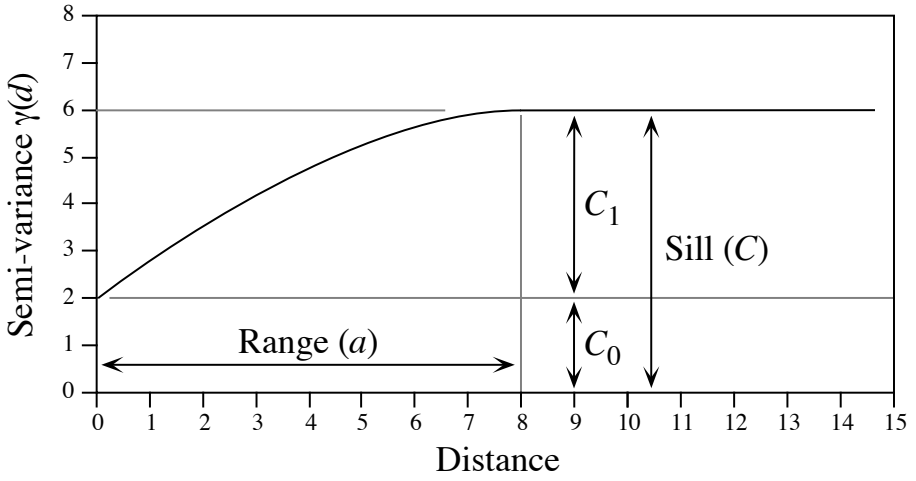


Figure 13.7 Spherical variogram model showing characteristic features: nugget effect ($C_0 = 2$ in this example), spatially structured component ($C_1 = 4$), sill ($C = C_0 + C_1 = 6$), and range ($a = 8$).

The equations usually found in the geostatistical literature look a bit different, but they correspond to the same calculations and give the same results:

$$\gamma(d) = \frac{1}{2W(d)} \sum_{i=1}^{W(d)} (y_i - y_{i+d})^2 \quad \text{or} \quad \gamma(d) = \frac{1}{2W(d)} \sum_{(h,i) \mid d_{hi} \approx d}^{W(d)} (y_h - y_i)^2$$

Both of these expressions mean that pairs of values are selected to be at distance d of each other; there are $W(d)$ such pairs for any given distance class d. The condition $d_{hi} \approx d$ means that distances may be grouped into distance classes, placing in class d the individual distances d_{hi} that are approximately equal to d. In directional variograms (below), d is a directional measure of distance, i.e. taken in a specified direction only. The semi-variance function is often called the variogram in the geostatistical literature. When computing a variogram, one assumes that the spatial correlation function applies to the entire surface under study (intrinsic stationarity, Subsection 13.1.1).

Generally, variograms tend to level off at a *sill* which is equal to the variance of the variable (Fig. 13.7); the presence of a sill implies that the data are second-order stationary. The distance at which the variance levels off is referred to as the *range* (parameter a); beyond that distance, the sampling units are not spatially correlated. The discontinuity at the origin (non–zero intercept) is called the *nugget effect*; the geostatistical origin of the method transpires in that name. It corresponds to the local variation occurring at scales finer than the sampling interval, such as sampling error, fine-scale spatial variability, and measurement error. The nugget effect is represented

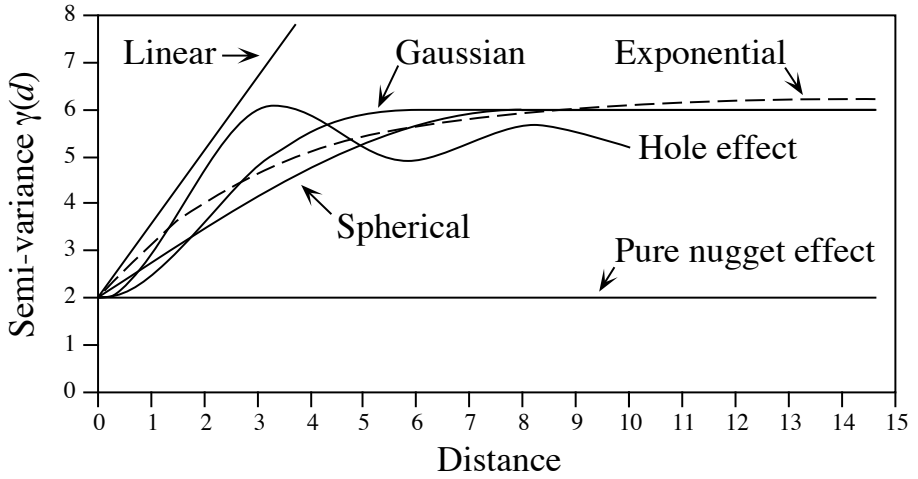


Figure 13.8 Commonly used variogram models.

by the error term ε_{ij} in spatial structure model 2 (eq. 1.2) of Subsection 1.1.1. It describes a portion of variation which is not autocorrelated, or is autocorrelated at a scale finer than can be detected by the sampling design. The parameter for the nugget effect is C_0 and the spatially structured component is represented by C_1; the sill, C, is equal to $C_0 + C_1$. The *relative nugget effect* is $C_0/(C_0 + C_1)$.

Although a sample variogram is a good descriptive summary of the spatial contiguity of a variable, it does not provide all the semi-variance values needed for kriging (Subsection 13.2.2). A model must be fitted to the sample variogram; the model will provide values of semi-variance for all the intermediate distances. The most commonly used models are the following (Fig. 13.8):

- Spherical model: $\gamma(d) = C_0 + C_1\left[1.5\frac{d}{a} - 0.5\left(\frac{d}{a}\right)^3\right]$ if $d \le a$; $\gamma(d) = C$ if $d > a$;
- Exponential model: $\gamma(d) = C_0 + C_1\left[1 - \exp\left(-3\,\frac{d}{a}\right)\right]$;
- Gaussian model: $\gamma(d) = C_0 + C_1\left[1 - \exp\left(-3\,\frac{d^2}{a^2}\right)\right]$;

• Hole effect model: $\gamma(d) = C_0 + C_1\left[1 - \frac{\sin(ad)}{ad}\right]$. An equivalent form is $\gamma(d) = C_0 + C_1\left[1 - \frac{a'\sin(d/a')}{d}\right]$ where $a' = 1/a$. $(C_0 + C_1)$ represents the value of γ towards which the dampening sine function tends to stabilize. This equation would adequately model a variogram of the periodic structures in Fig. 13.5a-b (variograms only differ from Geary's correlograms by the scale of the ordinate);

• Linear model: $\gamma(d) = C_0 + bd$ where b is the slope of the variogram model. A linear model with sill is obtained by adding the specification: $\gamma(d) = C$ if $d \geq a$;

• Pure nugget effect model: $\gamma(d) = C_0$ if $d > 0$; $\gamma(d) = 0$ if $d = 0$. The latter part applies to a point estimate. In practice, observations have the size of the sampling grain (Section 13.0); the error at that scale is always larger than 0.

Other less-frequently encountered variogram models are described in geostatistics textbooks. A model is usually chosen on the basis of the known or assumed process having generated the spatial structure. Several models may be added up to fit any particular sample variogram. Parameters may be fitted by weighted least squares; the weights are functions of the distance and the number of pairs in each distance class (Cressie, 1991); in practice, variograms are often fitted by visual estimation. Fitting a variogram model requires that the hypothesis of intrinsic stationarity be satisfied (Subsection 13.1.1).

Anisotropy

As mentioned at the beginning of Subsection 13.1.2, anisotropy is present in data when the spatial correlation function is not the same for all geographic directions considered (David, 1977; Isaaks & Srivastava, 1989). In *geometric anisotropy*, the variation to be expected between two sites distant by d in one direction is equivalent to the variation expected between two sites distant by $b \times d$ in another direction. The range of the variogram changes with direction while the sill remains constant. In a river for instance, the kind of variation expected in phytoplankton concentration between two sites 5 m apart across the current may be the same as the variation expected between two sites 50 m apart along the current even though the variation can be modelled by spherical variograms with the same sill in the two directions. Constant b is called the *anisotropy ratio* ($b = 50/5 = 10$ in the river example). This is equivalent to a change in distance units along one of the axes. The anisotropy ratio may be represented by an ellipse or a more complex figure on a map, its axes being proportional to the variation expected in each direction. In *zonal anisotropy*, the sill of the variogram changes with direction while the range remains constant. An extreme case is offered by a strip of land. If the long axis of the strip is oriented in the direction of a major environmental gradient, the variogram may correspond to a linear model (always increasing) or to a spherical model with a sill larger than the nugget effect, whereas the variogram in the direction perpendicular to it may show only random variation without spatial structure with a sill equal to the nugget effect.

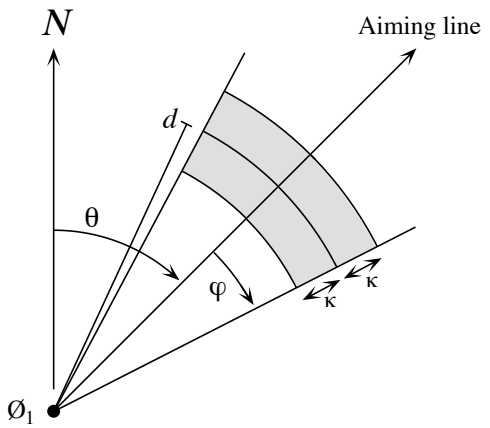


Figure 13.9 Search parameters for pairs of points in directional variograms and correlograms. From an observed study site $Ø_1$, an aiming line is drawn in the direction determined by angle θ (usually by reference to the geographic north, indicated by *N* in the figure). The angular tolerance parameter φ determines the search zone (grey) laterally whereas parameter κ sets the tolerance along the aiming line for each distance class *d*. Points within the search window (in gray) are included in the calculation of $I(d)$, $c(d)$ or $\gamma(d)$.

Directional variogram and correlogram

Directional variograms and correlograms may be used to determine whether anisotropy (defined in Subsection 13.1.2) is present in data; they may also be used to describe anisotropic surfaces or to account for anisotropy in kriging (Subsection 13.2.2). A direction of space is chosen (i.e. an angle θ, usually by reference to the geographic north) and a search is launched for the pairs of points that are within a given distance class *d* in that direction. There may be few such pairs perfectly aligned in the aiming direction, or none at all, especially when the observed sites are not regularly spaced on the map. More pairs can usually be found by looking within a small neighbourhood around the aiming line (Fig. 13.9). The neighbourhood is determined by an angular tolerance parameter φ and a parameter κ that sets the tolerance for distance classes along the aiming line. For each observed point $Ø_h$ in turn, one looks for other points $Ø_i$ that are at distance $d \pm \kappa$ from it. All points found within the search window are paired with the reference point $Ø_h$ and included in the calculation of semi-variance or spatial correlation coefficients for distance class *d*. In most applications, the search is bi-directional, meaning that one also looks for points within a search window located in the direction opposite (180°) the aiming direction. Isaaks & Srivastava (1989, their Chapter 7) propose a way to assemble directional measures of semi-variance into a single table and produce a contour map that describes the anisotropy in the data, if any; Rossi *et al*. (1992) have used the same approach for directional spatial correlograms.

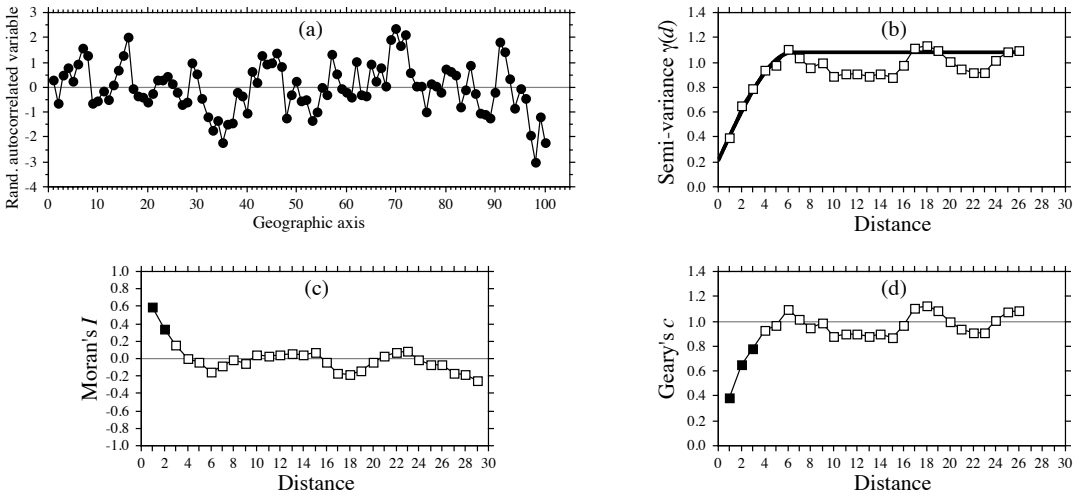


Figure 13.10 (a) Series of 100 equispaced random, spatially autocorrelated data. (b) Sample variogram, with spherical model superimposed (heavy line). Abscissa: distances between points along the geographic axis in (a). (c-d) Spatial correlograms. Dark squares: spatial correlation statistics that remain significant after progressive Bonferroni correction ($\alpha = 0.05$); white squares: non-significant values.

Numerical example. An artificial data set was produced containing random autocorrelated data (Fig. 13.10a). The data were generated using the turning bands method (David, 1977; Journel & Huijbregts, 1978); random normal deviates were autocorrelated following a spherical model with a range of 5. The sample variogram (without test of significance) and spatial correlograms (with tests) are shown in Fig. 13.10b-d. In this example, the data were standardized during data generation, so that the denominator of eq. 13.2 was 1; therefore, the sample variogram and Geary's correlogram were identical. The variogram suggests a spherical model with a range of 6 units and a small nugget effect (Fig. 13.10b).

Besides the description of spatial structures, variograms are used for several other purposes in spatial analysis. In Subsection 13.2.2, they will be the basis for interpolation by kriging. In addition, structure functions (variograms, spatial correlograms) may prove extremely useful to help determine the grain size of the sampling units and the sampling interval to be used in a survey, based upon the analysis of a pilot study. They may also be used to perform change-of-scale operations and predict the type of spatial correlation and variance that would be observed if the grain size of the sampling design were different from that actually used in a field study (Bellehumeur *et al*., 1997).

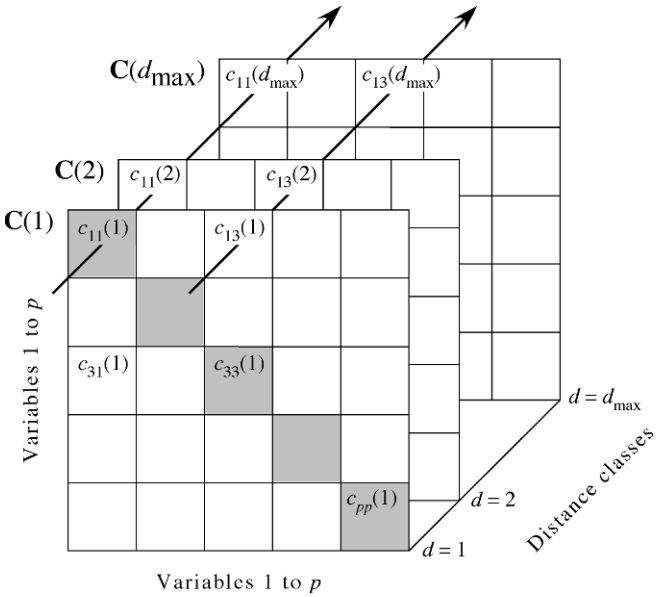


Figure 13.11 Representation of a variogram matrix **C** containing the information from all variograms and cross-variograms. **C** is composed of separate variance-covariance matrices **C**(d), each of size ($p \times p$), corresponding to one of the distance classes d. Redrawn from Wagner (2003).

4 — Multivariate variogram

Consider a multivariate matrix **Y** with n rows (sites) and p columns, e.g. species presence-absence or abundance data. A variogram $\gamma_j(d)$ for a single variable j is computed using eq. 13.10. The cross-variogram $\gamma_{jk}(d)$ between two variables j and k is now defined as follows (Isaaks & Srivastava, 1989):

$$\gamma_{jk}(d) = \frac{1}{2W(d)} \sum_{(h,i)\,|\,d_{hi} \approx d}^{W(d)} (y_{jh} - y_{ji})\,(y_{kh} - y_{ki}) \qquad \textbf{(13.11)}$$

It partitions the covariance between two variables among the distance classes d.

Each variogram and cross-variogram can be seen as a vector containing values computed for different distance classes; the largest distance class is labelled $d_{\max}$. For a multivariate response matrix **Y** of size ($n \times p$), a variogram is produced for each of the p variables and there is a cross-variogram for each of the $p(p-1)/2$ pairs of variables. These vectors can be assembled in a distance-dependent cubic symmetric variance-covariance matrix called the *variogram matrix* **C** (Myers, 1997; Fig. 13.11) with elements $c_{ij}(d) = \gamma_{jk}(d)$ (eq. 13.11). The arrows in the figure show the values

Variogram matrix

$c_{11}(d)$ used to draw the variogram $\gamma_{11}(d)$ of variable 1 and the values $c_{13}(d)$ used to draw the cross-variogram $\gamma_{13}(d)$ crossing variables 1 and 3.

Matrix **C** contains a series of square variance-covariance matrices **C**(d). Each matrix **C**(d) is of size ($p \times p$) because it is computed among the p descriptors; it contains the information for one of the distance classes d of each variogram and cross-variogram. The variance-covariance matrix $\mathbf{S_Y}$ of the p-dimensional matrix **Y** is the weighted sum of the **C**(d) matrices, showing that the set of **C**(d) matrices represents an additive decomposition of the total variance-covariance matrix $\mathbf{S_Y}$ among the distance classes d. The weights in that sum are the number of pairs of points used to compute the values in each distance class divided by the total number of pairs of points.

In order for the variances of the variables in data matrix **Y** to be additive, these must be in the same physical dimensions or standardized. This question was discussed in the first paragraph of Subsection 9.1.5. The variogram matrix can be used to plot several graphs (Wagner, 2003):

- The empirical variogram of variable j is obtained by plotting the diagonal elements $c_{jj}(d)$ (e.g. the values along the left-hand arrow in Fig. 13.11) against distances d.

- The empirical cross-variogram of variables j and k is obtained by plotting the non-diagonal elements $c_{jk}(d)$ (e.g. the values along the right-hand arrow in Fig. 13.11) against distances d.

Multivariate variogram

- Sum the diagonal elements (gray squares in Fig. 13.11) in each matrix **C**(d). Since the sum of the diagonal elements of **S** is the total variance in **Y** and the matrices **C**(d) decompose **S**, a plot of these sums against distances d is the *multivariate variogram* decomposing the total variance in **S**. An example is given in Ecological application 13.1b. Furthermore, Wagner (2003) showed that for species presence-absence data, a plot of these sums against distances d is an empirical *variogram of complementarity*, meaning the variogram of the dissimilarity in species composition. These sums are direct measures of species turnover between sites located at distances d; a higher sum of variances indicates larger differences among the sites separated by that distance than for other distances where the among-site sum of variances is lower.

- As shown in Section 4.1, the sum of all values in matrix **S** is equal to the variance of a new variable, **y**, computed as the sum by rows of all variables in **Y**. Because the matrices **C**(d) represent a decomposition of **S** among the distance classes d, one can sum all elements of each matrix **C**(d) and plot these sums against distances d to obtain a variogram of **y**. If **Y** contains species abundance data, the graph is a variogram of the total number of individuals at the sites, which can in some cases be interpreted as the total yield or the carrying capacity of the sites. If **Y** comprises species presence-absence data, a variogram of species richness is obtained (Wagner, 2003).

The statistics in multivariate variograms can be tested for significance using Mantel tests (Wagner, 2004). The tests used in function ***mso()*** of VEGAN in R, which

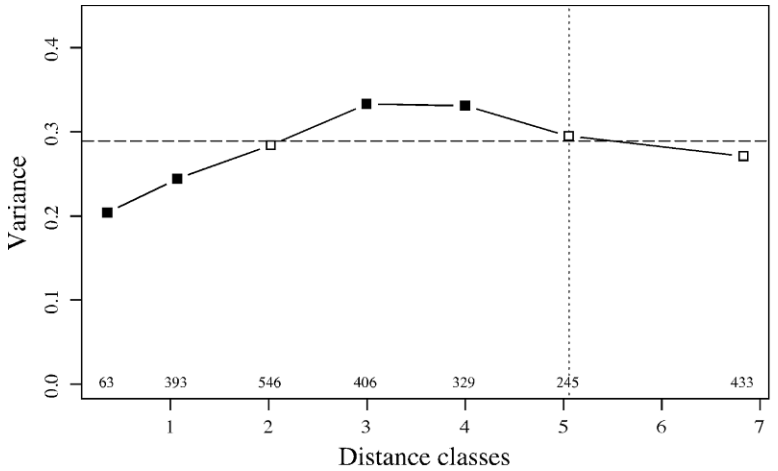


Figure 13.12 Multivariate variogram of the Hellinger-transformed and detrended mite data, computed using function ***mso()***. Dark squares: variances with p-values significant at the 5% level, after Bonferroni correction for 7 simultaneous tests. Dashed horizontal line: total variance in the data. Vertical dotted line: half the maximum number of classes; the last point, to the right of that line, includes all remaining pairs of sites and should not be interpreted. Values written above the abscissa: number of pairs involved in the calculation of each statistic.

are based on the matrix of squared distances, are identical to those used in the Mantel correlogram (Subsection 13.1.6).

Ecological application 13.1b

The oribatid mite data of Borcard & Legendre (1994), analysed in Ecological application 11.5, are used here to compute a multivariate variogram. Prior to analysis, the mite data were Hellinger-transformed (eq. 7.69) and detrended along the north-south axis of the study area to meet the stationarity assumption. Function ***mso()*** of the R VEGAN package was used to compute the variogram; see Section 13.6.

The results are shown in Fig. 13.12. The interval size of the distance classes was the distance that kept all points connected in a dbMEM analysis; this is the threshold distance (*thresh*) of Section 14.1, 1.01119 m. The horizontal line in Fig. 13.12 is the total variance in the data. It is also the weighted sum of the variances (sums of the diagonal elements) of the $\mathbf{C}(d)$ matrices over the different distance classes. Because the sum of the weights is 1, as explained in the description of the method, the dashed line is located at the weighted mean of the multivariate variogram values and can be used as a reference for their visual assessment.

The p-values were Bonferroni-corrected for 7 simultaneous tests. The variogram displays significant spatial correlation; it may correspond to a spherical or a hole model. These data will be further analysed by multiscale ordination in Section 14.4.

5 — Spatial covariance, semi-variance, correlation, cross-correlation

This subsection examines the relationships between spatial covariance, semi-variance and correlation (including cross-correlation), under the assumption of second-order stationarity, leading to the concept of cross-correlation. The assumption of second-order stationarity (Subsection 13.1.1) may be restated as follows:

- The first moment (mean of values i) of the variable has a constant and finite value:

$$\mathrm{E}[y_i] = \frac{1}{n}\sum_{i=1}^{n} y_i = m_i \qquad \textbf{(13.12)}$$

and its value does not depend on the position in the study area.

- The second moment (spatial covariance, numerator of eq. 13.1) of the variable exists (i.e. the variogram has a finite sill value):

$$C(d) = \left[\frac{1}{W(d)}\sum_{(h,i)\,|\,d_{hi}\approx d}^{W(d)} y_h y_i\right] - m_h m_i \qquad \textbf{(13.13)}$$

$$C(d) = \mathrm{E}[y_h y_i] - m^2 \quad \text{for } h, i \,|\, d_{hi} \approx d \qquad \textbf{(13.14)}$$

$h, i \,|\, d_{hi} \approx d$ means that the pairs of points h and i used to compute covariance $C(d)$ are at distances d_{hi} that are approximately equal to d. The values of $C(d)$ depend only on d and on the orientation of the distance vectors, not on their positions in the study area.

To understand eq. 13.13 as a measure of covariance, imagine the elements of the various pairs y_h and y_i written in two columns as if they were two variables. The equation for the covariance (eq. 4.4) may be written as follows, using a final division by n instead of $(n-1)$ (maximum-likelihood estimate of the covariance, which is standard in geostatistics):

$$s_{y_h y_i} = \frac{\sum y_h y_i}{n} - \frac{\sum y_h}{n}\frac{\sum y_i}{n} = \frac{\sum y_h y_i}{n} - m_h m_i$$

The overall variance ($\mathrm{Var}[y_i]$, with division by n instead of $n-1$) also exists since it is the covariance calculated for $d = 0$:

$$\mathrm{Var}[y_i] = \mathrm{E}[y_i - m_i]^2 = C(0) \qquad \textbf{(13.15)}$$

When computing the semi-variance, one only considers pairs of observations distant by d. Equations 13.9 and 13.10 are re-written as follows:

$$\gamma(d) = 0.5\ \mathrm{E}[y_h - y_i]^2 \quad \text{for } h, i \,|\, d_{hi} \approx d \qquad \textbf{(13.16)}$$

A few lines of algebra obtain the following formula:

$$\gamma(d) = \frac{\sum y_i^2 - \sum y_h y_i}{W(d)} = C(0) - C(d) \quad \text{for } h, i \mid d_{hi} \approx d \qquad \textbf{(13.17)}$$

Two properties are used in the derivation of eq. 13.17 from eq. 3.16: (1) $\sum y_h = \sum y_i$, and (2) the variance (Var[y_i], eq. 13.15) can be estimated using any subset of the observed values if the hypothesis of second-order stationarity is verified.

The correlation is the covariance divided by the product of the standard deviations. For a spatial process, the (auto)correlation is written as follows:

$$r(d) = \frac{C(d)}{s_h s_i} = \frac{C(d)}{\mathrm{Var}[y_i]} = \frac{C(d)}{C(0)} \qquad \textbf{(13.18)}$$

The right-hand formula is Moran's I (eq. 13.1). Consider the formula for Geary's c (eq. 13.2), which is the semi-variance divided by the overall variance (ignoring the fact that the variance in eq. 13.2 is computed with division by $n-1$ instead of n). The following derivation

$$c(d) = \frac{\gamma(d)}{\mathrm{Var}[y_i]} = \frac{C(0) - C(d)}{C(0)} = 1 - \frac{C(d)}{C(0)} = 1 - r(d)$$

shows that Geary's c is one minus the coefficient of spatial (auto)correlation (ignoring again the division by $n-1$ instead of n). In a graph, the semi-variance and Geary's c coefficient have exactly the same shape (e.g. Fig. 13.10, b and d); only the ordinate scales may differ if Var[y_i] is not 1. An autocorrelogram plotted using $r(d)$ has the exact reverse shape as a Geary correlogram. The important conclusion is that the plots of semi-variance, covariance, Geary's c coefficient, and $r(d)$, are equivalent to characterize spatial structures under the hypothesis of second-order stationarity (Bellehumeur & Legendre, 1998).

Cross-covariances may also be computed from eq. 13.13, using values of *two different variables* observed at locations distant by d (Isaaks & Srivastava, 1989). Equation 13.18 leads to a formula for cross-correlation that may be used to plot cross-correlograms; the construction of the cross-correlation statistic is the same as for time series (eq. 12.9). With transect data, the result is similar to that of eq. 12.9. However, the programs designed to compute spatial cross-correlograms do not require the data to be equispaced, contrary to programs for time-series analysis. The theory is presented by Rossi *et al*. (1992), as well as applications to ecology.

Ecological application 13.1c

A survey was conducted on a homogeneous sandflat in the Manukau Harbour, New Zealand, to identify the scales at which spatial heterogeneity could be detected in the distribution of adult and juvenile bivalves (*Macomona liliana* and *Austrovenus stutchburyi*), as well as indications of

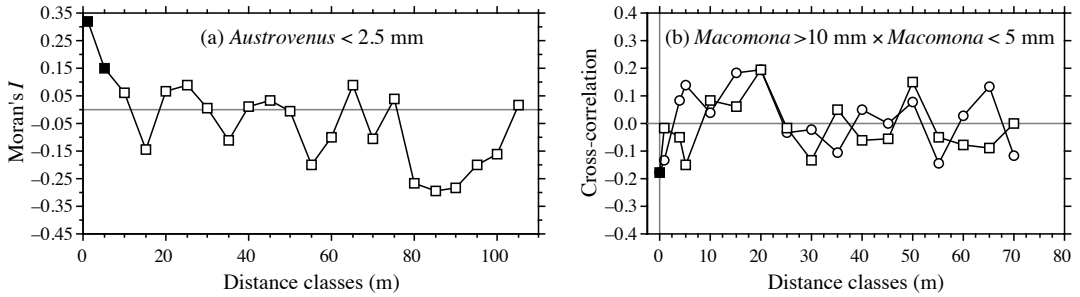


Figure 13.13 (a) Spatial autocorrelogram for juvenile *Austrovenus* densities. (b) Cross-correlogram for adult-juvenile *Macomona* interactions, folded about the ordinate: circles = positive lags, squares = negative lags. Dark symbols: correlation statistics that are significant after progressive Bonferroni correction ($\alpha = 0.05$). Redrawn from Hewitt *et al*. (1997).

adult-juvenile interactions within and between species. The results were reported by Hewitt *et al*. (1997); see also Ecological application 13.2. Sampling was conducted along transects established at three sites located within a 1-km^2 area; there were two transects at each site, forming a cross. This way, there were transects perpendicular to the direction of tidal flow, and others parallel. Sediment cores (10 cm diam., 13 cm deep) were collected using a nested sampling design; the basic design was a series of cores 5 m apart, but additional cores were taken 1 m from each of the 5-m-distant cores. This design provided several comparisons in the short distance classes (1, 4, 5, and 6 m). Using transects instead of rectangular areas allowed relatively large distances (150 m) to be studied, given the allowable sampling effort. Nested sampling designs have also been advocated by Fortin *et al*. (1989) and by Bellehumeur & Legendre (1998).

Spatial correlograms were used to identify scales of variation in bivalve concentrations. The Moran correlogram for juvenile *Austrovenus*, computed for the three transects perpendicular to the direction of tidal flow, displayed significant spatial correlation at distances of 1 and 5 m (Fig. 13.13a). The same pattern was found in the transects parallel to tidal flow. Figure 13.13a also indicates that the range of influence of spatial correlation was about 15 m. This was confirmed by plotting bivalve concentrations along the transects: LOWESS smoothing of the graphs (Subsection 10.3.8) showed patches of about 25-30 m in diameter (Hewitt *et al*., 1997, their Figs. 3 and 4).

Cross-correlograms were computed to detect signs of adult-juvenile interactions. In the comparison of adult (> 10 mm) to juvenile *Macomona* (< 5 mm), a significant negative cross-correlation was identified at 0 m in the direction parallel to tidal flow (Fig. 13.13b); correlation was not significant for the other distance classes. As in time series analysis, the cross-correlation function is not symmetrical; the correlation obtained by comparing values of $\mathbf{y}_1$ to values of $\mathbf{y}_2$ located at distance d on their right is not the same as when values of $\mathbf{y}_2$ are compared to values of $\mathbf{y}_1$ located at distance d on their right, except for $d = 0$. In Fig. 13.13b, the cross-correlogram is folded about the ordinate (compare to Fig. 12.9). Contrary to time series analysis, it is not useful in spatial analysis to discuss the direction of lag of a variable with respect to the other unless one has a specific hypothesis to test.

6 — Multivariate Mantel correlogram

Sokal (1986) and Oden & Sokal (1986) found an ingenious way to compute a correlogram for multivariate data, using the normalized Mantel statistic r_M and test of significance (Subsection 10.5.1). This method is useful, in particular, to describe the spatial structure of species assemblages.

The principle is to quantify the ecological relationships among sampling sites by means of a matrix **Y** of multivariate similarities or distances (using, for instance, coefficients S_{17} or D_{14} in the case of species abundance data), and compare **Y** to a *model matrix* **X** (Subsection 10.5.1), which is different for each geographic distance class (Fig. 13.14).

• For distance class 1 for instance, pairs of neighbouring stations (that belong to the first class of geographic distances) are coded 1, whereas the remainder of matrix **X**(1) contains zeros. A first Mantel statistic (r_{M1}) is calculated between **Y** and **X**(1).

• The process is repeated for the other distance classes d, building each time a model-matrix **X**(d) and recomputing the normalized Mantel statistic. Matrix **X**(d) may contain 1's for pairs that are in the given distance class, or the code value for that distance class (d) (as in Fig. 13.14), or any other value different from zero; all coding methods lead to the same value of the normalized Mantel statistic r_M.

The Mantel statistics, plotted against distance classes, produce a multivariate correlogram. Each value is tested for significance in the usual way, using either permutations or Mantel's normal approximation (Box 10.2). Computation of standardized Mantel statistics assumes second-order stationarity. Borcard & Legendre (2012) have shown that for univariate data, the tests of significance in a Mantel correlogram computed on the matrix of *squared* Euclidean distances was equivalent to the tests in a Geary's c correlogram. Using numerical simulations, they also showed that the power of the test in Mantel correlograms was high for multivariate data.

A multivariate correlogram can be computed with function ***mantel.correlog()*** in R; see Section 13.6. If the calculation is based upon a *squared* Euclidean distance matrix, the Mantel test results in the multivariate correlogram are identical to the Mantel test results computed by the multivariate variogram function ***mso()***, provided that the distance classes are the same (Borcard & Legendre, 2012). As in the case of univariate correlograms (above), one is advised to use some form of correction for multiple testing (Box 1.3) before interpreting multivariate correlograms and variograms.

Numerical example. Consider again the 10 sampling sites of Fig. 13.4. Assume that species assemblage data were available and produced similarity matrix **S** of Fig. 13.14. Matrix **S** played here the role of $\mathbf{D_Y}$ in the computation of Mantel statistics. Were the species data autocorrelated? Distance matrix **D**, already divided into 6 classes in Fig. 13.4, was recoded into a series of model matrices **X**(d) (d = 1, 2, etc.). In each of these, the pairs of sites that were in the given distance class received the value d, whereas all other pairs received the value 0. Mantel statistics were computed between **S** and each of the **X**(d) matrices in turn; positive and significant Mantel

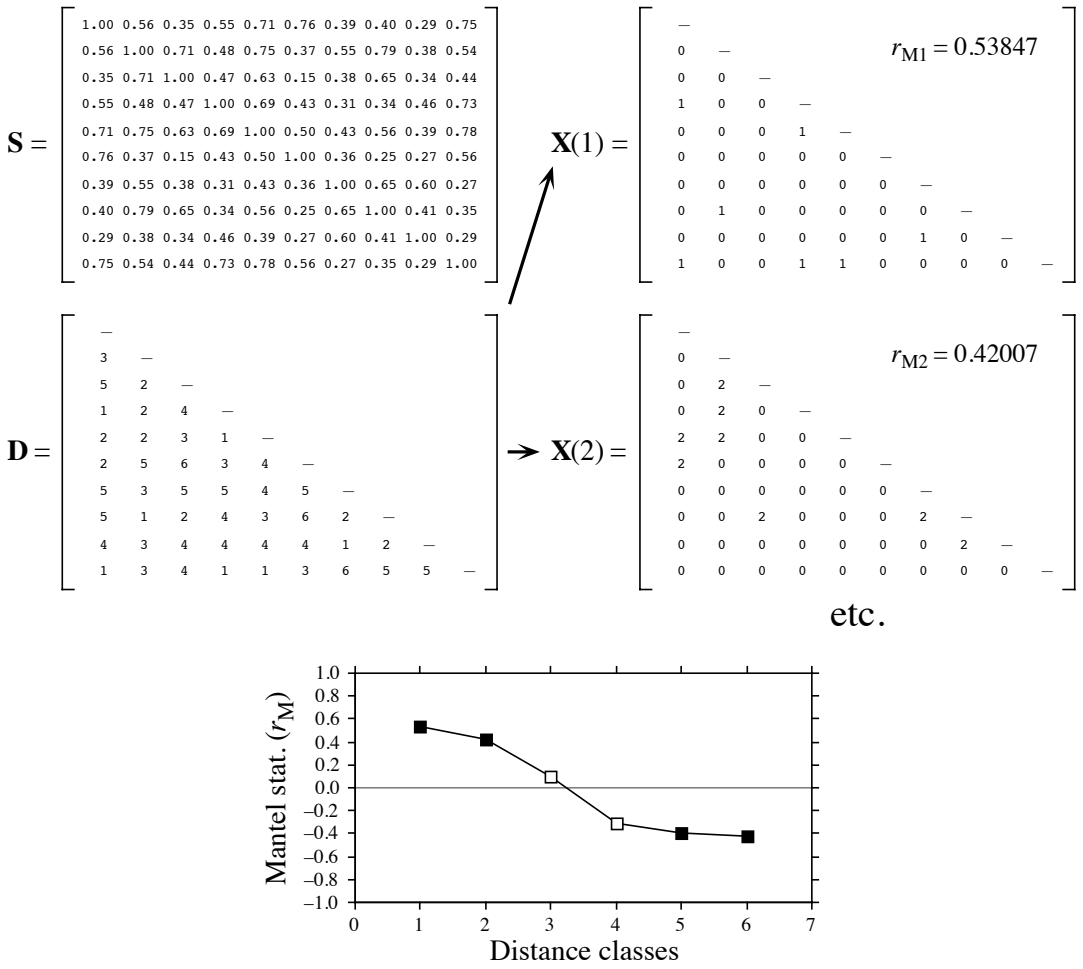


Figure 13.14 Construction of a Mantel correlogram for a similarity matrix **S** ($n = 10$ sites). The matrix of geographic distance classes **D**, from Fig. 13.4d, gives rise to model matrices **X**(1), **X**(2), etc. for the various distance classes d. These are compared, in turn, to matrix **S** using standardized Mantel statistics (r_{Md}). Dark squares in the correlogram: Mantel statistics that are significant after progressive Bonferroni correction ($\alpha = 0.05$).

statistics indicate positive spatial correlation in the present case. The statistics were tested for significance using 999 permutations and plotted against distance classes d to form the Mantel correlogram. The progressive Bonferroni method was used to account for multiple testing because interest was primarily in detecting spatial correlation in the first distance classes.

Before computing the Mantel correlogram, one must assume that the condition of second-order stationarity is satisfied. This condition is more difficult to explain in the case of

multivariate data; it means essentially that the surface is uniform in (multivariate) mean, variance and covariance at broad scale. The correlogram illustrated in Fig. 13.14 suggests the presence of a gradient. If the condition of second-order-stationarity is satisfied, this means that the gradient detected by this analysis is a part of a larger, autocorrelated spatial structure. This was called a "false gradient" in the numerical example of Subsection 13.1.2.

When $\mathbf{D_Y}$ is a similarity matrix and distance classes are coded as described above, positive Mantel statistics correspond to positive spatial correlation; this is the case in the numerical example. When the values in $\mathbf{D_Y}$ are distances instead of similarities, or if the 1's and 0's are interchanged in matrix $\mathbf{X}$, the signs of all Mantel statistics are changed. One should always specify whether positive spatial correlation is expressed by positive or negative values of the Mantel statistics when presenting Mantel correlograms. Mantel correlograms have been computed for real data by Legendre & Fortin (1989), Le Boulengé *et al*. (1996), and Fortin & Dale (2005).

13.2 Maps

The most basic step in spatial pattern analysis is the production of maps displaying the spatial distributions of values of the variable(s) of interest. Furthermore, maps are essential to help interpret spatial structure functions (Section 13.1).

Several methods are available in mapping programs. The final product of modern computer programs may be a contour map, a mesh map (such as Figs. 13.15b and 13.18b), a raised contour map, a shaded relief map, and so on. The present section is not concerned with the graphic representation of maps, but instead with the ways mapped values are obtained. Spatial interpolation methods have been reviewed by Lam (1983).

Geographic information systems (GIS) are widely used nowadays, especially by geographers and increasingly by ecologists, to manage complex data corresponding to points, lines, and surfaces in space. The present section is not an introduction to these complex systems. It only aims at presenting the most widespread methods for mapping univariate data (i.e. a single variable y). The spatial analysis of multivariate data (multivariate matrix $\mathbf{Y}$) is deferred to Sections 13.3 to 13.5.

Beware of non-additive variables such as pH, logarithms of counts of organisms, diversity measures, and the like (Subsection 1.4.2). Maps of such variables, produced by trend-surface analysis or interpolation methods, should be interpreted with caution because the interpolated values of such variables only make sense by reference to sampling units of the same size as those used in the original sampling design. Block kriging (Subsection 13.2.2) for blocks representing surfaces or volumes that differ from the grain of the observed data does not make sense for non-additive variables.

1 — Trend-surface analysis

Trend-surface analysis is the oldest method for producing smoothed maps. In this method, estimates of the variable at given locations are not obtained by interpolation, as in the methods presented in Subsection 13.2.2, but through a regression model calibrated over the entire study area.

In 1914, W. S. Gosset, writing under the pseudonym Student, proposed to express observed values as a polynomial function of time and mentioned that it could be done for spatial data as well. This is also one of the most powerful tools of spatial pattern analysis, and certainly the easiest to use. The objective is to express a response variable y as a nonlinear function of the geographic coordinates X and Y of the sampling sites where the variable was observed:

$$y = f(\mathrm{X}, \mathrm{Y})$$

In many cases, a polynomial of X and Y with cross-product terms is used; trend-surface analysis is then an application of polynomial regression (Subsection 10.3.4) to spatially-distributed data. For example a relatively complex, but smooth surface might be fitted to a variable using a third-order polynomial with 10 parameters (b_0 to b_9):

$$\hat{y} = f(\mathrm{X}, \mathrm{Y}) = b_0 + b_1\mathrm{X} + b_2\mathrm{Y} + b_3\mathrm{X}^2 + b_4\mathrm{XY} + b_5\mathrm{Y}^2 + b_6\mathrm{X}^3 + b_7\mathrm{X}^2\mathrm{Y} + b_8\mathrm{XY}^2 + b_9\mathrm{Y}^3 \quad \textbf{(13.19)}$$

Note the distinction between the response variable y, which may represent a physical or biological variable, and the Cartesian geographic coordinate Y. Using polynomial regression, trend-surface analysis produces an equation that is linear in its parameters, although the response of y to the explanatory variables in matrix $\mathbf{X} = [\mathrm{X},\mathrm{Y}]$ may be nonlinear. If variables y, X and Y have been centred on their respective means prior to model fitting, the model has an intercept of 0 by construct; hence parameter b_0 does not have to be fitted and it can be removed from the model.

Numerical example. The data from Table 10.6 are used here to illustrate the method of trend-surface analysis. The dependent variable of the analysis, y, is Ma, which was the log-transformed ($\log_e(x + 1)$) concentrations of aerobic heterotrophic bacteria growing on marine agar at salinity of 34 psu. The explanatory variables are the X and Y geographic coordinates of the sampling sites (Fig. 13.15a). The steps of the calculations are the following:

• Centre the geographic coordinates on their respective means. The reason for centring X and Y is given in Subsection 10.3.4; the amount of variation explained by a trend-surface equation is not changed by a translation (centring) of the spatial coordinates across the map.

• Determine the order of the polynomial equation to be used. A first-degree regression equation of Ma as a function of the geographic coordinates X and Y alone would only represent the linear variation of Ma with respect to X and Y; in other words, a flat surface, possibly sloping with respect to X, Y, or both. With the present data, the first-degree regression equation was not significant (R^2 = 0.02), meaning that there was no significant linear geographic trend to be described in the data. A regression equation incorporating the second-degree monomials (X^2, XY and Y^2) together with X and Y would be appropriate to model a surface presenting a single

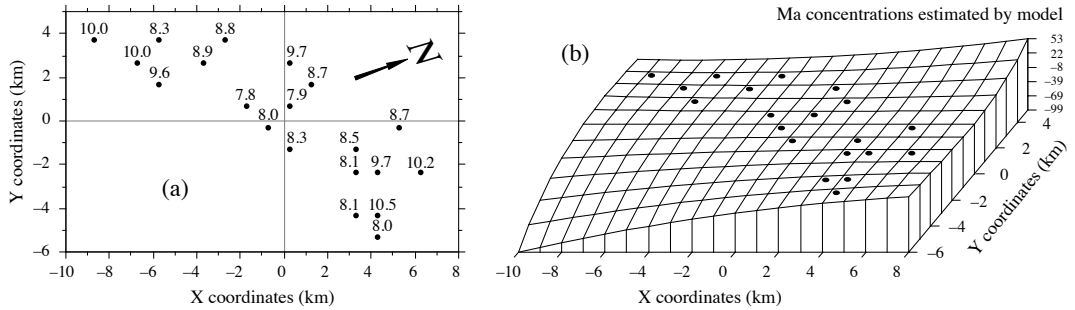


Figure 13.15 Variable Ma (log-transformed concentrations of aerobic heterotrophic bacteria growing on marine agar at salinity of 34 psu) at 20 sites in the Thau coastal lagoon, France, on 25 October 1988. (a) Map of the sampling sites with respect to arbitrary geographic coordinates X and Y. The observed values of Ma, from Table 10.6, are also shown. The *N* arrow points to the north. (b) Trend-surface map; the vertical axis gives the values of Ma estimated by the polynomial regression equation. Dots represent the sampling sites.

large bump or trough. Again, this did not seem to be the case with the present data since the second-degree equation was not significant ($R^2 = 0.39$). An equation incorporating the third-degree, fourth-degree, etc. terms would be able to model structures of increasing complexity and refinement. The cost, however, is a loss of degrees of freedom for every new monomial in the equation; trend-surface analysis using high-order equations thus requires a large number of observed sampling sites. In the present example, the polynomial was limited to the third degree, for a total of 9 terms; this is a large number of terms, considering that the data only contained 20 sampling sites.

• Using the values of coordinates X and Y, calculate the terms of the third-degree polynomial, by combining variables X and Y as follows: X^2, X×Y, Y^2, X^3, X^2×Y, X×Y^2, Y^3. Alternatively, one could compute a third degree orthogonal polynomial of the geographic coordinates. Ordinary and orthogonal polynomials can both be computed by function ***poly()*** in R (Section 13.6).

• Compute the multiple regression equation. The model obtained using all 9 regressors had $R^2 = 0.87$, but several of the partial regression coefficients were not significant.

• Remove nonsignificant terms. The linear terms may be important to express a linear gradient; the quadratic and cubic terms may be important to model more complex surfaces. Nonsignificant terms should not be left in the model, except when they are required for comparison purpose. Nonsignificant terms were removed one by one (backward elimination, Subsection 10.3.3) until all terms (monomials) in the polynomial equation were significant. The resulting trend-surface equation was highly significant ($R^2 = 0.81$, $p < 0.0001$):

$$\hat{y} = 8.13 - 0.16\,XY - 0.09\,Y^2 + 0.04\,X^2Y + 0.14\,XY^2 + 0.10\,Y^3$$

Remember, however, that tests of significance are too liberal with autocorrelated data, due to the non-independence of residuals, with the consequence that nonsignificant relationships are declared significant too often (Subsection 1.1.2).

• Lay out a regular grid of points (X', Y') and, using the regression equation, compute forecasted values ($\hat{y}'$) for these points. Plot a map (Fig. 13.15b) using the file with (X', Y', and $\hat{y}'$). Values estimated by a trend-surface equation at the study sites do not coincide with the values observed at these sites; regression is not an exact interpolator, contrary to kriging (Subsection 13.2.2).

Different features could be displayed by rotating the figure. The orientation chosen in Fig. 13.15b does not clearly show that the values along the long axis of the Thau lagoon are smaller near the centre than at the ends. It displays, however, the wavy structure of the data from the lower left-hand to the upper right-hand corner, which is roughly the south-to-north direction. The figure also clearly indicates that one should refrain from interpreting extrapolated data values, i.e. values located outside the area that has actually been sampled. In the present example, the values forecasted by the model in the lower left-hand and the upper right-hand corners (–99 and +53, respectively) are meaningless for log bacterial concentrations. Within the area where real data are available, however, the trend-surface model provides a good visual representation of the broad-scale spatial variation of the response variable.

Examination of the residuals is essential to make sure that the model is not missing some salient feature of the data. If the trend-surface model has extracted all the spatially-structured variation of the data, given the scale of the study, residuals should look random when plotted on a map and a correlogram of residuals should be non-significant. With the present data, residuals were small and did not display any recognizable spatial pattern.

A cubic trend-surface model is often appropriate with ecological data. Consider an ecological phenomenon that starts at the mean value of the response variable *y* at the left-hand border of the sampled area, increases to a maximum, then goes down to a minimum, and comes back to the mean value at the right-hand border. The amount of space required for the phenomenon to complete a full cycle — whatever the shape it may take — is its extent (Section 13.0). Using trend-surface analysis, such a phenomenon would be correctly modelled by a third-degree trend surface equation.

The degree of the polynomial that is appropriate to model a phenomenon is partly predictable. If the extent is of the same order as the size of the study area (say, in the X direction), the phenomenon will be correctly modelled by a polynomial of degree 3, which has two extreme values, a minimum and a maximum. If the extent is larger than the study area, a polynomial of degree less than 3 is sufficient; degree 2 if there is only one maximum, or one minimum, in the sampling window; and degree 1 if the study area is limited to the increasing, or decreasing, portion of the phenomenon. Conversely, if the scale of the phenomenon controlling the variable is smaller than the study area, more than two extreme values (minima and maxima) will be found, and a polynomial of order larger than 3 is required to model it correctly. The same reasoning applies to the X and Y directions when using a polynomial combining the X and Y geographic coordinates. So, using a polynomial of degree 3 acts as a filter: it is a way of looking for phenomena that are of the same extent, or larger, than the study area.

An assumption must be made when using the method of trend-surface analysis: that all observations form a single statistical population, subjected to one and the same generating process, and can consequently be modelled using a single polynomial equation of the geographic coordinates. Evidence to that effect may be available prior

to the analysis. When that is not the case, the hypothesis of homogeneity may be supported by examining the regression residuals (Subsection 10.3.1). When there are indications that values in different regions of the geographic space obey different processes (e.g. different geology, action of currents or wind, or influence of other physical variables), the study area should be divided into regions, to be modelled by separate trend-surface equations.

Polynomial regression, as used in the numerical example above, is a good first approach to fitting a model to a surface when the shape to be modelled is unknown, or known to be simple. In some instances, however, it may not provide a good fit to the data; trend-surface analysis must then be conducted using nonlinear regression (Subsection 10.3.6), which requires that an appropriate numerical model be provided to the estimation program. Consider the example of the effect of some human-generated environmental disturbance at a site, the indicator variable being the number of species. The response, in that case, is expected to be stronger near the impacted site, tapering off as one gets farther away from it.

Assuming that data were collected along a transect (a single geographic coordinate X) and that the impacted site was near the centre of the transect, a polynomial equation would not be appropriate to model an inverse-squared-distance diffusion process (Fig. 13.16a). An equation of the form:

$$\hat{y} = b_0 + \frac{b_1 X^2}{b_2 X^2 + 1}$$

would provide a much better fit (Fig. 13.16b). The minimum of that equation is b_0; this value occurs when $X = 0$. The maximum, b_1/b_2, is reached asymptotically as X becomes large in either the positive or negative direction. For data collected in different directions around the impacted site, a nonlinear trend-surface equation with similar properties would be of the form:

$$\hat{y} = b_0 + \frac{b_1 X^2 + b_2 Y^2}{b_3 X^2 + b_4 Y^2 + 1}$$

where X and Y are the coordinates of the sites in geographic space.

Trend-surface analysis is appropriate for describing broad-scale spatial trends in data, but it does not produce accurate fine-grained maps of the spatial variation of a variable. Other methods described in Chapter 14 allow researchers to model variation at finer scales. In some studies, the broad-scale trend itself is of interest; this is the case in the numerical example above and in Ecological application 13.2. In other situations, and especially in studies that cover large geographic expanses, the broad-scale trend may be already known and understood; researchers interested in geographic variation patterns may want to conduct analyses on detrended data, i.e. data from which the

Detrending

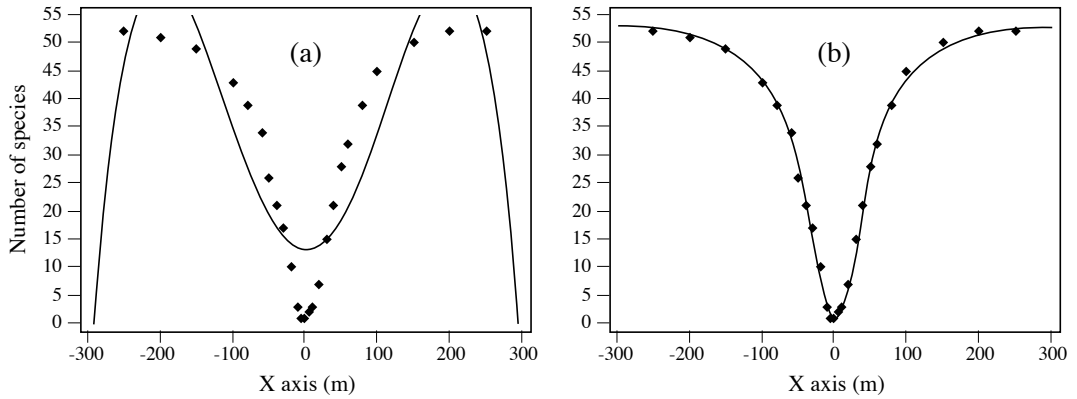


Figure 13.16 (a) Artificial data representing the number of species around the site of an environmental disturbance (located at X = 0) are not well-fitted by a 4th-order polynomial equation of the X coordinates ($R^2 = 0.7801$). (b) They are well-fitted by the following inverse-squared-distance diffusion equation: $\hat{y} = 1 + [0.0213X^2 / (0.0004X^2 + 1)]$ ($R^2 = 0.9975$).

broad-scale trend has been removed. Detrending a variable may be achieved by computing the residuals from a trend-surface equation of sufficient order, as in time-series analysis (Section 12.2).

If there is replication at each geographical observation point, it is possible to perform a test of goodness-of-fit of a trend-surface model (Draper and Smith, 1981; Legendre & McArdle, 1997). By comparing the observed error mean square after fitting the trend surface to the error mean square estimated by the among-replicate within-location variation, one can test if the model fits the data properly. The latter variation is computed from the deviations from the means at the various locations; it is the residual mean square of an ANOVA testing for differences among locations. When the trend surface goes through the expected values at the various locations, these two error mean squares are not much different, and their F-ratio does not significantly differ from 1. If, on the contrary, the fitted surface does not follow the major features of the variation among locations, the deviations of the data from the fitted trend-surface values are larger than the residual within-location variation. The F-statistic is then significantly larger than 1, indicating that the trend surface is misrepresenting the variation among locations.

Numerical example. Consider the artificial data in Fig. 13.17. Variable X represents a geographic axis along which sampling has taken place at 6 sites with replication. Variable y was constructed using equation $y = 2.5X - 0.3X^2 + \varepsilon$, where ε is a random standard normal deviate [N(0,1)]. A quadratic trend-surface model of X was fitted to the data. The residual mean square, or "error mean square after fitting the trend surface", was $MS_1 = 0.84909$ ($\nu = 27$). An analysis of variance was conducted on y using the grouping of data into 6 sites as the classification

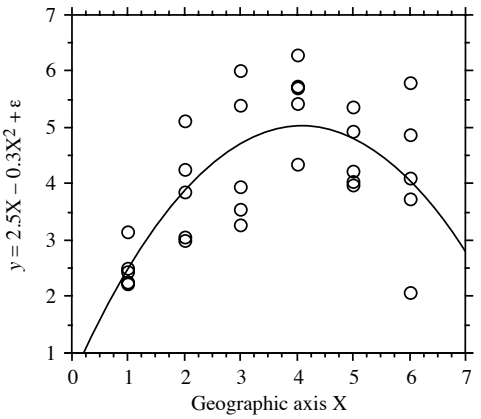


Figure 13.17 Artificial data representing sampling along a geographic axis X with 5 replicates at each site; $n = 30$. The F-test of goodness-of-fit indicates that the trend-surface equation $\hat{y} = 0.562 + 2.184X - 0.267X^2$ ($R^2 = 0.4899$) fits the data properly.

criterion. The residual mean square obtained from the ANOVA was $MS_2 = 0.87199$ ($\nu = 24$). The ratio of these two mean squares gave an F-statistic:

$$F = \frac{MS_1}{MS_2} = \frac{0.84909}{0.87199} = 0.97374$$

which was tested against $F_{\alpha=0.05(27,\,24)} = 1.959$. The F-statistic was not significantly different from 1 ($p = 0.530$), which indicated that the model fitted the data properly.

The trend-surface analysis was recomputed using a linear model of X. The model obtained was $\hat{y} = 3.052 + 0.316X$ ($R^2 = 0.1941$). MS_1 in this case was 1.29358 ($\nu = 28$). The F-ratio $MS_1/MS_2 = 1.29358/0.87199 = 1.48348$. The reference value was $F_{0.05(28,\,24)} = 1.952$. The probability associated with the F-ratio, $p = 0.165$, indicated that this model still fitted the data, which were constructed to contain a linear term ($2.5X$ in the construction equation) as well as a quadratic trend (term $-0.3X^2$), but the fit was poorer than with the quadratic polynomial model, which was capable of accounting for both the linear and quadratic trends.

This numerical example shows that trend-surface analysis may be applied to data collected along a transect; the "trend surface" is one-dimensional in that case. The numerical example at the end of Subsection 10.3.4 is another example of a trend-surface analysis of a dependent variable, salinity, with respect to a single geographic axis (Fig. 10.9). Trend-surface analysis may also be used to model data in three-dimensional geographic space (geographic coordinates X, Y and Z, where Z is either altitude or depth), or with one of the dimensions representing time. Section 13.5 will show how the analysis may be extended to a multivariate dependent data matrix **Y**.

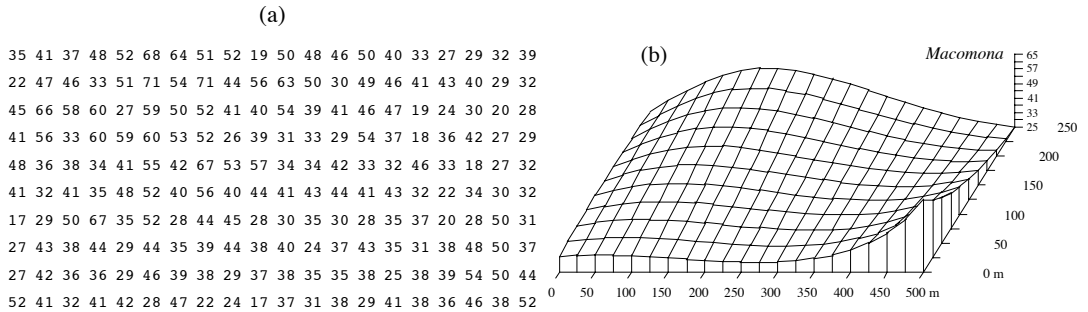


Figure 13.18 *Macomona* > 15 mm at 200 sites in Manukau Harbour, New Zealand, on 22 January 1994. (a) Actual counts at sampling sites in 200 regular grid cells; in the field, sites were not perfectly equispaced. (b) Map of the trend-surface equation explaining 32% of the spatial variation in the data. The values estimated from the trend-surface equation (log-transformed data) were back-transformed to raw counts before plotting. Modified from Legendre *et al*. (1997).

Haining (1987) described alternative methods for estimating the parameters of a trend-surface model when the residuals are spatially autocorrelated; in that case, least-squares estimation of the parameters is inefficient and standard errors as well as tests of significance are biased. Haining's methods allow one to recognize three components of spatial variation corresponding to the site, local, and regional scales, respectively.

Ecological application 13.2

A survey was conducted at 200 locations within a fairly homogeneous 12.5 ha rectangular sandflat area in Manukau Harbour, New Zealand, to identify factors that controlled the spatial distributions of the two dominant bivalves, *Macomona liliana* Iredale and *Austrovenus stutchburyi* (Gray), and to look for evidence of adult-juvenile interactions within and between species. Results were reported by Legendre *et al*. (1997). Most of the broad-scale spatial structure detected in the bivalve counts (two species, several size classes) was explained by the physical and biological variables. Results of principal component analysis and spatial regression modelling suggested that different factors controlled the spatial distributions of adults and juveniles. Larger size classes of both species displayed significant spatial structures, with physical variables explaining some but not all of this variation; the spatial patterns of the two species differed, though. Smaller organisms were less strongly spatially structured; virtually all of their spatial structure was explained by physical variables.

Highly significant trend-surface equations were found for all bivalve species and size classes (log-transformed data), indicating that the spatial distributions of the organisms were not random, but highly organised at the scale of the study site. The trend-surface models for smaller animals had much smaller coefficients of determination (R^2 = 0.10-0.20) than for larger animals (R^2 = 0.30-0.55). The best models, i.e. those with the highest R^2, were for the *Macomona* > 15 mm and *Austrovenus* > 10 mm. The coefficients of determination were consistently higher for *Austrovenus* than for *Macomona*, despite the fact that *Macomona* were usually far more

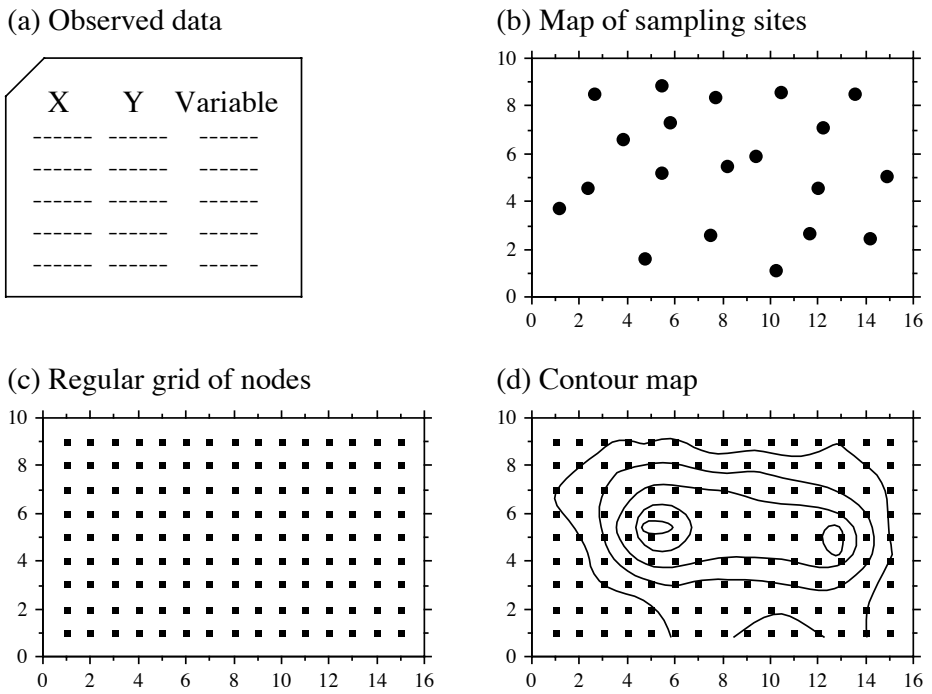


Figure 13.19 Summary of the interpolation procedure.

numerous than *Austrovenus*. A map illustrating the trend-surface equation is presented for the largest *Macomona* size class (Fig. 13.18); the field counts are also shown for comparison.

2 — Interpolated maps

In the family of interpolated map methods, the value of the variable at a point location on a map is estimated by local interpolation, using only the observations available in the vicinity of the point of interest. In this respect, interpolation mapping differs from trend surface analysis (Subsection 13.2.1), where estimates of the variable at given locations were not obtained by interpolation, as in the present subsection, but through a statistical model whose parameters were estimated from all observations in the study area. Figure 13.19 illustrates the principle of interpolation mapping. A regular grid of nodes (Fig. 13.19c) is defined over the area that contains the study sites $Ø_i$ (Fig. 13.19a, b). Interpolation assigns a value to each point of the grid. This is the single most important step in mapping. Following that, results may be represented in the form of contours (e.g. Fig. 13.19d) with or without colours or shades, or three-dimensional constructs such as Fig. 13.18b.

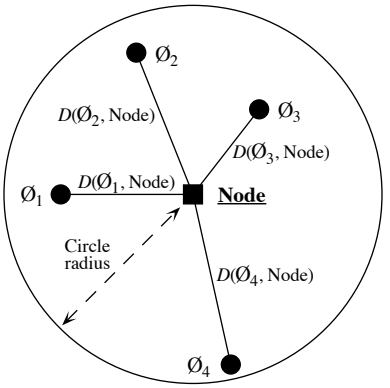


Figure 13.20 To estimate the value at a grid node (square), draw a search circle around it and consider the observed points ($Ø_i$) found within the circle. Observed points are separated from the node by distances $D(Ø_i, \text{Node})$.

Estimating the value corresponding to each grid node may be done in different ways. Different interpolation methods may produce maps that look different; this is also the case when using different parameters with a same method (e.g. different exponents in inverse-distance weighting).

The most simple rule would be to give, to each node of the grid, the value of the observation which is the closest to it. The end result is a division of the map into Voronoï polygons (Subsection 13.3.1) displaying a "zone of influence" drawn around each observation. Another simple solution consists in dividing the map into Delaunay triangles (Subsection 13.3.1). There is an observed value y_i at each site $Ø_i$. A triangular portion of plane, adjusted to the points $Ø_i$ that form the vertices (corners) of a Delaunay triangle, provides interpolated values for all points inside the triangle. Maps obtained using these solutions are shown in Chapter 11 of Isaaks & Srivastava (1989).

Alternatively, one may draw a "search circle" (or an ellipsoid for anisotropic data) around each grid node (Fig. 13.20). The radius of the circle may be determined in either of two ways. (1) One may fix a minimum number of observed points that must be included in the interpolation for each grid node; or (2) one may use the "distance of influence of the process" found by correlogram or variogram analysis (Section 13.1). The estimation procedure is repeated for each node of the grid. Several methods of interpolation may be used.

• Mean. — Consider all the observed study sites found within the circle; assign the mean of these values to the grid node. This method does not produce smooth maps; discontinuities in neighbouring grid node values occur as observed points move in or out of the search circle.

• Inverse-distance weighting. — Consider the observation sites found within the circle and calculate a weighted mean value, using the formula:

$$\hat{y}_{\text{Node}} = \sum_i w_i y_i \qquad \textbf{(13.20)}$$

where y_i is the value observed at point $\emptyset_i$ and weight w_i is the inverse of the distance (D) from point $\emptyset_i$ to the grid node to be estimated. The inverse distances, to some power k, are scaled by the sum of the weights for all points $\emptyset_i$ in the estimation, so as to produce values that are consistent with the values observed at points $\emptyset_i$ (unbiasedness condition):

$$w_i = \left(\frac{1}{D(\emptyset_i, \text{Node})^k} \right) \Big/ \sum_i \frac{1}{D(\emptyset_i, \text{Node})^k} \qquad \textbf{(13.21)}$$

A commonly-used exponent is $k = 2$. This corresponds, for instance, to the decrease in energy of waves dispersing across a two-dimensional surface. The greater the value of k, the less influence distant data points have on the value assigned to the grid node. This method produces smooth values over the grid of nodes. The range of estimated values is smaller than the range of observed data so that, contrary to trend-surface analysis (Fig. 13.15b), inverse-distance weighting does not produce meaningless values in the parts of the map beyond the area that was actually sampled. When the observation sites $\emptyset_i$ do not form a regular or nearly regular grid, however, this interpolation method may generate features in maps that have little to do with reality. As a consequence, inverse-distance weighting is not recommended in that situation.

• Weighted polynomial fitting. — In this method, a trend-surface equation (Subsection 13.2.1) is adjusted to the observed data points within the search circle, weighting each observation $\emptyset_i$ by the inverse of its distance (using some appropriate power k) to the grid node to be estimated. A first or second-order polynomial equation is usually used. The value estimated by the polynomial equation for the coordinates of a grid node is denoted z_{Node}. This method suffers from the same problem as inverse distance weighting with respect to observation sites $\emptyset_i$ that do not form a regular or nearly regular grid of points.

Kriging

• Kriging. — This is the mapping tool in the toolbox of geostatisticians. The method was named by Matheron after the South African geostatistician D. G. Krige, who was the first to develop formal solutions to the problem of estimating ore reserves from sampling (core) data (Krige, 1952, 1966). Geostatistics was developed by Matheron (1962, 1965, 1970, 1971, 1973) and co-workers at the *Centre de morphologie mathématique* of the *École des Mines de Paris*. Geostatistics comprises the estimation of variograms (Subsection 13.1.3), kriging, validation methods for kriging estimates, and simulations methods for geographically distributed ("regionalized") data. Major textbooks have been written by former students of Matheron: David (1977) and Journel & Huijbregts (1978). Other useful references are Clark (1979), Rendu (1981),

Verly *et al*. (1984), Armstrong (1989), Isaaks & Srivastava (1989), and Cressie (1991). Applications to environmental sciences and ecology have been discussed by Gilbert & Simpson (1985), Robertson (1987), Armstrong *et al*. (1989), Legendre & Fortin (1989), Soares *et al*. (1992), and Rossi *et al*. (1992). Geostatistical methods can be implemented using the software library of Deutsch & Journel (1992).

As in inverse-distance weighting (eq. 13.20), the estimated value for any grid node is computed as:

$$\hat{y}_{\text{Node}} = \sum_i w_i y_i$$

The chief difference between kriging and inverse-distance weighting is that, in kriging, the weights w_i applied to the points $\emptyset_i$ used in the estimation are not standardized inverses of the distances to some power k. Instead, the weights are based upon the covariances (semi-variances, eqs. 13.9 and 13.10) read on a variogram model (Subsection 13.1.3). They are found by linear estimation, using the equation:

$$\mathbf{C} \cdot \mathbf{w} = \mathbf{d}$$

$$\begin{bmatrix} c_{11} & \dots & c_{1n} & 1 \\ . & \dots & . & 1 \\ . & \dots & . & 1 \\ c_{n1} & \dots & c_{nn} & 1 \\ 1 & \dots & 1 & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ . \\ . \\ w_n \\ \lambda \end{bmatrix} = \begin{bmatrix} d_1 \\ . \\ . \\ d_n \\ 1 \end{bmatrix} \qquad \textbf{(13.22)}$$

where **C** is the covariance matrix among the n points $\emptyset_i$ used in the estimation, i.e. the semi-variances corresponding to the distances separating the various pair of points, provided by the variogram model; **w** is the vector of weights to be estimated (with the constraint that the sum of weights must be 1); and **d** is a vector containing the covariances between the various points $\emptyset_i$ and the grid node to be estimated. This is where a variogram model becomes essential; it provides the weighting function for the entire map and is used to construct matrix **C** and vector **d** for each grid node to be estimated. Element λ in vector **w** is a Lagrange parameter (as in Section 4.4) introduced to minimize the variance of the estimates under the constraint $\Sigma w_i = 1$ (unbiasedness condition). The solution to this linear system is obtained by matrix inversion (Section 2.8):

$$\mathbf{w} = \mathbf{C}^{-1}\,\mathbf{d} \qquad \textbf{(13.23)}$$

Vector **d** plays a role similar to the weights in inverse-distance weighting since the covariances in vector **d** decrease with distance. Using covariances, the weights are statistical in nature instead of geometrical.

Kriging takes into account the grouping of observed points $Ø_i$ on the map. When two points $Ø_i$ are close to each other, the value of the corresponding coefficient c_{ij} in matrix **C** is high; this contributes to lowering their respective weights w_i. In this way, the redundancy of information introduced by dense groups of sampling sites is taken into account.

When anisotropy is present, kriging can use two, four, or more variogram models computed for different geographic directions and combine their estimates when calculating the covariances in matrix **C** and vector **d**. In the same way, when estimation is performed for sampling sites in a volume, a separate variogram can be used to describe the vertical spatial variation. Kriging is the best interpolation method for data that are not on a regular grid or display anisotropy. The price to pay is increased mathematical complexity during interpolation.

Among the interpolation methods, kriging is the only one that provides a measure of the error variance for each value estimated at a grid node. For each grid node, the error variance, called *ordinary kriging variance* (s^2_{OK}), is calculated as follows (Isaaks & Srivastava, 1989), using vectors **w** and **d** from eq. 13.22:

$$s^2_{OK} = \mathrm{Var}[y_i] - \mathbf{w}'\mathbf{d} \qquad \textbf{(13.24)}$$

where $\mathrm{Var}[y_i]$ is the maximum-likelihood estimate of the variance of the observed values y_i (eq. 13.15). Equation 13.24 shows that s^2_{OK} only depends on the variogram model and the local density of points, and not on the values observed at points $Ø_i$. The ordinary kriging variance may be used to construct confidence intervals around the grid node estimates at some significance level α, using eq. 13.4. It may also be mapped directly. Regions of the map with large values s^2_{OK} indicate that more observations should be made because sampling intensity was too low.

Kriging, as described above, provides point estimates at grid nodes. Each estimate actually applies to a "point" whose size is the same as the grain of the observed data. The geostatistical literature also describes how *block kriging* may be used to obtain estimates for blocks (i.e. surfaces or volumes) of various sizes. Blocks may be small, or a single block may cover the whole map if one wishes to estimate a resource over a whole area. As mentioned in the introductory remarks of the present section, only additive variables can be used in block kriging. Block kriging programs always assume that the variable is *intensive*, e.g. concentration of organisms (Subsection 1.4.2). For *extensive* variables, such as the number of individual trees, one must multiply the block estimate by the ratio (block size / grain size of the original data).

3 — Measures of fit

Different measures of fit may be used to determine how well an interpolated map represents the observed data. With most methods, some measure may be constructed of

the closeness of the estimated (i.e. interpolated) values $\hat{y}_i$ to the values y_i observed at sites $\emptyset_i$. Four easy-to-use measures are:

- The mean absolute error: $MAE = \frac{1}{n}\sum_i |y_i - \hat{y}_i|$;
- The mean squared error: $MSE = \frac{1}{n}\sum_i (y_i - \hat{y}_i)^2$;
- The Euclidean distance: $D_1 = \sqrt{\sum_i (y_i - \hat{y}_i)^2}$;
- The correlation coefficient (r) between values y_i and $\hat{y}_i$ (eq. 4.7). In the case of a trend-surface model, the square of this correlation coefficient is the coefficient of determination of the model.

In the case of kriging, the above measures of fit cannot be used because the estimated and observed values are equal at all observed sites $\emptyset_i$. The technique of cross-validation can be used instead (Isaaks & Srivastava, 1989, their Chapter 15). One observation, say $\emptyset_1$, is removed from the data set and its value is estimated using the remaining points $\emptyset_2$ to $\emptyset_n$. The procedure is repeated for $\emptyset_2$, $\emptyset_3$, …, $\emptyset_n$. One of the measures of fit described above may be used to measure the closeness of the estimated to the observed values. If replicated observations are available at each sampling site (a situation that does not often occur), the F-test of goodness-of-fit described in Subsection 13.2.1 can be used with all interpolation methods.

13.3 Patches and boundaries

Multivariate data may be condensed into spatially-constrained clusters. These may be displayed on maps, using different colours or shades. The present section explains how clustering algorithms can be constrained to produce groups of spatially contiguous sites; study of the boundaries between homogeneous zones is also discussed. Prior to clustering, one must state unambiguously which sites are neighbours in space; the most common solutions to this problem are presented in Subsection 13.3.1.

1 — Connection networks

When sampling has been conducted on a regular rectangular grid, neighbouring points may be linked using simple connecting schemes whose names are derived from the game of chess (Cliff & Ord, 1981): rook's (rectangular: Fig. 13.21a), bishop's (diagonal: Fig. 13.21b), or king's connections (also called queen's: both rectangular and diagonal, Fig. 13.21c). Sampling in staggered rows leads to connecting each point

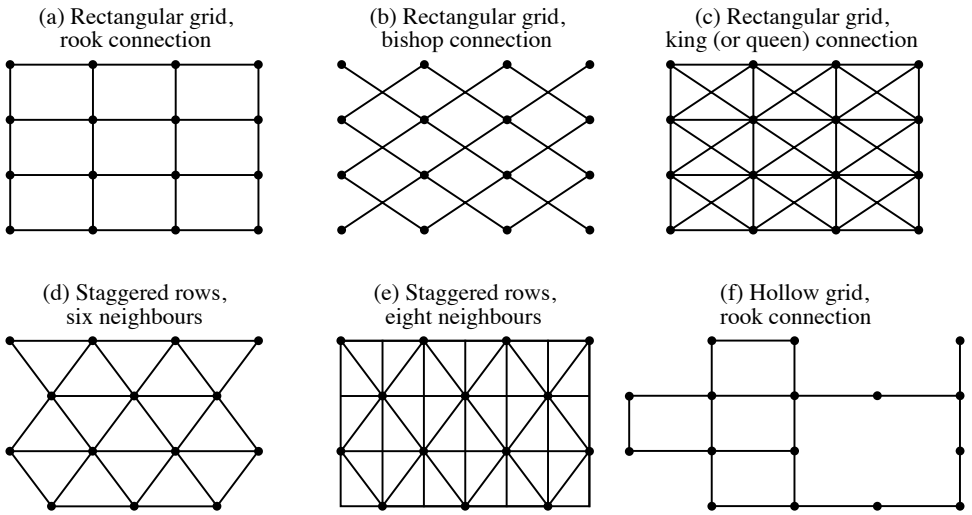


Figure 13.21 Connecting schemes for regular grids of points. See text.

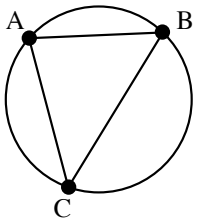
(except borders) to six (Fig. 13.21d) or eight neighbours (Fig. 13.21e). Algorithms may allow the construction of regular grids with missing points (Fig. 13.21f). When the objects represent irregularly-shaped land units covering a geographic area (e.g. types of ecosystems in a nature reserve), parcels sharing a common boundary are regarded as contiguous.

When the sites are positioned in an irregular manner, one can use geometric connecting schemes such as Delaunay triangulation, Gabriel graph, relative neighbourhood graph or minimum spanning tree, described below. There exists an inclusion relationship among these four connecting schemes: all edges that are members of a minimum spanning tree (MST) also obey the relative neighbourhood graph criterion; these are all members of a Gabriel graph, which in turn are all included in a Delaunay triangulation (Toussaint, 1980; Matula & Sokal, 1980; Gordon, 1996c):

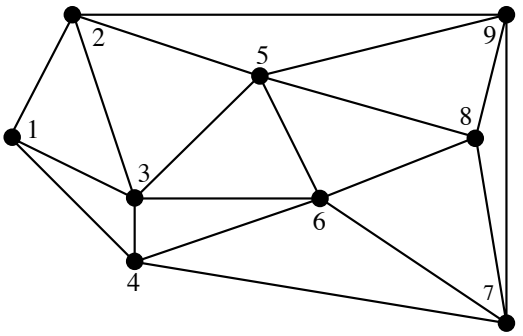
$$\text{MST} \subseteq \text{Relative neighbourhood graph} \subseteq \text{Gabriel graph} \subseteq \text{Delaunay triangulation}$$

Delaunay triangulation

• Delaunay triangulation. — The Delaunay triangulation criterion (Dirichlet, 1850; Upton & Fingleton, 1985) is illustrated in Fig. 13.22. For any triplet of points A, B and C, the three edges (i.e. lines) connecting these points are included in the triangulation if and only if the circumscribed circle (i.e. the circle passing through the three points; on the left in the figure) includes no other point. For example, the file of coordinates shown in the central part of the figure gives rise to the triangulation on the right. The triangulation is fully described by a list of pairs of points corresponding to its edges; this is how the information can be passed on to a computer program for space-constrained clustering (Subsection 13.3.2).



Point identifiers	Coordinates X	Y
1	0	3
2	1	5
3	2	2
4	2	1
5	4	4
6	5	2
7	8	0
8	7.5	3
9	8	5



19 edges form the Delaunay triangulation:

1–2	1–3	1–4	2–3	2–5	2–9	3–4
3–5	3–6	4–6	4–7	5–6	5–8	5–9
6–7	6–8	7–8	7–9	8–9		

Figure 13.22 Construction of a Delaunay triangulation for 10 points.

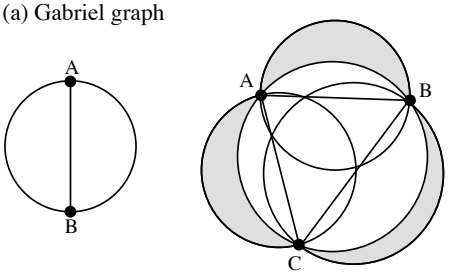
Long edges may be created at the outskirts of a set of points, simply because there is no other point located farther away in the sampling design; this is called a *border effect*. For example, edges 2–9 and 7–9 could have been removed from the triangulation in Fig. 13.22 by the presence of other points in the circumscribed circles of triangles (2, 5, 9) and (7, 8, 9) had the sampling extent been broader. Long peripheral edges can be removed by hand from the list, or by the computer algorithm.

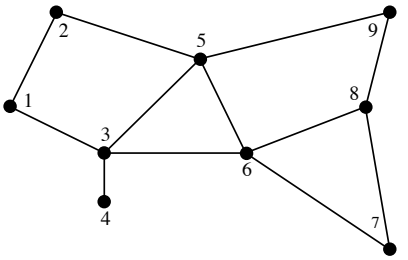
Gabriel graph

• Gabriel graph. — The Gabriel graph criterion (Gabriel & Sokal, 1969) differs from that of the Delaunay triangulation (Fig. 13.23a). Draw a line between two points A and B. This line is part of the Gabriel graph if and only if no other point C lies inside the circle *whose diameter is that line*. In other words, the edge between A and B is part of the Gabriel graph if $D^2(A, B) < D^2(A, C) + D^2(B, C)$ for all other points C in the study, where $D^2(A, B)$ is the square of the geographic distance between points A and B. Another way of expressing this criterion is the following: if CENTRE represents the middle point between A and B, the edge connecting A to B is part of the Gabriel graph if $D(A, B)/2 < D(\text{CENTRE}, C)$ for any other point C in the study.

The Gabriel graph in Fig. 13.23a is constructed for the same points as the Delaunay triangulation in Fig. 13.22. The 12 edges forming the Gabriel graph are a subset of the 19 edges of the Delaunay triangulation. Indeed, as shown by the sketch in the centre of the figure, the exclusion zone formed by the three circles corresponding to the Gabriel criterion (which have for diameters the edges A–B, B–C and A–C) may contain, in the shadowed areas outside the Delaunay circle (white inner circle), some points that the

(a) Gabriel graph

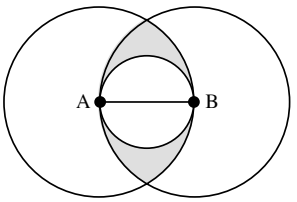


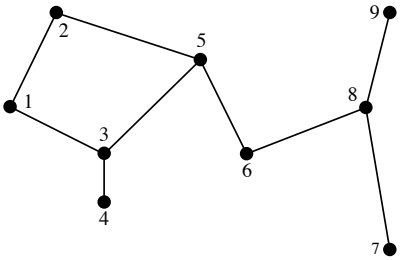


12 edges form the Gabriel graph:

1–2 1–3 2–5 3–4 3–5 3–6
5–6 5–9 6–7 6–8 7–8 8–9

(b) Relative neighbourhood graph





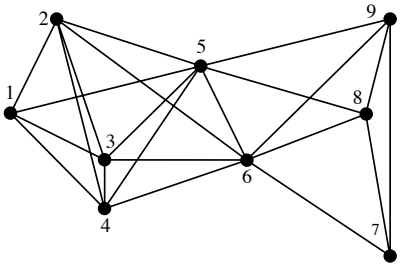
9 edges form the relative neighbourhood graph:

1–2 1–3 2–5 3–4 3–5 5–6
6–8 7–8 8–9

(c) Maximum distance graph

Criterion: $D \leq$ threshold

In this example, $D \leq 5$



22 edges form the maximum distance graph:

1–2 1–3 1–4 1–5 2–3 2–4 2–5 2–6
3–4 3–5 3–6 4–5 4–6 5–6 5–8 5–9
6–7 6–8 6–9 7–8 7–9 8–9

Figure 13.23 (a) Left: geometric criterion for the Gabriel graph. Centre: the zone of exclusion of the criterion, here for three points (grey zones + white inner circle), is larger than that of the Delaunay criterion (white inner circle). Right: graph for the example data, containing 12 edges. (b) Left: geometric criterion of the relative neighbourhood graph. The zone of exclusion of the criterion, here for two points (grey zones + white inner circle), is larger than that of the Gabriel criterion (white inner circle). Right: graph for the example data, containing 9 edges. (c) Left: criterion of the maximum distance graph. Right: graph for the example data with $D \leq 5$, with 22 edges.

Delaunay criterion circle does not exclude. This is why some edges that are authorized by the Delaunay criterion are excluded from the Gabriel graph.

Relative neighbourhood graph

• Relative neighbourhood graph. — The relative neighbourhood criterion is as follows (Toussaint, 1980; Fig. 13.23b). Draw a line between two points A and B. Draw a first circle centred over A and a second one centred over B, each one having the line from A to B as its radius. This line is part of the graph if no other point C in the study lies *inside the intersection of the two circles*. Points that fall on the circumference of one of the circles in the intersection zone do not count. In algebraic terms, the edge from A to B is part of the relative neighbourhood graph if and only if $D(A, B) \leq \max [D(A, C), D(B, C)]$ for all other points C in the study. For points forming an equilateral triangle, for instance, the three edges are included in the relative neighbourhood graph.

The relative neighbourhood graph in Fig. 13.23b is constructed for the same set of points as in Figs. 13.22 and 13.23a. The 9 edges forming the relative neighbourhood graph are a subset of the 12 edges of the Gabriel graph. Indeed, as shown by the sketch on the left of the figure, the exclusion zone at the intersection of the two circles corresponding to the relative neighbourhood criterion (which have for radius the edge A–B) may contain, in the shadowed zone outside the Gabriel circle (white inner circle), some points that the Gabriel criterion circle does not exclude. This is why some edges authorized by the Gabriel criterion are excluded from the relative neighbourhood graph.

Maximum distance graph

• Maximum distance graph. — Another strategy is to select a distance threshold and connect all points that are within that distance of each other. The result is called a maximum distance graph or an influence circle graph (Fig. 13.23c). One possible criterion to choose the distance threshold is to make it equal to the range of a variogram model (Fig. 13.7) computed for univariate (Subsection 13.1.3) or multivariate response data (Subsection 13.1.4).

Minimum spanning tree (MST)

• Minimum spanning tree (MST). — This tree connects the n points in the study with $(n - 1)$ edges. The sum of the weights (i.e. distances) of the edges used in the tree is minimum, meaning that it is smaller than or equal to the sum of the edge weights of any other tree connecting these n objects. Its construction is described at the end of Section 8.2; one way of obtaining it is to list the edges forming the primary connections of a single-linkage dendrogram. For points forming an equilateral triangle, for example, only two of the edges are included in the minimum spanning tree, whereas the three edges are included in a relative neighbourhood graph; the choice of the edge to leave out is arbitrary. The edges of a minimum spanning tree are either the same as, or a subset of, the edges of a relative neighbourhood graph of the same points. The minimum spanning tree for the example data set is shown in Fig. 14.3.

The list of connecting edges (Figs. 13.22 and 13.23) may be written out to a file. The file may be modified to take into account other information that researchers may have about the study area. For example, one may wish to eliminate edges that do not make sense in terms of gene flow because they cross unsuitable areas (e.g. a sea or a

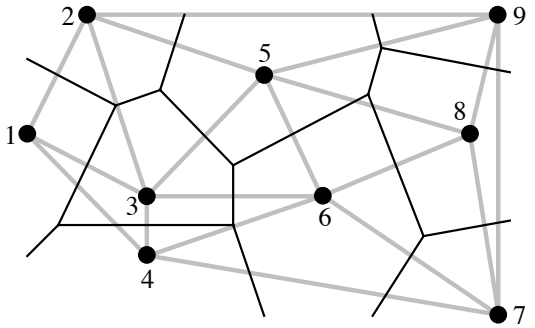


Figure 13.24 Delaunay triangulation (grey lines) and influence polygons (black lines) for the nine points of Fig. 13.22.

mountain range, in the case of terrestrial mammals). Or, one may wish to add connections that are potentially of interest although they do not imply first neighbours; for example, plants or animals may be able to cross water bodies (lake, sea) and settle in non-contiguous sites, which should be considered contiguous for the analysis because there is a direct path between them. Users of constrained clustering methods should not hesitate to modify lists of connections obtained from geometric criteria such as described above, to make the list of edges a better description of potential flow among sites, given the problem under study.

Influence polygons

It is sometimes interesting to determine the geometric zone of influence of each point on a map. The zone of influence of a point A includes all the other points of the surface that are closer to A than to any other point in the study. The zones of influence so defined have the shape of polygons, also called tiles, tessellae, or tesserae (singular: tessella or tessera). The resulting picture is called a mosaic or tessellation (adjective: tessellated); it is also referred to as a Dirichlet tessellation (1850), Voronoï polygons (1909), or Thiessen polygons (1911), from the names of the authors who described these mathematical structures.

Polygons are easily constructed from a Delaunay triangulation (Fig. 13.24). Draw the perpendicular bisector of each segment in the triangulation; the crossing points of the bisectors delimit the polygons (tiles). Computer algorithms may be used to calculate the surface area of each polygon, at least those that are closed; peripheral tiles may be open. Upton & Fingleton (1985) and Isaaks & Srivastava (1989) propose various applications of tessellations to spatial analysis.

2 — Space-constrained clustering

The delineation of clusters of contiguous objects has been discussed in Section 12.6 for time series and spatial transects. The method of chronological clustering, in

particular, was described in Subsection 12.6.4; it proceeds by imposing to a clustering algorithm a constraint of contiguity along the time series. Constraints of contiguity have been applied to spatial clustering by several authors, including Lefkovitch (1978, 1980), Monestiez (1978), Lebart (1978), Roche (1978), Perruchet (1981) and Legendre & Legendre (1984c). In the present subsection, it is generalized to two- or three-dimensional spatial data and to spatio-temporal data.

Constrained clustering differs as follows from its unconstrained counterpart:

- Unconstrained clustering methods (Chapter 8) only use the information in the similarity or distance matrix computed among the objects. In hierarchical methods, a local criterion is optimized at each step; in all methods included in the Lance and Williams general model, for instance, the objects or groups clustered at each step are those with the smallest fusion distance or the largest fusion similarity. In partitioning methods, a global criterion is optimized; in *K*-means, for instance, the algorithm looks for *K* groups that feature the smallest sum of within-group sums-of-squares E_K^2;

- Constrained clustering methods take into account more information than the unconstrained approaches. In the case of spatial or temporal contiguity, the only admissible clusters are those that obey the contiguity relationship. Spatial contiguity may be described by one of the connecting schemes of Subsection 13.3.1. The criterion to be optimized during clustering is relaxed to give priority to the constraint of spatial contiguity. It is no surprise, then, that a constrained solution may be less optimal than its unconstrained counterpart in terms of the clustering criterion, e.g. E_K^2. This is balanced by the fact that the resulting clusters are likely to more readily interpretable.

It is fairly easy to modify clustering algorithms to incorporate a constraint of spatial contiguity (Fig. 13.25). As an example, consider the clustering methods included in the Lance and Williams general agglomerative model (Subsection 8.5.9). At the beginning of the clustering process, the vector of group membership has each object in a different group (Fig. 13.25, right). Proceed as follows:

1. Compute a distance matrix (**D**) among objects using the non-geographic information. Turn it into a similarity matrix **S** using one of the equations of Subsection 7.2.1. This transformation will make step 3 of the procedure possible.

2. Choose a connecting scheme (Subsection 13.3.1) and produce a list of connection edges as in Figs. 13.22 and 13.23. Read in the file of edges and transform it into a *contiguity matrix* containing 1's for connected sites and 0's elsewhere.

3. Compute the Hadamard product of these two matrices, i.e. their product element by element (Section 2.5). The resulting matrix contains similarity values in the cells where the contiguity matrix contained 1's, and 0's elsewhere.

4. The largest similarity value in the matrix resulting from step 3 determines the next pair of objects or groups (*h* and *i*) to be clustered. Modify the vector of group

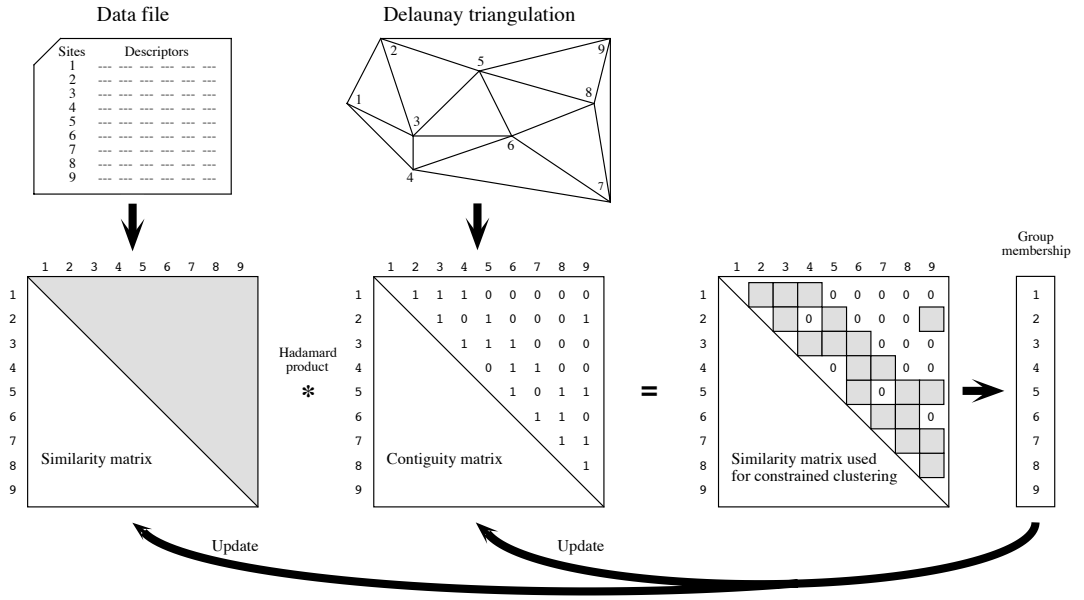


Figure 13.25 Summary of the spatially-constrained clustering procedure for methods included in the Lance and Williams general clustering model. The vector of group membership is represented on the right; at the start of the clustering process, each object is in a different group (numbers 1 to 9 in the example). Locations of the points are the same as in Fig. 13.22.

membership (right of the figure), giving the same group label to all members of former groups *h* and *i*.

5. Update the similarity matrix using eq. 8.12.

6. Update also the contiguity matrix. All objects that were neighbours to *h* are now also neighbours to *i* and vice versa.

7. Go back to step 3. Iterate until all objects are members of a single group.

8. Determine the most informative number of clusters, either by visual inspection of space-constrained clustering maps, or after calculating one of the indices mentioned at the end of Section 8.8 (available in R function ***clustIndex()*** of package CCLUST). Among those indices, the Calinski-Harabasz criterion was recommended by Gordon (1999) for constrained clustering. Pawitan & Huang (2003) proposed a permutation procedure to test the significance of successive partition levels in constrained clustering. Cross-validation seems another promising way of identifying the most informative partition in constrained clustering.

Ferligoj & Batagelj (1982) showed that the introduction of relational constraints (e.g. spatial contiguity) may occasionally produce reversals with any of the hierarchical clustering methods included in the Lance & Williams algorithm (Subsection 8.5.9], except complete linkage. Additional constraints may be added to the algorithm, for example to limit the size or composition of any group (Gordon, 1996c). *K*-means partitioning algorithms (Section 8.8) can also be constrained by the contiguity matrix shown in Fig. 13.25.

Space-constrained clustering is useful in a variety of situations. Here are some examples.

- In many studies, there are compelling reasons to force the clusters to be composed of contiguous sites; for instance, when delineating ecological regions, administrative units, or resource distribution networks.

- One may wish to relate the results of clustering to geographically-located potential causal factors that are known to be spatially autocorrelated, e.g. geological data.

- One may wish to cluster sites based upon environmental variables, using a constraint of spatial contiguity, in order to design a stratified biological sampling program to study community composition.

- To test the hypothesis that neighbouring sites are ecologically similar, one may compare unconstrained and constrained clustering solutions using the modified Rand index (Subsection 8.12.2). De Soete *et al*. (1987) give other examples where such comparisons may help test hypotheses in the fields of molecular evolution, psycholinguistics, cognitive psychology and evolution of languages.

- Constrained solutions are less variable than unconstrained clustering results, which may differ in major ways among clustering methods. Indeed, the constraint of spatial contiguity reduces the number of possible solutions and forces different clustering algorithms to converge onto largely similar clusters (Legendre *et al*., 1985).

Constrained clustering can also be used for three-dimensional or spatio-temporal sampling designs (e.g. Planes *et al*., 1993). As long as the three-dimensional or spatio-temporal contiguity of the observations can be accurately described as a file of edges as in Figs. 13.22 and 13.23, constrained clustering programs have no difficulty in computing the solution; the only difficulty is the representation of the results as three-dimensional or spatio-temporal maps. Higher-dimensional extensions of the geometric connecting schemes presented in Subsection 13.3.1 are available in the literature. In addition, space-constrained clustering can be used to detect discontinuities in spatial transects or time series, a topic that has been discussed in Section 12.6.

Legendre (1987b) suggested a way of introducing spatial proximity into clustering algorithms which is less stringent than the methods described above. The method consists in weighting the values in the ecological similarity or distance matrix by some function of the geographic distances among points, before clustering. The idea was

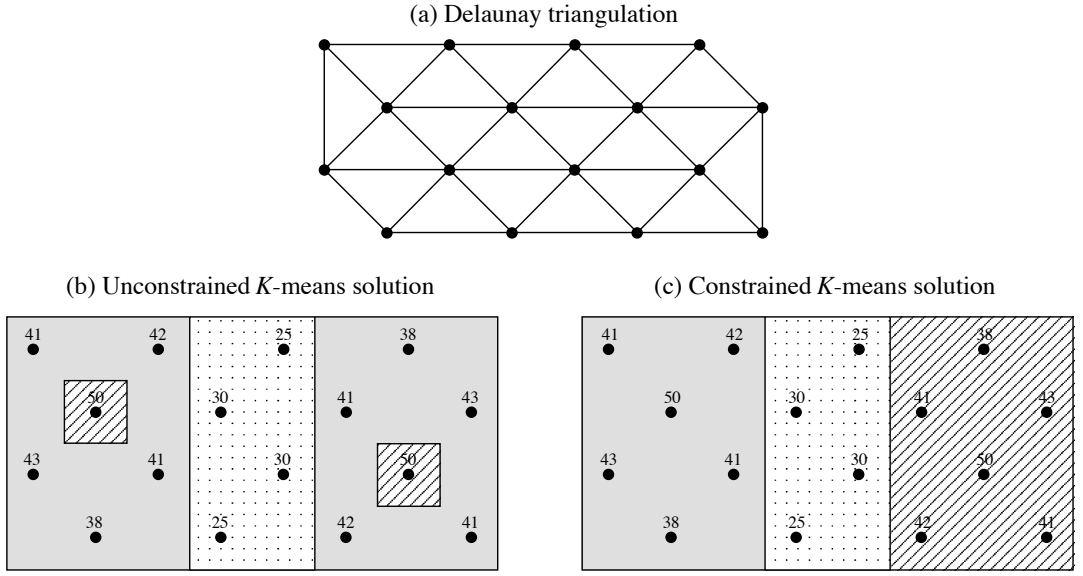


Figure 13.26 Numerical example showing the difference between (b) unconstrained and (c) constrained clustering solutions. (a) Delaunay triangulation with 35 edges, which were used as constraint in (c). The values of the artificial variable are given in panels (b) and (c); the three groups obtained by unconstrained and constrained *K*-means are identified by shadings.

implemented by Bourgault *et al*. (1992) who proposed to use a multivariate variogram or covariogram as spatial weighting function prior to clustering. Large ecological distances between sites that are close in space are downweighted to some extent by this procedure. It is then easier for clustering algorithms to incorporate somewhat diverging sites into neighbourhood clusters. Oliver & Webster (1989) suggested to use a univariate variogram for the same purpose. Constrained classification methods were reviewed by Gordon (1996c, 1999) and algorithms were surveyed by Murtagh (1985). Formal aspects were discussed by Ferligoj & Batagelj (1982, 1983). Generalized forms of constrained clustering were described by De Soete *et al*. (1987).

Numerical example. An artificial set of 16 sites was constructed to represent staggered-row sampling of a distribution with two peaks. From the geographic positions of the sites, a Delaunay triangulation (35 edges) was computed (Fig. 13.26a). A single variable was attributed to the sites. For three groups, the unconstrained *K*-means solution has a sum of within-group sums-of-squares $E_K^2 = 53$ (Fig. 13.26b). The constrained *K*-means solution, for three groups, has a value $E_K^2 = 188$ (Fig. 13.26c) which is higher than that of the unconstrained solution, for reasons explained above. The two partitions are interesting in different ways. The unconstrained solution identifies sites with similar values, whereas the constrained solution brings out the two regions with high values plus a region with lower values forming a valley between the peaks.

Space-constrained clustering has been applied to a variety of ecological situations. Applications to two-dimensional map data are found in Legendre & Legendre (1984c), Legendre & Fortin (1989), Legendre *et al*. (1989), Lapointe & Legendre (1994), and Fortin & Dale (2005, their Section 4.1.2). Two examples of application of space-constrained clustering to community composition data surveyed on a geographic surface and along a transect, respectively, are available in the documentation file of function ***constrained.clust()*** of the R package CONST.CLUST; see Section 13.6. Users are invited to run these examples. A space-constrained clustering map of the data of Table 13.2 (bacterial data from the Thau coastal lagoon, France) is shown in Fig. 13.28b. Applications to transect and stratigraphic data (sediment cores) were mentioned in Subsection 12.6.5 and 12.8.

3 — Ecological boundaries

Detection of boundaries is the complementary problem to the detection of homogeneous regions of space. Boundaries appear on maps as a by-product of constrained clustering, for example. Most methods of clustering delineate groups even in gradient situations; a boundary between groups does not have to correspond to a sharp discontinuity in the data. Other methods have been developed that focus on boundary elements; these methods do not aim at completely isolating regions of space.

For univariate or multivariate transect data, boundaries can be detected using the methods described in Section 12.6. Detection of boundaries of various sorts on maps is more complex. This is a well-studied topic in the field of image analysis; it has been reviewed by Davis (1975), Peli & Malah (1982) and Huang & Tseng (1988); see also Hobbs & Mooney (1990). The present section briefly summarizes the efforts made to detect boundaries in ecological data sets, using a technique called *wombling*, and to statistically assess their significance. Readers are referred to Section 4.2 (*Boundary delineation*) of the book of Fortin & Dale (2005) for details; several examples are also presented in that book.

Wombling

Wombling is a technique for detecting zones of rapid spatial change in a set of regionalized variables. It was developed by Womble (1951) and Barbujani *et al*. (1989) for gene frequencies and morphological measurements, and refined by Fortin and co-authors (Oden *et al*., 1993; Fortin, 1994, 1997; Fortin & Drapeau, 1995; Fortin *et al*., 1996) with emphasis on ecological data. The original form of wombling (*lattice wombling*) could only be applied to quantitative variables observed at sites forming a regular, rectangular grid of points. Recent developments include *categorical wombling* for qualitative variables (Oden *et al*., 1993) and *triangulation wombling* for sites linked by a Delaunay triangulation which do not necessarily correspond to a regular sampling grid (Fortin, 1994). The latter is a frequent situation in ecology.

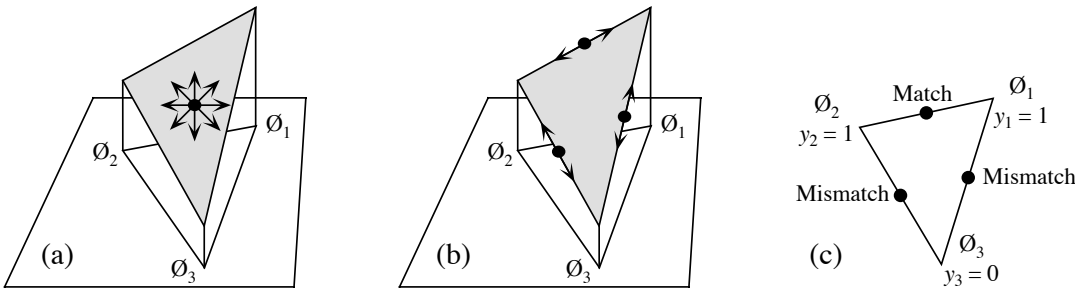


Figure 13.27 Examples of initial calculations on spatial elements (triplets of points) for three different boundary detection methods. Consider three points $Ø_i$, with coordinates (X_i, Y_i), forming a Delaunay triangle. (a, b) Variable y, measured at sites $Ø_i$, is quantitative (values shown as heights). (a) Method 1: find the direction of maximum slope of the triangle; allocate this slope value to the triangle centroid (dot). (b) Method 2: compute slopes along the edges connecting adjacent sites; allocate the values to the edge mid-points (dots). (c) Method 3: for a qualitative variable y (with 2 states in this example), adjacent sites are compared in terms of matches (0-0 or 1-1) or mismatches (0-1); allocate the matches and mismatches to the edge mid-points (dots).

A boundary is delineated on a map by linking adjacent points where the variable shows high rates of change (Fortin, 1994). Triangulation wombling (Fig. 13.27a) proceeds as follows:

- Link the observed sites by a Delaunay triangulation (Subsection 13.3.1).

- Consider a quantitative variable measured at three sites $Ø_i$ forming a Delaunay triangle. Each site has geographic coordinates (X_i, Y_i) and an observed value y_i. The plane to be fitted to these points is a linear function $y = f(X, Y) = b_0 + b_1X + b_2Y$ whose parameters can be computed by matrix inversion (Section 2.8):

$$\begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 1 & X_1 & Y_1 \\ 1 & X_2 & Y_2 \\ 1 & X_3 & Y_3 \end{bmatrix}^{-1} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

- Find the direction of maximum slope of the triangle. The slope varies with the direction considered (arrows in Fig. 13.27a). Using the b coefficients calculated above, the *maximum slope* of the triangle is:

$$m = \sqrt{\left[\frac{\partial f(X, Y)}{\partial X}\right]^2 + \left[\frac{\partial f(X, Y)}{\partial Y}\right]^2} = \sqrt{b_1^2 + b_2^2} \qquad \textbf{(13.25)}$$

and the *angle* from the X coordinate axis is given by $\tan^{-1}(b_2/b_1)$. Note that angles are in radians in R. Allocate this value of slope (m) to the centroid of the triangle, which is the point with coordinates:

$$[X, Y]_{centroid} = \left[\frac{X_1 + X_2 + X_3}{3}, \frac{Y_1 + Y_2 + Y_3}{3}\right] \quad \textbf{(13.26)}$$

- If several variables are considered (i.e. several species), calculate the mean slope ($\bar{m}$) of the variables at the centroid of each Delaunay triangle.

- Create an ordered list of the slope values. Starting at the top of the list (highest slopes), mark the corresponding triangle centroids on the map; they become *boundary elements*. Going down the list, mark a predetermined proportion of the slopes (e.g. 10%), or go down to a preselected value of slope. Other strategies are possible, e.g. going down the list to the value of the mean plus one or two standard deviations.

Boundary

- A *boundary* is delineated by linking contiguous boundary elements. A single boundary element unlinked to other elements may be seen as a small boundary.

An alternative would be to compute the slopes of the edges between adjacent sites (Fig. 13.27b). For univariate data, the rate of change would simply be the absolute value of the difference between values at sites $Ø_h$ and $Ø_i$: $|y_h - y_i|$. For multivariate data, any of the distance functions of Chapter 7 could be used. The disadvantage of this method is that slopes calculated along the edges of the Delaunay triangle do not have the same value as the *maximum* slope of the triangle, computed by eq. 13.25. To alleviate this problem, Dufrêne & Legendre (1991) calculated multivariate distances in four directions between pixels of a map; for each pixel, they used the largest of the distances to delineate boundaries.

Computation of statistics along the edges between adjacent sites is the option used in categorical wombling, which is appropriate for species presence-absence data. The basic statistic is to record a match or a mismatch between adjacent observed sites (Fig. 13.27c). For multivariate qualitative data, one may count both the positive and negative matches and embed this number into one of the symmetrical binary coefficients of Subsection 7.3.1 (e.g. the simple matching coefficient); for species presence-absence data, one may count the positive matches only and embed this number into one of the asymmetrical binary coefficients of Subsection 7.3.2 (e.g. the Jaccard coefficient).

Tests of significance, based on permutations (Subsection 1.2.2), have been proposed by Fortin and co-authors (Oden *et al*., 1993; Fortin, 1994, 1997; Fortin & Drapeau, 1995; Fortin *et al*., 1996) to answer the following questions:

- Are the boundaries found by this analysis similar to random boundaries in terms of the number of separate boundaries, their maximum or mean lengths, or other boundary or graph-theoretic statistics?

• Are the boundaries found by wombling the same as borders stated by hypothesis, or found by clustering methods, or obtained using different data for the same locations?

These papers also present applications of the method to real and simulated data. A computer program for wombling is commercially available (BOUNDARYSEER, Table 13.3). An R package is also available (Section 13.6).

4 — Dispersal

Individuals, populations, and communities often cross ecological boundaries; such crossings occur on different time scales. The routes taken by species when they invade a territory after a perturbation event (long-term, e.g. glaciation; short term, e.g. pollution) is a question of interest in biogeographic analysis. Dispersal routes may be easier to identify if, as a first step in the analysis, one delineates regions that are largely homogeneous in species composition. Regions may be delimited using prior hypotheses, by unconstrained or constrained cluster analysis, or using boundary detection methods.

Legendre & Legendre (1984c) developed coefficients to measure the likelihood of species dispersal between geographically contiguous regions, for species presence-absence or abundance data. The assumptions of these coefficients are that the species arrived by migration, and the past dispersal has left traces in present-day distributions. For presence-absence data, adjacent regions $\mathbf{x}_1$ and $\mathbf{x}_2$ can be compared using the same quantities a, b, and c as in the similarity coefficients of Subsection 7.3.1 and 7.3.2: a is the number of species that two regions have in common; b is the number of species found in $\mathbf{x}_1$ but not in $\mathbf{x}_2$; c is the number of species found in $\mathbf{x}_2$ but not in $\mathbf{x}_1$. The *combination* of the following indications is evidence for species dispersal from region $\mathbf{x}_1$ to $\mathbf{x}_2$:

• the number of species common to the two zones is high, i.e. a is large;

• b is substantially larger than c. Conversely, c larger than b would support the hypothesis of dispersal from $\mathbf{x}_2$ to $\mathbf{x}_1$.

The basic form of the *coefficient of species dispersal direction* (*DD*) is thus $a(b - c)$. To make the values of the coefficient comparable for faunas with different richness, each term is standardized by dividing it by the richness of the fauna or flora of the two regions combined:

$$DD_1(\mathbf{x}_1 \rightarrow \mathbf{x}_2) = \frac{a}{(a+b+c)} \frac{(b-c)}{(a+b+c)} \qquad \textbf{(13.27)}$$

When this coefficient is positive, it measures the likelihood that species dispersed from $\mathbf{x}_1$ to $\mathbf{x}_2$. A negative sign indicates that, if dispersal occurred, species migrated from $\mathbf{x}_2$ to $\mathbf{x}_1$ instead.

The asymmetric portion of this coefficient may be tested for significance using a McNemar test. Under the null hypothesis of no asymmetry (H_0: $b = c$), the test statistic

$$X^2_P = \frac{(|b-c|-1)^2}{(b+c)} \qquad \textbf{(13.28)}$$

is distributed as χ^2 with one degree of freedom. The value –1 subtracted in the numerator is Edwards' (1948) correction for continuity; this is the correction used by function ***mcnemar.test()*** in R. The test may be one-tailed if one has specific hypotheses about the direction of dispersal; otherwise, a two-tailed test is used.

The log-linear form of the McNemar statistic is (Sokal & Rohlf, 1995):

$$X^2_W = 2\,(b\,\log_e b + c\,\log_e c - (b+c)\log_e[\,(b+c)/2]\,)/q \qquad \textbf{(13.29)}$$

where q is the Williams (1976) correction for continuity:

$$q = 1 + \frac{1}{2\,(b+c)}$$

Equation 13.29 provides a more powerful test than the classical McNemar equation (eq. 13.28). If any of the values b or c is 0, the corresponding term ($x \log_e x$) is 0 since $\lim_{x \to 0} (x \log_e x) = 0$ (Section 6.5).

The first part of the DD_1 equation is easily recognized as the Jaccard coefficient of similarity (eq. 7.10). One may prefer to give double weight to the number of common species, a, as in the coefficient of Sørensen (eq. 7.11):

$$DD_2(\mathbf{x}_1 \rightarrow \mathbf{x}_2) = \frac{2a}{(2a+b+c)}\,\frac{(b-c)}{(a+b+c)} \qquad \textbf{(13.30)}$$

Two other forms of the coefficient use species abundance data instead of presence-absence:

$$DD_3(\mathbf{x}_1 \rightarrow \mathbf{x}_2) = \frac{W\,(A-B)}{(A+B-W)^2} \qquad \textbf{(13.31)}$$

and

$$DD_4(\mathbf{x}_1 \rightarrow \mathbf{x}_2) = \frac{2W}{(A+B)}\,\frac{(A-B)}{(A+B-W)} \qquad \textbf{(13.32)}$$

where W, A, and B are as in the Steinhaus similarity coefficient (eq. 7.24). Coefficient DD_4 gives double weight to the abundances of the species in common and is thus the counterpart of DD_2, whereas DD_3 gives these species single weight, as in DD_1. These two coefficients take the following indications as evidence for dispersal from $\mathbf{x}_1$ to $\mathbf{x}_2$:

- the number of species common to the two zones and their abundances are high, i.e. W is large;

• *A* is substantially larger than *B*; *B* larger than *A* would produce a negative coefficient, indicating possible dispersal from $\mathbf{x}_2$ to $\mathbf{x}_1$.

Legendre & Legendre (1984c) used *DD* coefficients and tests of significance to identify plausible routes taken by freshwater fishes to reinvade the Québec peninsula after the last glaciation. Borcard *et al*. (1995) used the same method in a finer-scale study, showing possible patterns of migration of Oribatid mites between zones of an exploited peat bog in the Swiss Jura. At broader scale, Bachraty *et al*. (2009) used *DD* coefficients to identify possible faunal dispersal pathways between adjacent deep-sea hydrothermal provinces of the world ocean.

13.4 Unconstrained and constrained ordination maps

Subsection 13.3.2 has shown how maps for multivariate data can be produced by clustering methods; these maps display discontinuous zones. For continuous representation of quantitative variables, however, the techniques of Section 13.2 can only produce maps for single variables. The present section shows how continuously-varying maps can be obtained for multivariate data sets through ordination methods. The relationship between univariate or multivariate structure functions (Section 13.1) and maps has been stressed in the introductory paragraph of Section 13.2.

The simplest method consists in analysing a data table with one of the ordination methods of Chapter 9 and map the first, or the first few ordination axes. For example:

• Decompose the variation of a (sites × species) presence-absence or abundance table into successive ordination axes, using PCA, CA, PCoA, or nMDS (Chapter 9).

• Consider the ordination of sites along the first axis. This axis is a new, synthetic quantitative variable describing the variation among sites. Associate it to the (X, Y) geographic coordinates of the sites. Produce a map using one of the methods described in Section 13.2. An example of such a map is given in Fig. 9.15 for correspondence analysis axis I of a vegetation data table.

• Repeat the operation, producing maps for ordination axes II, III, etc. as long as interesting or significant spatial variation can be detected. Univariate correlograms of the successive ordination axes (Subsection 13.1.1), with tests of significance, may be used as criterion for deciding which of the ordination axes should be mapped.

Simple ordination analysis leaves it to chance to find spatially-structured components of variation. One may decide instead to look directly for such components, by forcing the analysis to bring out axes of variation that are related to the X and Y coordinates, or combinations of X and Y into a spatial polynomial equation. The spatial polynomial is constructed as in Subsection 13.2.1. (In Chapter 14, the X and Y coordinates will be used to derive spatial eigenfunctions that will replace the

Table 13.2 Data from Table 10.6. There are two bacterial response variables (Bna and Ma, forming matrix **Y**) and three environmental variables (NH_4, phaeopigments, and bacterial production, forming matrix **X**). Five spatial variables (X^2, X^3, X^2Y, XY^2, and Y^3, included in matrix **W**) were derived from the X and Y coordinates, reported in the table, obtained by PCA rotation of the geographic coordinates of Table 10.6. The variables are described in more detail in Numerical example 1 of Subsection 10.3.5.

Station No.	Bna y_1	Ma y_2	NH_4 x_1	Phaeo. *a* x_2	Prod. x_3	X after PCA rotation	Y after PCA rotation
1	4.615	10.003	0.307	0.184	0.274	–9.4173	–1.2516
2	5.226	9.999	0.207	0.212	0.213	–7.1865	–1.0985
3	5.081	9.636	0.140	0.229	0.134	–5.8174	–1.4528
4	5.278	8.331	1.371	0.287	0.177	–6.8322	0.2706
5	5.756	8.929	1.447	0.242	0.091	–4.6014	0.4238
6	5.328	8.839	0.668	0.531	0.272	–4.2471	1.7929
7	4.263	7.784	0.300	0.948	0.460	–1.8632	–0.2848
8	5.442	8.023	0.329	1.389	0.253	–0.4940	–0.6391
9	5.328	8.294	0.207	0.765	0.235	0.8751	–0.9934
10	4.663	7.883	0.223	0.737	0.362	–0.1398	0.7300
11	6.775	9.741	0.788	0.454	0.824	–1.1546	2.4534
12	5.442	8.657	1.112	0.395	0.419	0.2145	2.0992
13	5.421	8.117	1.273	0.247	0.398	4.9824	–2.0562
14	5.602	8.117	0.956	0.449	0.172	3.9676	–0.3328
15	5.442	8.487	0.708	0.457	0.141	3.4602	0.5289
16	5.303	7.955	0.637	0.386	0.360	6.3515	–2.4105
17	5.602	10.545	0.519	0.481	0.261	5.8441	–1.5488
18	5.505	9.687	0.247	0.468	0.450	4.8293	0.1746
19	6.019	8.700	1.664	0.321	0.287	4.6762	2.4054
20	5.464	10.240	0.182	0.380	0.510	6.5527	1.1894

spatial polynomial in the same type of analysis as described here.) Ordination analysis of a species data table, constrained to be related to a spatial polynomial, can be carried out by canonical analysis (Chapter 11). Canonical analysis then becomes an extension to multivariate data tables of the method of trend surface analysis. The method will be described with the help of a numerical example. Another example (vegetation data) is found in Legendre (1990).

Numerical example. Bacterial data from the Thau coastal lagoon are used again here (Tables 10.6 and 13.2). The response data include Bna and Ma bacteria in the present example. No significant linear trend was present in the response data. To facilitate mapping, the X and Y geographic coordinates of the sites were rotated by principal component analysis (PCA using the covariance matrix; scaling type 1 was used); Table 13.2 gives the rotated coordinates. A third-degree polynomial of these new X and Y coordinates was created (Subsection 13.2.1) and subjected to forward selection in order to select the spatial monomials that significantly contributed to the explanation of the Bna and Ma bacterial variables. The following five terms of

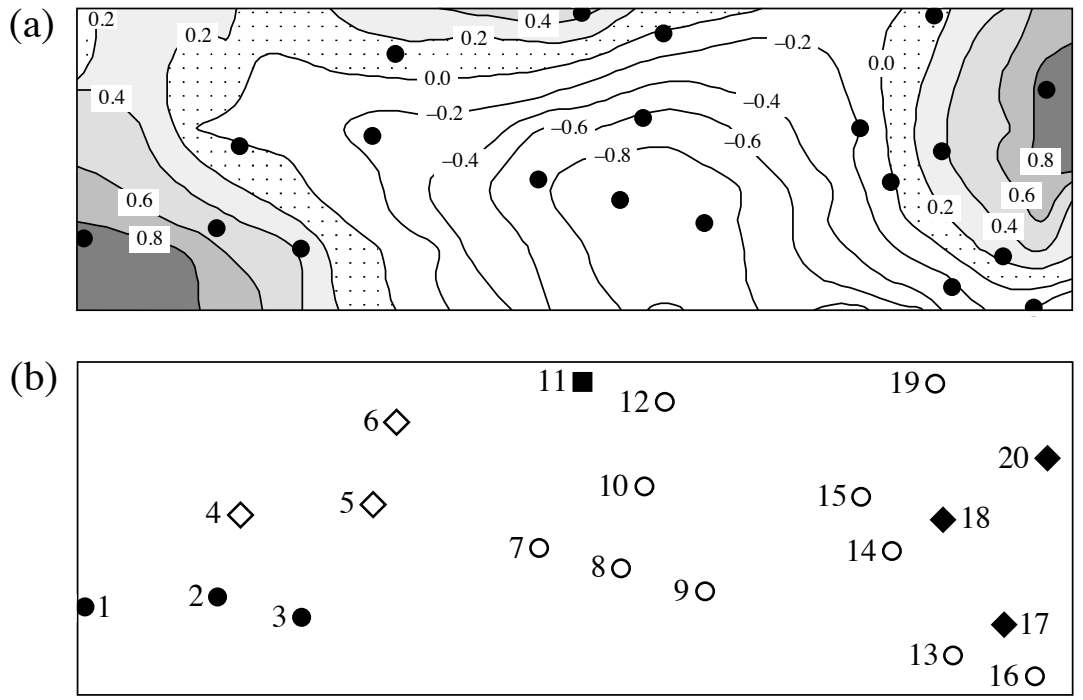


Figure 13.28 (a) Fitted site scores map of the first canonical axis of the bacterial variables in the Thau coastal lagoon constrained by a spatial polynomial. (b) Space-constrained clustering map of the bacterial data, 5 groups (symbols). Dots are the 20 sampling sites, with identification numbers in map b. The north-south direction is nearly parallel to the vertical axis of the maps; compare with the point positions in Fig. 13.15a.

the spatial polynomial were retained by the selection procedure: X^2, X^3, X^2Y, XY^2 and Y^3. Spatial eigenfunction analysis, described in Chapter 14, would not have produced a good spatial model with these data because the 20 sites are too irregularly spaced on the map (Fig. 13.28).

Redundancy analysis (RDA, Section 11.1) produced two canonical axes ($\lambda_1 = 0.622$, $\lambda_2 = 0.111$) because there were two dependent variables only (Bna, Ma) (see Table 11.1). The canonical relationship accounted for $R^2_a = 0.639$ of the variation in the bacterial data; it was globally significant ($p = 0.001$); so was the first canonical eigenvalue. 81% of the variance of Ma was expressed along axis I, but only 7% of the variance of Bna. The second canonical eigenvalue was not significant at $\alpha = 0.05$ ($p = 0.123$) although 42% of the variance of Bna was expressed on this axis. There are two non-canonical axes representing the non-spatially-structured variation of the response variables; they represent 16% and 10% of the variation of matrix **Y**, respectively. Canonical axis I differs from the first principal component of matrix **Y**: that axis would express the variation in the response variables (Bna, Ma) without the constraint of being a linear combination of the spatial monomials.

For axis I, the fitted site scores from matrix **Z** (eq. 11.18) were mapped (Fig. 13.28a) by kriging (Subsection 13.2.2) using program OKB2D of the GSLIB library (Deutsch & Journel, 1992); an all-directional spherical variogram models was fitted to the empirical variograms prior to kriging (Subsection 13.1.3). The trend surface equation that produced the fitted site scores for the 20 sites is:

$$\widehat{\text{Axis I}} = 1.0526X^2 + 1.1881X^3 - 0.5225X^2Y - 0.7674XY^2 + 0.7167Y^3$$

The spatial monomials were standardized before computing this equation.

Interpretation of the map is rather simple in this example: examination of Table 13.2 shows that the sites with the highest scores along canonical axis I (i.e. sites 1-3, 11, 17, 18 and 20, located in grey areas of Fig. 13.28a) possessed the highest concentrations of aerobic heterotrophic bacteria growing on marine agar (variable Ma), which was the variable dominating axis I. These sites also formed three separate and clearly identified groups in the space-constrained clustering map (Fig. 13.28b) produced using function ***constrained.clust()*** using the Ward.D2 algorithm (Subsection 8.5.8); see Section 13.6. The clustering level, 5 groups, displayed in the map was selected by cross-validation.

Thioulouse *et al*. (1995) proposed a different approach to mapping, which combines connection networks, decomposition of the variation into local and global components, eigenvalue decomposition, and mapping. The neighbouring relationships among sites are represented by a connection network (e.g. Delaunay triangulation for a homogeneous two-dimensional sampling area, or neighbouring relationships for sites along a river network) which is translated into a contiguity matrix **M** (Fig. 13.25). **M** is standardized to **P** by division by the total number of pairs of neighbours. A diagonal matrix **D** describes the degree of connectedness of the sites. Using matrices **P** and **D**, the authors define principal component and correspondence analysis for the total, local, and global components of variation; each fraction is decomposed into orthogonal axes, which may be mapped to facilitate interpretation. Their paper presents applications to simulated and real ecological data (bird survey).

13.5 Spatial modelling through canonical analysis

The significance of spatial heterogeneity for the functioning of ecosystems was discussed in Section 1.1. Models of ecosystem processes may fail to correctly model the spatial variation of communities (i.e. beta diversity, Subsection 6.5.3) if they do not include the spatial organization of the populations and communities among the models' predictor variables. This can be achieved by explicitly incorporating spatial predictors in ecological models of community composition data (or other response data matrices) using canonical analysis. The method consists in modelling the variation of the variables of interest across the study area as a linear combination of the environmental variables *and* some function of the geographic coordinates of the sites.

In this approach, one is interested in explicitly identifying the effect of spatial structures, singling it out from other environmental effects, through methods discussed in previous chapters and sections: partial regression analysis (univariate, Subsection 10.3.5) or partial canonical analysis (multivariate, Section 11.1.6) on the one hand; trend surface analysis (univariate, Subsection 13.2.1) and spatially constrained ordination maps (multivariate, Section 13.4) on the other hand. In its basic form, the analysis considers three data sets: **Y** contains the response variables; **X** is the set of explanatory environmental variables; **W** is the set of explanatory spatial variables. In the present section, which introduces the method, **W** contains a polynomial of the geographic coordinates of the sites. In Chapter 14, spatial eigenfunctions will replace the spatial polynomial. Variation partitioning (univariate, Section 10.3.5, or multivariate, Section 11.1.11) produces synthetic presentations of the results. Variation partitioning can actually accommodate more than one environmental and one spatial matrix — up to four with the presently available version of function ***varpart()*** in R.

There are two motivations, described in more detail in Section 14.1.4, to carry out spatial modelling of response matrix **Y** by variation partitioning. The first focuses on fraction [a] of Fig. 10.10, in cases where one wishes to control for spatial correlation in the analysis of species-environment relationships. The second corresponds to situations where both the spatial and non-spatial structures of the explanatory variables **X** are of interest to explain the variation of **Y**, in which case fractions [a], [b] and [c] can all be interpreted. In this approach, any structure identified in the response data is considered to indicate the presence of some process generating it. Mapping fraction [c] of the variation may help generate hypotheses about the spatial process or processes responsible for the observed residual spatial pattern.

The variation partitioning approach described in the present section is essentially correlative. It differs from the analysis of variance, which estimates the variation associated with well-defined effects in structured sampling or manipulative (i.e. controlled) experiments. In the initial stages of ecological research, correlative methods are routinely used to sort out hypotheses centring on broad correlative patterns among groups of variables, before specific hypotheses can be experimentally tested. In particular, the analysis presented in this section allows researchers to consider different groups of explanatory variables (environmental, spatial, or temporal) and examine their capacity to explain patterns in the multivariate response variables (species or others) that are of interest in a study; it further allows one to measure the degree of overlap that exists among these groups of explanatory variables with regard to that capacity (Anderson & Gribble, 1998). The correlations brought out by the analyses are only interpretable insofar as hypotheses can be formulated about the processes that may have generated the observed patterns. This approach is related to regression (Section 10.3) and path analysis (Section 10.4), in which a large number of plausible relationships may be hypothesized and sorted out by statistical analysis. The method is illustrated by a numerical example.

Numerical example. The data of Table 13.2 (Thau coastal lagoon) are reanalysed here. In the numerical example of Section 13.4, the variable selection procedure retained the following terms of the spatial polynomial: X^2, X^3, X^2Y, XY^2, and Y^3; the same terms are used in the present example. The following variation partitioning table (R^2 and R^2_a columns) was obtained using function ***varpart()*** of VEGAN. The values in the three columns to the right were obtained by RDA and partial RDA.

Fractions of variation	Proportion of var. of Y (R^2)	Adjusted R^2	Probability (999 perm.)	Canonical λ_1	Probability (999 perm.)
[a + b]	0.450	0.347	0.005*	0.359	0.025*
[b + c]	0.734	0.639	0.001*	0.622	0.001*
[a + b + c]	0.784	0.628	0.001*	0.632	0.001*
[a]		–0.011≈0	0.549	0.042	0.561
[b]		0.358	-----	-----	-----
[c]		0.281	0.011*	0.304	0.004*
Residuals = [d]		0.372	-----	-----	-----
[a + b + c + d]		1.0000			

In the table, significant fractions at level $\alpha = 0.05$ are identified by asterisks. Eigenvalues of the first canonical axis (canonical λ_1) are reported as fractions of the total variance of **Y**. None of the second canonical axes were significant; since each analysis only produced two canonical eigenvalues in this example, the portions of variation corresponding to λ_2 are the differences between columns 2 and 5 from the left. Fraction [b] is not an independently-calculated component of the variation; hence, it cannot be tested for significance nor decomposed into canonical axes (see Méot *et al*., 1998, for alternative solutions).

Variation partitioning decomposed the total explained variation [a + b + c], expressed as R^2_a, into a significant environmental component [a + b] and a significant component [c] that estimated the spatially-structured variation of **Y** not explained by the environmental variables. The table shows that [a + b], which is the variation of **Y** explained by the environmental variables, was entirely spatially structured; [a] is a small negative value that must be interpreted as a zero. The partitioning results are represented by a Venn diagram in Fig. 13.29.

Figure 13.30 shows maps of the fitted site scores from matrix **Z** (eq. 11.18) of the first canonical axis of fraction [a + b + c] and of its two components, [a + b] and [c]. These maps were obtained by kriging (Subsection 13.2.2) using program OKB2D of the GSLIB library (Deutsch & Journel, 1992); all-directional spherical variogram models were fitted to the empirical variograms prior to kriging (Subsection 13.1.3). While the adjusted proportions of variation of [a + b] and [c] add up to that of [a + b + c] (0.347 + 0.280 = 0.627), this is not the case for the proportions of variation represented by the first canonical axes: $\lambda_{1[a+b]} + \lambda_{1[c]} \neq \lambda_{1[a+b+c]}$. This is because the partition of fraction [a + b + c] into canonical axes is done independently of the partitions of [a + b] or [c]. As a consequence, maps of a given axis of variation (e.g. axis I of the various fractions, mapped in Fig. 13.30) do not exactly add up with this method; they only add up approximately.

The adjusted fraction [b + c] (0.639 of the bacterial data variation) is the one extracted by canonical analysis in the numerical example of Section 13.4 (same data); Fig. 13.28a maps this

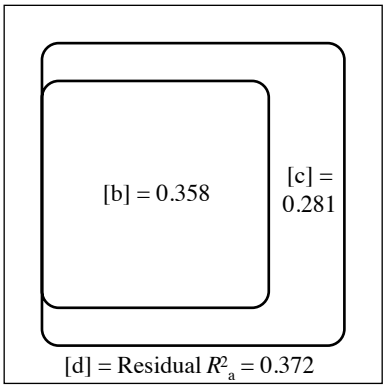


Figure 13.29 Venn diagram illustrating the variation partitioning results for the numerical example. The rounded rectangle surface areas approximate the relative fraction sizes with respect to the size of the outer rectangle, which represents the total variation in the response data. The fractions are identified by letters [b] to [d]; [a] is not shown because it is approximately zero. The values next to the identifiers are adjusted R^2 (R^2_a).

fraction of the variation. In this example, the map of axis I of fraction [b + c] (Fig. 13.28a) is very similar to the map of axis I of [a + b + c] (Fig. 13.30) because [a] is close to zero.

The maps of axis I of fraction [a + b + c] (63% of the variation in the bacterial variables) and [a + b] (36%) are quite similar, whereas the map of axis I of fraction [c] (33% of the variation) is quite different. The trend surface equation that produced the fitted site scores for the 20 sites is:

$$\widehat{\text{Axis I of [c]}} = 1.8017X^2 + 2.2817X^3 - 1.0809X^2Y - 1.3064XY^2 + 1.5563Y^3$$

In this equation, the spatial variables are residuals of the standardized terms of the spatial polynomial after controlling for the effect of the three environmental variables. Examination of the map of fraction [c] suggests a hypothesis for the origin of this fraction of variation, i.e. a marine influence, which had not been included among the explanatory variables in the analysis. Indeed, the negative values on the map form a plume originating at the connections of the Thau lagoon with the sea and extending westwards. To "explain away" fraction [c], i.e. to make it become non-significant, another analysis could be conducted that would include variables quantifying the marine influence on the stations of the lagoon among the environmental variables. Such variables could, for example, be obtained from a hydrodynamic model of the lagoon.

Interpretation of the fractions is described in Subsection 14.1.4. Applications of this method cover a wide range of ecological problems. Here is a selected list of fields and papers: palaeoecology (Zeeb *et al*., 1994; see also Ecological application 10.3b), stream monitoring (Passy, 2007), vegetation (Heikkinen & Birks, 1996; Bjorholm *et al*., 2005), periphyton (Cattaneo *et al*., 1993), protozoa (Buttler *et al*., 1996), zooplankton (Pinel-Alloul, 1995), aquatic macroinvertebrates (Pinel-Alloul *et al*., 1996), fish (Rodríguez & Magnan, 1995), birds (Bersier & Meyer, 1994; Gordo *et al*., 2007), and mammal conservation (Burbidge *et al*., 2008).

Variation partitioning has been applied to more than two explanatory data sets. (1) Pinel-Alloul *et al*. (1995) tested the hypothesis that biotic and abiotic factors, as well as spatial structuring, explained together the broad-scale spatial heterogeneity of zooplankton assemblages among lakes. The explanatory variables comprised abiotic

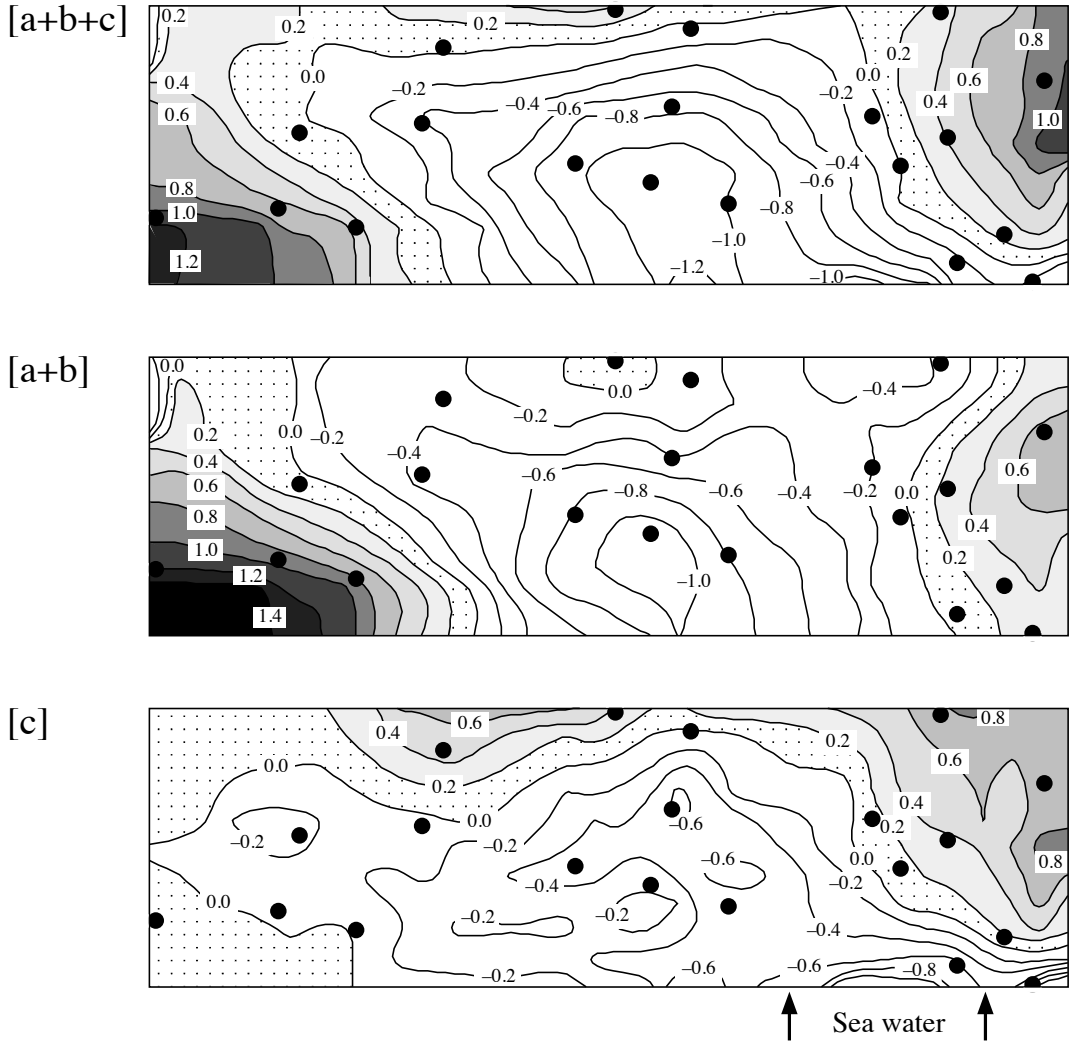


Figure 13.30 Bacterial variables: map of the fitted site scores of the first canonical axes of three fractions of the variation: top [a + b + c], middle [a + b], bottom [c]. Dots represent the 20 sampling sites. North is nearly parallel to the vertical axis of the maps. Compare with Fig. 13.28a, which represents fraction [b + c]. Arrows at the base of map [c], labelled "sea water", indicate the positions of connections of the Thau coastal lagoon with the adjacent Mediterranean Sea.

(physics and chemistry, morphometry) and biotic factors (phytoplankton and fish assemblages); these factors were analysed separately and together, using four approaches described in the paper. (2) Quinghong & Bråkenhielm (1995) explained the spatial patterns of epiphytic green algae and lichens using climatic, pollution, and

Table 13.3 Computer programs implementing the methods of surface pattern analysis discussed in Chapter 13. The list is not exhaustive.

BOUNDARYSEER Detection of boundaries (wombling) and space-constrained clustering

BioMedware Inc., 3526 W. Liberty, Suite 100, Ann Arbor, Michigan 48103, USA. http://www.biomedware.com/

GEOEAS Variogram, kriging; contour mapping

Developed by US Environmental Protection Agency. Available from ACOGS, P.O. Box 44247, Tucson, Arizona 85733-4247, USA. ftp://math.arizona.edu/incoming/unix.geoeas/

GSLIB Geostatistical software library

Geostatistical methods described in the Deutsch & Journel (1992) book. http://www.gslib.com/

ISATIS Variogram, kriging; contour mapping.

Geovariances, 49bis avenue Franklin-Roosevelt, F-77212 Avon Cedex, France. http://www.geovariances.com/en/software-ru2

SAAP Spatial autocorrelograms (Moran's I and Geary's c)

Written by D. Wartenberg, Department of Environmental and Community Medicine, Robert Wood Johnson Medical School, Piscataway, New Jersey, USA. Distributed by Exter Software: http://www.exetersoftware.com

SPACESTAT Directional and omnidirectional spatial variograms; several types of kriging; interpolated maps; global and local spatial correlation statistics; geographically weighted regression

BioMedware Inc., 3526 W. Liberty, Suite 100, Ann Arbor, Michigan 48103, USA; http://www.biomedware.com.

SURFER Kriging; other interpolation methods; contour mapping

RockWare Inc., 2221 East St. #101, Golden, Colorado 80401, USA. http://www.rockware.com/

geographic variables. They showed how to isolate the seven components of variation resulting from crossing three sets of explanatory variables. (3) Anderson & Gribble (1998) applied the variation partitioning method to three matrices of explanatory variables representing the environmental, spatial and temporal components, respectively. Using this approach, they were able to resolve the confounding of space and time that is often encountered when sampling is conducted over a long time period because of the large size of the surveyed area.

13.6 Software

Most of the methods described in this chapter cannot be implemented using the major statistical packages. Table 13.3 lists some of the software available commercially or from researchers. The list is not exhaustive.

The R language offers functions for all methods described in this chapter:

1. Structure functions. — Function ***correlog()*** of NCF: Moran's *I* spatial correlograms computed as described in Subsection 13.1.1 (two-tailed tests). ***sp.correlogram()*** of SPDEP: spatial correlograms based on Moran's *I*, Geary's *c*, or the spatial correlation function (eq. 13.18), based on a list of connection edges (two-tailed tests).

In package NCF, functions ***Sncf()***, ***Sncf.srf()***, ***Sncf2D()***, ***spline.correlog()*** and ***spline.correlog2D()*** estimate spline correlograms in various ways for univariate or multivariate spatial data, including directional correlograms for anisotropic data. Confidence envelopes are computed through bootstrapping. Function ***lisa()*** computes LISA statistics (local Moran) for single response variables. A randomization test is used to test for significance of the LISA statistics.

Functions ***variog()*** of GEOR, ***Variogram()*** of NLME, ***vario()*** of PASTECS, and ***est.variogram()*** of SGEOSTAT compute empirical variograms. A multivariate variogram of **Y** with permutation test is computed by ***mso()*** of VEGAN, after a PCA of **Y** by ***rda()***. A multivariate correlogram is computed by ***mantel.correlog()*** of VEGAN.

2. Maps. — Function ***lm()*** of STATS can be used to compute trend-surface analysis of univariate response data. ***eyefit()*** of GEOR adjusts variogram models to data. ***krige.conv()*** in GEOR carries out conventional kriging. Among the functions available with the Borcard *et al*. (2011) book, ***sr.value()*** draws bubble maps; values of a variable are represented either by circles of different sizes or by circles with shades of grey. Function ***s.value()*** of ADE4 also draws bubble maps using squares instead of circles.

3. Patches and boundaries. — Function ***constrained.clust()*** of package CONST.CLUST* carries out constrained hierarchical clustering along a time or spatial series or on a geographic surface, with cross-validation of the results. Wombling analysis for the estimation of boundaries (Subsection 13.3.3) is computed by package WOMBSOFT. Coefficients of dispersal direction, described in Subsection 13.3.4, are available in function ***bgdispersal()*** of VEGAN.

4. Unconstrained and constrained ordination maps. — R functions are listed in Section 9.5 for unconstrained and in Section 11.7 for constrained ordination.

5. Spatial modelling through canonical analysis. — Functions for canonical analysis and variation partitioning are described in Section 11.7

6. Miscellaneous methods. — Function ***geoXY()*** in SODA transforms latitude-longitude (LatLon) data to flat Cartesian coordinates. ***lm()*** in STATS can be used to carry out spatial detrending of univariate or multivariate data. ***poly()*** in STATS computes ordinary or orthogonal polynomials.

* Available on the Web page http://numericalecology.com/rcode.

Chapter

14 *Multiscale analysis: spatial eigenfunctions*

14.0 Introduction to multiscale analysis

Tobler's (1970) first law of geography states that "Everything is related to everything else, but near things are more related than distant things". This law is at the core of spatial analysis and modelling in geography and related fields such as biogeography, community ecology, population biology, landscape ecology and landscape genetics.

Spatial eigenfunction analysis is a family of methods for multiscale analysis of univariate or multivariate response data. Based on the theory introduced in Section 1.1 concerning the origin of spatial structures in ecosystems, these methods draw upon several developments discussed earlier in this book: distances (Chapter 7), principal coordinate analysis (Section 9.3), multiple regression modelling (Section 10.3.3), redundancy analysis (Section 11.1), variation partitioning (Sections 10.3.5 and 11.1.11), and the concept of scale in spatial patterns (Section 13.0).

The expression *spatial eigenfunction analysis* was proposed by Griffith & Peres-Neto (2006) for the whole family of methods where eigenvectors of spatial configuration matrices are computed and used as predictors in linear models, including the full range of general and generalized linear models*. The expression proposed by these authors covers both the early methods developed by geographers to analyse binary spatial connection matrices (Garrison & Marble, 1964; Gould, 1967; Tinkler, 1972; Griffith, 1996; these methods are briefly described in Subsection 14.2.2) and the more recent methods that take into account the distances among localities and are described in the present chapter. Extension of spatial eigenfunction analysis to time series is straightforward; all methodological developments in this chapter labelled "spatial" could readily be changed to "temporal".

* These two forms of linear models are described on the Web pages http://en.wikipedia.org/wiki/General_linear_model and http://en.wikipedia.org/wiki/Generalized_linear_model.

Multiscale spatial analysis is used to answer questions like the following:

• Description: What are the spatial scales of variation of the [univariate or multivariate] response data under study? What are the spatial patterns at these scales?

• Explanation (in the sense of Subsection 10.2.1): What are the processes that explain (meaning *account for*) the spatial variation of the response data at different scales? Indeed, different processes may affect (or be associated with) that spatial variation at different scales.

• For communities of organisms, beta diversity is the spatial variation in community composition among sites (Subsection 6.5.3). How does beta diversity relate to different types of processes at different scales, for example environmental control (Subsection 1.1.1, model 1) and neutral processes (Subsection 1.1.1, model 2)?

Space

In multiscale spatial analysis, the variation in response data is analysed with respect to variables (eigenfunctions) representing geographic variation, which may be divided into submodels corresponding to different spatial scales; see the following sections. Of course, space *per se* cannot be considered as an explanation of ecological variability. The spatial variables used as explanatory variables in analyses are proxies for real environmental or ecological explanatory variables. The spatial proxies serve to quantify and dissect the spatial variation present in the response data. Part of that variation can then be attributed to some of the potentially explanatory variables that are available for analysis, the remainder being considered as spatial variation that remains to be explained. Subsection 14.1.4 discusses different hypotheses that can be invoked to explain such variation.

Trend-surface analysis described in Subsection 13.2.1 and used for spatial modelling in Section 13.5 is a rather crude method. A model that simply uses the spatial coordinates is sufficient to model a flat surface; a quadratic model (coordinates to the powers 1 and 2) can represent a bowl or saddle shape; a cubic model (powers 1, 2 and 3) has one more bend in each geographic direction. To model fine structures would require a polynomial equation with more monomials than there are objects, which would render the method useless in practice for data analysis.

Given the lack of efficient methods to model multiscale spatial structures until the late 1990's and the beginning of the 21st century, researchers were then looking for appropriate approaches. Ideally, the modelling matrix should contain mutually orthogonal vectors; because of that property, these vectors could be combined into submodels, corresponding to different spatial scales, that would be linearly independent of one another in variation partitioning. The following sections describe modelling methods that meet these expectations: dbMEM analysis (formerly called PCNM analysis, Section 14.1); generalized Moran's eigenvector maps (MEM, Section 14.2); asymmetric eigenvector maps (AEM) developed to model the effects of directional physical processes (Section 14.3); and multiscale ordination (Section 14.4). Section 14.5 describes derived methods of spatial analysis based on MEM, and

Section 14.6 is a rejoinder that shows how these methods can help answer questions involving the multiscale analysis of beta diversity. Section 14.7 lists R functions available to carry out the calculations.

14.1 Distance-based Moran's eigenvector maps (dbMEM)

The positions of study sites on a map are identified by their spatial coordinates. Subsection 13.2.1 has shown how geographic coordinates can be used in a form of geographic modelling known as spatial trend-surface analysis. In that method, the coordinates are used, either directly (i.e. without transformation) or in the form of a polynomial function, to model univariate or multivariate response data through multiple regression (Subsection 13.2.1) or canonical analysis (Section 13.5).

Another way to look at the relative positions of the study sites is to compute geographic distances among them and write these to a geographic distance matrix $\mathbf{D}_{\text{Geo}}$. A simple form of spatial eigenfunction analysis is obtained by modifying the geographic distance matrix $\mathbf{D}_{\text{Geo}}$, as explained below, and computing spatial eigenfunctions by eigen-decomposition. That form was described in three papers by Borcard & Legendre (2002), Borcard *et al*. (2004) and Legendre & Borcard (2006), where the spatial eigenfunctions were called *Principal Coordinates of Neighbour Matrices* (PCNM). They are a special class of a family of eigenfunctions called *Moran's eigenvector maps* (MEM) described in Section 14.2. The formerly called PCNM eigenfunctions are actually MEM eigenfunctions based on simple geographic distances; they are now called *distance-based MEM*, abbreviated dbMEM.

PCNM
MEM
dbMEM

The construction of dbMEM eigenfunctions is described in Subsection 14.1.1. After their construction, spatial eigenfunctions can be used in linear models of the response data in the same way as polynomials of geographic coordinates: they become explanatory variables in multiple linear regression when analysing univariate response data, e.g. species richness at different sites, or in canonical analysis when studying the spatial variation of multivariate response data, e.g. community composition. They can also be used as one of the explanatory matrices in variation partitioning aimed at analysing a matrix of response data from two or more angles. These aspects are illustrated through ecological applications presented in Subsection 14.1.3.

Ecologists who are not familiar with multiscale eigenfunction analysis may consider looking at the examples and ecological applications of Subsections 14.1.2 and 14.1.3 first, before coming back to study the algorithm in detail.

1 — Algorithm

The steps involved in the construction of dbMEM eigenfunctions are the following:

• Compute a matrix of geographic distances by applying the Euclidean distance function to a set of Cartesian geographic coordinates. Latitude-longitude data can be transformed into flat Cartesian coordinates using function ***geoXY()*** (Section 14.7). Alternatively, for sites covering a large geographic area on the Earth's surface, geodesic distances can be computed. The end product of this first step is a square symmetric matrix of distances among sites.

• Choose a distance threshold called '*thresh*' to truncate the geographic distances, separating them in two groups: small and large distances. This way of proceeding was inspired by the division of distances into distance classes in Mantel correlograms (Subsection 13.1.6). How to determine the value of the threshold is described below. The distances smaller than or equal to the threshold are kept as they are in the modified distance matrix $\mathbf{D}_{\text{trunc}}$. The distances larger than the threshold are replaced by an arbitrary large distance. $\mathbf{D}_{\text{trunc}}$ is a truncated distance matrix because the distances larger than *thresh* have been removed and replaced by a large constant value. The value arbitrarily used in computer software is 4 times the value of the threshold. Any value larger than 4 would serve equally well the purpose of distorting the distance matrix and allowing spatial eigenfunctions to be computed from it, with little change to the numerical results.

• In dbMEM, the diagonal values of the distance matrix, which were originally zeros, are replaced by the value $(4 \times thresh)$[*]. This change on the diagonal of $\mathbf{D}_{\text{trunc}}$ indicates that a site is not connected to itself; this is also the case in the computation of Moran's *I* coefficients in correlograms.

• Compute a principal coordinate analysis (PCoA, Section 9.3) of $\mathbf{D}_{\text{trunc}}$ producing eigenvalues and eigenvectors. If PCoA were computed from $\mathbf{D}_{\text{Geo}}$ instead, the relative positions of the sites would be recovered in a two-dimensional ordination of the points; so there would be two positive eigenvalues (or three for sites representing a large area on the Earth's surface), and all the other eigenvalues would be 0. Here, PCoA is applied to the distorted (truncated) matrix $\mathbf{D}_{\text{trunc}}$. The surprising consequence is the production of $(n-1)$ eigenvalues different from 0 and $(n-1)$ corresponding eigenvectors, instead of two. Some of the eigenvalues are positive, some are negative. The examples below will show the balance between positive and negative eigenvalues, and what they mean. In the calculation of eigenvectors, the signs along any one eigenvector can be switched among software or computer platforms, because the sign

[*] In classical PCNM eigenfunctions, the diagonal values of $\mathbf{D}_{\text{trunc}}$ are 0, indicating that a site is connected to itself. The eigenvalues of classical PCNM analysis are larger than those of dbMEM analysis by a constant equal to $(4 \times thresh)^2/2$, so that there are artificially more positive eigenvalues in the PCNM than in the dbMEM solutions, but the eigenvectors, which are the spatial eigenfunctions, are identical to those computed by the dbMEM procedure described here.

of the first element of each eigenvector (+ or –) is assigned arbitrarily, as explained in Chapters 2 and 9.

• The principal coordinates, which are the spatial eigenfunctions, represent (or *model*) together the multiscale distance relationships among the sites. They can be used in the same ways as any other types of explanatory variables: they can be mapped (examples are shown below); or used as explanatory variables in linear modelling (multiple linear regression, generalized linear models, canonical analysis, etc.); or used in variation partitioning. Subsets of them can be selected by linear model selection procedures.

Moran's eigenvectors

The eigenfunctions computed in this way are called *Moran's eigenvectors* because their eigenvalues are equal to Moran's I coefficients of spatial correlation (eq. 13.1) computed for these eigenfunctions using the pairs of sites that remain connected after truncation, divided by a constant (Dray *et al*., 2006). In the case of a linear transect with equispaced points (example developed below), about half the eigenvectors have positive Moran's I and model positive spatial correlation, and the other half have negative Moran's I and model negative spatial correlation at short range. In most ecological studies, only the eigenvectors with positive Moran's I are used to model the spatial correlation in data, but the eigenvectors with negative Moran's I are also available to model the response data. In studies of territorial animals, for example, the eigenfunctions with negative Moran's I allow researchers to test hypotheses formulated to explain the negative spatial correlation among the study sites.

How should the truncation threshold be chosen? The method for choosing the value of the threshold, *thresh*, derives from the observation that MEM eigenfunctions display variation across the full set of sites under study if the sites form a connected graph in matrix $\mathbf{D}_{\text{trunc}}$, meaning that there is chain of connections made of distances smaller than or equal to *thresh* linking all sites; the concept of G_c-chain is explained in Section 8.2. If there are, say, two groups of points (sites) with no connection between the groups, the variation within each group is modelled by subsets of dbMEM eigenfunctions that do not vary in the other group. This observation suggests the following method: to ensure that all points are modelled by the same set of eigenfunctions, choose the value of *thresh* in such a way that the distances smaller than or equal to *thresh* in $\mathbf{D}_{\text{Geo}}$ form a G_c-chain linking all points in a connected graph. Disconnected pairs are identified in matrix $\mathbf{D}_{\text{trunc}}$ by distances equal to $(4 \times thresh)$. This leads to the following recommended method:

• Create a minimum spanning tree (MST, Sections 8.2 and 13.3.1) linking all points (sites) in the study. Identify the length of the largest edge in the chain forming the MST. An illustration is provided with Numerical example 3 below.

• Set *thresh* equal to the length of the largest edge in the MST, or any other value of the user's choice that is larger than that value. See the examples below. In practice, choosing a *thresh* value equal to the longest edge forming the MST maximizes the number of eigenfunctions that model positive spatial correlation. Although it is compatible with the definition of dbMEM eigenfunctions, choosing for *thresh* a value

larger than that reduces the number of eigenfunctions that model positive spatial correlation. The first few eigenvectors remain unchanged, but the following ones are changed. Hence it is recommended to routinely use the smallest possible truncation level for *thresh*, i.e. the value provided by the MST.

It may be interesting and appropriate in some studies to analyse together two or several disconnected groups of sites. In that case, separate sets of dbMEM eigenfunctions should be generated before analysing the spatial structure of response data observed at these sites, and assembled in a staggered matrix of eigenfunctions that will allow a single analysis to be conducted for the sites belonging to the separate groups. This is a far better practice than trying to create a single set of dbMEM eigenfunctions by choosing a large *thresh* value. Indeed, that would compromise the resolution of the analysis by affecting the fine-scale dbMEM eigenfunctions. Function ***create.MEM.model()*** described in Section 14.7 allows users to generate eigenfunctions for this type of analysis. An application is found in a study of the spatial metacommunity architecture of zooplankton in groups of pools located in separate valleys of the High Andes in Bolivia (Declerck *et al*., 2011).

A MEM scalogram is a diagram representing the proportion of variance (R^2) explained by the MEM eigenfunctions ordered along the abscissa by decreasing eigenvalues (Legendre & Borcard, 2006). For univariate response data **y**, the scalogram can display the Pearson correlations, the regression coefficients, or the absolute values of the *t*-statistics associated with the regression coefficients computed between **y** and the MEM eigenfunctions. Because the MEM eigenfunctions are orthogonal to one another, the partial regression coefficients obtained in a multiple regression of **y** on all MEM eigenfunctions are equal to the simple regression coefficients between **y** and each MEM eigenfunction in turn. For a multivariate response matrix **Y**, the ordinate of the scalogram can display either the R^2 explained by the various MEM eigenfunctions or the associated *F*-statistics. A *t*-value scalogram displays *t*-statistics obtained by multiple regression of a single response variable on the MEM eigenfunctions. An example of *t*-value scalogram is shown in Fig. 14.5 (Ecological application 14.1a).

2 — Numerical examples

This subsection examines numerical examples of dbMEM eigenfunctions produced for different sampling designs.

Numerical example 1. Consider a transect with 50 equally-spaced sites. Only the positions of the points along the transect are required for the generation of the inter-point distance matrix $\mathbf{D}_{\text{Geo}}$ and the calculation of dbMEM eigenfunctions; the point positions were represented by the integers 1 to 50 for the calculations, but they could have been given by any other series of 50 equally-spaced values. There were 49 non-zero eigenvalues and 49 corresponding eigenvectors produced. The eigenvalues, ranging from the largest ($\lambda_1 = 14.9$) to the smallest ($\lambda_{49} = -15.0$), ordered the eigenfunctions by wavelengths, from broad scale to fine scale. Twenty-four eigenvectors had positive eigenvalues and Moran's *I*, and 25 eigenvectors had negative eigenvalues and Moran's *I*.

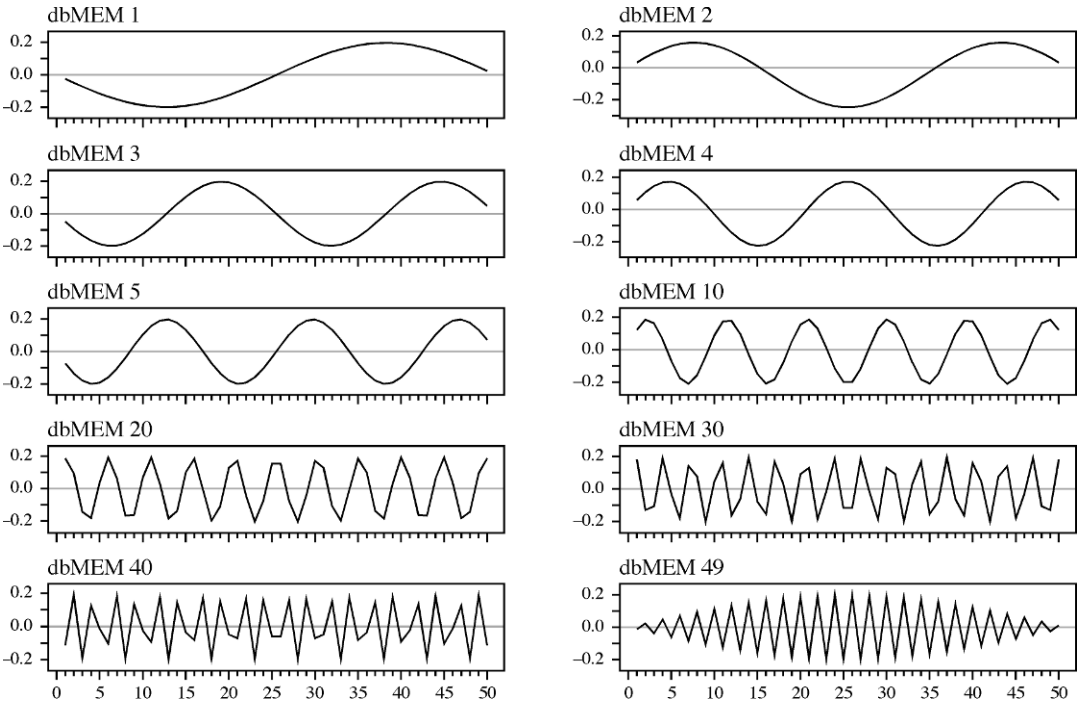


Figure 14.1 Graphs of ten of the 49 dbMEM eigenfunctions that represent the spatial variation along a transect with 50 equally-spaced points. Abscissa, from left to right: sites 1 to 50. Ordinates: values along the dbMEM eigenfunctions.

Figure 14.1 shows some of the 49 eigenfunctions that were produced. The first seven eigenfunctions illustrated in the figure (dbMEM 1 to 5, 10, 20) belong to the group modelling positive spatial correlation. dbMEM 30, 40 and 49 displays negative spatial correlation at small distances; indeed, these eigenfunctions have negative Moran's I values. The values along each eigenfunction form a sine wave. Signs may be reverted along any one eigenfunction because different software, or the same software running on different computer platforms, can produce switched signs along any of the eigenvectors in PCoA. The wavelengths of eigenfunctions are rarely integer multiples of the sampling interval. Because the eigenfunctions are only computed and plotted at the 50 sampling points, interference generates amplitude-modulation (AM) waves in panels dbMEM 30, 40 and 49, although the eigenfunctions are perfectly correlated with sine waves.

Guénard *et al*. (2010, eq. 3) showed that for a transect containing n regularly-spaced points and with sampling interval s, the wavelength λ_i of the sine wave corresponding to the eigenfunction with rank i is:

$$\lambda_i = 2\frac{(n+s)}{(i+1)} \qquad \textbf{(14.1)}$$

For example, for a 50-point transect as the one used to compute Fig. 14.1, the complete sine wave of dbMEM 1 has a wavelength of 51 units (eq. 14.1), compared to the length of the transect which is 50 – 1 = 49 inter-point units, assuming sampling points with negligible size. The following eigenfunctions form sine waves of shorter wavelengths, as predicted by eq. 14.1. The last eigenfunction shown in Fig. 14.1, dbMEM 49, has a wavelength of 2.04; hence it is not in phase with the set of points that are spaced by 1 unit and a false wave that has the length of the series modulates the amplitude of the sine wave in the envelope of the eigenfunction graph (amplitude modulation, or AM, wave). When the points are irregularly-spaced along a transect, the sine waves are deformed, but one can still recognize the eigenfunctions that model broader-scaled and finer-scaled phenomena. Examples are given by Borcard *et al*. (2004, Appendix C).

Numerical example 2. Surveys of permanent forest plots[*], as well as many field experiments, are conducted on regular grids of points. A regular 12 × 8 grid (96 points) was generated for the present example to illustrate dbMEM eigenfunctions on regular grids. The eigenfunctions were computed for that grid using a *thresh* value of 1, which corresponded to the distance between adjacent points in the horizontal and vertical directions. There were a total of 95 eigenvalues and corresponding eigenfunctions. The first 48 eigenfunctions modelled positive spatial correlation[†] and the last 47 modelled negative spatial correlation.

Figure 14.2 shows maps of ten of the 48 eigenfunctions modelling positive spatial correlation. The patterns alternate between vertical, horizontal, and diagonal contrasts. The two central horizontal lines of points in the map of dbMEM 1 have the same large sine wave shape as dbMEM 1 in Fig. 14.1: the values go from 0 (grey) on the left, to negative values (white), to 0 (grey) in the centre of the lines, to positive (black), and back to 0 (grey) on the right.

For a square regular grid of points, many of the pairs of successive eigenfunctions produced during dbMEM generation have multiple eigenvalues (Section 2.10, third property), creating situations of circularity. Different software, or different computer platforms using the same software, may produce pairs of eigenvectors that are rotated differently. Note, however, that these pairs of eigenfunctions explain together the same amount of variation in the data, whatever the rotation. In any case, the patterns on some eigenfunction maps are symmetric while for pairs of multiple eigenvalues, successive maps display the same pattern with a 90º rotation. Borcard *et al*. (2011, Fig. 7.4) showed spatial eigenfunction maps for a square grid of points.

Numerical example 3. Nine points were used to illustrate the construction of connection networks in Figs. 13.22 to 13.24. These points are used again here to compute dbMEM

---

[*] See note on the CTFS permanent forest plots in Subsection 6.5.3 and Ecological application 14.1b where one of those forest plots is analysed.

[†] The expected value of Moran's I under the null hypothesis regarding the absence of spatial correlation is a small negative value, $E(I) = -(n-1)^{-1}$ (eq. 13.6). Sometimes, as in the present example, it happens that a dbMEM eigenfunction has a small negative Moran's I value that is larger than $E(I)$, and thus also has a small negative eigenvalue. In the example, eigenfunction 48 had a Moran's I value of –0.00948, which is larger than $E(I)$ = –0.01053; it was counted among those that modelled positive spatial correlation.

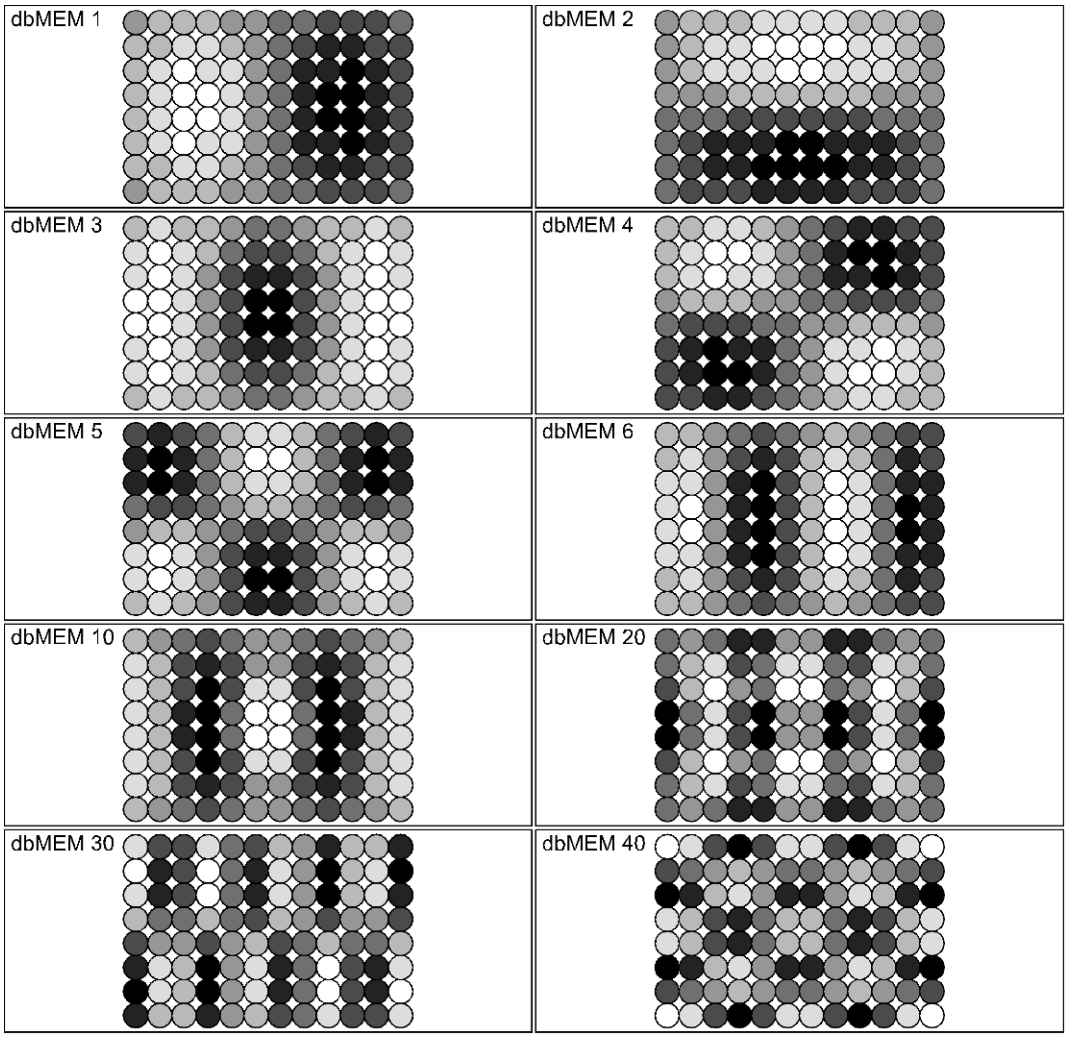


Figure 14.2 Bubble maps showing ten of the 48 dbMEM eigenfunctions that display positive spatial correlation. Shades of grey represent the values in each eigenvector, from white (largest negative value) to black (largest positive value). Signs may be reverted in the construction of the eigenvectors with no consequence for the analysis; reverted signs would interchange black and white in the panels. The maps were produced using function ***sr.value()*** (Section 13.6).

eigenfunctions. Figure 14.3 shows the minimum spanning tree computed from the spatial coordinates of the points found in Fig. 13.22. The longest edge along the tree has a length of 3.04 units; that length was used as the *thresh* value for the computation of the truncated distance matrix $\mathbf{D}_{\text{trunc}}$. Figure 14.4 shows maps of the eight eigenfunctions produced by dbMEM

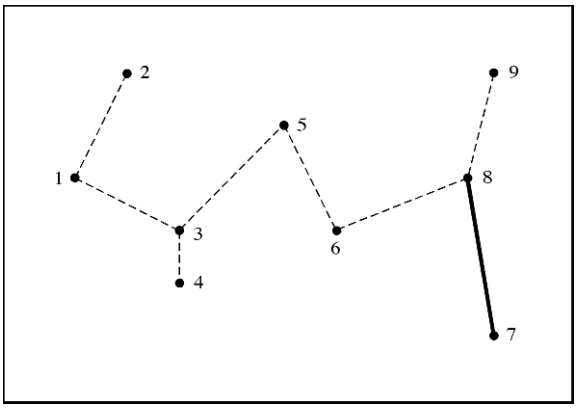


Figure 14.3 Minimum spanning tree of the nine points from Fig. 13.22. The longest edge, between points 7 and 8, is in bold; its length (3.04) is used as the *thresh* value. The dashed edges are shorter.

decomposition: four display positive spatial correlation and have positive eigenvalues and Moran's I, and four display negative spatial correlation and have negative eigenvalues and Moran's I. The map of dbMEM 1, for example, displays a gradient between negative values (white) on the left and positive values (black) on the right. dbMEM 2 contrasts the centre of the map (white circles) with the left and right edges (dark grey and black circles). The last maps have circles with contrasting shades side by side, picturing negative spatial correlation at short range.

An important methodological point concerns the analysis of data that embed a linear spatial gradient. A linear gradient can be modelled by dbMEM eigenfunctions. Along a regular transect, for example, the even-numbered dbMEM eigenfunctions can be used together in a linear function to model a linear gradient almost perfectly, i.e. with a very high R^2. This is because the even-numbered eigenfunctions (1, 3, 5, etc.) are asymmetrically positioned with respect to the centre of the transect, as can be seen in Fig. 14.1. The odd-numbered eigenfunctions (2, 4, 6, etc.), on the contrary, have symmetrical shapes with respect to the centre of the transect and cannot be used to model a linear gradient. This being said, it is not good practice to use eigenfunctions to model a gradient. There are two reasons for this. Firstly, a gradient can be seen as a portion of a spatial structure that is much larger than the study area. Nothing will be learned by using half of the eigenfunctions to model such a structure, which can be modelled more simply by a linear function of the positions of the sites along a transect or their geographic coordinates on a surface. Secondly, if eigenfunctions are used to model a linear gradient, they cannot be used to model more interesting spatial structures. As a consequence, when the response data contain a linear gradient in one or two geographic dimensions, it is recommended to detrend them prior to MEM analysis (Subsection 13.2.1).

Detrend the data

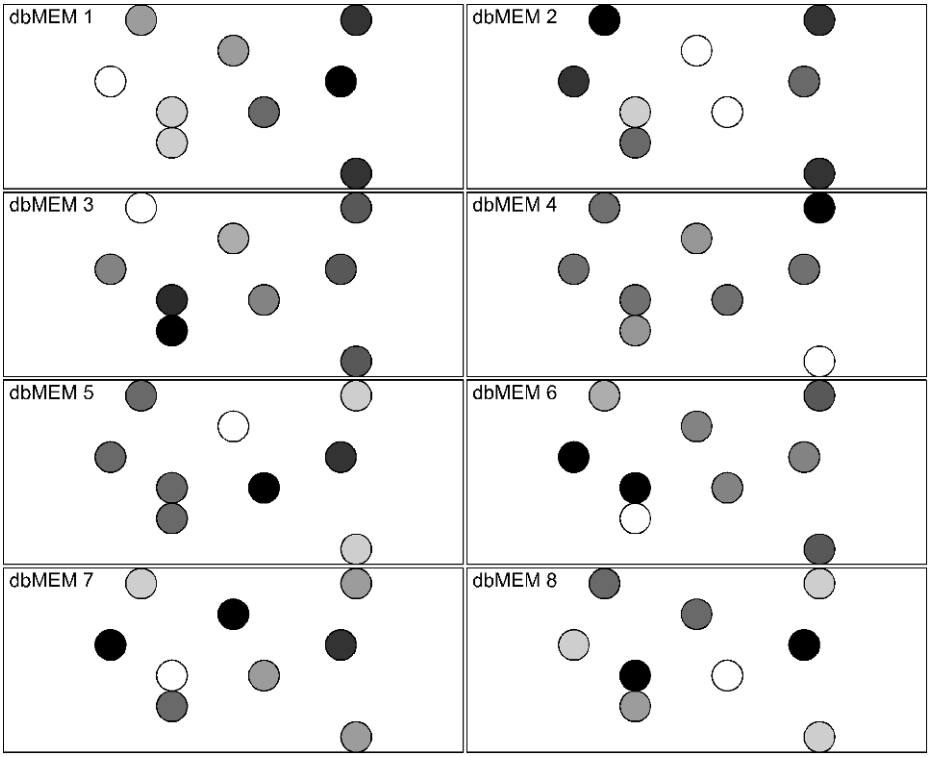


Figure 14.4 Bubble maps showing the eight dbMEM eigenfunctions for the nine points from Fig. 13.22. The first four model positive spatial correlation. Shades of grey: see Fig. 14.2.

Users of spatial eigenfunction analysis are warned against the temptation to interpret a single MEM eigenfunction that happens to fit well a spatial structure that can be observed in response data. Had the set of study points be offset with respect to the actual study area, for instance to the east or to the west, that MEM would probably not fit the response data so well, or not at all, and the structure would then be fitted by other eigenfunctions. The message here is that users should look for sets of MEM eigenfunctions, corresponding to a given spatial scale, that, together, fit response data fairly well, not individual eigenfunctions.

3 — Ecological applications

Many applications of spatial eigenfunction analysis of the dbMEM type are available in the ecological literature. Three of those are summarized here.

Ecological application 14.1a

This application illustrates dbMEM analysis for a single response variable along a transect of equally-spaced sampling sites. Borcard *et al*. (2004) reanalysed data originally collected and analysed by Tuomisto & Poulsen (2000) who surveyed the fern assemblages (32 species) along a 1300 m transect in the tropical forest, in the Huanta region of the Upper Amazonian River in northeastern Peru. The response variable examined in the present application is the abundance of the fern *Adiantum tomentosum* Klotzsch in 260 adjacent 5 × 5 m quadrats. Explanatory environmental variables were also collected along the transect. The objective of the study presented by Borcard *et al*. (2004) was to determine the spatial scales at which the abundance of the species was structured, and relate these scales to environmental variables that were hypothesized to affect the spatial distribution of the species.

In the analysis of the Huanta fern data, Borcard *et al*. (2004) used all 176 PCNM eigenfunctions that had positive eigenvalues. These corresponded to dbMEM eigenfunctions 1 to 129 that modelled positive spatial correlation plus the first 47 dbMEM that modelled negative spatial correlation. In the analysis recomputed for the present application, only the 129 dbMEM that modelled positive spatial correlation were used; this corresponds to present-day practice. The changes to the results are small as can be seen by comparing the present results to those shown in Appendix B of Borcard *et al*. (2004). A dbMEM model could also be developed based on the eigenfunctions that model negative spatial correlation; such a model, if significant, would display avoidance phenomena.

Prior to analysis against the dbMEM and environmental variables, the abundance data were square-root transformed to make the distribution of abundances more symmetrical, albeit not strictly normal. There was no significant spatial trend in the data, so no detrending was carried out. The dbMEM eigenfunctions were computed along the transect; the integers 1 to 260 represented quadrat positions along the transect. Among the 129 dbMEM that had positive Moran's I values and thus modelled positive spatial correlation, 26 significant eigenfunctions ($p \leq 0.05$) were identified by forward selection (Section 11.1.10, paragraph 7). In the scalogram (Fig. 14.5), these eigenfunctions are identified by black symbols. These same 26 eigenfunctions were also significant in a multiple regression of the response variable against the 129 dbMEM eigenfunctions.

Together, these 26 significant eigenfunctions formed a descriptive model with $R^2_a = 0.568$. They were divided into four submodels as shown in the scalogram (Fig. 14.5): the first seven dbMEM were called the very-broad-scale (VBS) submodel, the next seven the broad-scale (BS) submodel, the next six the medium-scale (MS) submodel, and the last six formed the fine-scale (FS) submodel. Natural clusters of significant values can be seen in the scalogram, especially for the first two groups; they helped divide the dbMEM into submodels. Because there is a single response variable in this application, i.e. the *Adiantum tomentosum* fern abundance, fitted values were computed by regression of the species data on the eigenfunctions forming each submodel. These fitted values are show in Fig. 14.6 b-e. All submodels were statistically significant.

The next step of the analysis was to identify the environmental variables that corresponded to the different submodels.

• The very-broad-scale (VBS) submodel was explained by six variables that were significant in forward selection ($R^2_a = 0.382$ with respect to the variance in the VBS fitted values): quadrat elevation (this variable was detrended against quadrat positions along the transect prior to the analysis), trees 3-7.5 cm dbh (dbh is diameter at breast height), trees 31.5-62.5 cm dbh, lianas 8-15 cm diameter, thickness of soil organic horizon, canopy height.

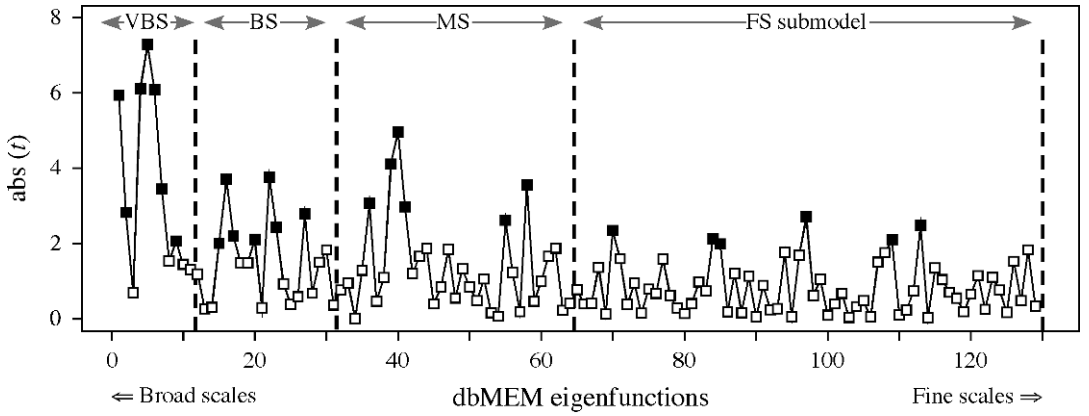


Figure 14.5 Scalogram of the fern *Adiantum tomentosum* multiscale structure along the Huanta transect. Abscissa: the 129 dbMEM eigenfunctions with positive Moran's *I*. Ordinate: absolute values of the *t*-statistics. The 26 eigenfunctions selected by forward selection ($p \leq 0.05$) are identified by black squares. Dashed lines indicate the divisions into the VBS, BS, MS and FS submodels.

• The broad-scale (BS) submodel was explained by a different linear model containing two of the same variables as in the VBS submodel ($R^2_a = 0.124$ with respect to the variance in the BS fitted values): quadrat elevation, and lianas 8-15 cm diameter.

• The medium-scale (MS) submodel was explained by two variables ($R^2_a = 0.064$ with respect to the variance in the MS fitted values): waterlogging (saturation of the soil by groundwater), which was not selected in the VBS and BS submodels, and canopy height.

• The fine-scale (FS) submodel was not explained by any of the environmental variables that were available for the analysis.

In multiscale analysis of ecological data, one often finds that the fine-scale submodel is not explained by the available environmental variables, although the eigenfunctions composing it are significant. One remains uncertain as to the interpretation to give to this observation: either the environmental variables that could explain the fine-scale variation have not been measured, or the fine-scale spatial structure displayed by the response data is not due to environmental control (Chapter 1, eq. 1.1) but represents autocorrelation (eq. 1.2) generated by the dynamics of the population (in the present example), or by community dynamics when studying community composition data. This question of interpretation is revisited in Subsection 14.1.4.

Variation partitioning (Subsections 10.3.5 and 11.1.11) was performed with respect to (1) the seven environmental variables selected above, (2) the 20 dbMEM eigenfunctions forming submodels VBS, BS and MS that are related to the environmental variables, and (3) the six dbMEM eigenfunctions forming the FS submodel. Figure 14.7a shows the partitioning results. Even though the two sets of dbMEM eigenfunctions were uncorrelated, subtraction of R^2_a coefficients artificially created small negative values in the fractions of the partition corresponding to intersections between the dbMEM submodels; these small values should be interpreted as zeros. This annoying problem can be corrected by creating a hierarchy among the

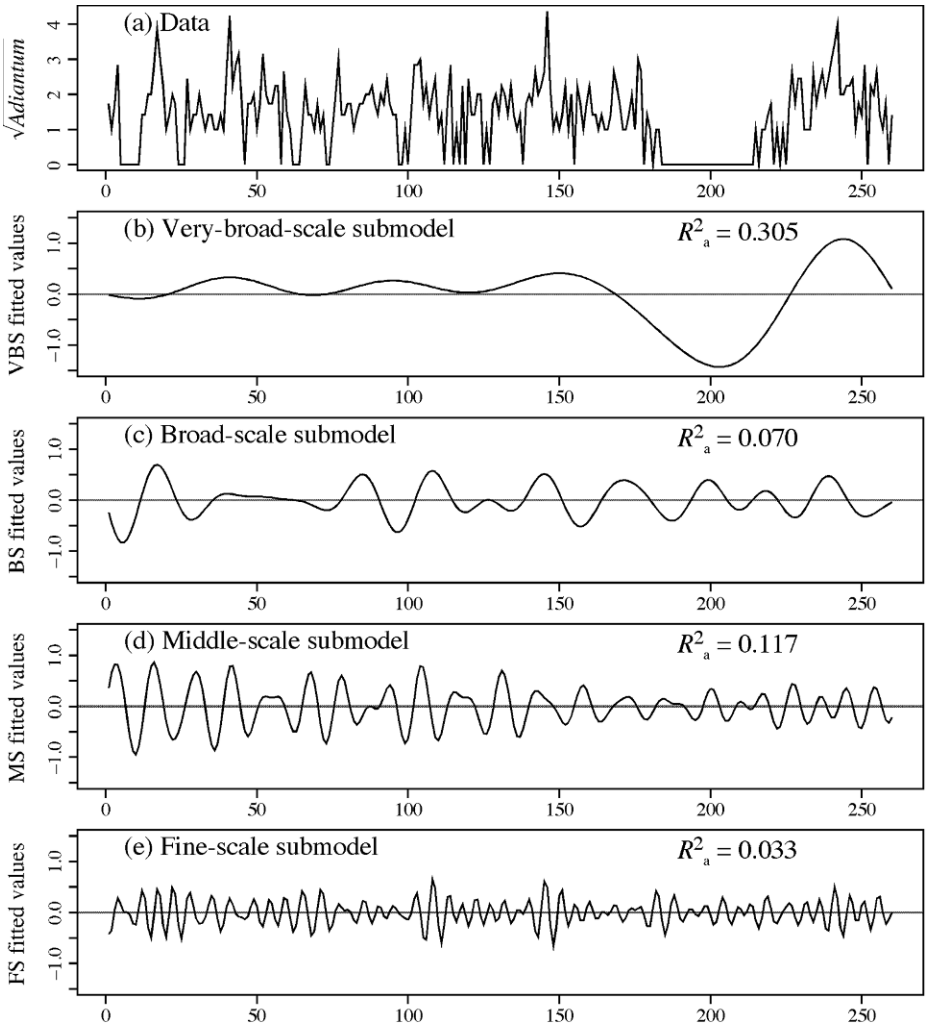


Figure 14.6 MEM analysis of the fern *Adiantum tomentosum* in the Huanta transect, Peru. Abscissa: quadrat positions along the transect. (a) Square-root transformed abundances. Zero values were observed in portions of the transect. (b) Fitted values of very-broad-scale (VBS) submodel, centred (seven dbMEM); (c) of broad-scale (BS) submodel (seven dbMEM); (d) of medium-scale (MS) submodel, (six dbMEM); (e) of fine-scale (FS) submodel (six dbMEM). The adjusted R^2 of each submodel is shown. The ordinate scale is the same in graphs b-e to emphasize differences in explained variation among submodels.

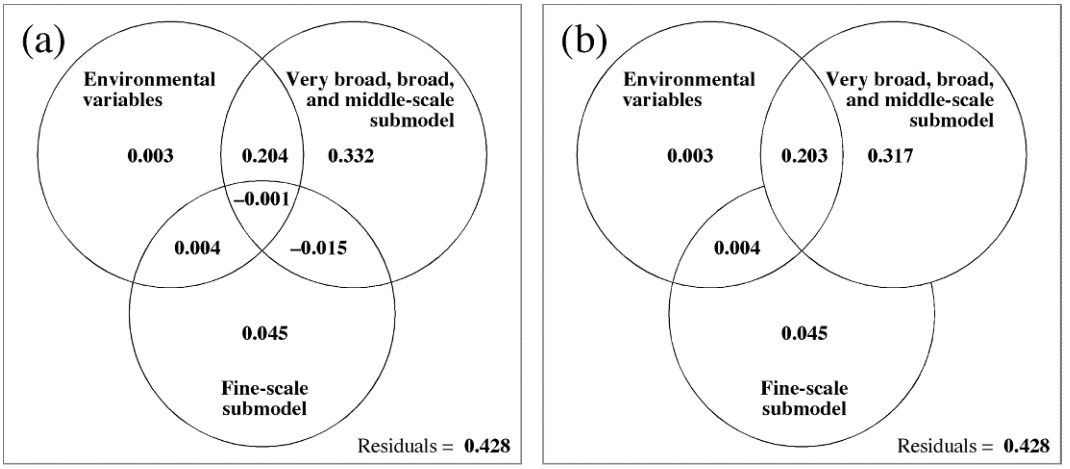


Figure 14.7 (a) Venn diagram presenting the variation partitioning results for the fern *Adiantum tomentosum* along the Huanta transect. The fraction values displayed are computed from adjusted R-squares (R^2_a). (b) Same partition with hierarchical partitioning, which keeps the two sets of dbMEM orthogonal (intersection fractions = 0). The diagram on the left is the one produced by function ***varpart()*** of the VEGAN package in R, onto which text was added.

eigenfunction submodels in the analysis, or by apportioning the R^2_a of the intersection fractions proportionally to the variation explained by each submodel (Legendre *et al*., 2012). In the present application, if one applies hierarchical partitioning and states that the (VBS, BS and MS) submodel has priority over the FS submodel, then the (VBS, BS and MS) submodel is served first in the variance resource and secures for itself the small negative fractions found in the intersection with the FS submodel. As a result, the intersection in explained variation between the two submodels has disappeared in Fig. 14.7b.

Nearly all the among-site fern variation explained by the environmental variables was also explained by the eigenfunctions, showing that it was spatially structured. The first three spatial submodels, grouped in the upper-right circle of Fig. 14.7b, accounted for about half (52%) of the among-site variation of fern abundances, and (0.203/0.520) ≈ 40% of that explained variation was shared with the environmental variables, pointing to an environmental control of that portion of variation. This left about 60% of that variation unexplained by the environmental variables. The fine-scale submodel explained much less of the fern abundance variation, but very little of that variation was shared with the environmental variables, suggesting that a different process was at work that generated this smaller amount of explained variation.

Ecological application 14.1b

This application illustrates dbMEM analysis for a multivariate response matrix on a regular grid positioned on a geographic surface (map). Data from a 24-ha permanent forest plot established and surveyed in 2005 in the Gutianshan National Nature Reserve in the Zhejiang Province of

China were analysed by Legendre *et al*. (2009) to determine how much of the spatial variation in species composition (beta diversity) was spatially structured, and of that, how much variation was related to the topography of the forest plot. The forest plot was fully surveyed, i.e. all 140676 trees with diameter at breast height (dbh) larger than 1 cm were tagged, identified to species, measured, and georeferenced. The trees belonged to 49 families and 159 species. The climate of the plot (29º15'N) is subtropical. The Gutianshan plots is a member of the CTFS network (see footnote in Subsection 6.5.3).

For the analysis, the forest plot was divided in the computer into cells of 20 × 20 m. The first part of the spatial analysis was based on those 600 cells. There were between 19 and 54 tree species per cell. Four topographic variables were available. Three of them (altitude, convexity of the cells, and slope) were developed into cubic polynomials to allow these variables to model nonlinear relationships with the tree species abundances, thus increasing their explanatory power; see polynomial regression, Subsection 10.3.4. The fourth variable, aspect, is a circular variable; it was transformed into two new variables, sin(aspect) and cos(aspect), which allowed their use in linear models.

339 PCNM eigenfunctions (now called dbMEM) were used in the analysis without any selection. Indeed, the R^2_a statistic obtained from the analysis of the community composition data by the full set of 339 PCNM ($R^2_a = 0.626$) was nearly identical to that obtained after forward selection of 179 eigenfunctions that were significant at the 0.05 level ($R^2_a = 0.625$). The 339 PCNMs included 200 functions with positive Moran's *I*, which modelled positive spatial correlation, and 139 with negative Moran's *I**. Among the first 180 eigenfunctions, nearly all significantly explained the tree community variation; these eigenfunctions represent broad to medium-scale variation.

The variation partitioning results (Fig. 14.8a) indicated that 63% of the among-cell variation (R^2_a) of the community composition (159 species) was spatially structured and explained by the 339 spatial eigenfunctions. Nearly half of that (0.278/0.626 = 44%) was also explained by the four topographic variables. Without surprise, nearly all the variation explained by the topographic variables was shared with the spatial eigenfunctions. The spatially-structured fraction of variation unexplained by the topographic variables ($R^2_a = 0.348$) could be related to unmeasured environmental variables, like soil chemistry, or it may have been generated by community dynamics, including neutral processes (Subsection 1.1.1).

For multivariate response data, variation partitioning is computed using RDA (Section 11.1) instead of multiple regression. The total variation of the community composition data by the table of topographic variables (fraction [a+b] of the Venn diagram in Fig. 10.10), and that of the joined tables of topographic and PCNM variables (fraction [a+b+c]), were partitioned into a number of orthogonal canonical axes by RDA. The table of fitted values corresponding to [c], the fraction uniquely explained by the PCNM variables, was computed by partial RDA, which also partitioned it into orthogonal canonical axes. Maps of the first canonical axis of each of these analyses are presented in Fig. 14.8b-d. Note that these maps are not additive, i.e. the values on the map of [a+b+c] are not equal to the values on the map of [a+b] plus those on the map of [c], because the production of orthogonal axes is done separately by the three canonical analyses. The interpretation of these maps is not as straightforward as for univariate response

* In classical PCNMs, some of the eigenfunctions with negative Moran's *I* have positive eigenvalues, as explained in the footnote of Subsection 14.1.1. These eigenfunctions were not used when the analysis was repeated with only the 200 dbMEM eigenfunctions with positive Moran's *I*. That analysis produced nearly identical variation partitioning results as in Fig. 14.8a.

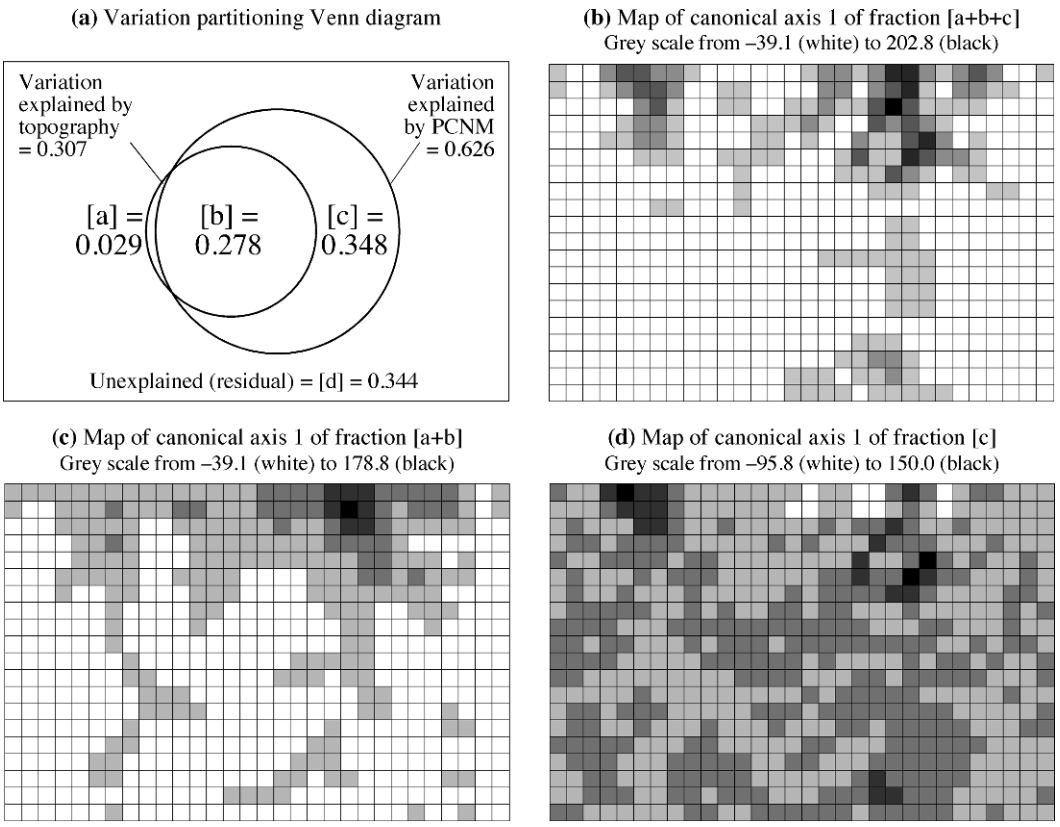
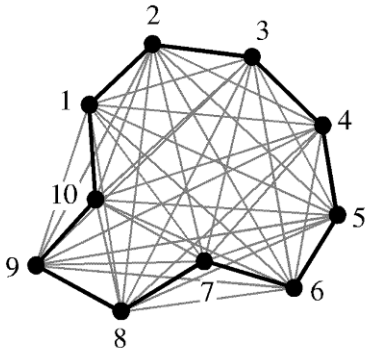


Figure 14.8 (a) Venn diagram: variation partitioning of the Gutianshan forest community composition with respect to topographic variables and spatial eigenfunctions. (b-d) The Gutianshan forest plot was divided into 600 cells of 20 × 20 m. Maps of canonical axis 1 of (b) fraction [a+b+c] (43% of the species variation, unadjusted R^2), (c) fraction [a+b] (24% of the species variation), (d) fraction [c] (14% of the species variation). Values in the cells are represented by shades of grey as shown above each map. Modified from Legendre *et al*. (2009).

data. These maps are useful, though, because they allow ecologists to visualize the spatial variation on separate canonical axes for each fraction of variation separately.

In addition to the 20 × 20 m cells, the forest plot was also divided in the computer into cells of sizes 10 × 10 m, 40 × 40 m, and 50 × 50 m. These four cell sizes allowed divisions of the 24-ha plot into cells of equal sizes that added up to the whole plot. The variation of the four resulting community tables was partitioned as described for the 20 × 20 m cell data. Comparison of the results showed that the effect of the topographic variables (fraction [a+b]) increased with cell size whereas the variation uniquely modelled by the eigenfunctions ('pure' spatial variation, fraction [c]) decreased. Hence the perceived balance between environmental and spatial effects varies with the size of the sampling units ('grain size' in Section 13.0).

(a) Sites along the shoreline



(b) $\mathbf{D}_{\text{trunc}}$ matrix

Sites	1	2	3	4	5	6	7	8	9	10
1	—									
2	*1*	—								
3	4	*1*	—							
4	4	4	*1*	—						
5	4	4	4	*1*	—					
6	4	4	4	4	*1*	—				
7	4	4	4	4	4	*1*	—			
8	4	4	4	4	4	4	*1*	—		
9	4	4	4	4	4	4	4	*1*	—	
10	*1*	4	4	4	4	4	4	4	*1*	—

Figure 14.9 Example of construction of dbMEM eigenfunctions for a loop sampling design. (a) Ten sites located along the shoreline of a lake. The lines representing the distances between neighbouring sites along the shore are in bold. (b) Truncated distance matrix $\mathbf{D}_{\text{trunc}}$: the neighbouring sites are at distance of 1 in this sketchy example; non-neighbouring sites (grey lines in panel a) receive values $4 \times thresh$, where $thresh = 1$. Diagonal values (= 4) not shown. Redrawn from Brind'Amour *et al*. (2005).

Ecological application 14.1c

This application illustrates dbMEM analysis of multivariate response data in a non-standard sampling situation. Freshwater fish were censused by snorkelling at 90 sites in the littoral zone of a small lake; the data were analysed by Brind'Amour *et al*. (2005). The littoral zone of a lake forms a loop instead of a transect. The same situation occurs when sampling the beach around an island or the ecotone* around a forest patch. To compute dbMEM eigenfunctions in such a situation, one must proceed as follows (Fig. 14.9):

• Calculate the distances between adjacent sites along the sampling loop (e.g. the shoreline in the Brind'Amour *et al*., 2005, fish data). The largest of these distances provides the *thresh* value.

• Construct matrix $\mathbf{D}_{\text{trunc}}$: keep the original distances between the neighbouring sites as they are in the distance matrix. Recode the distances between non-neighbouring sites to $4 \times thresh$. The value of *thresh* is 1 in the simplified example of Fig. 14.9; in a real study, it would be the largest distance between adjacent sites. Values 1 are found in the subdiagonal row of matrix $\mathbf{D}_{\text{trunc}}$. An additional value 1 between sites 10 and 1 closes the loop. The diagonal of $\mathbf{D}_{\text{trunc}}$ receives values of $4 \times thresh$ to produce dbMEM eigenfunctions (Subsection 14.1.1).

* An ecotone is a transition area between two adjacent and different patches of landscape, such as forest and grassland.

For the 7-species fish community found in the small study lake, Brind'Amour *et al*. (2005) used RDA (Section 11.1) to compute dbMEM models corresponding to four scales: very broad, broad, middle, and fine scale. The community data were Hellinger-transformed (Section 7.7) prior to RDA. For interpretation, the first canonical axis of matrix **Z** (eq. 11.18) corresponding to each submodel was related by multiple regression to an explanatory matrix of environmental variables.

Besides the *Adiantum tomentosum* variable studied in Ecological application 14.1a, three other ecological data sets were analysed by Borcard *et al*. (2004): marine zooplankton collected along transects in Guadeloupe, chlorophyll *a* in a brackish lagoon in southern France, and the oribatid mite data also used in Ecological application 11.5. Other examples are: Jones *et al*. (2008: fern community composition across 1045 circular plots in an old growth rain forest in Costa Rica); Léonard *et al*. (2008: macrophyte community composition in 232 quadrats sampled along 24 transects in a fluvio-lacustrine underwater landscape); Arias-González *et al*. (2008: fish and coral community composition among reefs, and among reef types within reefs, in the western Caribbean Sea); Declerck *et al*. (2011: application briefly described at the end of Subsection 14.1.1); Bellchambers *et al*. (2011: distribution of spider conch in the Cocos Islands); Astorga *et al*. (2011: spatial variation of macroinvertebrate species richness in headwater streams at two spatial scales in Finland); Andersen *et al*. (2011: environmental control and spatial structures in peatland vegetation); Legendre & Birks (2012: analysis of fossil diatom assemblages (139 taxa) in a sediment core from south-western Scotland covering the past 10000 years).

Applications are also found in the field of hydrology. — Lacey *et al*. (2007) and Roy *et al*. (2010): modelling turbulent flow in a river in Québec; Noorduijn *et al*. (2010): water table response to alley farming of trees in Western Australia; Ali *et al*. (2010): soil moisture patterns in relation to hydrometeorological variables and topography in a forested catchment in Québec.

4 — Interpretation of the fractions

In simple regression or canonical analysis modelling, one is interested in interpreting the variation of the response data (vector **y** or matrix **Y**) that is accounted for by the explanatory variables (matrix **X**) according to a model of causal relationships stated prior to the analysis. That model may be formulated loosely or stated quite precisely. The fraction of variation explained by the model is estimated by the adjusted coefficient of determination (R^2_a) in multiple regression (eq. 10.21) and in RDA (eq. 11.5). The residual variance is assumed to be a random error component.

There may be two reasons for decomposing the variation of response vector **y** or matrix **Y** into additive components through the variation partitioning approach (Subsections 10.3.5 and 11.1.11). Fractions [a], [b] and [c] referred to in the following paragraphs are shown in Fig. 10.10.

• If the spatial structure is considered to be a source of false correlations, i.e. that are not indicative of causal relationships, fraction [b] measures the interference of the spatial variables with the analysis of the relationship between **Y** and **X**. Fractions [b] and [c] should not be interpreted separately in that case, although one may still be interested in modelling the spatial structure of **Y** (fraction [b + c]) in a different analysis. As explained in Subsection 14.5.3, corrected tests of significance of fraction [a], which measures the effect of the **X** variables on **Y**, are obtained by incorporating MEM eigenfunctions as covariables in partial RDA to control for spatial correlation.

• If both the spatial and non-spatial structures of the explanatory variables are considered causal to the spatial variation of **Y**, fraction [a + b] estimates the amount of variation of **Y** explained by **X**. In such a case, the residuals of the analysis of **Y** by **X** are assumed to contain two identifiable fractions: [c], which is spatially structured, and [d], which is the random error component. A test of significance allows one to determine, at some confidence level α, if [c] accounts for a significant fraction of the variation of **Y**. When that is the case, one should try to interpret fraction [c]. The next step is to "explain away" fraction [c] if possible. In other words, one should try to make fraction [c] fade away by adding explanatory variables (if available) to matrix **X** and recomputing the model. Mapping the site scores for the significant canonical axes of fraction [c] may help identify the processes responsible for this fraction of variation (Borcard & Legendre, 1994). Section 1.1 has shown, however, that spatial structures found in communities may originate from neutral processes of population and community dynamics (eq. 1.2) in addition to spatial dependence on environmental factors (eq. 1.1), so that not all the spatial variation in **Y** may be explainable by **X**.

In statistical analysis, causality, if invoked, resides in the hypotheses of the researcher (Subsection 4.5.4). The objective of causal statistical modelling is to assess how much of the observed variation can be explained by a consistent body of hypotheses (i.e. a set of compatible hypotheses). Problems of interpretation may occur, however, when important causal factors are left out of the model. The amount of variation of **Y** explained by the model may be small and, if these factors are causally anterior to both the variables in **Y** and some of the variables in set **X**, false correlations may appear in the model; this is also the case in path analysis (Section 10.4).

In community analysis, researchers are faced with a multiplicity of potential causal agents acting at a variety of spatial and temporal scales, thus creating a network of interactions that may be difficult to untangle. Section 13.0 mentioned three general models often invoked to explain community variation: the environmental control model (ECM), the biotic control model (BCM), and historical dynamics (HD). The latter refers to past natural events, such as isolation by geographic barriers and disturbances of various kinds (e.g. storms, forest fires, volcanic eruptions, landslides), and to anthropogenic causes such as agriculture, logging, mining, and constructions of various sizes (Plate 14.1, p. 906). These factors are usually not explicitly represented by variables in **X**. Some of these events may be traced by researchers (e.g. tornadoes, forest fires, logging, past agricultural plots) and explicitly included in a second round of modelling, while others cannot and may only be invoked in general terms to account

Table 14.1 Causal factors invoked to explain the fractions of variation [a] to [d] of Fig. 10.10, assuming that matrix **W** contains spatial eigenfunctions. The table focuses on the correlations between environmental variables (matrix **X**) and community composition (matrix **Y**). The following hypotheses are invoked: environmental control model (ECM), biotic control model (BCM), historical dynamics (HD), and spatial autocorrelation. Bullets: factors explicitly stated in the model; asterisks: factors not explicitly spelled out. Arrows: causal relationships. Modified from Borcard & Legendre (1994).

Fraction	Causal factors	Process	Causal model[1]
[a]	• Non-spatially-structured component of environmental or biotic factors	ECM BCM	E → C
	* Non-spatially-structured environmental or biotic factors not included in the analysis	ECM BCM	F → E, F → C
	* Historical events without spatial structure at the scale of the study	HD	
[b]	• Spatially-structured component of biotic or environmental factors included in the analysis	ECM BCM	E → C
	* Spatially-structured environmental or biotic factors not included in the analysis	ECM BCM	F → E, F → C
	* Spatially-structured historical events	HD	
	* Spatial autocorrelation in **X** and **Y**	Spatial autocorrelation	E ↻, C ↻
[c]	* Spatially-structured environmental or biotic factors not included in the analysis	ECM BCM	F → C
	* Spatially-structured historical events	HD	
	* Spatial autocorrelation in matrix **Y**	Spatial autocorrelation	C ↻
[d]	* Environmental or biotic factors not included in analysis and not spatially structured at scale of study	ECM BCM	
	* Historical events not included in analysis and not spatially structured at scale of study	HD	
	• Random variation, sampling error, etc.	Noise	

[1] C: community structure (matrix **Y**)
E: factor explicitly represented by explanatory variable(s) in the analysis (in matrix **X**)
F: factor not represented by explanatory variable(s) in the analysis

for community variation. Table 14.1 summarizes the interpretation of the fractions of variation of Fig. 10.10, assuming that matrix **W** contains spatial eigenfunctions. The following examples refer to factors that may intervene to explain community variation in a temperate forest; they illustrate the statements found in Table 14.1.

[a] The environmental and biotic factors that are explicitly represented by variables in matrix **X** and generate the variation explained by fraction [a] have either local effects or spatial variation at scales finer than those detected by the spatial eigenfunction model. Besides these factors, local variation in unobserved soil chemistry data or other environmental variables may affect the community structure (matrix **Y**) as well as the explanatory variables found in **X**, a case that would lead to covariation between **X** and **Y** (false correlation). In addition, localized infestation by pest insects may have occurred in the past, leaving variation at some sites in the forest that persisted throughout the years; such a historical event may also have left traces in the variables included in **X**, for instance a higher content in soil organic matter due to larger amounts of dead wood deposits, leading to causal or non-causal correlations.

[b] The environmental and biotic factors that are explicitly represented by variables in **X** often have spatial variation, detectable by the spatial eigenfunction model, which may explain part of the variation of the forest community. Besides these factors, spatial variation in unobserved environmental factors may affect the community structure (matrix **Y**) as well as the explanatory variables found in **X**, a case that would lead to covariation between **X** and **Y** (false correlation). In addition, past occupation of the territory under study by agriculture may have left spatially-structured variation in the forest community; it may also have left traces in the measured soil variables of matrix **X**, leading to causal or non-causal correlations. Spatial correlation in both the response and explanatory variables may cause covariation between matrices **Y** and **X**, hence inflating fraction [b].

[c] Part of the spatial structure of the community may be caused by environmental or biotic factors that were not included in the analysis; for instance, a soil moisture gradient or the effects of grazers may not have been measured. A windstorm may have occurred in the past, creating a clearing in the forest that was later recolonized and has left a detectable broad-scale spatial structure in the forest community. In other types of communities, competition within or among species may play an important role but may have been left unmeasured. Neutral community processes such as growth and reproduction are also a major source of spatial autocorrelation (eq. 1.2), which is responsible for part of the spatially-structured variation observed in communities; it cannot be explained by external factors.

[d] This fraction represents the unexplained variation of matrix **Y** that either is not spatially structured, or has spatial structure at scales finer than those detected by the spatial eigenfunction model. Some of that variation may perhaps be explained by factors that have not been included in the analysis, such as local patches of grazers. If these explanatory variables had been included in **X**, that variation would have contributed to fraction [a]. The remainder is random local variation, which may be referred to as *local innovation*, and sampling error.

The above examples illustrate the fact that, in some cases, trying to increase the fraction of explained variation by incorporating more environmental variables into the model is doomed to failure. Fraction [c], which may represent an important proportion

of the unexplained variation, can often only be explained by neutral population or community-based spatial processes (e.g. reproduction, biotic interactions) or by past events that can sometimes be documented, but remain often unknown to the investigator.

Partitioning the spatial variation of communities into components and mapping them allow researchers to find interesting correlations that are consistent with models of causal relationships. It also allows one to quantify and map fraction [c], which measures by how much preconceived models fall short of accounting for all the spatial variation in data. The same type of analysis can be conducted on time series. Ecologists can use insights obtained by analysing fraction [c] to formulate better ecological hypotheses, which they may undertake to test by going back to the field to collect new data (Borcard & Legendre, 1994; Section 13.5, numerical example).

14.2 Moran's eigenvector maps (MEM), general form

MEM

Dray *et al*. (2006) developed a general algebraic formulation for the construction of *Moran's eigenvector maps*, or MEM eigenfunctions. They observed that the computation of dbMEM (Section 14.1) actually involved two different types of information (Fig. 14.10). The first type is found in connectivity matrix **B**, which contains information about the presence or absence of connections between points. The graph edges representing the connections in Fig. 4.10a are coded in binary form in matrix **B**: the value is 1 if there is a black edge between two points, and 0 otherwise. The second type of information is found in the edge weighting matrix **A**, which contains the values, or weights, that are placed on the edges of the graph. In this generalized approach, matrices **B** and **A** are constructed separately; **B** can contain any type of graph and **A** any type of weights.

1 — Algorithm described through an example

This generalization leads to a new algorithm for the construction of dbMEM eigenfunctions, different from that of Section 14.1.1, which will now be described. Following the description of the new algorithm for dbMEM, Subsection 14.2.2 will explain how to obtain MEM eigenfunctions that differ from dbMEM. In the numerical example that follows (next paragraph), the algorithm is described with *distances* in matrix **A**. In the paragraph after the numerical example, the algorithm is redescribed with *similarities* in matrix **A**.

Numerical example 3 (continued). Let us consider again the data (nine points) used in Numerical example 3 of Subsection 14.1.2. These points were used to illustrate different types of connection networks (graphs) in Figs. 13.22 to 13.24. The point coordinates are shown in Fig. 13.22, and Fig. 14.10a displays lines representing the distances among the points.

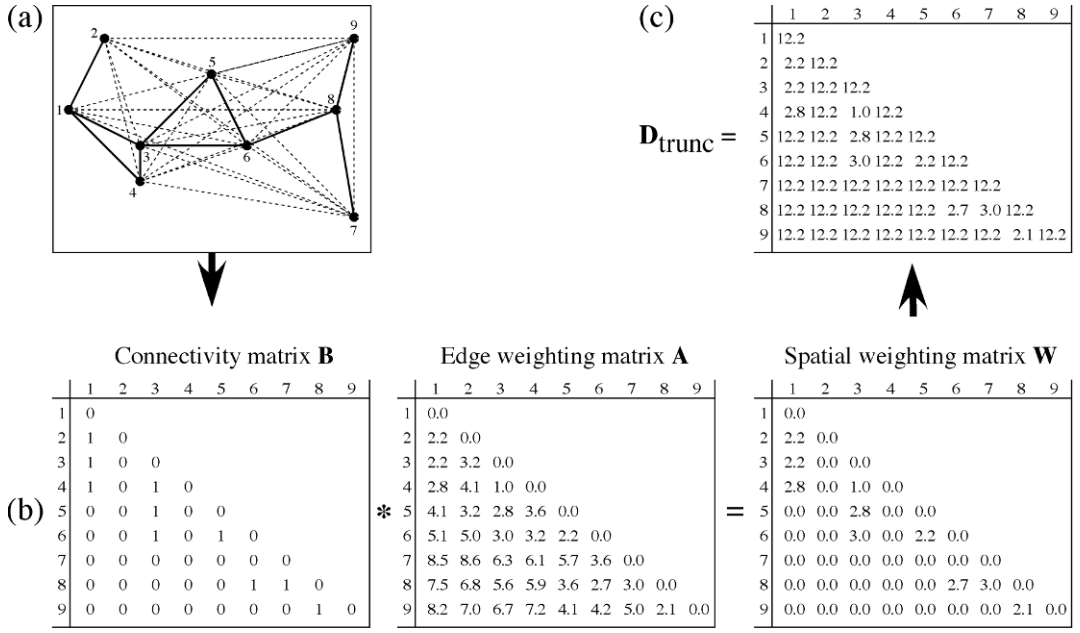


Figure 14.10 Generalized MEM eigenfunctions are the principal coordinates of matrix $\mathbf{D}_{\text{trunc}}$. (a) Graph of the nine sites of Numerical example 3, drawn on a map. The 10 distances that are equal to or shorter than the longest edge of the minimum spanning tree (Fig. 14.3) are in black, those that will be discarded are represented by dashed grey lines. The largest distance in black in the graph, 3.04, is the *thresh* value. It is found between sites 7 and 8. (b) Connectivity matrix **B** where the 10 edges in black in panel (a) are represented by values 1, the other edges by 0. The weight matrix **A** can contain any set of weights of interest. In this example, it contains the distances among points. Matrix **W** is the result of the Hadamard product (represented by *) of **B** and **A**. (c) Matrix $\mathbf{D}_{\text{trunc}}$ is obtained by replacing the zero values in **W** by 4 times *thresh*: $4 \times 3.04 = 12.16$, rounded to 12.2 in the figure. The diagonal contains values $4 \times$ *thresh*, indicating that a site is not connected to itself. For clarity, only the lower triangular and diagonal portions of the matrices are represented; all four matrices are symmetric.

• The 10 edges in black in Fig. 14.10a are equal to or shorter than the longest edge of the minimum spanning tree (Fig. 14.3), which has a length of 3.04 units. That value is chosen as the *thresh* value, as in Numerical example 3 of Subsection 14.1.2.

Connectivity matrix **B**

• Construct the binary *connectivity matrix* **B** (Fig. 14.10b, left). In that matrix, the 10 distances equal to or smaller than *thresh* are represented by 1 (connection edges present) and those larger than *thresh* are coded 0 (no connection).

Edge weighting matrix **A**

• Construct the *edge weighting matrix* **A** (Fig. 14.10b, centre). The weights can be any set of values that are appropriate for the problem at hand, representing the *difficulty of exchange* between points. In the present example, the inter-point distances are used as weights.

Spatial weighting matrix **W**

• Compute the Hadamard product (Section 2.5) of **B** and **A** to obtain the *spatial weighting matrix* **W** (Fig. 14.10b, right). The 10 distances in matrix **A** that correspond to edges coded by 1 in **B** are preserved in **W**; the other distances are replaced by zeros. Note that the diagonal contains zeros; these values will be changed in the next step.

• Construct matrix $\mathbf{D}_{\text{trunc}}$ (Fig. 14.10c). It is obtained by replacing the zero values in **W** by 4 times the *thresh* value. The fact that the zeros on the diagonal are replaced by $4 \times thresh$ indicates that a site is not connected to itself.

• Compute the principal coordinates of matrix $\mathbf{D}_{\text{trunc}}$ by PCoA (Section 9.3): they are the MEM eigenfunctions. For the present example where inter-point distances were used in matrix A, the MEM eigenfunctions obtained are identical to the dbMEM eigenfunctions of Numerical example 3; they are plotted in Fig. 14.4. This example shows that dbMEM are but a special type of MEM eigenfunctions. Other values than inter-point distances will be used in the next subsection. Because matrix $\mathbf{D}_{\text{trunc}}$ has non-zero values on the diagonal, the function used to compute PCoA must not assume that the diagonal contains zeros*.

In the Dray *et al*. (2006) paper, the algorithm is described in terms of *similarities* instead of distances. To obtain dbMEM eigenfunctions identical to those of Section 14.1, the edge weighting similarity matrix $\mathbf{A} = [a_{ij}]$ is computed as

$$a_{ij} = 1 - \left(\frac{d_{ij}}{4 \times thresh} \right)^2$$

and the inter-point distances are used as the d_{ij} values. The a_{ij} are similarities representing the *ease of communication* of matter, energy or information between points. **A** is multiplied (Hadamard product) with connectivity matrix **B** (where $b_{ij} = 1$ when $d_{ij} \leq thresh$) to produce **W**, which contains zeros for all excluded distances. The diagonal elements are also zero, indicating that a site is not similar, or connected, to itself. This is still a similarity matrix since, among the non-zero entries, larger values indicate more strongly connected pairs of points. Matrix **W** is subjected to PCoA computed for a similarity matrix, as described in Subsection 9.3.3. The computation steps are the following: (1) skip eq. 9.40, (2) centre the similarity matrix (eq. 9.41 or 9.42), and (3) proceed with eigen-decomposition of the centred matrix. The eigenvalues are the same as those computed from $\mathbf{D}_{\text{trunc}}$ in the previous paragraph, to within a multiplicative constant. The eigenvectors normalized to 1 are identical to those of the previous paragraph when they are also normalized to 1. They are the dbMEM eigenfunctions.

MEM eigenfunctions with positive eigenvalues are usually presented scaled to lengths of $\sqrt{\lambda_k}$ since they are the result of principal coordinate analysis. MEM with

* The computation of principal coordinates analysis for matrices with non-zero diagonals is available in function ***pcoa.all()*** of package PCNM (Section 14.7). That function also contains an option that allows users to output the principal coordinates corresponding to negative eigenvalues. These eigenvectors are outputted as they are computed by function ***eigen()***, i.e. normalized to lengths of 1. They are not scaled to lengths of $\sqrt{\lambda_k}$.

negative eigenvalues must be presented scaled to lengths of 1, however, because scaling them to lengths of $\sqrt{\lambda_k}$ would produce complex vectors in which each value would have a real and an imaginary part. When they are used as explanatory variables in regression or canonical analysis, including variation partitioning, eigenfunctions scaled to any value have the same explanatory power and produce the exact same R^2. This is why scaling them to lengths of 1 instead of $\sqrt{\lambda_k}$ is legitimate in MEM analysis, where the eigenfunctions are used as explanatory variables in regression or canonical analysis.

2 — *Different types of MEM eigenfunctions*

In spatial/landscape studies in ecology, study sites can be depicted as linked by different types of relationships (spatial neighbouring, exchange routes, mutual effects, etc.), which are represented by lines drawn on maps. Landscape analysis is also of interest in genetics, evolution, epidemiology, anthropology, demography, economics, and related fields. *Graphs* are schematic representations of relationships among sites. *Graph theory* is the mathematical study of graphs, which are structures used to model pairwise relations between the objects of interest in a study. In graph theory, the sites are called *nodes* and the lines are called *edges*. Urban *et al*. (2009) and Dale & Fortin (2010) reviewed the use of graphs in ecology. Graph theory was briefly used in Section 8.2.

Graph

Nodes, edges

Dale & Fortin (2010) explained the difference between aspatial (non-spatial) and spatial graphs. Examples of aspatial graphs are food webs and atoms linked by chemical bonds to form molecules (atom positions in these models are chemical relationships, not measured positions). Among the spatial graphs, the authors distinguished between planar spatial graphs, which are used in the present section to create the binary connectivity matrix **B**, and directed spatial graphs, which will be of interest in Section 14.3. They reviewed the use of graphs in landscape studies in ecology, evolution, genetics, and epidemiology.

Spatial, aspatial graph

Matrix **B** can be constructed using different types of graphs like those described in Subsection 13.3.1 for regular grids or irregularly-spaced points on a map. The present subsection describes how to obtain different types of MEM eigenfunctions by modifying the contents of matrices **B** and **A**, followed by computation of eigenfunctions as explained in Section 14.2.1. The main categories are the following:

1. *Binary MEM eigenfunctions*. — In some problems, only the presence of connections among sites matters. Differentiated values on the edges may be of no interest, or cannot be determined, so that the weights in matrix **A** are all equal. This type of spatial eigenfunctions was developed by statistical geographers; the literature on this subject was reviewed by Tinkler (1972) and Griffith (1996). Griffith & Peres-Neto (2006) called this type of methods *topology-based spatial filtering* because computation of the eigenfunctions is based on the topology of the connection network described by the spatial connectivity matrix **B**. In applications of spatial eigenfunction analysis by Tinkler (1972) and other authors, a transport network composed of

Binary MEM

locations (nodes) and routes linking them (edges) was represented by a binary connectivity matrix, and the structure of the network was characterized by the eigenvectors of that matrix, called eigenfunctions. In Griffith (1996), urban census areas that were spatially adjacent were connected in matrix **B** (called **C** by Griffith) and non-adjacent areas were not connected. After centring matrix **B** (eq. 9.41 or 9.42), eigenfunctions were computed; they characterized the spatial relationships among the census areas. In all these applications, inter-point distances were not taken into consideration.

Transformed distances

2. *Transforming the geographic distances recorded in* ***A***. — In some problems, the geographic distance does not produce the best set of explanatory eigenfunctions, and more efficient sets of eigenfunctions are found by using some non-linear function of the geographic distances. Dray *et al*. (2006) proposed three families of functions for nonlinear transformation of the geographic distances into similarities, which are used in the SPACEMAKER package. Different exponent values can be tested in turn to find the transformation of distances that produces the best model for a matrix of response data*.

• Linear function f_1:

$$s_{ij} = 1 - \frac{d_{ij}}{\max(d_{ij})} \qquad \textbf{(14.2)}$$

where d_{ij} is the geographic distance between points i and j. Division by $\max(d_{ij})$ ensures that the similarities are in the range [0, 1]. f_1 does not change the modelling capacity of the resulting MEM eigenfunctions: in an analysis based on a matrix $[d_{ij}' = d_{ij}/\max(d_{ij})]$, the eigenvalues are changed compared to an analysis based on matrix $[d_{ij}]$, but the eigenvectors scaled to lengths 1 are the same. f_1 is presented here as a reference to help in understanding f_2.

• Concave-down function f_2:

$$s_{ij} = 1 - \left(\frac{d_{ij}}{\max(d_{ij})}\right)^{\alpha} \qquad \textbf{(14.3)}$$

With $\alpha > 0$, function f_2 operates a non-linear transformation of distances d_{ij}. The similarity s_{ij} decreases as d_{ij} increases, but more rapidly for larger d_{ij} values. When $\alpha = 1$, f_2 is the same as f_1. To appreciate the shape of the transformation, readers can compute this function for d_{ij} values from 1 to 20, using a positive integer larger than 1 for α, and plot the results. Fractional positive values of exponents α produce concave-up transformations where the similarity s_{ij} decreases less rapidly for larger d_{ij} values.

* Function ***test.W()*** of the SPACEMAKER package allows users to automatically test the effect of different exponents of functions f_2 and f_3, or any other function that transforms the distances before they are included in matrix **A**, and select the one that produces the model with the lowest value of AIC_c (eq. 10.23). A tutorial (Dray, 2010) is provided with the package.

• Concave-up function f_3:

$$s_{ij} = 1 / d_{ij}^{\beta} \quad \textbf{(14.4)}$$

where β is a positive real number. The similarity s_{ij} decreases as d_{ij} increases, but less rapidly for larger d_{ij} values.

Other transformations of geographic distances to similarities can be devised and applied to data. In their search for the best predictive model for the oribatid mite data also used in Ecological applications 11.5 and 14.4, Dray *et al*. (2006) tried in turn functions f_1, f_2 with $\alpha = 2$ to 10, and f_3 with $\beta = 1$ to 10. A similarity matrix **S** was computed for each of these functions and exponent values. **S** was used as the edge weighting matrix **A** to compute the spatial weighting matrix **W**, which was decomposed into MEM eigenfunctions. After selection of the best MEM submodel in each case, the set of MEM eigenfunctions that produced the lowest value of AIC_c (eq. 10.23) was retained as the best spatial model of the response data. Details are presented in Ecological application 14.2a hereunder.

When f_1 or f_2 provides the best predictive model, the following method can be used: compute a minimum spanning tree for matrix $\mathbf{D} = [1 - s_{ij}]$ to identify the value of *thresh*, then replace $\max(d_{ij})$ by $(4 \times thresh)$ in the denominators of transformation equations f_1 and f_2; for example, f_1: $s_{ij} = 1 - (d_{ij}/(4 \times thresh))$. This alternative method, which uses *thresh*, would be in line with the similarity-based procedure for dbMEM described in Subsection 14.2.1.

Geographic resistance

3. *Using other measures of geographic resistance as weights in* ***A***. — The edge weighting matrix **A** can be generalized further by using measures of landscape resistance that are not based on transformations of the geographic distances. Dale & Fortin (2010) give several examples. One of them is the analysis of animal movement, which can be based on estimates of the attractiveness of patches for the species under study, or on estimates of movements obtained from field observations. Other aspects that can be used to construct matrix **A** are transport models, landscape connectivity, least-cost paths, and multiple paths forming corridors. Applications of these measures of resistance in landscape ecology and genetics have yet to be fully explored.

The MEM eigenfunctions obtained in these different ways can all be grouped into spatial submodels corresponding to different spatial scales, which can be analysed by RDA, mapped, and interpreted separately. Dray *et al*. (2006) remind users that the choice of a spatial weighting matrix **W** in the similarity approach, or $\mathbf{D}_{trunc}$ in the distance-based approach, is a critical step in spatial eigenfunction analysis. While the structures modelled by eigenfunctions obtained using different **A** matrices are fairly similar for regular sampling designs like regular grids, they may differ greatly for irregular distributions of sites. The authors recommend a pragmatic approach where several solutions are explored and the one that explains the response data best is retained, unless ecological theory suggests a specific way for the construction of **A**, e.g. knowledge of propagule dispersal processes. In linear modelling, model efficiency can be estimated by the AIC_c coefficient (eq. 10.23) associated with the model.

Ecological application 14.2a

An example of MEM analysis with selection of the geographic distances leading to the best model is included in the Dray (2010) tutorial document about MEM eigenfunction analysis. It was reproduced, with additional comments, in Borcard *et al*. (2011, Subsection 7.4.3.2). The application involves the oribatid mite data already used in Ecological application 11.5. In the present application, the detrended and Hellinger-transformed abundances of the 35 mite morphospecies were used together with the spatial coordinates of the 70 soil cores from which the mites were extracted. Geographic distances were computed among the core positions. For reference, dbMEM analysis (Subsection 14.1) of the same data produced a spatial model containing eight MEM eigenfunctions that explained the mite data with $R^2_a = 0.24$. Can one find a better spatial model of the mite data?

The first attempt was based on a Delaunay triangulation with binary weights (1 for presence of an edge, otherwise 0) in the edge weighting matrix **A**. Forward selection of MEM eigenfunctions was carried out by function ***ortho.AIC()*** of the SPACEMAKER package (Section 14.7). That function computes AIC_c for successive models containing orthogonal explanatory variables ordered by their contributions to R^2. The best model, i.e. the one with the lowest AIC_c, contained seven MEM variables. Its AIC_c value was –94.2.

In an attempt to find a better model, the edges of the Delaunay triangulation were weighted using function f_2 described above. Values of exponent α from 1 to 10 were investigated (f_2 with $\alpha = 1$ is function f_1), and for each exponent, forward selection was performed. The best among these models turned out to be one obtained with $\alpha = 1$. The value of AIC_c was –95.5, lower than with binary weights; the model contained six MEM variables.

In a further attempt to find a better model, connectivity matrices **B** were constructed based on inclusion circles of increasing radii around each point. A multivariate variogram (Subsection 13.1.4) showed that spatial correlation was maximum at a distance of about 4 m. Since the shortest distance that kept all points connected in a minimum spanning tree was *thresh* = 1.01 m, the distances to be investigated covered the range [1.01 m, 4 m]. Ten circle sizes were created. For each value, the inter-point distances smaller than or equal to the stated distance value were coded by 1 in matrix **B**. There were no weights in matrix **A**. The connection network with distances smaller than or equal to 2.01 produced the best model; it had an AIC_c coefficient of –100.6. The model contained five MEM variables.

The search for the best model was broadened by combining different distance thresholds, as in the previous paragraph, with weighting the edges in matrix **A** using function f_2 with different values of exponent α. The connection network with distances ≤ 2.67 produced the best model when combined with exponent $\alpha = 3$. That model had $AIC_c = -102.7$; it included seven MEM eigenfunctions.

The R^2_a statistic increased along that search. It was 0.20 for the two models based on Delaunay triangulations, 0.21 for the model that included all distances ≤ 2.01 but no weights, and 0.29 for the model that included all distances ≤ 2.67 weighted by f_2 with exponent $\alpha = 3$. That last model, which involved seven MEM eigenfunctions, led to a RDA that produced two significant axes representing two orthogonal spatial models, whose values can be plotted on maps of the 70 soil cores, as suggested in the R script provided by Borcard *et al*. (2011). Careful empirical selection of the spatial weighting matrix **W** produced a much better model, in terms of AIC_c and R^2_a, than the initial model. The final model is parsimonious in terms of the number of MEM eigenfunctions that are included.

Ecological application 14.2b

Landeiro *et al*. (2011) compared MEM eigenfunctions derived from geographic distances and from distances along a river network, called watercourse distances, in an ecological study of a stream network in the Ducke Reserve in Central Amazon, Brazil. The eigenfunctions derived from watercourse distances provided stronger explanations (higher R^2_a) of community variation among sampling sites located along the network than eigenfunctions derived from geographic distances, for both real (fish, caddisfly) and simulated data.

14.3 Asymmetric eigenvector maps (AEM)

In fluid ecosystems (water, air), distributions of the organisms that form communities are often driven in part by directional physical processes. These include water currents in rivers, lakes and oceans, prevailing wind along mountainsides, river networks, and glaciations at historical time scales. Spatial modelling by MEM eigenfunctions was not designed to take into account such directional processes. A different type of spatial eigenfunction analysis, called *asymmetric eigenvector maps* (AEM), was developed by Blanchet *et al*. (2008a, 2009) to model this type of phenomenon.

Directional process

1 — Algorithm described through an example

The algorithm for the construction of AEM eigenfunctions representing directional physical processes is described here through an example.

Numerical example 4. Let us consider the river network shown in Fig. 14.11a, the same as in Fig. 1.14. The network contains eight nodes (N1 to N8); six of them are lakes, the other two are river junction points. For the sake of the example, let us assume that sampling has taken place at these eight nodes. One could study fish communities found either in the six lakes, or at all nodes. Even if no data are available at the root of the network, a non-sampled node 0 is added there to unite the network into a single structure. AEM modelling could be used to test the hypothesis that the communities are spatially related through a process of colonization of the network from node 0.

The river network can be coded as shown in Fig. 14.11b:

Nodes-by-edges matrix

• The nodes are the rows of the coding table (matrix **E**), the columns are the edges. In matrix **E**, an edge located between node *i* and node 0 is coded 1 in the row corresponding to node *i*; the other edges are coded 0. The coding is the same as in Fig. 1.14. For a denser connection network, as in Fig. 1 of Blanchet *et al*. (2009), matrix **E** can often have many more columns than rows.

• A vector of weights $\mathbf{w} = [w_j]$ may be applied to the edges. The weights may be based upon geographic distances, estimates from a flow model, current speed, and so on. They represent the ease of communication of matter, energy or information among the sites. The weight values must be structured as similarities between nodes, not as distances, so that larger weights produce higher values in the matrix $\mathbf{E}_w$ resulting from the scalar product of **E** by the diagonal matrix of

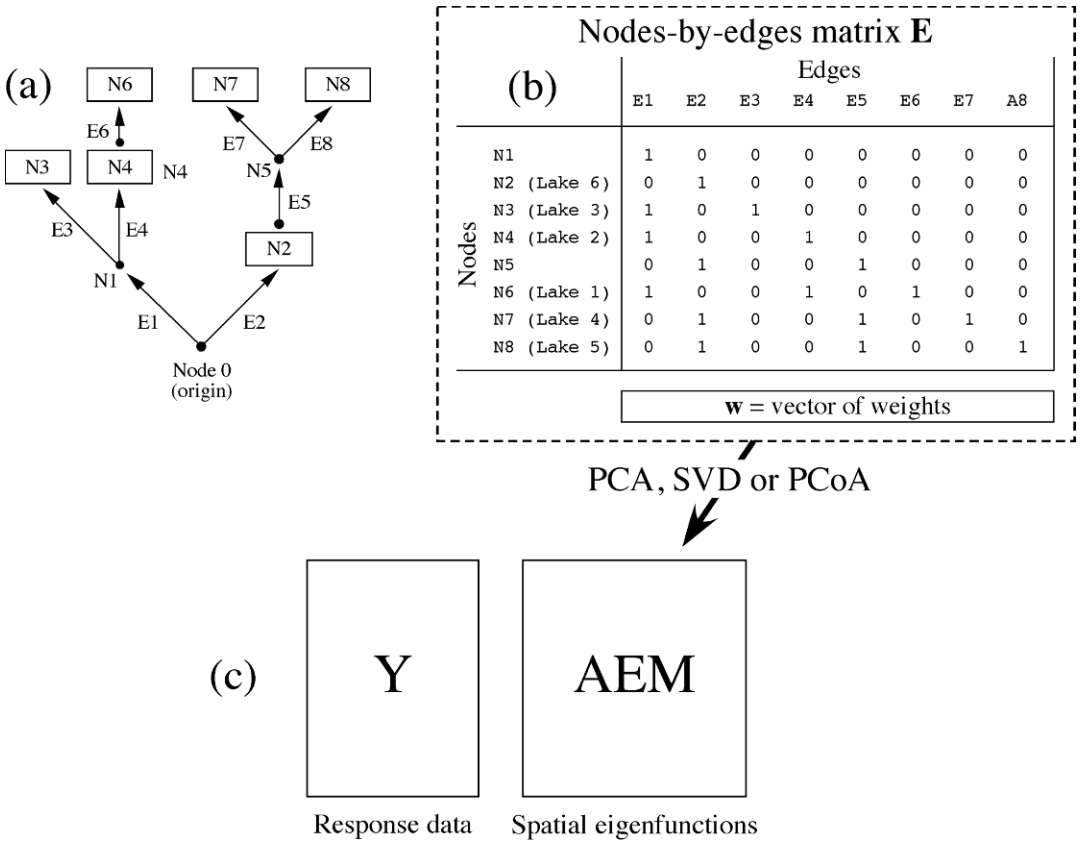


Figure 14.11 Schematic of AEM analysis. (a) River network from Fig. 1.14. The N_i are nodes and the E_j are edges. The nodes in boxes are lakes, labelled Lake 1 to Lake 6 in Fig. 1.14. The node at the origin of the network is labelled Node 0. (b) Nodes-by-edges matrix **E**. PCA of matrix **E** or $\mathbf{E}_w$ (weighted), or SVD of $\mathbf{E}_c$ (centred) or $\mathbf{E}_{wc}$ (weighted and centred), or PCoA of a distance matrix computed from **E** or $\mathbf{E}_w$, produces the table of AEM spatial eigenfunctions, which is used in (c) to analyse the variation of response matrix **Y** by RDA.

weights $\mathbf{D}(w_j)$. Matrix $\mathbf{E}_w$ has the same number of rows and columns as **E**. The initially chosen weights may be reworked through weighting functions f_1, f_2 and f_3 of Subsection 14.2.2 in order to transform them in an optimal way. Several examples of the search for optimal weights are found in the applications presented by Blanchet *et al*. (2011).

The weighted or unweighted matrix, $\mathbf{E}_w$ or **E**, is subjected to principal component analysis (PCA); matrices **F** or **G** of the PCA output contain the AEM eigenfunctions. Alternatively, matrix **E** can be centred by columns, producing $\mathbf{E}_c$, and then subjected to singular value decomposition (SVD). A third computational approach is to compute a Euclidean distance matrix (D_1, eq. 7.32) among the rows of $\mathbf{E}_w$ or **E**, followed by principal coordinate analysis

(PCoA, Section 9.3). The three forms of decomposition produce equivalent AEM spatial eigenfunctions.

PCA produces matrices **F** or **G** of object scores (eqs. 9.4 and 9.14), SVD produces matrix **V** (eq. 2.31), and PCoA produces a matrix of eigenvectors rescaled as principal coordinates. All four matrices can be used interchangeably as the explanatory matrix in the next step of the analysis because they only differ by the scaling of their vectors. The singular values obtained by SVD can be transformed into the eigenvalues of the PCA, as shown in Subsection 9.1.9; the relationship between the eigenvalues of the PCoA of a Euclidean distance matrix (D_1) and those of the PCA computed on the same original data has been shown in Subsection 9.3.2. Contrary to MEM where positive and negative eigenvalues are found, all AEM eigenvalues are positive or null because PCA eigenvalues cannot be negative (Section 9.1).

Contrary to MEM analysis, the response data should not be detrended (Subsection 13.2.1) prior to AEM analysis. The reason is that a gradient structure is a logical consequence of a directional forcing process, and detrending would remove a portion of its expected signature. This makes the comparison of the results of MEM and AEM modelling difficult. An example in Blanchet *et al*. (2011) shows how to resolve that problem by introducing a model corresponding to the spatial trend in the variation partitioning operation involving the MEM and AEM models.

For the example, if response data (e.g. fish community composition) were only available for the six lakes, the rows of **E** corresponding to the six lakes could be selected for PCA, SVD or PCoA instead of the whole matrix **E**. The number of columns would remain the same, although columns containing only zeros, if present, could be discarded before decomposition since they would have no effect on the AEM eigenfunctions.

The matrix of AEM eigenfunctions can now be used as explanatory data in multiple regression analysis to explain the variation of univariate response data **y**, or in RDA to explain the variation of multivariate response data **Y** (Fig. 14.11c).

Moran's *I* coefficients (eq. 13.1) can be computed for each AEM eigenfunction* (Blanchet *et al*., 2011); this allows users to separate the eigenfunctions modelling positive and negative spatial correlation, and select one of the two groups for analysis. The eigenfunctions of the selected group can then be divided into submodels for multiscale analysis. The whole matrix of AEM, or the submodels, can be used jointly with other matrices of explanatory variables in variation partitioning. The latter form of analysis may involve environmental explanatory data. It can also involve MEM eigenfunctions, which model non-directional spatial processes; an example is found in Ecological application 14.1a. Before variation partitioning, a hierarchy can be established among the AEM submodels to avoid the production of un-interpretable non-zero fractions of shared variation between the orthogonal submodels; else, the intersection fractions can be apportioned proportionally to the variation explained by each submodel (Legendre *et al*., 2012); see Fig. 14.7b.

For the example (Fig. 14.11), AEM eigenfunctions were computed with edges having equal weights of 1, i.e. all $w_j = 1$. The first four eigenfunctions had Moran's *I* larger than the expected value $E(I)$ and thus modelled positive spatial correlation. Bubble maps are shown in Fig. 14.12;

* Moran's *I* coefficients can be computed using function ***moran.I.multi()*** of the AEM package; see Section 14.7.

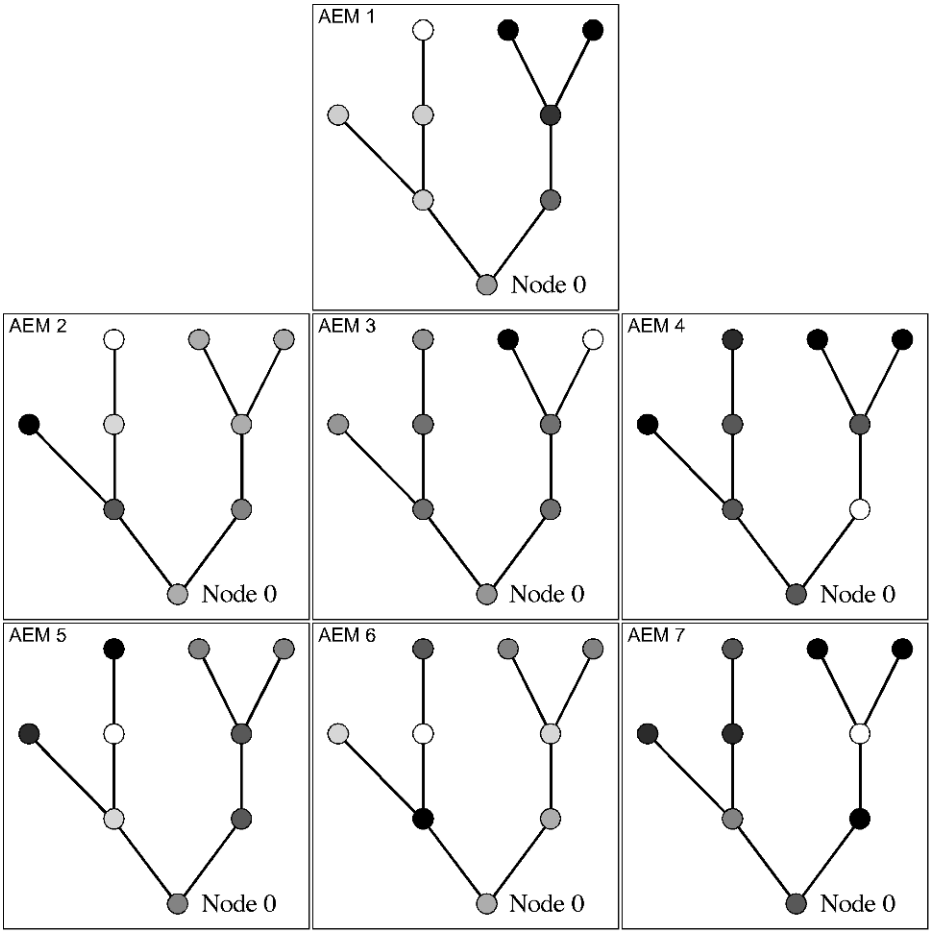


Figure 14.12 Bubble maps of the seven AEM eigenfunctions corresponding to the example in Fig. 14.11. Shades of grey: see Fig. 14.2.

the eight nodes are drawn on top of the river network. AEM 1 displays a double gradient showing the directional process corresponding to the structure of the river network: node shades go from gray (node 0) to white in the left branch and from gray to black in the right branch. AEM 2 differentiates the extreme nodes of the left branch, black versus white; the nodes of the right branch are near the mean value (0, gray) for this AEM. In a similar way, AEM 3 differentiates the extreme nodes of the right branch, black versus white; all the other nodes are zero (gray) for this AEM. AEM 4 displays a concave-up shape with low values near the origin of the network and high values at the tips: the white dot (N2) has the lowest value and from that point, values increase in the right branch to maximum at N7 and N8, and similarly in the left branch to an intermediate value at N6 and a maximum value at N3. The last three AEMs, 5, 6 and 7, show quick successions of black and white dots in the left or right branch; they model

negative spatial correlation on the river network, as shown by their Moran's I smaller than $E(I)$. This set of AEM eigenfunctions is well suited to model directional processes along the network with differentiation of the two main branches.

Time series

Time series represent a form of directional stochastic process (Section 12.0). To emphasize the directional nature of the process influencing the data, AEM analysis, which was designed to take trends into account, should be applied to the non-detrended series. MEM analysis can be applied to data series that were detrended to remove the directional component (recommendation of Subsection 14.1.2). Detrended palaeoecological sediment core data, for example, could be studied by MEM analysis.

2 — Ecological applications

Three applications of AEM analysis drawn from the ecological literature are summarized hereunder.

Ecological application 14.3a

A fully developed ecological application of AEM analysis is presented by Blanchet *et al*. (2008a). It concerns the diet composition of brook trout, *Salvelinus fontinalis*, collected in 42 lakes in the river network of the Mastigouche Reserve in Québec, Canada. The diet composition of 20 trouts per lake was divided into nine functional prey categories. The data were analysed to test the hypothesis that diet composition was determined at least in part by the structure of the river network, which captured the geomorphological differences among the lakes in different portions of the river network as well as genetic differences among the trout populations, which migrated from lake to lake along the network.

The spatial modelling was conducted in different ways: (1) based on the lake geographic coordinates (polynomial of the coordinates, dbMEM analysis); (2) using the same coding of the nodes as in Magnan *et al*. (1994; coding briefly described in Subsection 1.5.7); and (3) using a coding of the edges as described in Fig. 14.11. Then, the edges of the network were analysed in three different ways: (3.1) RDA with forward selection was conducted on the nodes-by-edges matrix (matrix **E** in Fig. 14.11); (3.2) the edges were used to compute distances along the river network (watercourse distances), assuming that all edges had the same resistance value along the network, and these distances were transformed into MEM eigenfunctions which were used in RDA with forward selection; (3.3) the edges coded into matrix **E** were used to compute AEM eigenfunctions, assuming again that all edges had the same w_j values, and the AEMs were used in RDA with forward selection.

The analyses based on edges outperformed the analyses based on geographic coordinates of the lakes and on nodes of the network in terms of R^2_a. The non-directional MEM model based on watercourse distances produced $R^2_a = 0.56$ whereas the directional AEM model had $R^2_a = 0.64$. For these data, all the spatial variation explained by the MEM model was also explained by the AEM model, and the latter explained a small but significant extra fraction of variation ($R^2_a = 0.08$) that was not explained by the symmetric MEM model*.

* This additional result was computed and provided by F. Guillaume Blanchet, University of Alberta, Edmonton. It is not found in the Blanchet *et al*. (2008a) paper.

Three other application of AEM analysis are found in Blanchet *et al*. (2011). The applications include the distribution of a crustacean (*Atya*) in a river, bacterial production in a fluvial lake, and the distribution of the copepodite stages of a crustacean on the Atlantic Ocean shelf. In each of these applications, the AEM, MEM, and dbMEM modelling results were compared. These applications show that AEM and MEM analyses can often be used together to identify the directional and non-directional components in the spatial structure of ecological data; the overlap between the fractions of explained variation by the two models is often high.

Ecological application 14.3b

Gray & Arnott (2011) studied zooplankton dispersal among 45 lakes in the Killarney Provincial Park, Ontario, Canada, that were heavily impacted by acidification before and during the 1970's. Zooplankton data were available for 1972, 1990 and 2005.

Spatial modelling and variation partitioning were carried out to determine the relative importance of spatial relationships and environmental control in models explaining the variation in zooplankton community structure over the 45 lakes. The objective was to determine if dispersal was an important determinant of the structure of recovering zooplankton communities over time. Three spatial models were compared: a non-directional spatial dbMEM model based on overland distances, a stream dispersal MEM model using watercourse distances among lakes (as in Ecological application 14.2b), and asymmetric AEM models describing overland dispersal by wind among lakes in the predominant spring and summer wind direction. The analysis was applied to data collected in 1972 (acidified) and in 1990 and 2005 (recovery period).

Analysis of the community structure across the 45 lakes was carried out for the three study periods separately with respect to environmental (physical and chemical) data and the three spatial models, for a total of nine variation partitioning results. The environmental variables had greater effect on zooplankton community structure than the spatial models, yet these had significant effects. In 1972 (highly acidified lakes), the symmetric spatial model had greater effect than the stream dispersal and directional models, but during the following study periods (chemical recovery), the effect of the directional (wind-driven) dispersal model grew and, in 2005, it explained more of the community variation than the symmetric and stream dispersal models. The directional model explained a new portion of the community variation that was not shared with the environmental variables and was larger than that explained by the other two models. The authors concluded that limitation of zooplankton dispersal was an important obstacle to lake recovery after acidification.

Ecological application 14.3c

Sharma *et al*. (2011) assembled a database on fish community composition (100 species), lake morphology, water quality, climatic conditions, and hydrological connectivity for 9885 lakes in Ontario, Canada. The authors compared dbMEM and AEM models through variation partitioning to determine if the spatial patterns could have been produced by human-mediated or natural modes of dispersal. Examination of the relative roles of spatial structure and environmental conditions showed that most of the explained variation of native species assemblages was governed by environmental conditions. Non-native fish assemblage composition could be related to human-mediated dispersal, showing that the ecological processes that underlaid biogeographical patterns differed for native and non-native fish species.

14.4 Multiscale ordination (MSO)

Are the environmental variables responsible for the spatial correlation observed in the response matrix, for example in community composition data? If so, at which distance classes is that effect important? Is it for all scales (distance classes) or for some scales only? Wagner (2003, 2004) addressed these important questions by combining multivariate variograms (Subsection 13.1.4) with simple (Chapter 9) or canonical ordination (Chapter 11), in a form of analysis called *multiscale ordination* (MSO). MSO had originally been described by Noy-Meir & Anderson (1971) and Ver Hoef & Glenn-Lewin (1989) for the analysis of the covariation among species in blocks of different sizes. That form of analysis required a continuous sampling design composed of adjacent or regularly-spaced quadrats along a transect or on a lattice. Wagner (2003, 2004) generalized the method to regular or irregular sampling designs using the geostatistical framework. Spatial eigenfunctions can be incorporated as covariables into MSO; an ecological application will illustrate the interest of doing that.

Wagner's multiscale ordination, described in the present section, is based on multivariate variograms (Subsection 13.1.4) and on ordination methods. It can be carried out using PCA, CA, RDA, CCA, or the partial versions of RDA and CCA. The variables in the multivariate data matrix $\mathbf{Y}$ must all have the same physical dimensions. If they do not, they need to be standardized before MSO based on PCA or RDA (Subsection 9.1.5) because the variances of individual variables must be summed to obtain the variogram statistics (Subsection 13.1.4); in any case, the condition of dimensional homogeneity of the variables must also be fulfilled before computing PCA or RDA (Subsection 9.1.5). The data used in CA and CCA are frequencies (Section 9.2), which should not be standardized. In the context of simple ordination (PCA, CA), MSO partitions the variance of the ordination axes among distance classes. This exploratory analysis aims at identifying the ordination axes that display spatial structure and finding out if the spatial structure differs among axes. With the canonical ordination methods, the analysis can incorporate matrices of environmental variables and eigenfunctions (MEM or AEM), which makes it possible to determine if the spatial correlation in the data is due to induced spatial dependence (eq. 1.1, Chapter 1) or the presence of spatial autocorrelation in the response data (eq. 1.2).

In the present section, the different levels of analysis will be described for ordination methods of increasing complexity that preserve the Euclidean distance: PCA, RDA and partial RDA. Couteron & Ollier (2005) give details for methods that preserve the chi-square distance (CA, CCA, partial CCA). MSO with Euclidean-based methods is presented in detail here because these methods are flexible; by applying the transformations of Section 7.7 to community composition data, MSO through PCA and RDA can preserve a variety of distances, including the chi-square.

MSO can be computed on non-stationary data (see Ecological application 14.4), but one must be aware of the fact that the calculation of confidence intervals in variograms requires stationarity, a problem that can be solved by detrending the data

(Subsection 13.2.1). With non-stationary data, the confidence intervals are too broad and thus the tests for spatial correlation are too conservative.

The MSO algorithm is implemented as follows in Wagner's function ***mso()*** of the VEGAN package in R.

1. *MSO with simple ordination (PCA).* — Consider a multivariate matrix **Y** with n rows (sites) and p columns (e.g. Hellinger-transformed species presence-absence or abundance data). The spatial coordinates of the sites along the transect or on the surface must also be provided to the function for the calculation of the distance classes.

• Compute the variogram matrix of **Y** and draw a multivariate variogram, as explained in Subsection 13.1.4. The variogram will indicate the presence of spatial correlation in **Y**, if any, pointing out the distance classes where spatial correlation is significant.

• Compute a PCA of **Y**; it produces eigenvalues λ_k and eigenvectors $\mathbf{u}_k$. Eigenvalue λ_k can be partitioned among distance classes d by computing $\lambda_k(d) = \mathbf{u}_k' \, \mathbf{C}(d) \, \mathbf{u}_k$ for each distance class. A plot of $\lambda_k(d)$ against distances d is a variogram of PCA axis k. It shows the spatial correlation structure of the variation represented by that axis.

After a PCA, function ***mso()*** only produces a multivariate variogram of **Y**. The variogram of PCA axes is not presently computed by function ***mso()***. One can use a regular univariate variogram function (Section 13.6) to compute a variogram of the k^{th} principal component, which is found in matrix **F** of the PCA results (eq. 9.4).

2. *MSO with canonical ordination (RDA).* — If the multivariate variogram indicates significant spatial correlation, one may test the hypothesis that explanatory (e.g. environmental) data can explain that spatial structure. When it is the case, the spatial structure can be attributed to induced spatial dependence (eq. 1.1). That point can be addressed by MSO based on canonical ordination, using an explanatory matrix **X** to model the spatial variation of **Y**.

• Carry out a canonical ordination by RDA to analyse **Y** using explanatory data **X** that are thought (hypothesized) to drive the observed spatial variation.

• Compute a multivariate variogram of matrix $\hat{\mathbf{Y}}$ shown in Fig. 11.2. The variogram decomposes the *explained variance* of the RDA model among the distance classes. The variogram statistics can be plotted in a graph. For the first analysis carried out in Ecological application 14.4 (below), the variogram statistics decomposing the explained variation into distance classes are represented by circles in Fig. 14.13a.

• Compute a multivariate variogram of the matrix of *residuals* (Fig. 11.2, lower left matrix) as well as the tests significance of the variogram statistics. Plot the variogram statistics in the graph (squares in Fig. 14.13a). Is there significant spatial correlation remaining in the residual data after fitting the explanatory variables? If not, this result indicates that the explanatory variables **X** explain the multiscale spatial structure well. If significant spatial correlation remains in some distance classes, either there are other

explanatory variables (environmental or historical, Subsection 14.1.4) that may explain that variation but were not included in explanatory matrix **X**, or the remaining spatial correlation is true spatial autocorrelation in the data (eq. 1.2).

• Compute a multivariate variogram of **Y** as described in Subsection 13.1.4. Compute the confidence intervals of the variogram statistics using parametric standard errors. Plot these confidence intervals in the graph (two continuous lines in Fig. 14.13a). The confidence intervals are computed under the second-order stationarity assumption (Subsection 13.1.1).

• Add together the variogram statistics of the explained and residual variation. Plot these sums in the graph (crosses in Fig. 14.13a) and compare them to the confidence intervals plotted in the preceding step (previous bullet). The confidence intervals of the variogram of **Y** provide a diagnostic tool: if the empirical variogram of the explained plus residual variation remains entirely within the confidence envelope, this indicates that the linear relationship between **Y** and the explanatory variables does not vary significantly with scale. In the absence of significant spatial correlation in the residuals, values found outside the confidence envelope may indicate that the relationship between **Y** and the explanatory variables is scale-dependent, or that there may be a mismatch between the scale at which the predictors **X** were measured and the scale of the response of **Y** to these factors (Wagner & Fortin, 2005). The roles of the many mechanisms that may be responsible for the mismatch are still poorly understood and should be the subject of further studies, e.g. by numerical simulations.

If one finds indications of scale-dependent relationships, one can look for the cause either through data analysis or by testing specific hypotheses. For example, one could use MEM eigenfunctions to model the response data at different scales, then regress the fitted values or canonical axes on explanatory data **X** in order to identify the source of the spatial correlations. An example of this approach is presented in Ecological application 14.1a. Another approach is to select MEM eigenfunctions that are significantly related to **Y** and use them as covariables in MSO based on partial RDA; see item 3 below. An example in found in the part of Ecological application 14.4 that is illustrated in Fig. 14.13b.

MSO can be computed for non-stationary data, but beware: because the confidence intervals are then too broad, they may not evidence values of the empirical variogram of the explained plus residual variation that lie outside the confidence envelope of the multivariate variogram.

Although the variogram of the explained plus residual variation is not identical to the variogram of **Y**, the weighted sums of the values in these two variograms are equal to the total variance in **Y**. The weights in these sums are the number of pairs of points used to compute the values in each distance class divided by the total number of pairs of points.

3. *MSO with partial canonical ordination (partial RDA).* — If an explicit MEM or AEM spatial model is available to account for spatial correlation at scales of interest in the study, the MEM or AEM model can be used as matrix of covariables **W** in partial RDA (Subsection 11.1.6).

• Carry out a partial canonical ordination to analyse **Y** using explanatory data **X** and covariables **W**. The analysis must be computed using the first calculation method of Subsection 11.1.6: residuals of both **Y** and **X** are computed with respect to **W**, obtaining $\mathbf{Y}_{\text{res}|\mathbf{W}}$ and $\mathbf{X}_{\text{res}|\mathbf{W}}$, before the canonical analysis of $\mathbf{Y}_{\text{res}|\mathbf{W}}$ by $\mathbf{X}_{\text{res}|\mathbf{W}}$. This is the method used to compute partial RDA in function ***rda()*** of VEGAN (Table 11.5).

• Compute a multivariate variogram of the *variance explained* by the partial canonical analysis, i.e. the variance of $\mathbf{Y}_{\text{res}|\mathbf{W}}$ explained by $\mathbf{X}_{\text{res}|\mathbf{W}}$. This variogram can be computed either on the matrix of fitted values of that analysis or on the canonical axes (matrix **Z** in Fig. 11.2), with identical results. It decomposes the variance explained by the partial RDA among the distance classes. Plot these values in a graph; they are represented by circles in Fig. 14.13b for Ecological application 14.4.

• Compute a multivariate variogram of the *residual variation* after fitting matrices **W** and **X**; compute also the tests of significance of the variogram statistics. Plot these values in the graph (squares in Fig. 14.13b). Is there significant spatial correlation remaining in the residual data after fitting the explanatory variables? If not, this indicates that the explanatory matrices **X** and **W** explain the multiscale spatial structure well.

• Compute a multivariate variogram of $\mathbf{Y}_{\text{res}|\mathbf{W}}$ as described in Subsection 13.1.4 and compute the confidence intervals on that variogram using parametric standard errors. Plot these confidence intervals in the graph (two continuous lines in Fig. 14.13b).

• Add together the variogram statistics of the explained and residual variation. Plot these sums in the graph (crosses in Fig. 14.13b) and compare them to the confidence intervals already plotted. The confidence intervals of the variogram of $\mathbf{Y}_{\text{res}|\mathbf{W}}$ provide a diagnostic tool: if the empirical variogram of the explained plus residual variation remains entirely within the confidence envelope, this indicates that the linear relationship between $\mathbf{Y}_{\text{res}|\mathbf{W}}$ and explanatory matrix $\mathbf{X}_{\text{res}|\mathbf{W}}$ does not vary with scale.

MSO provides diagnostics about the interpretation of spatial correlation at different scales. Under the assumption that all relevant environmental factors have been measured (and measured at the relevant spatial scales), MSO allows researchers to distinguish between induced spatial dependence and true spatial autocorrelation.These assumptions are difficult to check in practice. It also allows users to assess the presence of significant correlation in residuals, as well as the scale-invariance of the species-environment correlation. How to interpret MSO results is illustrated in Ecological application 14.4.

Ecological application 14.4

The oribatid mite data of Borcard & Legendre (1994) were used in Ecological application 11.5 to compute a multivariate variogram. The analysis reported here was developed by Borcard *et al*. (2011) to illustrate different facets of MSO analysis. The mite data were Hellinger-transformed prior to analysis. For the variograms, the interval size of the distance classes was the distance that kept all points connected in a dbMEM analysis; this is the threshold distance (*thresh* = 1.0112 m) computed in Ecological application 14.2a.

In a first set of analyses, a MSO variogram was computed, without detrending, after a RDA of the mite data modelled by several environmental variables. These variables were: substrate density, water content of the soil, substrate types (7 classes), shrub density (3 classes) and microtopography of soil surface (2 classes). The MSO plot is shown in Fig. 14.13a. The confidence interval envelope is that of the multivariate variogram of the Hellinger-transformed mite data, undetrended. Examination of the plot shows that the variogram of the explained plus residual variation (dashed line) increases monotonically, which is the signature of a strong linear gradient in the data. The variogram of the residuals, however, is flat and not significant, showing that the environmental variables account well for the gradient in the data. The dashed line is not within the confidence envelope of the multivariate variogram for distance classes 0, 1 and 4, despite the fact that the confidence intervals, computed using non-stationary data, were too broad. This indicates that the spatial structure of the mite data varied with scale.

In an attempt to determine if a MEM model would successfully control for the spatial structure in the species-environment relationship, MSO was computed from a partial RDA that controlled for a set of seven MEM eigenfunctions selected near the end of Ecological application 14.2a. The results of this second set of analyses are displayed in Fig. 14.13b. The variogram of the explained plus residual variation is now entirely within the confidence envelope of the variogram of $\mathbf{Y}_{\text{res}|\mathbf{W}}$ and the separate variograms of explained and residual variation are flat, showing that the MEM eigenfunctions successfully controlled for the spatial correlation of the mite data that was not well explained by the environmental variables. For these data, variation partitioning (Subsections 10.3.5, 11.1.11) showed that the seven selected MEM eigenfunctions accounted almost completely for the linear trend present in both the mite data and the environmental variables. Using the selected MEM as covariables in partial RDA had effectively detrended the mite data as well as the environmental variables because the mite and environmental data were structured by the same spatial trend (Borcard & Legendre, 1994).

In a third series of analyses, the Hellinger-transformed mite data were detrended along the north-south sampling axis to meet the stationarity assumption for the computation of confidence intervals. Will this analysis produce different results than above? A MSO variogram plot was computed after a RDA of the mite data modelled by the same environmental variables as above, except that they were also detrended along the north-south sampling axis, like the mite data. The MSO plot is shown in Fig. 14.13c. The confidence interval envelope is that of the multivariate variogram displayed in Fig. 13.12 where the mite data had also been detrended; the confidence intervals are thus not too broad. The variogram of the explained plus residual variation mostly remains within the confidence envelope, except for distance classes 1, 3 and 4 where a slight departure is observed; this indicate that the species-environment relationship varied slightly with scale. In the variogram of residual variation, none of the statistics forming the variogram are significant, showing that the spatial correlation in the detrended data is well explained by the detrended environmental variables. In summary, the MSO plot indicates that the environmental variables accounted well for the spatial structure in the data although the explained variation shows that the spatial structure varied with scale.

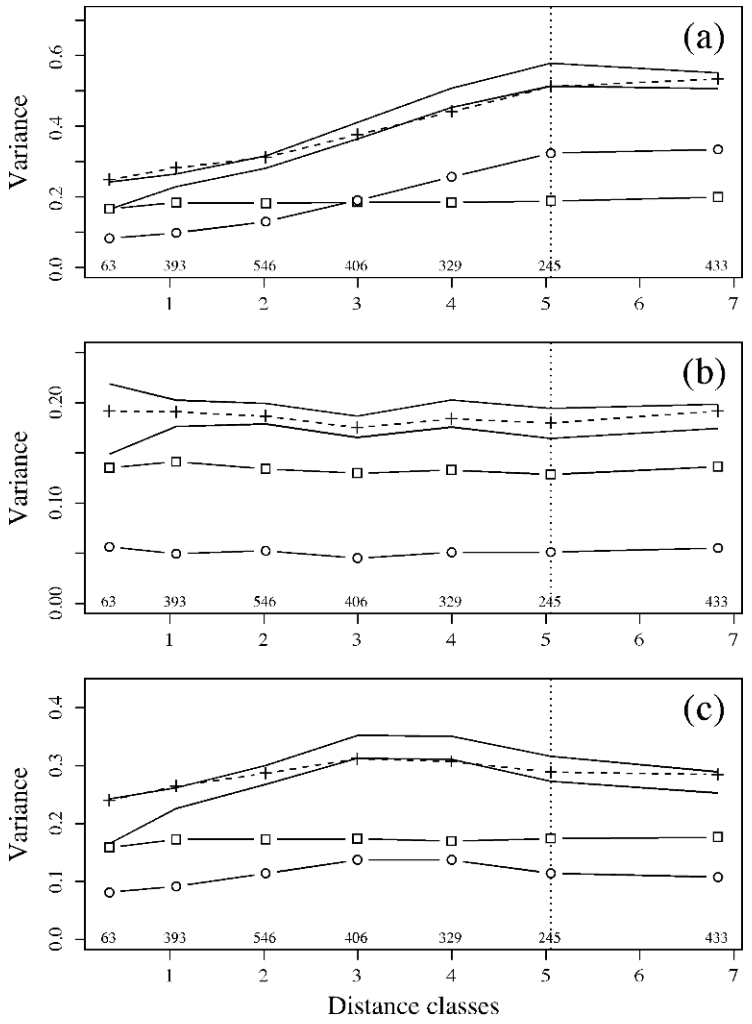


Figure 14.13 MSO plots of (a) the RDA of the Hellinger-transformed, undetrended mite data, analysed against environmental variables, (b) the partial RDA of the same matrices with MEM eigenfunctions as covariables, and (c) the RDA of the Hellinger-transformed and detrended mite data analysed against detrended environmental variables. Plots produced by function ***mso()***. Crosses: variograms of explained plus residual variation. Continuous lines delineate the confidence envelopes of the variograms of $\mathbf{Y}$ in (a), of $\mathbf{Y}_{res|\mathbf{W}}$ in (b), and of detrended $\mathbf{Y}$ in (c). Squares: variograms of residual variation; the squares are white, indicating that the variogram statistics are not significant in these examples. Circles: variograms of explained variation. Vertical dotted lines: half the maximum number of classes; the last points, to the right of these lines, include all remaining pairs of sites and should not be interpreted. Values written above the abscissas: number of pairs involved in the calculation of each statistic.

As in the second set of analyses, a MSO was computed from a partial RDA that controlled for the seven pre-selected MEM eigenfunctions. The results of this fourth set of analyses were almost identical to those shown in Fig. 14.13b: the variogram of the explained plus residual variation was now entirely within the confidence envelope of the variogram of $\mathbf{Y}_{\text{resid}|\mathbf{W}}$ and the separate variograms of explained and residual variation were flat, showing that the MEM eigenfunctions successfully controlled for the spatial correlation of the detrended mite data that was not well explained by the detrended environmental variables.

The results of this analysis could have differed from Fig. 14.13b if the mite data had been structured by a broad-scale spatial trend running in a different direction from that of the environmental data. Had that been the case, the trend in the undetrended response data **Y** shown in Fig. 14.13b would not be modelled by the undetrended environmental variables **X**. However, the fourth set of analyses described in the present paragraph, with **Y** and **X** having been both detrended, would not have been impaired by these differing trends.

The R code to run the MSO analyses reported in the present ecological application is found in Section 7.5.3 of Borcard *et al*. (2011).

14.5 Other eigenfunction-based methods of spatial analysis

This section describes additional statistical methods based on spatial eigenfunctions that were not covered in the previous sections.

1 — Space-time interaction

A commonly used approach to test hypotheses about natural or man-made environmental changes, including climate change, is to sample portions of ecosystems repeatedly over time. This type of sampling is usually done without replication of sites; in this way, the sampling effort can be spent on maximizing the expanse of space covered by the study. If the sampling sites and times are represented by dummy variables or Helmert contrasts, as in paragraphs 3 and 4 of Subsection 11.1.10, one can use canonical analysis to study the effect of the sites on species composition while controlling for the effect of time, and vice versa. An important limit of this approach is that the interaction between space and time cannot be estimated for lack of replicates. Assessing that interaction is, however, of great interest in such studies because a significant interaction would indicate that the spatial structure of the univariate or multivariate response data has changed through time, and conversely that the temporal variations differed significantly among the sites, thus indicating, for example, the signature of climate change on ecosystems.

Legendre *et al*. (2010) described a statistical method to *analyse the interaction* between the space (S) and time (T) factors *in space-time studies without replication*; the acronym of the method is STI (for *space-time interaction*). The method can be applied to multivariate response data, e.g. ecological community composition, through partial RDA. The method consists in representing the space and/or time factors by spatial and/or temporal eigenfunctions (MEM, Sections 14.1 and 14.2, or AEM,

STI

Section 14.3). It is not necessary to represent both space and time by eigenfunctions: for example, if there are many sites and only a few sampling times, e.g. 2 or 3, spatial relationships may be coded using spatial eigenfunctions and temporal relationships using dummy variables or Helmert contrasts. Coding the space and/or time factors by spatial and/or temporal eigenfunctions requires fewer coding variables than dummy variables or Helmert contrasts. The interaction can be represented by variables obtained by computing the Hadamard product of each eigenfunction that codes for space with each eigenfunction that codes for time. Enough degrees of freedom are saved to correctly estimate the residual fraction of variation and test the significance of the interaction term.

The above paper gives details about the computation method. The R package STI is available to carry out the calculations (Section 14.7). The paper also contains two applications to real species assemblage data: an analysis of Trichoptera (insects, 56 species) emerging from a stream and captured in 22 emergence traps during 100 days, grouped into 10 consecutive 10-day periods, and a study of four surveys conducted between 1982 and 1995 in the Barro Colorado Island permanent forest plot (315 species of trees). Another application is found in Laliberté *et al*. (2009) where tree seedling abundances at 40 sites along a transect in a temperate forest understory, monitored during a 9-year period, were analysed for space-time interaction. The analysis of spatio-temporal data is also discussed in Cressie & Wikle (2011).

2 — Multiscale codependence analysis

A causal relationship between an explanatory (**x**) and a response variable (**y**) across space implies that the two variables are correlated. When the correlation between **x** and **y** is not significant, the causal hypothesis must be abandoned. Conversely, a significant correlation can be interpreted as support of the causal hypothesis that **x** may have an effect on **y**. Given the multiscale nature of ecological processes, one may wonder at which scales **x** is an important predictor of **y**. The same question can be asked about pairs of variables forming a bivariate time series; for simplicity, the presentation here will focus on space.

MCA

Guénard *et al*. (2010) developed *multiscale codependence analysis* (MCA) to address the above question and test the significance of the correlations between two variables at different spatial scales. The method is based on spatial eigenfunctions, MEM or AEM, which correspond to different and identifiable spatial scales: indeed, a Moran's *I* statistic (eq. 13.1) can be computed for each eigenfunction. If the sampling is regular along a transect, eq. 14.1 can be used to determine the wavelegths of the k eigenfunctions, which are assembled in a matrix called **W**, of size $n \times k$. Correlation coefficients are computed between **y** and each of the k eigenfunctions, and written in a vector $\mathbf{r_{yW}}$ of length k. Similarly, correlation coefficients are computed between **x** and each of the k eigenfunctions, and written in a vector $\mathbf{r_{xW}}$. The Hadamard product of the two vectors, $\mathbf{r_{yW}}$ and $\mathbf{r_{xW}}$, is the vector of *codependence coefficients*, which reflect the strength of the **x**-**y** correlations at the different scales represented by the eigenfunctions in matrix **W**. Each codependence coefficient can be tested for

significance using a τ (tau) statistic obtained by computing the product of the *t*-statistics associated with the two correlation coefficients. The testing procedure is described in the paper. An R package is available for the calculations (Section 14.7).

In the above paper, the method was applied to model the river habitat of juvenile Atlantic salmon (parr). MCA showed that variables describing substrate composition of the river bed were the most influential predictors of parr abundance at the 0.4 – 4.1 km scales whereas mean channel depth was more influential at the 200 – 300 m scales. This example shows that when properly assessed, the multiscale structuring observed in nature may be used to refine our understanding of natural processes.

3 — Estimating and controlling for spatial structure in modelling

The examples and applications reported in Sections 14.1 to 14.3 show that spatial eigenfunctions can efficiently model all kinds of spatial structures in data. Can they be used to find a solution to the problem described in Subsection 1.1.2, that spatial correlation inflates the level of type I error in tests of species-environment relationships in regression and canonical analysis?

A species-environment relationship after controlling for spatial structure can be represented by fraction [a] in a Venn diagram (e.g. Figs. 10.10) showing the partitioning of the variation of response data, univariate **y** or multivariate **Y**, with respect to environmental (left circle) and spatial variables (right circle). A real example is shown in Fig. 14.7. Using numerical simulations, Peres-Neto & Legendre (2010) showed that spatial eigenfunctions provided an effective answer to the problem. Firstly, one must determine if the spatial component of **y** or **Y** is significant. This can be done by regression of **y**, or canonical analysis of **Y**, against all MEM spatial predictors, or by univariate (for **y**) or multivariate (for **Y**) variogram analysis. Secondly, if the spatial component is significant, one can select a subset of spatial predictors, and use the environmental (**X**) and the selected spatial predictors (covariables **W**) in a partial regression (for **y**, Subsection 10.3.5) or partial canonical analysis (for **Y**, Subsection 11.1.6).

For the analysis of community composition data, the authors found that a species-by-species forward selection procedure, described in their paper, was to be preferred to a global, community-based selection. In this method, eigenfunctions are selected for each species independently, and the union of the selected sets is used as the matrix of MEM covariables in canonical analysis. This provides an effective method of control for type I error in the assessment of species-environment relationships. The paper also showed that polynomial regressors (Subsection 13.2.1) did not produce tests of significance with correct levels of type I error.

The Peres-Neto & Legendre (2010) paper provides theoretical support to the effect observed in Ecological application 14.4, that MEM used as covariables in canonical analysis effectively controlled for the spatial correlation observed in the species-environment relationship in the first part of the analysis of the mite data.

14.6 Multiscale analysis of beta diversity

The present book is concerned with the analysis of multivariate ecological matrices, with special emphasis on community composition data. The book started with a difficult problem: Section 1.1 explained the origin of spatial structures in ecosystems, and showed that these structures may be generated by two types of processes, i.e. spatial dependence induced by environmental variables (eq. 1.1) and community dynamics that can generate spatial autocorrelation (eq. 1.2). How can one distinguish between these two families of processes? To do so, it is necessary to analyse community variation at multiple scales.

The focus of ecologists on natural communities of organisms stems from the fact that communities are the best response variable available to assess the effects of natural or man-made changes to the natural environment. Ecologists determine effects by studying how the members of natural communities, i.e. the species, react to changes, appraised through an appropriate sampling design. The difficulty of this approach is that species assemblages form multivariate data tables that cannot be analysed using simple univariate statistical methods. In addition, the presence of spatial structures in communities indicates that some processes have been at work that generated these structures. By relating natural communities to hypothesized causal factors, one can determine if changes observed in communities can be related to these assumed causes. In the face of climate changes and other major anthropogenic impacts, species act as dormant spies in ecosystems that ecologists can awaken to test hypotheses about the origins of changes, which is an essential step before remedial actions can be developed.

Biodiversity is a most important property of ecosystems because of its numerous services to humans, including aesthetic services (Section 6.5). The term biodiversity covers different components: taxonomic (most importantly at the species level), phylogenetic, genetic, ecological and cultural. An important aspect is the spatial organization of biodiversity, called beta diversity — the spatial variation of community composition through space. Some beta diversity studies focus on species turnover along well-identified environmental gradients. A more general concept, followed in this book, is the non-directional approach, where spatial variation of communities is studied not along selected gradients but over whole natural ecosystems. In the latter approach, the variance of a community composition matrix is a measure of beta diversity, and it was shown in this book that this variance can be analysed and decomposed using numerical methods that form a crescendo of power and refinement, from the description of multivariate (multi-species) structures in Chapters 8 and 9 to analyses carried out with respect to explanatory (e.g. environmental) variables in Chapter 11.

After developing concepts and methods of spatial analysis in Chapter 13, the present chapter has come back to the question stated in Section 1.1: how can one study the multiscale structure of communities? Methods based on spatial eigenfunctions

provided answers by drawing upon methods studied earlier in the book: distance measures (Chapter 7), principal coordinate analysis (Section 9.3), linear modelling by multiple regression (Section 10.3.3) and redundancy analysis (Section 11.1), variation partitioning (Sections 10.3.5 and 11.1.11), and the concept of scale in spatial patterns (Section 13.0).

There are already several methods available for multiscale ecological analysis, and more will undoubtedly be developed in the future. The spatial eigenfunction basis described in the present chapter is rich in possibilities, and it is left to the imagination of researchers, driven by ecological questions, to continue the development of derived methods. The help of mathematicians and statisticians will be welcome to provide solid mathematical foundations for these methods. These new methods will hopefully allow researchers to detect more clearly the signals arising from natural communities, which can be deciphered as messages from species spies in ecosystems, and used to manage the balance between human societies and the natural environment.

14.7 Software

R-language functions are available to compute all methods of analysis described in this chapter.

1. Distance-based Moran's eigenvector maps (dbMEM). — Package VEGAN contains functions ***pcnm()*** for construction of dbMEM eigenfunctions. Package PCNM, presently available on https://r-forge.r-project.org/R/?group_id=195, contains functions ***PCNM()*** and ***quickPCNM()*** for classical PCNM and dbMEM analysis. Among the functions available with the Borcard *et al*. (2011) book, ***create.MEM.model()*** allows users to generate a staggered matrix of dbMEM spatial eigenvectors corresponding to several groups of disconnected sites that are analysed together.

2. Moran's eigenvector maps (MEM), general form. — Package SPACEMAKER, presently available on https://r-forge.r-project.org/R/?group_id=195, contains functions for MEM spatial modelling. In particular, function ***test.W()*** computes and tests MEM eigenfunctions for various pre-constructed spatial weighting matrices **W** and selects the best spatial model using AIC.

3. Asymmetric eigenvector maps (AEM). — Package AEM, presently available on https://r-forge.r-project.org/R/?group_id=195, contains functions for AEM spatial modelling.

4. Multiscale ordination (MSO). — Function ***mso()*** is available in VEGAN for multiscale ordination.

5. Other eigenfunction-based methods of spatial analysis. — To study space-time interaction in surveys without replication: package STI is available on Ecological Archives E091-019-S1 (http://esapubs.org/archive/archive_E.htm, year 2010) as well as on the Web page http://sites.google.com/site/miqueldecaceres/software. Functions ***STImodels()*** and ***quickSTI()*** of that package are also found in package PCNM; see paragraph 1 above. Codependence analysis is available in package CODEP.

6. Miscellaneous methods. — Function ***geoXY()*** of SoDA transforms latitude-longitude (LatLon) data to flat Cartesian coordinates. Variation partitioning involving spatial eigenfunction submodels, with hierarchical partitioning or proportional apportioning of the shared fractions of variation, is found in R functions available in a supplement of the Legendre *et al*. (2012) paper, downloadable from the *ESA Ecological Archives* page http://esapubs.org/archive/archive_E.htm.

Package ADESPATIAL, presently under development, will contains functions to carry out the analyses described in Sections 14.1 to 14.3 of this chapter.

Plate 14.1 Historical dynamics (Section 14.1.4). In the ancient Gallic city of *Burdigala* (now Bordeaux, in southwestern France), the Romans built a Gallo-Roman town around 50 BC. At the beginning of the 2nd century AD, an oval-shaped amphitheatre (132 m × 111 m) was erected on the northwest outskirts of the city. It was later abandoned, probably during the 4th century, when the ancient city decreased in size and the amphitheatre was left out of the new city walls. At the end of the Middle Ages, the monument was progressively dismantled, in particular its western side. At the beginning of the 19th century, the inside of the amphitheatre was divided into lots were upper middle-class houses were built. The only standing portion nowadays is the northwestern porch (upper-right in the picture). The elliptic outline of the amphitheatre, now called Palais-Gallien, has been preserved in the city plan. It is still visible nowadays on aerial photographs of Bordeaux (44°50'52" N, 00°34'59" W). Information kindly provided by David Hourcade, Institut Ausonius (UMR 5607, *Centre national de la recherche scientifique/Université de Bordeaux 3*), France. Photo reproduced with permission of *Institut national de l'information géographique et forestière*, France (©IGN France, flight 1966_CDP6309_P_8000_5579).

References

References to cited works

Numbers in brackets: pages of the book where references are cited.

Aart, P. J. M. (van der) 1973. Distribution analysis of wolfspiders (Araneae, Lycosidae) in a dune area by means of principal component analysis. *Neth. J. Zool.* **23**: 266-329. [452]

Aart, P. J. M. (van der) & N. Smeenk-Enserink. 1975. Correlations between distributions of hunting spiders (Lycosidae, Ctenidae) and environmental characteristics in a dune area. *Neth. J. Zool.* **25**: 1-45. [410, 452, 487, 660, 661]

Ables, J. G. 1974. Maximum entropy spectral analysis. *Astron. Astrophys. Suppl.* **15**: 383-393. [765]

Addicott, J. F., J. M. Aho, M. F. Antolin, M. F. Padilla, J. S. Richardson & D. A. Soluk. 1987. Ecological neighborhoods: scaling environmental patterns. *Oikos* **49**: 340-346. [787]

Agresti, A. 2002. *Categorical data analysis. 2nd edition*. Wiley Series in Probability and Statistics, John Wiley & Sons, Hoboken, New Jersey. xv + 710 pp. [235]

Akaike, H. 1974. A new look at the statistical model identification. *IEEE Trans. Autom. Control* **19**: 716-723. [566]

Ali, G., A. Roy & P. Legendre. 2010. Spatial relationships between soil moisture patterns and topographic variables at multiple scales in a humid temperate forested catchment. *Water Resour. Res.* **46**, W10526, doi:10.1029/2009WR008804. [877]

Allan, J. D. 1975. Components of diversity. *Oecologia* **18**: 359-367. [260]

Allen, T. F. H., S. M. Bartell & J. F. Koonce. 1977. Multiple stable configurations in ordination of phytoplankton community change rates. *Ecology* **58**: 1076-1084. [769, 774]

Allen, T. F. H. & T. W. Hoekstra. 1991. Role of heterogeneity in scaling of ecological systems under analysis. 47-68 *in*: J. Kolasa & S. T. A. Pickett [eds.] *Ecological heterogeneity*. Springer-Verlag, New York. [786]

Allen, T. F. H. & T. B. Starr. 1982. *Hierarchy – Perspectives for ecological complexity*. Univ. of Chicago Press, Chicago. xvi + 310 pp. [9]

Alonso, D., R. S. Etienne & A. J. McKane. 2006. The merits of neutral theory. *Trends Ecol. Evol.* **21**: 451-457. [12]

Amanieu, M., P. Legendre, M. Troussellier & G.-F. Frisoni. 1989. Le programme Écothau: théorie écologique et base de la modélisation. *Oceanol. Acta* **12**: 189-199. [569, 574]

Anderberg, M. R. 1973. *Cluster analysis for applications*. Academic Press, New York. xiii + 359 pp. [386]

Andersen, R., M. Poulin, D. Borcard, R. Laiho, J. Laine, H. Vasander & E.-T. Tuittila. 2011. Environmental control and spatial structures in peatland vegetation. *J. Veg. Sci.* **22**: 878-890. [877]

Anderson, M. J. 1999. Distinguishing direct from indirect effects of grazers in intertidal estuarine assemblages. *J. Exp. Mar. Biol. Ecol.* **234**: 199-218. [648, 649]

Anderson, M. J. 2006. Distance-based tests for homogeneity of multivariate dispersions. *Biometrics* **62**: 245-253. [303, 656, 682]

Anderson, M. J., T. O. Crist, J. M. Chase, M. Vellend, B. D. Inouye, A. L. Freestone, N. J. Sanders, H. V. Cornell, L. S. Comita, K. F. Davies, S. P. Harrison, N. J. B. Kraft, J. C. Stegen & N. G. Swenson. 2011. Navigating the multiple meanings of β diversity: a roadmap for the practicing ecologist. *Ecol. Lett.* **14**: 19-28. [258, 260]

Anderson, M. J., K. E. Ellingsen & B. H. McArdle. 2006. Multivariate dispersion as a measure of beta diversity. *Ecol. Lett.* **9**: 683-693. [258, 285, 305, 327]

Anderson, M. J. & N. A. Gribble. 1998. Partitioning the variation among spatial, temporal and environmental components in a multivariate data set. *Aust. J. Ecol.* **23**: 158-167. [853, 857]

Anderson, M. J. & P. Legendre. 1999. An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. *J. Statist. Comput. Simulation* **62**: 271-303. [579, 651, 652]

Anderson, M. J. & T. J. Willis. 2003. Canonical analysis of principal coordinates: a useful method of constrained ordination for ecology. *Ecology* **84**: 511-525. [709]

Anderson, T. W. 1971. *The statistical analysis of time series*. John Wiley & Sons, New York. 704 pp. [750, 894]

Anderson, T. W. 2003. *An introduction to multivariate statistical analysis. 3rd edition*. Wiley-Interscience, Hoboken, New Jersey. xx + 721 pp. [143]

Angot, M. 1961. Analyse quantitative du cycle diurne de la production primaire dans le Pacifique subtropical près de la Nouvelle-Calédonie. *Bull. Inst. Océanogr. (Monaco)* (1200): 1-34. [753]

Anselin, L. 1995. Local Indicators of Spatial Association–LISA. *Geogr. Anal.* **27**: 93-115. [806]

Ardisson, P.-L., E. Bourget & P. Legendre. 1990. Multivariate approach to study species assemblages at large spatiotemporal scales: the community structure of the epibenthic fauna of the Estuary and Gulf of St. Lawrence. *Can. J. Fish. Aquat. Sci.* **47**: 1364-1377. [510, 511, 774]

Arfi, R., F. Blanc & D. Calmet. 1982. Étude d'impact en milieu marin: échantillonnage et traitement des données. 341-364 *in*: S. Frontier [ed.] *Stratégies d'échantillonnage en écologie*. Masson, Paris et Les Presses de l'Université Laval, Québec. [763, 764]

Arfi, R. & P. Dumas. 1990. Séries chronologiques: analyse spectrale de Fourier et par maximisation d'entropie – Présentation, simulations, applications. 105-126 *in*: S. Frontier [ed.] *Biométrie et océanographie*. Actes de Colloques, 10, IFREMER, Brest. [765]

Arias-González, J. E., P. Legendre & F. A. Rodríguez-Zaragoza. 2008. Scaling up beta diversity on Caribbean coral reefs. *J. Exp. Mar. Biol. Ecol.* **366**: 28-36. [877]

Armstrong, M. [ed.]. 1989. *Geostatistics. Vol. 1 and 2*. Kluwer Academic Publishers, Dordrecht. xxix + 491 pp., xvii + 546 pp. [832]

Armstrong, M., D. Renard & P. Berthou. 1989. *Applying geostatistics to the estimation of a population of bivalves*. ICES C. M. 1989/K37. 22 pp. [832]

Astorga, A., J. Heino, M. Luoto, M. & T. Muotka. 2011. Freshwater biodiversity at regional extent: determinants of macroinvertebrate taxonomic richness in headwater streams. *Ecography* **34**: 705-713. [877]

Bach, P., P. Legendre, M. Amanieu & G. Lasserre. 1992. Strategy of eel (*Anguilla anguilla* L.) exploitation in the Thau lagoon. *Estuar. Coast. Shelf Sci.* **35**: 55-73. [185, 186]

Bachraty, C., P. Legendre & D. Desbruyères. 2009. Biogeographic relationships among deep-sea hydrothermal vent faunas at global scale. *Deep-Sea Res. I* **56**: 1371-1378. [849]

Barbalat, S. & D. Borcard. 1997. Distribution of four beetle families (Coleoptera: Buprestidae, Cerambycidae, phytophagous Scarabaeidae and Lucanidae) in different forest ecotones in the Areuse Gorges (Neuchâtel, Switzerland). *Écologie* **28**: 199-208. [402]

Barbujani, G., N. L. Oden & R. R. Sokal. 1989. Detecting regions of abrupt change in maps of biological variables. *Syst. Zool.* **38**: 377-389. [844]

Bare, B. B. & D. W. Hann. 1981. Applications of ridge regression in forestry. *For. Sci.* **27**: 339-348. [564]

Barnes, H. 1952. The use of transformations in marine biological statistics. *J. Cons. Int. Explor. Mer* **18**: 61-71. [46]

Barrodale, I. & R. E. Erickson. 1980. Algorithms for least-square linear prediction and maximum entropy spectral analysis. Part I: Theory. *Geophysics* **45**: 420-432. [765]

Bartlett, M. S. 1938. Further aspects of the theory of multiple regression. *Proc. Camb. Phil. Soc.* **34**: 33-40. [682]

Bartlett, M. S. 1948. Internal and external factor analysis. *Brit. J. Psychol. Stat. Sect.* **1**: 73-81. [683]

Bartlett, M. S. 1950. Tests of significance in factor analysis. *Brit. J. Psychol. Stat. Sect.* **3**: 77-85. [448]

Bartlett, M. S. 1954. A note on the multiplying factors for various chi-squared approximations. *J. Roy. Statist. Soc. Ser. B* **16**: 296-298. [157]

Bartlett, M. S. 1978. Nearest neighbour models in the analysis of field experiments. *J. Roy. Statist. Soc. Ser. B* **40**: 147-174. [20]

Barton, D. E. & F. N. David. 1956. Some notes on ordered random intervals. *J. Roy. Statist. Soc. Ser. B* **18**: 79-94. [256]

Bates, D. M. & J. M. Chambers. 1992. Nonlinear models. Chapter 10 *in*: J. M. Chambers [ed.] *Statistical models in S*. Wadsworth & Brooks/Cole, Pacific Grove, California. [583]

Baum, B. R. 1992. Combining trees as a way of combining data for phylogenetic inference and the desirability of combining gene trees. *Taxon* **41**: 3-10. [417]

Beach, C. M. & J. G. MacKinnon. 1978. A maximum likelihood procedure for regression with autocorrelated errors. *Econometrica* **46**: 51-58. [20]

Beals, E. W. 1984. Bray-Curtis ordination: an effective strategy for analysis of multivariate ecological data. 1-55 *in*: A. MacFadyen & E. D. Ford [eds.] *Advances in Ecological Research, 14*. Academic Press, London. [334]

Beaugrand, G., M. Edwards & L. Legendre. 2010 Marine biodiversity, ecosystem functioning and carbon cycles. *Proc. Natl. Acad. Sci. USA* **107**: 10120-10124. [130]

Belbin, L. 1991. Semi-strong hybrid scaling, a new ordination algorithm. *J. Veg. Sci.* **2**: 491-496. [512]

Belbin, L. & C. McDonald. 1993. Comparing three classification strategies for use in ecology. *J. Veg. Sci.* **4**: 341-348. [383]

Bell, M. A. & P. Legendre. 1987. Multicharacter chronological clustering in a sequence of fossil sticklebacks. *Syst. Zool.* **36**: 52-61. [773]

Bellchambers, L. M., J. J. Meeuwig, S. N. Evans & P. Legendre. 2011. Modelling habitat associations of the common spider conch in the Cocos (Keeling) Islands. *Mar. Ecol. Prog. Ser.* **432**: 83-90. [877]

Bellehumeur, C. & P. Legendre. 1998. Multiscale sources of variation in ecological variables: modelling spatial dispersion, elaborating sampling designs. *Landscape Ecology* **13**: 15-25. [817, 818]

Bellehumeur, C., P. Legendre & D. Marcotte. 1997. Variance and spatial scales in a tropical rain forest: changing the size of sampling units. *Plant Ecol.* **130**: 89-98. [812]

Bendat, J. S. & A. G. Piersol. 1971. *Random data – Analysis and measurement procedures*. Wiley-Interscience, New York. xv + 407 pp. [754]

Beninca, E., J. Huisman, R. Heerkloss, K. D. Jöhnk, P. Branco, E. H. Van Nes, M. Scheffer & S. P. Ellner. 2008. Chaos in a long-term experiment with a plankton community. *Nature* **451**: 822-828. [2]

Benzécri, J. P. 1969. Statistical analysis as a tool to make patterns emerge from data. 35-60 *in*: S. Watanabe [ed.] *Methodologies of pattern recognition*. Academic Press, New York. [464]

Benzécri, J. P. and coll. 1973. *L'analyse des données. Tome I: La taxinomie. Tome II: L'analyse des correspondances*. Dunod, Paris. viii + 615, vii + 619 pp. [428, 464, 477, 483]

Bergmann, C. 1847. Ueber die Verhältnisse der Wärmeökonomie der Thiere zu ihrer Grösse. *Gottinger Studien* **3**: 595-708. [537]

Bernstein, B. B. & J. Zalinski. 1983. An optimum sampling design and power tests for environmental biologists. *J. Environ. Manag.* **16**: 35-43. [267]

Berryman, J. G. 1978. Choice of operator length for maximum entropy spectral analysis. *Geophysics* **43**: 1384-1391. [765]

Bersier, L.-F. & D. Meyer. 1994. Bird assemblages in mosaic forests: the relative importance of vegetation structure and floristic composition along the sucessional gradient. *Acta Œcologica* **15**: 561-576. [855]

Bertalanffy, L. 1968. *General system theory: foundations, development, applications*. Braziller, New York. xv + 289 pp. [221]

Besag, J. & P. Clifford. 1989. Generalized Monte Carlo significance tests. *Biometrika* **76**: 633-642. [21]

Beum, C. O. J. & E. G. Brundage. 1950. A method for analyzing the sociomatrix. *Sociometry* **13**: 141-145. [404]

Bezdek, J. C. 1987. Some non-standard clustering algorithms. 225-287 *in*: P. Legendre & L. Legendre [eds.] *Developments in numerical ecology.* NATO ASI Series, Vol. G-14. Springer-Verlag, Berlin. [338, 423]

Binet, D., M. Gaborit, A. Dessier & M. Roux. 1972. Premières données sur les copépodes pélagiques de la région congolaise. II. Analyse des correspondances. *Cah. O. R. S. T. O. M. Sér. Océanogr.* **10**: 125-137. [479]

Birks, H. J. B. 1993. Is the hypothesis of survival on glacial nunataks necessary to explain the present-day distributions of Norwegian mountain plants? *Phytocoenologia* **23**: 399-426. [582, 583]

Birks, H. J. B. 1996. Statistical approaches to interpreting diversity patterns in the Norwegian mountain flora. *Ecography* **19**: 332-340. [582, 583, 855]

Birks, H. J. B. 2010. Numerical methods for the analysis of diatom assemblage data. 23-54 *in*: J. P. Smol & E. F. Stoermer [eds.] *The diatoms – Applications for the environmental and earth sciences. 2nd edition*. Cambridge University Press, New York. [672]

Birks, H. J. B., H. A. Austin, N. E. Indrevaer, S. M. Peglar & C. Rygh. 1998. *An annotated bibliography of canonical correspondence analysis and related constrained ordination methods 1986-1996*. Available from H. J. B. Birks, Botanical Institute, University of Bergen, Allégaten 41, N-5007 Bergen, Norway. Also available from the WWWeb page http:/numericalecology.com/cca_bib/. [629, 670]

Birks, H. J. B., S. Juggins & J. M. Line. 1990a. Lake surface-water chemistry reconstructions from palaeolimnological data. 301-313 *in*: B. J. Mason [ed.] *The surface waters acidification programme*. Cambridge University Press, Cambridge. [671]

Birks, H. J. B., J. M. Line, S. Juggins, A. C. Stevenson & C. J. F. ter Braak. 1990b. Diatoms and pH reconstruction. *Phil. Trans. R. Soc. Lond. B* **327**: 263-278. [672]

Birks, H. J. B., A. F. Lotter, S. Juggins & J. P. Smol [eds.] 2012. *Tracking environmental change using lake sediments, Volume 5: Data handling and numerical techniques*. Springer, Dordrecht, The Netherlands. xi + 716 pp. [670]

Bishop, Y. M. M., S. E. Fienberg & P. W. Holland. 1975. *Discrete multivariate analysis – Theory and practice*. MIT Press, Cambridge, Mass. x + 557 pp. [230, 235, 241, 244, 765]

Bivand, R. 1980. A Monte Carlo study of correlation coefficient estimation with spatially autocorrelated observations. *Quaest. Geogr.* **6**: 5-10. [18]

Bjorholm, S., J.-C. Svenning, F. Skov & H. Balslev. 2005. Environmental and spatial controls of palm (Arecaceae) species richness across the Americas. *Global Ecol. Biogeogr.* **14**: 423-429. [855]

Bjørnstad, O. N. & W. Falck. 2001. Nonparametric spatial covariance functions: estimation and testing. *Environ. Ecol. Stat.* **8**: 53-70. [805]

Bjørnstad, O. N., N. C. Stenseth & T. Saitoh. 1999. Synchrony and scaling in dynamics of voles and mice in northern Japan. *Ecology* **80**: 622-637. [805]

Blackith, R. E. & F. O. Albrecht. 1959. Morphometric differences between the eye-stripe polymorphs of the red locust. *Scient. J. Roy. Coll. Sci.* **27**: 13-27. [695]

Blackith, R. E. & R. A. Reyment. 1971. *Multivariate morphometrics*. Academic Press, London. ix + 412 pp. [695]

Blanc, F., P. Chardy, A. Laurec & J.-P. Reys. 1976. Choix des métriques qualitatives en analyse d'inertie. Implications en écologie marine benthique. *Mar. Biol. (Berl.)* **35**: 49-67. [270]

Blanchet, F. G., P. Legendre & D. Borcard. 2008a. Modelling directional spatial processes in ecological data. *Ecol. Model.* **215**: 325-336. [888, 892]

Blanchet F. G., P. Legendre & D. Borcard. 2008b. Forward selection of explanatory variables. *Ecology* **89**: 2623-2632. [658]

Blanchet, F. G., P. Legendre & D. Borcard. 2009. Erratum to "Modelling directional spatial processes in ecological data" [Ecological Modelling 215 (2008): 325-336]. *Ecol. Model.* **220**: 82-83. [888]

Blanchet, F. G., P. Legendre, R. Maranger, D. Monti & P. Pepin. 2011. Modelling the effect of directional spatial ecological processes at different scales. *Oecologia* **166**: 357-368. [889, 890, 893]

Blashfield, R. K. & M. S. Aldenderfer. 1978. The literature on cluster analysis. *Multivar. Behav. Res*. **13**: 271-295. [340]

Bloom, S. A. 1981. Similarity indices in community studies: potential pitfalls. *Mar. Ecol. Prog. Ser*. **5**: 125-128. [312, 323]

Bloomfield, P. 1976. *Fourier analysis of time series – An introduction*. Wiley, New York. xiii + 258 pp. [714]

Bock, H. H. 1989. Probabilistic aspects in cluster analysis. 12-44 *in*: O. Opitz [ed.] Conceptual and numerical analysis of data. Springer-Verlag, Berlin. [415]

Bock, H. H. 1996. Probability models and hypotheses testing in partitioning cluster analysis. 377-453 *in*: P. Arabie, L. J. Hubert & G. De Soete [eds.] *Clustering and Classification*. World Scientific Publ. Co., River Edge, New Jersey. [415]

Boggs, P. T. & J. E. Rogers. 1990. Orthogonal distance regression. *Contemp. Math*. **112**: 183-194. [556]

Boltzmann, L. 1898. *Vorlesungen über Gastheorie, Vol. II*. J. A. Barth, Leipzig. [222]

Bonferroni, C. E. 1935. Il calcolo delle assicurazioni su gruppi di teste. 13-60 *in*: *Studi in onore del Professore Salvatore Ortu Carboni*. Roma. [23]

Borcard, D. 1996. Typologie des assemblages d'espèces d'Oribates (Acari, Oribatei) de la tourbière du Cachot (Jura suisse): espèces indicatrices ou groupements caractéristiques? *Bull. Soc. Neuchâtel. Sci. Nat*. **119**: 63-73. [402]

Borcard, D., W. Geiger & W. Matthey. 1995. Oribatid mite assemblages in a contact zone between a peat-bog and a meadow in the Swiss Jura (Acari, Oribatei): influence of landscape structures and historical processes. *Pedobiologia* **39**: 318-330. [849]

Borcard, D., F. Gillet & P. Legendre. 2011. *Numerical ecology with R*. Use R! series, Springer Science, New York. xi + 306 pp. [xvi, 33, 217, 396, 403, 409, 423, 424, 566, 656, 658-660, 695, 703, 773, 858, 866, 887, 898, 900, 904]

Borcard, D. & P. Legendre. 1994. Environmental control and spatial structure in ecological communities: an example using oribatid mites (Acari, Oribatei). *Environ. Ecol. Stat*. **1**: 37-53. [396, 658, 700, 815, 878, 879, 881, 898]

Borcard, D. & P. Legendre. 2002. All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecol. Model*. **153**: 51-68. [861]

Borcard, D. & P. Legendre. 2012. Is the Mantel correlogram powerful enough to be useful in ecological analysis? A simulation study. *Ecology* **93**: 1473-1481. [819]

Borcard, D., P. Legendre, C. Avois-Jacquet & H. Tuomisto. 2004. Dissecting the spatial structure of ecological data at multiple scales. *Ecology* **85**: 1826-1832. [861, 866, 870, 877]

Borcard, D., P. Legendre & P. Drapeau. 1992. Partialling out the spatial component of ecological variation. *Ecology* **73**: 1045-1055. [571, 658]

Borcard, D. & C. Vaucher-von Ballmoos. 1997. Oribatid mites (Acari, Oribatida) of a primary peat bog-pasture transition in the Swiss Jura mountains. *Écoscience* **4**: 470-479. [402]

Borgman, L. E. & W. F. Quimby. 1988. Sampling for tests of hypothesis when data are correlated in space and time. 25-43 *in*: L. H. Keith [ed.] *Principles of environmental sampling*. ACS Professional Reference Book. American Chemical Society. [21]

Bos, A. (van den) 1971. Alternative interpretation of maximum entropy spectral analysis. *IEEE Trans. Inf. Theory* **17**: 493-494. [764, 765]

Botta-Dukát, Z. 2005. Rao's quadratic entropy as a measure of functional diversity based on multiple traits. *J. Veg. Sci*. **16**: 533-540. [255]

Boudoux, M. & C.-H. Ung. 1979. Applications de la régression pseudo-orthogonale en recherche forestière. *Biom-Praxim*. **19**: 59-89. [564]

Boudreault, F. R., J. D. Dupont & C. Sylvain. 1977. Modèles linéaires de prédiction des débarquements de homard aux îles de la Madeleine (Golfe du Saint-Laurent). *J. Fish. Res. Board Can*. **34**: 379-383. [782]

Bourbaki, N. 1960. *Eléments d'histoire des mathématiques*. Hermann, Paris. 277 pp. [59]

Bourbeau, L., F. Ouellette & F. Pinard. 1984. *Le système BLOPS 2.0: un dictionnaire morphologique informatisé du français et sa logithèque*. Université de Montréal, Montréal. [228]

Bourgault, G., D. Marcotte & P. Legendre. 1992. The multivariate (co)variogram as a spatial weighting function in classification methods. *Math. Geol*. **24**: 463-478. [843]

Bowerman, B. L. & R. T. O'Connell. 1987. *Time series forecasting*. Duxbury Press, Boston. xi + 540 pp. [780]

Box, G. E. P. & D. R. Cox. 1964. An analysis of transformations. *J. Roy. Statist. Soc. Ser. B* **26**: 211-243. [49]

Box, G. E. P. & G. M. Jenkins. 1976. *Time series analysis – Forecasting and control. Revised edition*. Holden-Day, San Francisco. xxi + 575 pp. [20, 714, 780, 781]

Bray, J. R. & J. T. Curtis. 1957. An ordination of the upland forest communities of southern Wisconsin. *Ecol. Monogr.* **27**: 325-349. [xii, 12, 311, 582, 669, 707, 785]

Breiman, L., J. H. Friedman, R. A. Olshen & C. G. Stone. 1984. *Classification and regression trees*. Wadsworth International Group, Belmont, California. [406, 409]

Brill, R., H. Christanson, G. F. Estabrook, H. S. Fleming, B. Handley, P. Legendre, F. Ouellette, D. J. Rogers & M. Wirth. 1972. *Program CHARANAL. User's manual*. H. S. Fleming & S. G. Appan [eds.] Gulf Universities Research Corp., NASA/Mississippi Test Facility, Bay St-Louis. [378]

Brillinger, D. R. 1981. *Time series – Data analysis and theory. Expanded edition*. Holden-Day, San Francisco. xii + 540 pp. [714, 763]

Brillouin, L. 1956. *Science and information theory*. Academic Press, New York. 320 pp. [228, 253]

Brind'Amour, A., D. Boisclair, P. Legendre & D. Borcard. 2005. Multiscale spatial distribution of a littoral fish community in relation to environmental variables. *Limnol. Oceanogr.* **50**: 465-479. [876, 877]

Brock, G., V. Pihur, S. Datta & S. Datta. 2008. clValid: an R package for cluster validation. *J. Stat. Softw.* **25**(4). http://www.jstatsoft.org/v25/i04. [418, 424]

Bronson, R. 2011. *Schaum's outline of matrix operations. 2nd edition*. McGraw-Hill, New York. 240 pp. [59]

Brown, B. M. & J. S. Maritz. 1982. Distribution-free methods in regression. *Aust. J. Stat*. **24**: 318-331. [31]

Brown, M. B. 1976. Screening effects in multidimensional contingency tables. *Appl. Statist*. **25**: 37-46. [240]

Brown, T. A. 2006. *Confirmatory factor analysis for applied research*. Guilford, New York. xviii + 475 pp. [535]

Buckingham, E. 1914. On physically similar systems; illustrations of the use of dimensional equations. *Phys. Rev. (2nd series)* **4**: 345-376. [117]

Burbidge, A. A., N. L. McKenzie, K. E. C. Brennan, J. C. Z. Woinarski, C. R. Dickman, A. Baynes, G. Gordon, P. W. Menkhorst & A. C. Robinson. 2008. Conservation status and biogeography of Australia's terrestrial mammals. *Aust. J. Zool.* **56**: 411-422. [855]

Burg, J. P. 1967. *Maximum entropy spectral analysis*. Paper presented at the 37th Annual Meeting of Exploration Geophysics, October 31, Oklahoma City, Okla. [764, 765]

Burgman, M. A. 1987. An analysis of the distribution of plants on granite outcrops in southern Western Australia using Mantel tests. *Vegetatio* **71**: 79-86. [601]

Burgman, M. A. 1988. Spatial analysis of vegetation patterns in southern Western Australia: implications for reserve design. *Aust. J. Ecol.* **13**: 415-429. [601]

Burnham, K. P. & D. R. Anderson. 2002. *Model selection and multimodel inference: a practical information-theoretic approach. 2nd edition*. Springer-Verlag, New York. xxvi + 488 pp. [567]

Burt, C. 1952. Tests of significance in factor analysis. *Brit. J. Psychol. Stat. Sect.* **5**: 109-133. [448]

Buttler, A., B. G. Warner, P. Grosvernier & Y. Matthey. 1996. Vertical patterns of testate amoebae (Protozoa: Rhizopoda) and peat-forming vegetation on cutover bogs in the Jura, Switzerland. *New Phytol.* **134**: 371-382. [855]

Cadoret, L., P. Legendre, M. Adjeroud & R. Galzin. 1995. Répartition spatiale des Chaetodontidae dans différents secteurs récifaux de l'île de Moorea, Polynésie française. *Écoscience* **2**: 129-140. [487, 489, 670, 671]

Cailliez, F. 1983. The analytical solution of the additive constant problem. *Psychometrika* **48**: 305-308. [503]

Cailliez, F. & J.-P. Pagès. 1976. *Introduction à l'analyse des données*. Société de Mathématiques appliquées et de Sciences humaines, Paris. xxii + 616 pp. [39, 269, 505, 506]

Cain, A. J. & G. A. Harrison. 1958. An analysis of the taxonomist's judgement of affinity. *Proc. Zool. Soc. Lond.* **131**: 85-98. [44]

Calinski, T. & J. Harabasz. 1974. A dendrite method for cluster analysis. *Commun. Stat.* **3**: 1-27. [389]

Campbell, D. J. & E. Shipp. 1974. Spectral analysis of cyclic behaviour with examples from the field cricket *Teleogryllus commodus* (Walk.). *Anim. Behav.* **22**: 862-875. [758]

Campbell, V., P. Legendre & F.-J. Lapointe. 2009. Assessing congruence among ultrametric distance matrices. *J. Classif.* **26**: 103-117. [218]

Campbell, V., P. Legendre & F.-J. Lapointe. 2011. The performance of the Congruence Among Distance Matrices (CADM) test in phylogenetic analysis. *BMC Evol. Biol.* **11**: 64. http://www.biomedcentral.com/1471-2148/11/64. [218]

Carlier, A. & P. M. Kroonenberg. 1996. Decompositions and biplots in three-way correspondence analysis. *Psychometrika* **61**: 355-373. [269]

Carpenter, S. R. & J. E. Chaney. 1983. Scale of spatial pattern: four methods compared. *Vegetatio* **53**: 153-160. [789]

Carroll, J. D. 1987. Some multidimensional scaling and related procedures devised at Bell Laboratories, with ecological applications. 65-138 *in*: P. Legendre & L. Legendre [eds.] *Developments in numerical ecology*. NATO ASI Series, Vol. G-14. Springer-Verlag, Berlin. [512, 516]

Casgrain, P., P. Legendre, J.-L. Sixou & C. Mouton. 1996. A graph-theory method to establish serological relationships within a bacterial taxon, with example from *Porphyromonas gingivalis*. *J. Microbiol. Methods* **26**: 225-236. [68]

Cassie, R. M. & A. D. Michael. 1968. Fauna and sediments of an intertidal mud flat: a multivariate analysis. *J. Exp. Mar. Biol. Ecol.* **2**: 1-23. [450]

Cattaneo, A., P. Legendre & T. Niyonsenga. 1993. Exploring periphyton unpredictability. *J. North Am. Benthol. Soc.* **12**: 418-430. [855]

Cattell, R. B. 1952. *Factor analysis — An introduction and manual for the psychologist and social scientist*. Harper, New York. 462 pp. [266]

Cattell, R. B. 1966. The data box: its ordering of total resources in terms of possible relational systems. 67-128 *in*: R. B. Cattell [ed.] *Handbook of multivariate experimental psychology*. Rand McNally & Co., Chicago. [266, 267]

Cavalli-Sforza, L. L. & A. W. F. Edwards. 1967. Phylogenetic analysis: models and estimation procedures. *Evolution* **21**: 550-570. [301, 346]

Cazelles, B., M. Chavez, D. Berteaux, F. Ménard, J. O. Vik, S. Jenouvrier & N. C. Stenseth. 2008. Wavelet analysis of ecological time series. *Oecologia* **156**: 287-304. [766, 767]

Cazelles, B. & S. Hales. 2006. Infectious diseases, climate influences and nonstationarity. *PLoS Med.* **3**: 1212-1213 (e328). [767]

Chalmond, B. 1986. Régression avec résidus spatialement autocorrélés et recherche de la tendance spatiale. *Statist. Anal. Données* **11**: 1-25. [18]

Chambers, J. M. 1977. *Computational methods for data analysis*. Wiley, New York. xi + 268 pp. [589]

Chambers, J. M., W. S. Cleveland, B. Kleiner & P. A. Tukey. 1983. *Graphical methods for data analysis*. Wadsworth International Group, Belmont, California. xiv + 395 pp. [591]

Chambers, J. M. & B. Kleiner. 1982. Graphical techniques for multivariate data and for clustering. 209-244 *in*: P. R. Krishnaiah & L. N. Kanal [eds.] *Handbook of statistics. Vol. 2*. North-Holland Publ. Co., Amsterdam. [418]

Chao, A., R. L. Chazdon, R. K. Colwell & T.-J. Shen. 2005. A new statistical approach for assessing similarity of species composition with incidence and abundance data. *Ecol. Lett.* **8**: 148-159. [284, 285]

Chase, J. M., N. J. B. Kraft, K. G. Smith, M. Vellend & B. D. Inouye. 2011. Using null models to disentangle variation in community dissimilarity from variation in a-diversity. *Ecosphere* **2**: art24. doi:10.1890/ES10-00117.1. [294]

Chatfield, C. 1989. *The analysis of time series. 4th edition*. Chapman & Hall, London. xiii + 241 pp. [714]

Cheetham, A. H. & J. E. Hazel. 1969. Binary (presence-absence) similarity coefficients. *J. Paleontol.* **43**: 1130-1136. [270]

Chipman, J. S. 1979. Efficiency of least squares estimation of linear trend when residuals are autocorrelated. *Econometrica* **47**: 115-128. [20]

Chodorowski, A. 1959. Ecological differentiation of turbellarians in Harsz-Lake. *Pol. Arch. Hydrobiol.* **6**: 33-73. [249]

Cicéri, M.-F., B. Marchand & S. Rimbert. 1977. *Introduction à l'analyse de l'espace*. Collection de Géographie applicable. Masson, Paris. ix + 173 pp. [789]

Clark, I. 1979. *Practical geostatistics*. Elsevier Applied Sciences, London. xii + 129 pp. [831]

Clark, P. J. 1952. An extension of the coefficient of divergence for use with multiple characters. *Copeia* **1952**: 61-64. [306]

Clarke, K. R. 1988. Detecting change in benthic community structure. 131-142 *in*: R. Oger [ed.] *Proceedings of invited papers, fourteenth international biometric conference, Namur, Belgium*. Société Adolphe Quételet, Gembloux, Belgium. [608–610]

Clarke, K. R. 1993. Non-parametric multivariate analyses of changes in community structure. *Aust. J. Ecol.* **18**: 117-143. [608–610]

Clarke, K. R. & R. M. Warwick. 1994. *Change in marine communities – An approach to statistical analysis and interpretation*. Plymouth Marine Laboratory, Plymouth. 144 pp. [609, 610]

Cleveland, W. S. 1979. Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.* **74**: 829-836. [590]

Cleveland, W. S. 1985. *The elements of graphing data*. Wadsworth, Monterey, California. xii + 323 pp. [591]

Cliff, A. D. & J. K. Ord. 1973. *Spatial autocorrelation*. Pion, London. 178 pp. [20]

Cliff, A. D. & J. K. Ord. 1975. The comparison of means when samples consist of spatially autocorrelated observations. *Environment and Planning A* **7**: 725-734. [18]

Cliff, A. D. & J. K. Ord. 1981. *Spatial processes – Models and applications*. Pion, London. 266 pp. [12, 15, 18, 20, 790, 792, 793, 797-800, 804, 834]

Clifford, H. T. & D. W. Goodall. 1967. A numerical contribution to the classification of the Poaceae. *Aust. J. Bot.* **15**: 499-519. [349]

Clifford, H. T. & W. Stephenson. 1975. *An introduction to numerical classification*. Academic Press, New York. xii + 229 pp. [270]

Clifford, P., S. Richardson & D. Hémon. 1989. Assessing the significance of the correlation between two spatial processes. *Biometrics* **45**: 123-134. [18, 20]

Clint, M. & A. Jennings. 1970. The evaluation of eigenvalues and eigenvectors of real symmetric matrices by simultaneous iteration. *Computer J.* **13**: 76-80. [456, 479]

Clua, E., N. Buray, P. Legendre, J. Mourier & S. Planes. 2010. Behavioural response of sicklefin lemon sharks *Negaprion acutidens* to underwater feeding for ecotourism purposes. *Mar. Ecol. Prog. Ser.* **414**: 257-266. [318]

Cochran, W. G. 1954. Some methods for strengthening the common χ^2 tests. *Biometrics* **10**: 417-451. [230]

Cochran, W. G. 1977. *Sampling techniques. 3rd edition*. Wiley, New York. [21, 241]

Cochrane, D. & G. H. Orcutt. 1949. Application of least squares regression to relationships containing autocorrelated error terms. *J. Amer. Statist. Assoc.* **44**: 32-61. [20]

Cohen, J. 1988. *Statistical power analysis for the behavioral sciences. 2nd edition*. Lawrence Erlbaum Assoc., Publ., Hillsdale, New Jersey. xxi + 567 pp. [804]

Cole, L. C. 1949. The measurement of interspecific association. *Ecology* **30**: 411-424. [270]

Cole, L. C. 1957. The measurement of partial interspecific association. *Ecology* **38**: 226-233. [270]

Colebrook, J. M. & A. H. Taylor. 1984. Significant time scale of long-term variability in the plankton and the environment. *Rapp. P.-V. Réun. Cons. Int. Explor. Mer.* **183**: 20-26. [766]

Coleman, J. S. 1964. *Introduction to mathematical sociology*. The Free Press of Glencoe, Collier-Macmillan Ltd., New York. xiv + 554 pp. [68]

Colwell, R. K. 1974. Predictability, constancy and contingency of periodic phenomena. *Ecology* **55**: 1148-1153. [744]

Conover, W. J. 1980. *Practical nonparametric statistics. 2nd edition*. Wiley, New York. xiv + 493 pp. [584]

Cook, D. G. & S. J. Pocock. 1983. Multiple regression in geographical mortality studies, with allowance for spatially correlated errors. *Biometrics* **39**: 361-371. [20]

Cordier, B. 1965. *Sur l'analyse factorielle des correspondances*. Thèse de doctorat, Université de Rennes, France. [465]

Couteron, P. & S. Ollier, S. 2005. A generalized, variogram-based framework for multi-scale ordination. *Ecology* **86**: 828-834. [894]

Cox, D. R. 1957. Note on grouping. *J. Amer. Statist. Assoc*. **52**: 543-547. [241]

Cramér, H. 1946. *Mathematical methods of statistics*. Princeton Univ. Press. xvi + 575 pp. [1]

Crawley, M. J. 2007. *The R book*. Wiley, Chichester. viii + 942 pp. [33]

Cressie, N. A. C. 1991. *Statistics for spatial data*. Wiley, New York. xx + 900 pp. [810, 832]

Cressie, N. A. C. & C. K. Wikle. 2011. *Statistics for spatio-temporal data*. Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, New Jersey. xxii + 588 pp. [901]

Crowder, M. J. & D. J. Hand. 1990. *Analysis of repeated measures*. Chapman and Hall, London. [18]

Cryer, J. D. 1986. *Time series analysis*. PWS-Kent Pub. Co., Boston. xi + 286 pp. [780]

Cullis, B. R. & A. C. Gleeson. 1991. Spatial analysis of field experiments. An extension to two dimensions. *Biometrics* **47**: 1449-1460. [20]

Czekanowski, J. 1909. Zur Differentialdiagnose der Neandertalgruppe. *Korrespondenz-Blatt deutsch. Ges. Anthropol. Ethnol. Urgesch*. **40**: 44-47. [xv, 285, 304, 403]

Czekanowski, J. 1913. *Zarys metod statystycznych w zastosowaniu do antropologii*. Travaux de la Société des Sciences de Varsovie. III. Classe des sciences mathématiques et naturelles, no. 5. iv + 228 pp. [xv, 276, 285]

Daget, J. 1976. *Les modèles mathématiques en écologie*. Collection d'Écologie, No. 8. Masson, Paris. viii + 172 pp. [270]

Daget, P. 1980. Le nombre de diversité de Hill, un concept unificateur dans la théorie de la diversité écologique. *Acta Oecol. Oecol. Gen*. **1**: 51-70. [253, 255]

Dagnelie, P. 1960. Contribution à l'étude des communautés végétales par l'analyse factorielle. *Bull. Serv. Carte phytogéogr. B* **5**: 7-71, 93-195. [xii, 270]

Dagnelie, P. 1965. L'étude des communautés végétales par l'analyse statistique des liaisons entre les espèces et les variables écologiques: principes fondamentaux. *Biometrics* **21**: 345-361. [xii]

Dagnelie, P. 1975. *L'analyse statistique à plusieurs variables*. Les Presses agronomiques de Gembloux, Gembloux (Belgique). 362 pp. [193]

D'Agostino, R. B. 1971. An omnibus test of normality for moderate and large sample sizes. *Biometrika* **58**: 341-348. [191]

D'Agostino, R. B. 1972. Small sample probability points for the *D* test of normality. *Biometrika* **59**: 219-221. [191]

D'Agostino, R. B. 1982. Departures from normality, tests for. 315-324 *in*: S. Kotz & N. L. Johnson [eds.] *Encyclopedia of statistical sciences. Vol. 2*. Wiley, New York. [187, 188, 191]

Dale, M. R. T. 1999. *Spatial pattern analysis in plant ecology*. Cambridge University Press, Cambridge. x + 326 pp. [790]

Dale, M. R. T. & M.-J. Fortin. 2010. From graphs to spatial graphs. *Annu. Rev. Ecol. Evol. Syst.* **41**: 21-38. [884, 886]

Dale, M. R. T. & M. Mah. 1998 The use of wavelets for spatial pattern analysis in ecology. *J. Veg. Sci.* **9**: 805-814. [767]

Damsleth, E. & E. Spjøtvoll. 1982. Estimation of trigonometric components in time series. *J. Amer. Statist. Assoc.* **77**: 381-387. [752]

d'Aubigny, G. 2006. Dépendance spatiale et autocorrélation. 17-45 *in*: J.-J. Droesbeke, M. Lejeune & G. Saporta [eds.] *Analyse statistique de données spatiales*. Éditions TECHNIP, Paris. [793]

David, M. 1977. *Geostatistical ore reserve estimation*. Developments in Geomathematics, 2. Elsevier Scient. Publ. Co., Amsterdam. xix + 364 pp. [21, 797, 810, 812, 831]

Davies, P. T. & M. K.-S. Tso. 1982. Procedures for reduced-rank regression. *Appl. Statist.* **31**: 244-255. [649]

Davis, A. W. 1978. On the asymptotic distribution of Gower's m^2 goodness-of-fit criterion in a particular case. *Ann. Inst. Statist. Math. Part A* **30**: 71-79. [704]

Davis, L. S. 1975. A survey of edge detection techniques. *Computer Graphics Image Process* **4**: 248-270. [844]

Day, W. H. E. 1977. Validity of clusters formed by graph-theoretic cluster methods. *Math. Biosci.* **36**: 299-317. [393, 411]

Day, W. H. E. 1983. Distribution of distances between pairs of classifications. 127-131 *in*: J. Felsenstein [ed.] *Numerical taxonomy*. NATO ASI Series, Vol. G-1. Springer-Verlag, Berlin. [529]

Day, W. H. E. 1986. Analysis of quartet dissimilarity measures between undirected phylogenetic trees. *Syst. Zool.* **35**: 325-333. [529]

De'ath, G. 2002. Multivariate regression trees: a new technique for modeling species-environment relationships. *Ecology* **83**: 1105-1117. [406, 410, 411, 660]

de Boor, C. 1978. *A practical guide to splines*. Springer-Verlag, Berlin. xxiv + 392 pp. [589]

De Cáceres, M, & P. Legendre. 2008. Beals smoothing revisited. *Oecologia* **156**: 657-669. [334].

De Cáceres, M. & P. Legendre. 2009. Associations between species and groups of sites: indices and statistical inference. *Ecology* **90**: 3566-3574. [399, 424]

De Cáceres, M., P. Legendre & M. Moretti. 2010. Improving indicator species analysis by combining groups of sites. *Oikos* **119**: 1674-1684. [399, 402]

Declerck, S. A. J., J. S. Coronel, P. Legendre & L. Brendonck. 2011. Scale dependency of processes structuring metacommunities of cladocerans in temporary pools of High-Andes wetlands. *Ecography* **34**: 296-305. [864, 877]

de Fraga, R., A. P. Lima & W. E. Magnusson. 2011. Mesoscale spatial ecology of tropical snake assemblage: the width of riparian corridors in central Amazonia. *Herpetol. J.* **21**: 51-57. [519]

de Gruijter, J. J., D. J. Brus, M. F. P. Bierkens & M. Knotters. 2006. *Sampling for natural resource monitoring*. Springer-Verlag, Berlin. xiv + 332 pp. [6]

de Gruijter, J. J. & C. J. F. ter Braak. 1990. Model-free estimation from spatial samples: a reappraisal of classical sampling theory. *Math. Geol.* **22**: 407-415. [6]

De Neufville, R. & J. H. Stafford. 1971. *Systems analysis for engineers and managers*. McGraw-Hill, New York. xiii + 353 pp. [182]

De Soete, G., J. D. Carroll & W. S. DeSarbo. 1987. Least squares algorithms for constructing constrained ultrametric and additive tree representations of symmetric proximity data. *J. Classif.* **4**: 155-173. [842, 843]

Deevey, E. S. 1969. Coaxing history to conduct experiments. *BioScience* **19**: 40-43. [249]

Demers, S., P. E. Lafleur, L. Legendre & C. L. Trump. 1979. Short-term covariability of chlorophyll and temperature in the St. Lawrence Estuary. *J. Fish. Res. Board Can.* **36**: 568-573. [757]

Demers, S. & L. Legendre. 1981. Mélange vertical et capacité photosynthétique du phytoplancton estuarien (estuaire du Saint-Laurent). *Mar. Biol (Berl.)* **64**: 243-250. [750]

Denman, K. L. 1976. Covariability of chlorophyll and temperature in the sea. *Deep-Sea Res.* **23**: 539-550. [757]

Denman, K. L. 1977. Short-term variability in vertical chlorophyll structure. *Limnol. Oceanogr.* **22**: 434-441. [757]

Denman, K. L., A. Okubo & T. Platt. 1977. The chlorophyll fluctuation spectrum in the sea. *Limnol. Oceanogr.* **22**: 1033-1038. [757]

Denman, K. L. & T. Platt. 1975. Coherences in the horizontal distributions of phytoplankton and temperature in the upper ocean. *Mém. Soc. R. Sci. Liège, Ser. 6.* **7**: 19-30. [757, 761, 762]

Denman, K. L. & T. Platt. 1976. The variance spectrum of phytoplankton in a turbulent ocean. *J. Mar. Res.* **34**: 593-601. [757]

Dessier, A. & A. Laurec. 1978. Le cycle annuel du zooplancton à Pointe-Noire (RP Congo). Description mathématique. *Oceanol. Acta* **1**: 285-304. [768]

Deutsch, C. V. & A. G. Journel. 1992. *GSLIB – Geostatistical software library and user's guide*. Oxford University Press, New York. [832, 852, 854, 857]

Dévaux, J. & G. Millerioux. 1976a. Possibilité de l'utilisation de la cotation d'abondance de Frontier (1969) pour l'analyse multivariable des populations phytoplanctoniques. *C. R. Hebd. Séances Acad. Sci., Sér. D Sci. Nat.* **283**: 41-44. [36, 451]

Dévaux, J. & G. Millerioux. 1976b. Méthode d'estimation de la biomasse totale du phytoplancton à partir des nombres de cellules, issus d'une cotation d'abondance. *C. R. Hebd. Séances Acad. Sci., Sér. D Sci. Nat.* **283**: 927-930. [36]

Dévaux, J. & G. Millerioux. 1977. Sur la possibilité d'un calcul de la diversité spécifique de populations phytoplanctoniques à partir de dénombrements issus d'une cotation d'abondance. *C. R. Hebd. Séances Acad. Sci., Sér. D Sci. Nat.* **284**: 1569-1571. [36, 199]

Dice, L. R. 1945. Measures of the amount of ecologic association between species. *Ecology* **26**: 297-302. [276, 317]

Dickman, M. 1968. Some indices of diversity. *Ecology* **49**: 1191-1193. [250]

Dietz, E. J. 1983. Permutation tests for association between two distance matrices. *Syst. Zool.* **32**: 21-26. [598, 600]

Digby, P. G. N. & R. A. Kempton. 1987. *Multivariate analysis of ecological communities*. Chapman & Hall, London. viii + 206. pp. [68]

Diggle, P. J. 1990. *Time series – A biostatistical introduction*. Oxford University Press, New York. xi + 257 pp. [714]

Dirichlet, G. L. 1850. Über die Reduktion der positiven quadratischen Formen mit drei unbestimmten ganzen Zahlen. *Journal für die reine und angewandte Mathematik* **40**: 209-234. [835, 839]

Dixon, W. J. [ed.] 1981. *BMDP statistical software 1981*. Univ. California Press, Berkeley. x + 725 pp. [230, 231, 240, 241]

Dolédec, S. & D. Chessel. 1994. Co-inertia analysis: an alternative method for studying species-environment relationships. *Freshwater Biol*. **31**: 277-294. [696]

Dolédec, S., D. Chessel, C. J. F. ter Braak & S. Champely. 1996. Matching species traits to environmental variables: a new three-table ordination method. *Environ. Ecol. Stat*. **3**: 143-166. [614, 617, 619]

Downing, J. A. 1979. Aggregation, transformation, and the design of benthos sampling programs. *J. Fish. Res. Board Can*. **36**: 1454-1463. [50]

Draper, N. & H. Smith. 1981. *Applied regression analysis. 2nd edition*. Wiley, New York. xiv + 709 pp. [538, 568, 826]

Dray, S. 2010. *Moran's eigenvectors of spatial weighting matrices in R*. Tutorial included in R package SPACEMAKER. http://R-Forge.R-project.org/projects/sedar/ [885, 887]

Dray, S., D. Chessel & J. Thioulouse. 2003. Co-inertia analysis and the linking of the ecological data tables. *Ecology* **84**: 3078-3089. [696, 702]

Dray, S. & P. Legendre. 2008. Testing the species traits-environment relationships: the fourth-corner problem revisited. *Ecology* **89**: 3400-3412. [613, 616–620]

Dray, S., P. Legendre & P. R. Peres-Neto. 2006. Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). *Ecol. Model*. **196**: 483-493. [863, 881, 883, 885, 886]

Dray, S., R. Pélissier, P. Couteron, M.-J. Fortin, P. Legendre, P. R. Peres-Neto, E. Bellier, R. Bivand, F. G. Blanchet, M. De Cáceres, A.-B. Dufour, E. Heegaard, T. Jombart, F. Munoz, J. Oksanen, J. Thioulouse & H. H. Wagner. 2012. Community ecology in the age of multivariate multiscale spatial analysis. *Ecol. Monogr.* **82**: (in press). [9, 11]

Dufrêne, M. & P. Legendre. 1991. Geographic structure and potential ecological factors in Belgium. *J. Biogeogr.* **18**: 257-266. [846]

Dufrêne, M. & P. Legendre. 1997. Species assemblages and indicator species: the need for a flexible asymmetrical approach. *Ecological Monographs* **67**: 345-366. [398–401]

Dungan, J. L., J. N. Perry, M. R. T. Dale, P. Legendre, S. Citron-Pousty, M.-J. Fortin, A. Jakomulska, M. Miriti & M. S. Rosenberg. 2002. A balanced view of scaling in spatial statistical analysis. *Ecography* **25**: 626-640. [786, 787]

Dutilleul, P. 1990. *Apport en analyse spectrale d'un périodogramme modifié et modélisation des séries chronologiques avec répétitions en vue de leur comparaison en fréquence*. Doctoral Dissertation, Univ. Cath. Louvain, Louvain-la-Neuve, Belgium. vi + 304 pp. [751, 752]

Dutilleul, P. 1993a. Modifying the *t* test for assessing the correlation between two spatial processes. *Biometrics* **49**: 305-314. [18, 20]

Dutilleul, P. 1993b. Spatial heterogeneity and the design of ecological field experiments. *Ecology* **74**: 1646-1658. [21, 790]

Dutilleul, P. 1998. Incorporating scale in ecological experiments: data analysis. 387-425 *in*: D. L. Peterson & V. T. Parker [eds.] *Ecological scale – Theory and applications*. Columbia University Press, New York. [752]

Dutilleul, P. R. L. 2011. *Spatio-temporal heterogeneity — Concepts and analyses*. Cambridge University Press, Cambridge. xxii + 393 pp. [22, 714, 734, 752, 753, 759, 788, 791, 797]

Dutilleul, P. & P. Legendre. 1992. Lack of robustness in two tests of normality against autocorrelation in sample data. *J. Statist. Comput. Simulation* **42**: 79-91. [18, 187, 191]

Dutilleul, P. & P. Legendre. 1993. Spatial heterogeneity against heteroscedasticity: an ecological paradigm versus a statistical concept. *Oikos* **66**: 152-171. [22, 788–790]

Dutilleul, P., J. D. Stockwell, D. Frigon & P. Legendre. 2000. The Mantel test versus Pearson's correlation analysis: assessment of the differences for biological and environmental studies. *J. Agr. Biol. Envir. S.* **5**: 131-150. [603]

Dutilleul, P. & C. Till. 1992. Evidence of periodicities related to climate and planetary behaviors in ring-width chronologies of Atlas cedar (*Cedrus atlantica*) in Morocco. *Can. J. For. Res.* **22**: 1469-1482. [752]

Eckart, C. & G. Young. 1936. The approximation of one matrix by another of lower rank. *Psychometrika* **1**: 211-218. [103]

Edgington, E. S. 1995. *Randomization tests. 3rd edition*. Marcel Dekker, Inc., New York. xxii + 409 pp. [25–27, 30, 31]

Edwards, A. L. 1948. Note on the "correction for continuity" in testing the significance of the difference between correlated proportions. *Psychometrika* **13**: 185-187. [848]

Edwards, A. W. F. & L. L. Cavalli-Sforza. 1965. A method for cluster analysis. *Biometrics* **21**: 362-375. [379]

Efron, B. 1975. The efficiency of logistic regression compared to normal discriminant analysis. *J. Amer. Statist. Assoc.* **70**: 892-898. [708]

Efron, B. 1979. Bootstrap methods: another look at the jackknife. *Ann. Stat.* **7**: 1-26. [31]

Efron, B. & R. J. Tibshirani. 1993. *An introduction to the bootstrap*. Chapman & Hall, New York. xvi + 436 pp. [31]

Eisner, F. 1931. Das Wiederstands Problem. *in*: *3ième Congrès international de Mécanique appliquée*, Stockholm. [118]

Enright, J. T. 1965. The search for rhythmicity in biological time-series. *J. Theor. Biol.* **8**: 426-468. [740-743]

Escofier, B. & J. Pagès. 1994. Multiple factor analysis (AFMULT package). *Comput. Stat. Data An.* **18**: 121-140. [703]

Escofier-Cordier, B. 1969. L'analyse factorielle des correspondances. *Cah. Bur. Univ. Rech. Opér. Univ. Paris* **13**: 25-59. [464, 465]

Escoufier, Y. 1973. Le traitement des variables vectorielles. *Biometrics* **29**: 751-760. [699]

Estabrook, G. F. 1966. A mathematical model in graph theory for biological classification. *J. Theor. Biol.* **12**: 297-310. [344, 411, 412]

Estabrook, G. F. & B. Gates. 1984. Character analysis of the *Banisteriopsis campestris* complex (Malpighiaceae), using spatial autocorrelation. *Taxon* **33**: 13-25. [800]

Estabrook, G. F. & D. J. Rogers. 1966. A general method of taxonomic description for a computed similarity measure. *BioScience* **16**: 789-793. [280-282]

Eubank, R. L. 1988. *Spline smoothing and nonparametric regression*. Marcel Dekker, New York. xvii + 438 pp. [589]

Everitt, B. S. 1977. *The analysis of contingency tables*. Chapman and Hall, London. 128 pp. [245]

Everitt, B. S. 1980. *Cluster Analysis. 2nd edition*. Halsted Press, New York. 136 pp. [366, 386]

Ezekiel, M. 1930. Methods of correlation analysis. John Wiley and Sons, New York. 427 pp. [565, 566, 633]

Fager, E. W. 1957. Determination and analysis of recurrent groups. *Ecology* **38**: 586-595. [316, 318, 392, 393]

Fager, E. W. & J. A. McGowan. 1963. Zooplankton species groups in the North Pacific. *Science (Wash. D. C.)* **140**: 453-460. [318, 392, 393]

Faith, D. P. 1983. Asymmetric binary similarity measures. *Oecologia (Berl.)* **57**: 287-290. [277, 297]

Faith, D. P., P. R. Minchin & L. Belbin. 1987. Compositional dissimilarity as a robust measure of ecological distance. *Vegetatio* **69**: 57-68. [332, 512, 516]

Falconer, D. S. 1960. *Introduction to quantitative genetics*. Ronald Press Inc., New York. ix + 365 pp. [34]

Fasham, M. J. R. 1977. A comparison of nonmetric multidimensional scaling, principal component and reciprocal averaging for the ordination of simulated coenoclines and coenoplanes. *Ecology* **58**: 551-561. [519]

Fausett, L. 1994. *Fundamentals of neural networks – Architectures, algorithms and applications*. Prentice Hall, Englewood Cliffs, New Jersey. xvi + 461 pp. [423]

Fedriani, J. M., T. K. Fuller & R. M. Sauvajot. 2001. Does availability of anthropogenic food enhance densities of omnivorous mammals? An example with coyotes in southern California. *Ecography* **24**: 325-331. [243]

Ferligoj, A. & V. Batagelj. 1982. Clustering with relational constraint. *Psychometrika* **47:** 413-426. [842, 843]

Ferligoj, A. & V. Batagelj. 1983. Some types of clustering with relational constraints. *Psychometrika* **48**: 541-552. [843]

Ferriere, R., B. Cazelles, F. Cezilly & J.-P. Desportes. 1996. Predictability and chaos in bird vigilant behaviour. *Anim. Behav.* **52**: 457-472. [2]

Field, J. G. 1969. The use of the information statistic in the numerical classification of heterogeneous systems. *J. Ecol.* **57**: 565-569. [375]

Field, J. G., K. R. Clarke & R. M. Warwick. 1982. A practical strategy for analysing multispecies distribution patterns. *Mar. Ecol. Prog. Ser.* **8**: 37-52. [523]

Field, J. G. & F. T. Robb. 1970. Numerical methods in marine ecology. 2. Gradient analysis of rocky shore samples from False Bay. *Zool. Afr.* **5**: 191-210. [509, 510]

Fienberg, S. E. 1970. The analysis of multidimensional contingency tables. *Ecology* **5l**: 419-433. [243]

Fienberg, S. E. 1980. *The analysis of cross-classified categorical data. 2nd edition*. MIT Press, Cambridge, Mass. xiv + 198 pp. [230, 235, 240–242, 536]

Fingerut, J. T., C. A. Zimmer & R. K. Zimmer. 2003. Larval swimming overpowers turbulent mixing and facilitates transmission of a marine parasite. *Ecology* **84**: 2502-2515. [243]

Finn, J. D. 1974. *A general model for multivariate analysis*. Holt, Rinehart and Winston, New York. xiii + 423 pp. [556]

Fisher, R. A. 1935. *The design of experiments*. Oliver and Boyd, Edinburgh. 252 pp. [25]

Fisher, R. A. 1936. The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7:** 179-188. [674]

Fisher, R. A. 1940. The precision of discriminant functions. *Annals of Eugenics* **10**: 422-429. [464]

Fisher, R. A. 1954. *Statistical methods for research workers. 12th edition*. Oliver & Boyd, Edinburgh. 356 pp. [23, 27]

Fisher, R. A., A. S. Corbet & C. B. Williams. 1943. The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Anim. Ecol.* **12**: 42-58. [258]

Fisher, W. D. 1958. On grouping for maximum homogeneity. *J. Amer. Statist. Assoc.* **53**: 789-798. [383, 769]

Ford, E. D. 1976. The canopy of a Scots pine forest: description of a surface of complex roughness. *Agric. Meteorol.* **17**: 9-32. [807]

Fortier, L. & L. Legendre. 1979. Le contrôle de la variabilité à court terme du phytoplancton estuarien: stabilité verticale et profondeur critique. *J. Fish. Res. Board Can.* **36**: 1325-1335. [737]

Fortier, L., L. Legendre, A. Cardinal & C. L. Trump. 1978. Variabilité à court terme du phytoplancton de l'estuaire du Saint-Laurent. *Mar. Biol. (Berl.)* **46**: 349-354. [723]

Fortin, M.-J. 1994. Edge detection algorithms for two-dimensional ecological data. *Ecology* **75**: 956-965. [844-846]

Fortin, M.-J. 1997. Effects of data types on vegetation boundary delineation. *Can. J. For. Res.* **27**: 1851-1858. [844, 846]

Fortin, M.-J. & M. R. T. Dale. 2005. *Spatial analysis – A guide for ecologists*. Cambridge University Press, Cambridge. xiii + 365 pp. [11, 12, 14, 767, 790, 806, 821, 844]

Fortin, M.-J. & P. Drapeau. 1995. Delineation of ecological boundaries: comparison of approaches and significance tests. *Oikos* **72**: 323-332. [844, 846]

Fortin, M.-J., P. Drapeau & G. M. Jacquez. 1996. Quantification of the spatial co-occurrence of ecological boundaries. *Oikos* **77**: 51-60. [844, 846]

Fortin, M.-J., P. Drapeau & P. Legendre. 1989. Spatial autocorrelation and sampling design in plant ecology. *Vegetatio* **83**: 209-222. [818]

Fourier, J. 1822. Théorie analytique de la chaleur. *in*: G. Darboux [ed.] *1888, Oeuvres de Fourier. Tome 1*. Gauthier-Villars, Paris. [109]

François-Bongarçon, D. 1991. Geostatistical determination of sample variances in the sampling of broken gold ores. *CIM Bulletin* **84** (950): 46-57. [21]

Fréchet, A. 1990. Catchability variations of cod in the marginal ice zone. *Can. J. Fish. Aquat. Sci.* **47**: 1678-1683. [243]

Fréchette, M., & L. Legendre. 1982. Phytoplankton photosynthetic response to light in an internal tide dominated environment. *Estuaries* **5**: 287-293. [737, 738]

Freedman, D. & D. Lane. 1983. A nonstochastic interpretation of reported significance levels. *J. Bus. Econ. Statist.* **1**: 292-298. [651]

Freund, R. J. & P. D. Minton. 1979. *Regression methods – A tool for data analysis*. Statistics: Textbooks and Monographs, Vol. 30. Marcel Dekker Inc., New York. xi + 261 pp. [562]

Friedman, M. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Amer. Statist. Assoc.* **32**: 675-701. [216]

Frontier, S. 1969. Sur une méthode d'analyse faunistique rapide du zooplancton. *J. Exp. Mar. Biol. Ecol.* **3**: 18-26. [36]

Frontier, S. 1973. Evaluation de la quantité totale d'une catégorie d'organismes planctoniques dans un secteur néritique. *J. Exp. Mar. Biol. Ecol.* **12**: 299-304. [36]

Frontier, S. 1976. Étude de la décroissance des valeurs propres dans une analyse en composantes principales: comparaison avec le modèle du bâton brisé. *J. Exp. Mar. Biol. Ecol.* **25**: 67-75. [449]

Frontier, S. & F. Ibanez. 1974. Utilisation d'une cotation d'abondance fondée sur la progression géométrique, pour l'analyse en composantes principales en écologie planctonique. *J. Exp. Mar. Biol. Ecol.* **14**: 217-224. [36, 451]

Frontier, S. & D. Viale. 1977. Utilisation d'une cotation d'abondance mise au point en planctologie pour l'évaluation des troupeaux de cétacés en mer. *J. Rech. Océanogr.* **2**: 15-22. [36]

Fry, J. C., N. C. B. Humphrey & T. C. Iles. 1981. Time-series analysis for identifying cyclic components in microbiological data. *J. Appl. Bacteriol.* **50**: 189-224. [714, 754]

Furnas, G. W. 1984. The generation of random, binary unordered trees. *J. Classif.* **1**: 187-233. [30]

Gabbott, P. A. & V. N. Larman. 1971. Electrophoretic examination of partially purified extracts of *Balanus balanoides* containing a settlement inducing factor. 143-153 *in*: D. J. Crisp [ed.] *Fourth european marine biology symposium*. Cambridge Univ. Press, Cambridge. [805]

Gabriel, K. R. 1971. The biplot graphic display of matrices with application to principal component analysis. *Biometrika* **58**: 453-467. [443]

Gabriel, K. R. 1982. Biplot. 263-271 *in*: S. Kotz & N. L. Johnson [eds.] *Encyclopedia of statistical sciences*. Vol. 1. Wiley, New York. [443]

Gabriel, K. R. & R. R. Sokal. 1969. A new statistical approach to geographic variation analysis. *Syst. Zool.* **18**: 259-278. [836]

Galiano, E. F. 1982. Pattern detection in plant populations through the analysis of plant-to-all-plants distances. *Vegetatio* **49**: 39-43. [789]

Galton, F. 1889. *Natural inheritance*. Macmillan & Co., London. ix + 259 pp. [539]

Galzin, R. & P. Legendre. 1987. The fish communities of a coral reef transect. *Pac. Sci.* **41**: 158-165. [774]

Garland, T. Jr. 1983. The relation between maximal running speed and body mass in terrestrial mammals. *J. Zool. (Lond.)* **199**: 157-170. [122]

Garrison, W. L. & D. F. Marble. 1964. Factor analytic study of the connectivity of a transportation matrix. *Papers and Proceedings, Regional Science Association* **12**: 231-238. [859]

Gauch, H. G. Jr. 1982. *Multivariate analysis in community ecology*. Cambridge University Press, Cambridge. x + 298 pp. [483, 486]

Gauch, H. G. Jr., R. H. Whittaker & T. R. Wentworth. 1977. A comparative study of reciprocal averaging and other ordination techniques. *J. Ecol.* **65**: 157-174. [479]

Gause, G. F. 1935. *Vérification expérimentale de la théorie mathématique de la lutte pour la vie*. Actual. Sci. Ind. no 277. Hermann Éditeur, Paris. [478]

Gauss, K. F. 1809. *Theoria motus corporum coelestium in sectionibus conicis solem ambientium*. Frid. Perthes et I. H. Besser, Hamburg. [541]

Geary, R. C. 1954. The contiguity ratio and statistical mapping. *Incorp. Statist.* **5**: 115-145. [793]

Geffen, E., M. J. Anderson & R. K. Wayne,. 2004. Climate and habitat barriers to dispersal in the highly mobile gray wolf. *Mol. Ecol.* **13**: 2481-2490. [649]

Gentle, J. E. 2007. *Matrix algebra: theory, computations, and applications in statistics*. Springer Texts in Statistics, Springer Science, New York. xxii + 528 pp. [59]

Getis, A. & B. Boots. 1978. *Models of spatial processes – An approach to the study of point, line and area patterns*. Cambridge Univ. Press, Cambridge. [789, 790]

Gifi, A. 1990. *Nonlinear multivariate analysis*. John Wiley & Sons, Chichester. xx + 579 pp. [39]

Gilbert, R. O. & J. C. Simpson. 1985. Kriging for estimating spatial pattern of contaminants: potential and problems. *Environ. Monit. Assess.* **5**: 113-135. [832]

Gini, C. 1912. *Variabilità e mutabilità: contributo allo studio delle distribuzioni e delle relazioni statistiche*. Studi Economico-Giuridici, Facoltà di Giurisprudenza della Regia Università di Cagliari, anno III, parte II, Cuppini, Bologna, 156 pp. *Reprinted in*: E. Pizetti & T. Salvemini [eds.] 1955. *Memorie di metodologica statistica*. Libreria Eredi Virgilio Veschi, Roma. [xv, 254]

Gittins, R. 1985. *Canonical analysis – A review with applications in ecology*. Springer-Verlag, Berlin. 351 pp. [630, 681, 691]

Glansdorff, P. & I. Prigogine. 1971. *Structure, stabilité et fluctuations*. Masson, Paris. 288 pp. [37]

Gokhale, D. V. & S. Kullback. 1978. *The information in contingency tables*. Marcel Dekker Inc., New York. x + 365 pp. [235]

Goldstein, M. & W. R. Dillon. 1978. *Discrete discriminant analysis*. John Wiley & Sons, New York. x + 186 pp. [532]

Goodall, D. W. 1954. Objective methods for the classification of vegetation. III. An essay in the use of factor analysis. *Aust. J. Bot.* **2**: 304-324. [xii, 425, 483]

Goodall, D. W. 1964. A probabilistic similarity index. *Nature (Lond.)* **203**: 1098. [288]

Goodall, D. W. 1966a. A new similarity index based on probability. *Biometrics* **22**: 882-907. [288, 289]

Goodall, D. W. 1966b. Deviant index: a new tool for numerical taxonomy. *Nature (Lond.)* **210**: 216. [380]

Goodall, D. W. 1974. A new method for the analysis of spatial pattern by random pairing of quadrats. *Vegetatio* **29**: 135-146. [800]

Goodman, L. A. & W. H. Kruskal. 1954. Measures of association for cross classifications. *J. Amer. Statist. Assoc.* **49**: 732-764. [270]

Goodman, L. A. & W. H. Kruskal. 1959. Measures of association for cross classifications. II. Further discussion and references. *J. Amer. Statist. Assoc.* **54**: 123-163. [270]

Goodman, L. A. & W. H. Kruskal. 1963. Measures of association for cross classifications. III. Approximate sampling theory. *J. Amer. Statist. Assoc.* **58**: 310-364. [270]

Gordo, O., J. J. Sanz & J. M. Lobo. 2007. Environmental and geographical constraints on common swift and barn swallow spring arrival patterns throughout the Iberian Peninsula. *J. Biogeogr.* **34**: 1065-1076. [855]

Gordon, A. D. 1973. Classification in the presence of constraints. *Biometrics* **29**: 821-827. [773]

Gordon, A. D. 1994. Identifying genuine clusters in a classification. *Comput. Statist. Data Anal.* **18**: 561-581. [415, 416, 611]

Gordon, A. D. 1996a. Hierarchical classification. 65-121 *in*: P. Arabie, L. J. Hubert & G. De Soete [eds.] *Clustering and Classification*. World Scientific Publ. Co., River Edge, New Jersey. [369, 415, 416]

Gordon, A. D. 1996b. Null models in cluster validation. 32-44 *in*: W. Gaul & D. Pfeifer [eds.] *From data to knowledge*. Springer-Verlag, Berlin. [415, 416]

Gordon, A. D. 1996c. A survey of constrained classification. *Comput. Statist. Data Anal.* **21**: 17-29. [835, 842, 843]

Gordon, A. D. 1999. *Classification, 2nd edition*. Monographs on Statistics and Applied Probability, 82. Chapman and Hall/CRC, London. x + 256 pp. [841, 843]

Gordon, A. D. & H. J. B. Birks. 1972. Numerical methods in Quaternary palaeoecology. I. Zonation of pollen diagrams. *New Phytol.* **71**: 961-979. [773]

Gordon, A. D. & H. J. B. Birks. 1974. Numerical methods in Quaternary palaeoecology. II. Comparison of pollen diagrams. *New Phytol.* **73**: 221-249. [773]

Gorelick R. & S. M. Bertram. 2010. Multi-way multi-group segregation and diversity indices. *PLoS ONE* **5**: e10912. doi:10.1371/journal.pone.0010912 [243]

Gosselin, M., L. Legendre, S. Demers, J.-C. Therriault & M. Rochet. 1986. Physical control of the horizontal patchiness of sea-ice microalgae. *Mar. Ecol. Prog. Ser.* **29**: 289-295. [596]

Gould, P. R. 1967. On the geographical interpretation of eigenvalues. *T. I. Brit. Geogr.* **42**: 53-92. [859]

Gower, J. C. 1966. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*. **53**: 325-338. [xii, 381, 446, 451, 492, 493, 497, 498, 517]

Gower, J. C. 1967. A comparison of some methods of cluster analysis. *Biometrics* **23**: 623-637. [348, 357, 360, 378, 380]

Gower, J. C. 1971a. A general coefficient of similarity and some of its properties. *Biometrics* **27**: 857-871. [278, 286]

Gower, J. C. 1971b. Statistical methods of comparing different multivariate analyses of the same data. 138-149 *in*: F. R. Hodson, D. G. Kendall & P. Tautu [eds.] *Mathematics in the archaeological and historical sciences*. Edinburgh University Press, Edinburgh. [611, 703, 704]

Gower, J. C. 1975. Generalized Procrustes analysis. *Psychometrika* **40**: 33-51. [611, 704]

Gower, J. C. 1982. Euclidean distance geometry. *Math. Scientist* **7**: 1-14. [492, 499, 500]

Gower, J. C. 1983. Comparing classifications. 137-155 *in*: J. Felsenstein [ed.] *Numerical taxonomy*. NATO ASI Series, Vol. G-1. Springer-Verlag, Berlin. 644 pp. [413]

Gower, J. C. 1984. Ordination, multidimensional scaling and allied topics. 727-781 *in*: W. Lederman [ed.] *Handbook of Applicable Mathematics*. Vol. VI: E. Lloyd [ed.] *Statistics*. Wiley, Chichester. [425, 428]

Gower, J. C. 1985. Measures of similarity, dissimilarity, and distance. 397-405 *in*: S. Kotz & N. L. Johnson [eds.] *Encyclopedia of statistical sciences. Vol. 5*. Wiley, New York. [270, 500]

Gower, J. C. 1987. Introduction to ordination techniques. 3-64 *in*: P. Legendre & L. Legendre [eds.] *Developments in numerical ecology*. NATO ASI Series, Vol. G14. Springer-Verlag, Berlin. [428, 512, 611]

Gower, J. C. 1990. Three dimensional biplots. *Biometrika* **77**: 773-785. [443]

Gower, J. C. & P. Legendre. 1986. Metric and Euclidean properties of dissimilarity coefficients. *J. Classif.* **3**: 5-48. [270, 276, 296, 323, 501]

Gower, J. C., S. Lubbe & N. le Roux. 2011. *Understanding biplots*. Wiley, Chichester. xi + 463 pp. [443]

Gower, J. C. & G. J. S. Ross. 1969. Minimum spanning trees and single linkage cluster analysis. *Appl. Statist.* **18**: 54-64. [345, 523]

Grace, J. B. 2006. *Structural equation modeling and natural systems*. Cambridge University Press, Cambridge. xii + 365 pp. [593]

Gray, D. K. & S. E. Arnott. 2011. Does dispersal limitation impact the recovery of zooplankton communities damaged by a regional stressor? *Ecol. Appl.* **21**: 1241-1256. [893]

Graybill, F. A. 2001. *Matrices with applications in statistics. 2nd edition*. Duxbury Press, Pacific Grove, California. xi + 461. [59]

Green, P. G. & J. D. Carroll. 1976. *Mathematical tools for applied multivariate analysis*. Academic Press, New York. 376 pp. [60]

Green, R. H. 1979. *Sampling design and statistical methods for environmental biologists*. John Wiley & Sons, New York. xi + 257 pp. [21, 241, 267]

Greenacre, M. J. 1983. *Theory and applications of correspondence analysis*. Academic Press, London. xi + 364 pp. [428, 464]

Greenacre, M. J. 2007. *Correspondence analysis in practice. 2nd edition*. Chapman & Hall / CRC Press, Boca Raton, Florida. xiii + 280 pp. [464]

Greenacre, M. J. 2010. *Biplots in practice*. Fundación BBVA, Bilbao. 237 pp. [443]

Greenberg, J. H. 1956. The measurement of linguistic diversity. *Language* **32**: 109-115. [254]

Greig-Smith, P. 1983. *Quantitative plant ecology. 3rd edition*. University of California Press, Berkeley. xiv + 359 pp. [319, 390]

Griffith, D. A. 1978. A spatially adjusted ANOVA model. *Geogr. Anal.* **10**: 296-301. [20]

Griffith, D. A. 1987. *Spatial autocorrelation – A primer.* Association of American Geographers, Washington, D. C. 10 + 82 pp. [20, 790]

Griffith, D. A. 1988. *Advanced Spatial Statistics*. Kluwer, Dordrecht. xiv + 273 pp. [15, 18]

Griffith, D. A. 1996. Spatial autocorrelation and eigenfunctions of the geographic weights matrix accompanying geo-referenced data. *Can. Geogr.* **40**: 351-367. [859, 884, 885]

Griffith, D. A. & P. R. Peres-Neto. 2006. Spatial modelling in ecology: the flexibility of eigenfunction spatial analyses. *Ecology* **87**: 2603-2613. [859, 884]

Grondona, M. O. & N. Cressie. 1991. Using spatial considerations in the analysis of experiments. *Technometrics* **33**: 381-392. [20]

Guedj, D. 1999. *La méridienne*. Robert Laffont, Paris. 302 pp. [142]

Guedj, D. (translated by A. Goldhammer). 2001. *The measure of the world: a novel*. University of Chicago Press, Chicago. 310 pp. [142]

Guénard, G., P. Legendre, D. Boisclair & M. Bilodeau. 2010. Multiscale codependence analysis: an integrated approach to analyze relationships across scales. *Ecology* **91**: 2952-2964. [865, 901]

Guille, A. 1970. Bionomie benthique du plateau continental de la côte catalane française. II. Les communautés de la macrofaune. *Vie Milieu* **21**: 149-280. [406]

Günther, B. 1975. Dimensional analysis and theory of biological similarity. *Physiol. Rev.* **55**: 659-699. [122, 141]

Haberman, S. J. 1973. The analysis of residuals in cross-classified tables. *Biometrics* **29**: 205-220. [244]

Hahsler, M., K. Hornik & C. Buchta. 2008. Getting things in order: an introduction to the R package seriation. *J. Stat. Softw.* **25**: 1-34. [403]

Haining, R. 1987. Trend-surface models with regional and local scales of variation with an application to aerial survey data. *Technometrics* **29**: 461-469. [828]

Haining, R. 1990. *Spatial data analysis in the social and environmental sciences*. Cambridge University Press, Cambridge. xxi + 409 pp. [18]

Haining, R. 2003. Spatial data analysis – Theory and practice. Cambridge University Press, Cambridge. xx + 432 pp. [18]

Hair, J. F., W. C. Black, B. J. Babin & R. E. Anderson. 2010. *Multivariate data analysis, 7th edition*. Prentice Hall, New Jersey. xxviii + 785 pp. [143]

Hajdu, L. J. 1981. Geographical comparison of resemblance measures in phytosociology. *Vegetatio* **48**: 47-59. [322]

Hájek, J. 1969. *A course in nonparametric statistics*. Holden-Day, San Francisco. viii + 184 pp. [201]

Hall, A. V. 1965. The peculiarity index, a new function for use in numerical taxonomy. *Nature (Lond.)* **206**: 952. [380]

Hall, P., N. I. Fisher & B. Hoffmann. 1994. On the nonparametric estimation of covariance functions. *Ann. Stat.* **22**: 2115-2134. [805]

Hall, P. & P. Patil. 1994. Properties of nonparametric estimators of autocovariance for stationary random fields. *Probability Theory and Related Fields* **99**: 399-424. [805]

Hann, B. J., P. R. Leavitt & P. S. S. Chang. 1994. Cladocera community response to experimental eutrophication in Lake 227 as recorded in laminated sediments. *Can. J. Fish. Aquat. Sci.* **51**: 2312-2321. [773]

Harrington, D. 2009. *Confirmatory factor analysis*. Oxford University Press, New York. ix + 122 pp. [535]

Harris, R. E. & W. A. G. Charleston. 1977. An examination of the marsh microhabitats of *Lymnaea tomentosa* and *L. columella* (Mollusca: Gastropoda) by path analysis. *N. Z. J. Zool.* **4**: 395-399. [596, 597]

Hartigan, J. A. 1975. *Cluster algorithms*. John Wiley & Sons, New York. xiii + 349 pp. [386]

Harvey, A. C. 1981. *The econometric analysis of time series*. Wiley, New York. xi + 384 pp. [20]

Harvey, A. C. & G. D. A. Phillips. 1979. Maximum likelihood estimation of regression models with autoregressive-moving average disturbances. *Biometrika* **66**: 49-58. [20]

Hatcher, A. & C. A. Frith. 1985. The control of nitrate and ammonium concentrations in a coral reef lagoon. *Coral Reefs* **4**: 101-110. [124]

Hatcher, B. G., J. Imberger & S. V. Smith. 1987. Scaling analysis of coral reef systems: an approach to problems of scale. *Coral Reefs* **5**: 171-181. [124]

Hatheway, W. H. 1971. Contingency-table analysis of rain forest vegetation. 271-313 *in*: G. P. Patil, E. C. Pielou & W. E. Waters [eds.] *Statistical ecology*. Vol. 3: *Many species populations, ecosystems, and systems analysis*. Pennsylvania State University Press, University Park and London. [464]

Hawkins, D. M. & D. F. Merriam. 1973. Optimal zonation of digitized sequential data. *Math. Geol.* **5**: 389-395. [769]

Hawkins, D. M. & D. F. Merriam. 1974. Zonation of multivariate sequences of digitized geologic data. *Math. Geol.* **6**: 263-269. [769]

Hawksworth, F. G., G. F. Estabrook & D. J. Rogers. 1968. Application of an information theory model for character analysis in the genus *Arceuthobium* (Viscaceae). *Taxon* **17**: 605-619. [234]

He, F. & P. Legendre. 1996. On species-area relations. *Am. Nat.* **148**: 719-737. [327]

He, F. & P. Legendre. 2002. Species diversity patterns derived from species-area models. *Ecology* **83**:1185-1198. [327]

He, F., P. Legendre, C. Bellehumeur & J. V. LaFrankie. 1994. Diversity pattern and spatial scale: a study of a tropical rain forest of Malaysia. *Environ. Ecol. Stat.* **1**: 265-286. [586, 786, 788]

He, F., P. Legendre & J. V. LaFrankie. 1996. Spatial pattern of diversity in a tropical rain forest of Malaysia. *J. Biogeogr.* **23**: 57-74. [251, 327, 586]

He, F., P. Legendre & J. V. LaFrankie. 1997. Distribution patterns of tree species in a Malaysian tropical rain forest. *J. Veg. Sci.* **8**: 105-114. [586]

Healy, M. J. R. 1984. The use of R^2 as a measure of goodness of fit. *J. R. Statist. Soc. A* 147: 608-609. [566]

Hecky, R. E. & P. Kilham. 1988. Nutrient limitation of phytoplankton in freshwater and marine environments. A review of recent evidence on the effects of enrichment. *Limnol. Oceanogr.* **33**: 796-822. [8]

Heikkinen, R. K. & H. J. B. Birks. 1996. Spatial and environmental components of variation in the distribution patterns of subarctic plant species at Kevo, N. Finland. A case study at the meso-scale level. *Ecography* **19**: 341-351. [855]

Helmus, M.,R., T. J. Bland, C. K. Williams & A. R. Ives. 2007 Phylogenetic measures of biodiversity. *Am. Nat.* **169**: E68-E83. [255]

Heo, M. & K. R. Gabriel. 1998. A permutation test of association between configurations by means of the RV coefficient. *Commun. Stat. Simul. Comput.* **27**: 843-856. [700]

Hermy, M. 1987. Path analysis of standing crop and environmental variables in the field layer of two Belgian riverine forests. *Vegetatio* **70**: 127-134. [596]

Hewitt, J. E., P. Legendre, B. H. McArdle, S. F. Thrush, C. Bellehumeur & S. M. Lawrie. 1997. Identifying relationships between adult and juvenile bivalves at different spatial scales. *J. Exp. Mar. Biol. Ecol.* **216**: 77-98. [745, 818]

Hill, A. V. 1950. The dimensions of animals and their muscular dynamics. *Sci. Prog.* **38**: 209-230. [119, 121]

Hill, M. O. 1973a. Diversity and evenness: a unifying notation and its consequences. *Ecology* **54**: 427-432. [250, 251, 254, 255]

Hill, M. O. 1973b. Reciprocal averaging: an eigenvector method of ordination. *J. Ecol.* **61**: 237-249. [456, 464, 478, 479, 491]

Hill, M. O. 1974. Correspondence analysis: a neglected multivariate method. *Appl. Statist.* **23**: 340-354. [428, 464, 469, 478]

Hill, M. O. 1979a. *TWINSPAN – A FORTRAN program for arranging multivariate data in an ordered two-way table by classification of the individuals and attributes*. Section of Ecology and Systematics, Cornell University, Ithaca, New York. iv + 49 pp. [381, 382, 397]

Hill, M. O. 1979b. *DECORANA – A FORTRAN program for detrended correspondence analysis and reciprocal averaging*. Section of Ecology and Systematics, Cornell University, Ithaca, New York. 52 pp. [485]

Hill, M. O. & H. G. Gauch Jr. 1980. Detrended correspondence analysis, an improved ordination technique. *Vegetatio* **42**: 47-58. [483, 485, 486]

Hirschfeld, H. O. 1935. A connection between correlation and contingency. *Proc. Camb. Phil. Soc.* **31**: 520-524. [464]

Hobbs, R. J. & H. A. Mooney [eds.] 1990. *Remote sensing of biosphere functioning*. Springer-Verlag, New York. 350 pp. [844]

Hochberg, Y. 1988. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **75**: 800-802. [23]

Hocking, R. R. 1976. The analysis and selection of variables in linear regression. *Biometrics* **32**: 1-49. [561]

Hoerl, A. E. 1962. Application of ridge analysis to regression problems. *Chem. Eng. Prog.* **58**: 54-59. [563, 564]

Hoerl, A. E. & R. W. Kennard. 1970a. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*. **12**: 55-67. [563]

Hoerl, A. E. & R. W. Kennard. 1970b. Ridge regression: applications to nonorthogonal problems. *Technometrics* **12**: 69-82. [563]

Hollander, M. & D. A. Wolfe. 1973. *Nonparametric statistical methods*. Wiley, New York. [583]

Holm, S. 1979. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6**: 65-70. [23]

Hope, A. C. A. 1968. A simplified Monte Carlo significance test procedure. *J. Roy. Statist. Soc. Ser. B* **30**: 582-598. [27, 294]

Hotelling, H. 1931. The generalization of Student's ratio. *Ann. Math. Statist.* **2**: 360-378. [304]

Hotelling, H. 1933. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **24**: 417-441, 498-520. [xv, 430]

Hotelling, H. 1936. Relations between two sets of variates. *Biometrika* **28**: 321-377. [690]

Hotelling, H. & M. R. Pabst. 1936. Rank correlation and tests of significance involving no assumption of normality. *Ann. Math. Statist.* **7**: 29-43. [212]

Huang, J. S. & D. H. Tseng. 1988. Statistical theory of edge detection. *Comp. Vis. Graph. Image Proc.* **43**: 337-346. [844]

Hubálek, Z. 1982. Coefficients of association and similarity, based on binary (presence-absence) data: an evaluation. *Biol. Rev.* **57**: 669-689. [323]

Hubbell, S. P. 2001. *The unified neutral theory of biodiversity and biogeography.* Princeton University Press, Princeton, New Jersey. xiv + 375 pp. [12, 598, 607]

Hubert, L. J. & P. Arabie. 1985. Comparing partitions. *J. Classif.* **2**: 193-218. [413]

Hubert, L. J. & F. B. Baker. 1977. The comparison and fitting of given classification schemes. *J. Math. Psychol.* **16**: 233-253. [417]

Hubert, L. J. & R. G. Golledge. 1981. A heuristic method for the comparison of related structures. *J. Math. Psychol.* **23**: 214-226. [606]

Hudon, C., E. Bourget & P. Legendre. 1983. An integrated study of the factors influencing the choices of the settling site of *Balanus crenatus* cyprid larvae. *Can J. Fish. Aquat. Sci.* **40**: 1186-1194. [805]

Hudon, C. & Lamarche, G. 1989. Niche segregation between American lobster *Homarus americanus* and rock crab *Cancer irroratus*. *Mar. Ecol. Prog. Ser.* **52**: 155-168. [601]

Huet, S., E. Jolivet & A. Messéan. 1992. *La régression non-linéaire – Méthodes et applications en biologie*. Institut National de la Recherche Agronomique, Paris. 247 pp. [583]

Huntley, H. E. 1967. *Dimensional analysis*. Dover, New York. 153 pp. [109]

Hurlbert, S. H. 1971. The nonconcept of species diversity: a critique and alternative parameters. *Ecology* **52**: 577-586. [249–251, 254, 256]

Hurlbert, S. H. 1984. Pseudoreplication and the design of ecological field experiments. *Ecol. Monogr.* **54**:187-211. [8, 143, 535, 785]

Hurley, J. R. & R. B. Cattell. 1962. The Procrustes program: producing direct rotation to test a hypothesized factor structure. *Behav. Sci.* **7**: 258-262. [703]

Hurvich, C. M. & C.-L. Tsai. 1993. A corrected Akaike information criterion for vector autoregressive model selection. *J. Time Ser. Anal.* 14: 271-279. [566]

Hutchinson, G. E. 1957. Concluding remarks. *Cold Spring Harbor Symp. Quant. Biol.* **22**: 415-427. [4, 12, 258, 271, 478, 785]

Hutchinson, G. E. 1965. *The ecological theater and the evolutionary play.* Yale Univ. Press, New Haven. xiii + 139 pp. [4, 258]

Ibanez, F. 1971. Effet des transformations des données dans l'analyse factorielle en écologie planctonique. *Cah. Océanogr.* **23**: 545-561. [450]

Ibanez, F. 1972. Interprétation de données écologiques par l'analyse des composantes principales: écologie planctonique de la mer du Nord. *J. Cons. Cons. Int. Explor. Mer* **34**: 323-340. [452]

Ibanez, F. 1973. Méthode d'analyse spatio-temporelle du processus d'échantillonnage en planctologie, son influence dans l'interprétation des données par l'analyse en composantes principales. *Ann. Inst. Océanogr. (Paris)* **49**: 83-111. [448]

Ibanez, F. 1981. Immediate detection of heterogeneities in continuous multivariate, oceanographic recordings. Application to time series analysis of changes in the bay of Villefranche sur Mer. *Limnol. Oceanogr.* **26**: 336-349. [772]

Ibanez, F. 1982. L'échantillonnage en continu en océanographie. 365-384 *in*: S. Frontier [ed.] *Stratégies d'échantillonnage en écologie*. Collection d'Écologie, No. 17. Masson, Paris et les Presses de l'Université Laval, Québec. [781]

Ibanez, F. 1984. Sur la segmentation des séries chronologiques planctoniques multivariables. *Oceanol. Acta* **7**: 481-491. [769]

Ibanez, F. & G. Seguin. 1972. Etude du cycle annuel du zooplancton d'Abidjan. Comparaison de plusieurs méthodes d'analyse multivariable: composantes principales, correspondances, coordonnées principales. *Invest. Pesq.* **36**: 81-108. [464, 479]

Ifrah, G. 1981. *Histoire universelle des chiffres*. Éditions Seghers, Paris. 568 pp. [67]

Iman, R. L. & W. J. Conover. 1979. The use of the rank transformation in regression. *Technometrics* **21**: 499-509. [36]

Iman, R. L. & W. J. Conover. 1983. *A modern approach to statistics*. Wiley, New York. xxiii + 497 pp. [584]

Ipsen, D. C. 1960. *Units, dimensions and dimensionless numbers*. McGraw-Hill, New York. xii + 236 pp. [109]

Isaaks, E. H. & R. M. Srivastava. 1989. *Applied geostatistics*. Oxford Univ. Press, New York. xix + 561 pp. [810, 811, 813, 817, 830, 832–834, 839]

Jaccard, P. 1900. Contribution au problème de l'immigration post-glaciaire de la flore alpine. *Bull. Soc. Vaudoise Sci. Nat.* **36**: 87-130. [xiv, 275]

Jaccard, P. 1901. Etude comparative de la distribution florale dans une portion des Alpes et du Jura. *Bull. Soc. Vaudoise Sci. nat.* **37**: 547-579. [275]

Jaccard, P. 1908. Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaudoise Sci. nat.* **44**: 223-270. [275]

Jackson, D. A. 1993. Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. *Ecology* **74**: 2204-2214. [448, 449]

Jackson, D. A. 1995. PROTEST: a PROcrustean randomization TEST of community environment concordance. *Écoscience* **2**: 297-303. [612, 704]

Jackson, D. A. & K. M. Somers. 1989. Are probability estimates from the permutation model of Mantel's test stable? *Can. J. Zool.* **67**: 766-769. [31]

Jackson, D. A. & K. M. Somers. 1991a. The spectre of 'spurious' correlations. *Oecologia* **86**: 147-151. [43]

Jackson, D. A. & K. M. Somers. 1991b. Putting things in order: the ups and downs of detrended correspondence analysis. *Am. Nat.* **137**: 704-712. [486, 487]

Jackson, D. A., K. M. Somers & H. H. Harvey. 1992. Null models and fish communities: evidence of nonrandom patterns. *Am. Nat.* **139**: 930-951. [391]

Jackson, R. C. & T. J. Crovello. 1971. A comparison of numerical and biosystematic studies in *Haplopappus*. *Brittonia* **23**: 54-70. [523, 524]

Jain, A. K. & R. C. Dubes. 1988. *Algorithms for clustering data*. Prentice Hall, Englewood Cliffs, New Jersey. xiv + 320 pp. [346, 366, 369, 377, 383, 386, 415]

Jambu, M. & M.-O. Lebeaux. 1983. *Cluster analysis and data analysis*. Elsevier-North-Holland, Amsterdam. xxiv + 898 pp. [349]

Jardine, N. & R. Sibson. 1968. The construction of hierarchic and non-hierarchic classifications. *Comput. J.* **11**: 177-184. [393]

Jardine, N. & R. Sibson. 1971. *Mathematical taxonomy*. Wiley, London. xviii + 286 pp. [393]

Jenkins, G. M. & D. G. Watts. 1968. *Spectral analysis and its applications*. Holden-Day, San Francisco. xviii + 525 pp. [714, 729]

Jenkins, S. H. 1975. Food selection by beavers. A multidimensional contingency table analysis. *Oecologia* (Berl.) **21**: 157-173. [243]

Jolicoeur, P. 1959. Multivariate geographical variation in the wolf *Canis lupus* L. *Evolution* **13**: 283-299. [675]

Jolicoeur, P. 1973. Imaginary confidence limits of the slope of the major axis of a bivariate normal distribution: a sampling experiment. *J. Amer. Statist. Assoc.* **68**: 866-871. [546]

Jolicoeur, P. 1975. Linear regression in fishery research: some comments. *J. Fish. Res. Board Can.* **32**: 1491-1494. [550]

Jolicoeur, P. 1990. Bivariate allometry: interval estimation of the slopes of the ordinary and standardized normal major axes and structural relationship. *J. Theor. Biol.* **144**: 275-285. [548, 550, 552, 553]

Jolicoeur, P. & J. E. Mosimann. 1960. Size and shape variation in the painted turtle. A principal component analysis. *Growth* **24**: 339-354. [443]

Jolicoeur, P. & J. E. Mosimann. 1968. Intervalles de confiance pour la pente de l'axe majeur d'une distribution normale bidimensionnelle. *Biom-Praxim.* **9**: 121-140. [548, 551]

Jones, M. M., H. Tuomisto, D. Borcard, P. Legendre, D. B. Clark & P. C. Olivas. 2008. Explaining variation in tropical plant community composition: influence of environmental and spatial data quality. *Oecologia* **155**: 593-604. [877]

Jones, R. H. 1964. Prediction of multivariate time series. *J. Appl. Meteorol.* **3**: 285-289. [782]

Josse, J., J. Pagès & F. Husson. 2008. Testing the significance of the RV coefficient. *Comput. Stat. Data An.* **53**: 82-91.[700]

Journel, A. G. & C. J. Huijbregts. 1978. *Mining geostatistics*. Academic Press, London. x + 600 pp. [37, 38, 812, 831]

Jumars, P. A., D. Thistle & M. L. Jones. 1977. Detecting two-dimensional spatial structure in biological data. *Oecologia (Berl.)* **28**: 109-123. [794]

Kaplan, D. 2009. *Structural equation modeling: foundations and extensions. 2nd edition*. Sage Publ., Thousand Oaks, California. xi + 255 pp. [593]

Kaufman, L. & P. J. Rousseeuw. 1990. *Finding groups in data – An introduction to cluster analysis*. J. Wiley & Son, New York. xiv + 342 pp. [365, 384, 386, 424]

Kedem, B. 1980. *Binary time series*. Lecture notes in pure and applied mathematics, Vol. 39. Marcel Dekker, New York. ix + 140 pp. [766]

Keitt, T. H. & D. L. Urban. 2005 Scale-specific inference using wavelets. *Ecology* **86**: 2497-2504. [767]

Kempthorne, O. 1952. *The design and analysis of experiments*. Robert E. Krieger Publ. Co., Huntington, N. Y. xix + 631 pp. [6]

Kendall, D. G. 1971. Seriation from abundance matrices. 215-252 *in*: F. R. Hodson, D. G. Kendal & P. Tautu [eds.] *Mathematics in the archaeological and historical sciences*. Edinburgh Univ. Press, Edinburgh. [483]

Kendall, D. G. 1988. Seriation. 417-424 *in*: S. Kotz & N. L. Johnson [eds.] *Encyclopedia of statistical sciences. Vol. 8*. John Wiley & Sons, New York. [403]

Kendall, M. G. 1938. A new measure of rank correlation. *Biometrika* **30**: 81-93. [413]

Kendall, M. G. 1948. *Rank correlation methods*. Charles Griffin & Co., London. vii + 160 pp. [206, 212, 213]

Kendall, M. G. 1976. *Time series*. Charles Griffin & Co., London. ix + 197 pp. [725]

Kendall, M. G. & B. Babington Smith. 1939. The problem of *m* rankings. *Ann. Math. Stat.* **10**: 275-287. [213, 216]

Kendall, M. G. & W. R. Buckland. 1960. *A dictionary of statistical terms. 2nd edition*. Oliver and Boyd, Edinburgh. xi + 575 pp. [234]

Kendall, M. G. & J. K. Ord. 1990. *Time series. 3rd edition*. Edward Arnold, Sevenoaks, Kent. x + 296 pp. [714, 720, 721, 724, 750, 756]

Kendall, M. G. & A. Stuart. 1963. *The advanced theory of statistics. Vol. 1. 2nd edition*. Charles Griffith & Co., London. xii + 433 pp. [363]

Kendall, M. G. & A. Stuart. 1966. *The advanced theory of statistics. Vol. 3*. Hafner Publ. Co., New York. ix + 552 pp. [546, 695]

Kendall, M. G., A. Stuart & J. K. Ord. 1983. *The advanced theory of statistics. Vol. 3. 4th edition*. Charles Griffin & Co., London. x + 780 pp. [714, 725]

Kenkel, N. C. & L. Orlóci. 1986. Applying metric and nonmetric multidimensional scaling to ecological studies: some new results. *Ecology* **67**: 919-928. [487]

Kent, M. & P. Coker. 1992. *Vegetation description and analysis – A practical approach*. John Wiley & Sons, New York. x + 363 pp. [382]

Kerlinger, F. N. & E. J. Pedhazur. 1973. *Multiple regression in behavioral research*. Holt, Rinehart and Winston, Inc. New York. x + 534 pp. [571]

Kermack, K. A. & J. B. S. Haldane. 1950. Organic correlation and allometry. *Biometrika* **37**: 30-41. [547]

Kierstead, H. & L. B. Slobodkin. 1953. The size of water masses containing plankton blooms. *J. Mar. Res*. **12**: 141-147. [125, 126, 136]

Kim, W.-S., H.-T. Huh, J.-G. Je & K.-N. Han. 2003. Evidence of two-clock control of endogenous rhythm in the Washington clam, *Saxidomus purpuratus*. *Mar. Biol*. **142**: 305-309. [766]

Kline, R. B. 2011. *Principles and practice of structural equation modeling, 3rd edition*. Guilford Press, New York. xvi + 427 pp. [593]

Kluge, A. G. & J. S. Farris. 1969. Quantitative phyletics and the evolution of anurans. *Syst. Zool*. **18**: 1-32. [53]

Koch, G. G. & D. B. Gillings. 1983. Inference, design based vs. model based. 84-88 *in*: S. Kotz & N. L. Johnson [eds.] *Encyclopedia of statistical sciences*. *Vol. 4*. Wiley, New York. [6]

Kolasa, J. & S. T. A. Pickett [eds.] 1991. *Ecological heterogeneity*. Springer-Verlag, New York. xi + 332 pp. [22]

Kolasa, J. & C. D. Rollo. 1991. Introduction: The heterogeneity of heterogeneity: A glossary. 1-23 *in*: J. Kolasa & S. T. A. Pickett [eds.] *Ecological heterogeneity*. Springer-Verlag, New York. [788, 789]

Koleff, P., K. J. Gaston & J. J. Lennon. 2003. Measuring beta diversity for presence-absence data. *J. Anim. Ecol*. **72**: 367-382. [260]

Kotz, S. & N. L. Johnson. 1982. Degrees of freedom. 293-294 *in*: S. Kotz & N. L. Johnson [eds.] *Encyclopedia of statistical science*. *Vol. 2*. Wiley, New York. [19]

Krackhardt, D. 1988. Predicting with networks: Nonparametric multiple regression analysis of dyadic data. *Soc. Networks* **10**: 359-381. [606]

Krige, D. G. 1952. A statistical analysis of some of the borehole values in the Orange Free State goldfield. *J. Chem. Metall. Min. Soc. S. Afr.* **53**: 47-70. [831]

Krige, D. G. 1966. Two-dimensional weighted moving average trend surfaces for ore evaluation. 13-38 *in*: *Proceedings of the symposium on mathematical statistics and computer applications in ore valuation*, Johannesburg. [831]

Kroonenberg, P. M. 1983. *Three-mode principal component analysis – Theory and applications*. DSWO Press, Leiden. 398 pp. [269]

Kroonenberg, P. M. 2008. *Applied multi-way data analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, New Jersey. xxi + 579 pp. [235, 269]

Kroonenberg, P. M. & Y. De Roo. 2010. *3WayPack: A program suite for three-way analysis*. The Three-Mode Company, Leiden University, Leiden. 179 pp. + computer programs. [269]

Kruskal, J. B. 1964a. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* **29**: 1-27. [512, 514]

Kruskal, J. B. 1964b. Nonmetric multidimensional scaling: a numerical method. *Psychometrika* **29**: 115-129. [512, 514, 516]

Kruskal, J. B. & M. Wish. 1978. *Multidimensional scaling*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-011. Sage Publications, Beverly Hills. 93 pp. [512, 516]

Kruskal, W. H. & F. Mosteller. 1988. Representative sampling. 77-81 *in*: S. Kotz & N. L. Johnson [eds.] *Encyclopedia of statistical sciences*. *Vol. 8*. Wiley, New York. [6]

Krylov, V. V. 1968. Species association in plankton. *Oceanology* (translated from *Okeanologiya*, in Russian) **8**: 243-251. [318, 392, 393]

Kryszczuk, K. & P. Hurley. 2010. Estimation of the number of clusters using multiple clustering validity indices. *Proceedings of the 9th International Workshop on Multiple Classifier Systems*, Cairo, April 2010 (MCS 2010): 114-123. [418]

Kulczynski, S. 1928. Die Pflanzenassoziationen der Pieninen. *Bull. Int. Acad. Pol. Sci. Lett. Cl. Sci. Math. Nat. Ser. B*, Suppl. II (1927): 57-203. [277, 286, 339, 403–405]

Kullback, S. 1959. *Information theory and statistics*. John Wiley & Sons, New York. xvii + 395 pp. [235, 682]

Kuusipalo, J. 1987. Relative importance of factors controlling the success of *Oxalis acetosella*: an example of linear modelling in ecological research. *Vegetatio* **70**: 171-179. [596]

Lacey, R. W. J., P. Legendre & A. Roy. 2007. Spatial-scale partitioning of in situ turbulent flow data over a pebble cluster in a gravel-bed river. *Water Resour. Res.* **43**, W03416, doi:10.1029/2006WR005044. [877]

Laliberté, E. 2008. Analyzing or explaining beta diversity: Comment. *Ecology* **89**: 3232-3237. [607]

Laliberté, E. & P. Legendre. 2010. A distance-based framework for measuring functional diversity from multiple traits. *Ecology* **91**: 299-305. [255]

Laliberté, E., A. Paquette, P. Legendre & A. Bouchard. 2009. Assessing the scale-specific importance of niches and other spatial processes on beta diversity: a case study from a temperate forest. *Oecologia* **159**: 377-388. [901]

Lam, N. S.-N. 1983. Spatial interpolation methods: a review. *Amer. Cartogr.* **10**: 129-149. [821]

Lambshead, P. J. D. & G. L. J. Paterson. 1986. Ecological cladistics. An investigation of numerical cladistics as a method for analysing ecological data. *J. Nat. Hist.* **20**: 895-909. [391]

Lance, G. N. & W. T. Williams. 1965. Computer programs for monothetic classification («association analysis»). *Comput. J.* **8**: 246-249. [378]

Lance, G. N. & W. T. Williams. 1966a. A generalized sorting strategy for computer classifications. *Nature (Lond.)* **212**: 218. [353, 367, 370]

Lance, G. N. & W. T. Williams. 1966b. Computer programs for hierarchical polythetic classification («similarity analyses»). *Comput. J.* **9**: 60-64. [372, 375]

Lance, G. N. & W. T. Williams. 1966c. Computer programs for classification. *Proc. ANCCAC Conference*, Canberra, May 1966, Paper 12/3. [306]

Lance, G. N. & W. T. Williams. 1967a. Mixed-data classificatory programs. I. Agglomerative systems. *Aust. Comput. J.* **1**: 15-20. [306]

Lance, G. N. & W. T. Williams. 1967b. Mixed-data classificatory programs. II. Divisive systems. *Aust. Comput. J.* **1**: 82-85. [377]

Lance, G. N. & W. T. Williams. 1967c. A general theory of classificatory sorting strategies. I. Hierarchical systems. *Computer J.* **9**: 373-380. [345, 351, 353, 357, 367, 370, 371]

Lance, G. N. & W. T. Williams. 1967d. A general theory of classificatory sorting strategies. II. Clustering systems. *Comput. J.* **10**: 271-277. [349, 385]

Lance, G. N. & W. T. Williams. 1968. Note on a new information-statistic classificatory program. *Comput. J.* **11**: 195. [378]

Lande, R. 1996. Statistics and partitioning of species diversity, and similarity among multiple communities. *Oikos* **76**: 5-13. [260]

Landeiro, V. L., W. E. Magnusson, A. S. Melo, H. M. V. Espírito-Santo & L. M. Bini. 2011. Spatial eigenfunction analyses in stream networks: do watercourse and overland distances produce different results? *Freshwater Biol*. **56**: 1184-1192. [888]

Langhaar, H. L. 1951. *Dimensional analysis and theory of models*. Wiley, New York. xi + 166 p. [109, 126]

Lapointe, F.-J. 1998. How to validate phylogenetic trees? A stepwise procedure. 71-88 *in*: C. Hayashi, N. Oshumi, K. Yajima, Y. Tanaka, H.-H. Bock & Y. Baba [eds.] *Data science, classification, and related methods*. Springer-Verlag, Tokyo. [415]

Lapointe, F.-J. & P. Legendre. 1990. A statistical framework to test the consensus of two nested classifications. *Syst. Zool*. **39**: 1-13. [529, 606]

Lapointe, F.-J. & P. Legendre. 1991. The generation of random ultrametric matrices representing dendrograms. *J. Classif*. **8**: 177-200. [529, 606]

Lapointe, F.-J. & P. Legendre. 1992a. A statistical framework to test the consensus among additive trees (cladograms). *Syst. Biol*. **41**: 158-171. [529, 607]

Lapointe, F.-J. & P. Legendre. 1992b. Statistical significance of the matrix correlation coefficient for comparing independent phylogenetic trees. *Syst. Biol*. **41**: 378-384. [529]

Lapointe, F.-J. & P. Legendre. 1994. A classification of pure malt Scotch whiskies. *Appl. Statist*. **43**: 237-257. [529, 844]

Lapointe, F.-J. & P. Legendre. 1995. Comparison tests for dendrograms: a comparative evaluation. *J. Class*. **12**: 265-282. [529]

Larntz, K. 1978. Small sample comparisons of exact levels for chi-square goodness-of-fit statistics. *J. Amer. Statist. Assoc*. **73**: 253-263. [230]

Larsen, D. R. & P. L. Speckman. 2004. Multivariate regression trees for analysis of abundance data. *Biometrics* **60**: 543-549. [406]

Laurec, A. 1979. *Analyse des données et modèles prévisionnels en écologie marine*. Ph.D. Thesis. Univ. Aix-Marseille. 405 pp. + annexes. [763]

Laws, E. A. & J. W. Archie. 1981. Appropriate use of regression analysis in marine biology. *Mar. Biol. (Berl.)* **65**: 13-16. [555]

Lear, G., M. J. Anderson, J. P. Smith, K. Boxen & G. D. Lewis. 2008. Spatial and temporal heterogeneity of the bacterial communities in stream epilithic biofilms. *FEMS Microbiol. Ecol*. **65**: 463-473. [649]

Lebart, L. 1978. Programme d'agrégation avec containtes (C. A. H. contiguïté). *C. Anal. Données* **3**: 275-287. [840]

Lebart, L. & J. P. Fénelon. 1971. *Statistique et informatique appliquées*. Dunod, Paris. 426 pp. [309, 469]

Lebart, L., A. Morineau & J.-P. Fénelon. 1979. *Traitement des données statistiques – Méthodes et programmes*. Dunod, Paris. xiii + 510 pp. [36, 451]

Le Boulengé, É. 1972. État de nos connaissances sur l'écologie du rat musqué *Ondatra zibethica* L. *La Terre et La Vie* **26**: 3-37. [604]

Le Boulengé, É., P. Legendre, C. de le Court, P. Le Boulengé-Nguyen & M. Languy. 1996. Microgeographic morphological differentiation in muskrats. *J. Mammal*. **77**: 684-701. [604, 821]

Leclerc, B. & G. Cucumel. 1987. Consensus en classification*:* une revue bibliographique. *Math. Sci. Humaines* **100**: 109-128. [418]

Lefkovitch, L. P. 1976. Hierarchical clustering from principal coordinates: an efficient method for small to very large numbers of objects. *Math. Biosci.* **31**: 157-174. [381, 525]

Lefkovitch, L. P. 1978. Cluster generation and grouping using mathematical programming. *Math. Biosci.* **41**: 91-110. [840]

Lefkovitch, L. P. 1980. Conditional clustering. *Biometrics* **36**: 43-58. [840]

Legand, M. 1958. Variations diurnes du zooplancton autour de la Nouvelle-Calédonie. *O. R. S. T. O. M., Inst. Fr. Océanie Sect. Océanogr. Rapp. Sci.* (6): 1-42. [753]

Le Gendre, A. M. 1805. *Nouvelles méthodes pour la détermination des orbites des comètes*. Courcier, Paris. [541]

Legendre, L. 1971. Production primaire dans la Baie-des-Chaleurs (Golfe Saint-Laurent) *Nat. Can. (Qué.)* **98**: 743-773. [114]

Legendre, L. 1973. Phytoplankton organization in Baie des Chaleurs (Gulf of St-Lawrence). *J. Ecol.* **61**: 135-149. [257, 320]

Legendre, L. 1987a. Multidimensional contingency table analysis as a tool for biological oceanography. *Biol. Oceanogr.* **5**: 13-28. [238, 239, 242]

Legendre, L. 2004. *Scientific research and discovery: Process, consequences and practice*. Excellence in Ecology, 16 (O. Kinne, Ed.). International Ecology Institute, Oldendorf-Luhe. xxix + 235 pp. [8, 124, 130]

Legendre, L. 2008a. *Scientific research and discovery: Process, consequences and practice*. Edition électronique. Excellence in Ecology, 16 (O. Kinne, Ed.). International Ecology Institute, Oldendorf-Luhe. xv + 157 pp. http://www.int-res.com/book-series/excellence-in-ecology/ee16/ [8]

Legendre, L., M. Aota, K. Shirasawa, M. J. Martineau & M. Ishikawa. 1991. Crystallographic structure of sea ice along a salinity gradient and environmental control of microalgae in the brine cells. *J. Mar. Syst.* **2**: 347-357. [596, 688, 690]

Legendre, L. & S. Demers. 1984. Towards dynamic biological oceanography and limnology. *Can. J. Fish. Aquat. Sci.* **41**: 2-19. [757]

Legendre, L., S. Demers & D. Lefaivre. 1986. Biological production at marine ergoclines. 1-29 *in*: J.-C. Nihoul [ed.] *Marine interfaces ecohydrodynamics*. Elsevier, Amsterdam. [789]

Legendre, L., M. Fréchette & P. Legendre. 1981. The contingency periodogram: a method of identifying rhythms in series of nonmetric ecological data. *J. Ecol.* **69**: 965-979. [744, 745, 747]

Legendre, L., R. G. Ingram & Y. Simard. 1982. Aperiodic changes of water column stability and phytoplankton in an Arctic coastal embayment, Manitounuk Sound, Hudson Bay. *Nat. Can. (Qué.)* **109**: 775-786. [246]

Legendre, L. & P. Legendre. 1978. Associations. 261-272 *in*: A. Sournia [ed]. *Phytoplankton manual. Monographs on oceanographic Methodology, Vol. 6*. UNESCO, Paris. [389]

Legendre, L. & P. Legendre. 1979a. *Écologie numérique. 1. Le traitement multiple des données écologiques*. Masson, Paris et les Presses de l'Université du Québec. xiv + 197 pp. [xii]

Legendre, L. & P. Legendre. 1979b. *Écologie numérique. 2. La structure des données écologiques*. Masson, Paris et les Presses de l'Université du Québec. viii + 254 pp. [xii]

Legendre, L. & P. Legendre. 1983a. *Numerical ecology.* Developments in environmental modelling, 3. Elsevier Scientific Publ. Co., Amsterdam, The Netherlands. xvi + 419 pp. [xii]

Legendre, L. & P. Legendre. 1983b. Partitioning ordered variables into discrete states for discriminant analysis of ecological classifications. *Can. J. Zool.* **61**: 1002-1010. [241, 243]

Legendre, L. & P. Legendre. 1984a. *Écologie numérique. 2nd edition. 1. Le traitement multiple des données écologiques*. Masson, Paris et les Presses de l'Université du Québec. xv + 260 pp. [xii]

Legendre, L. & P. Legendre. 1984b. *Écologie numérique. 2nd edition. 2. La structure des données écologiques*. Masson, Paris et les Presses de l'Université du Québec. viii + 335 pp. [xii, 718, 719]

Legendre, P. 1976. An appropriate space for clustering selected groups of western North American *Salmo. Syst. Zool.* **25**: 193-195. [346, 523]

Legendre, P. 1987b. Constrained clustering. 289-307 *in*: P. Legendre & L. Legendre [eds.] *Developments in numerical ecology*. NATO ASI series, Vol. G-14. Springer-Verlag, Berlin. [777, 842]

Legendre, P. 1990. Quantitative methods and biogeographic analysis. 9-34 *in*: D. J. Garbary & R. G. South [eds.] *Evolutionary biogeography of the marine algae of the North Atlantic*. NATO ASI Series, Vol. G 22. Springer-Verlag, Berlin. [850]

Legendre, P. 1993. Spatial autocorrelation: Trouble or new paradigm? *Ecology* **74**: 1659-1673. [9, 183, 572, 606, 788, 802]

Legendre, P. 2000. Comparison of permutation methods for the partial correlation and partial Mantel tests. J. *Statist. Comput. Simul.* **67**: 37-73. [604]

Legendre, P. 2005. Species associations: the Kendall coefficient of concordance revisited. *J. Agr. Biol. Envir. S.* **10**: 226-245. [216, 217, 319, 390, 395–397, 700]

Legendre, P. 2008b. Model II regression user's guide, R edition. Available as a vignette of the R package LMODEL2. 14 pp. [552, 555]

Legendre, P. 2010. Coefficient of concordance. 164-169 *in*: N. J. Salkind [ed.] *Encyclopedia of Research Design, Vol. 1*. SAGE Publications, Inc., Los Angeles. [216]

Legendre, P. & M. J. Anderson. 1999. Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. *Ecol. Monogr.* **69**: 1-24. [504, 648]

Legendre, P. & A. Beauvais. 1978. Niches et associations de poissons des lacs de la Radissonie québécoise. *Nat. Can. (Qué.)* **105**: 137-158. [393, 394]

Legendre, P. & H. J. B. Birks. 2012. From classical to canonical ordination. 201-248 *in*: H. J. B. Birks, A. F. Lotter, S. Juggins & J. P. Smol [eds.] *Tracking Environmental Change using Lake Sediments, Volume 5: Data handling and numerical techniques*. Springer, Dordrecht, The Netherlands. [428, 670, 723, 763, 877]

Legendre, P. & D. Borcard. 2006. Quelles sont les échelles spatiales importantes dans un écosystème ? 425-442 *in*: J.-J. Droesbeke, M. Lejeune & G. Saporta [éds], *Analyse statistique de données spatiales*. Éditions TECHNIP, Paris. [861, 864]

Legendre, P., D. Borcard & P. R. Peres-Neto. 2005. Analyzing beta diversity: partitioning the spatial variation of community composition data. *Ecol. Monogr.* **75**: 435-450. [258, 602, 603, 607, 702]

Legendre, P., D. Borcard & P. R. Peres-Neto. 2008. Analyzing or explaining beta diversity: Comment. *Ecology* **89**: 3238-3244. [607]

Legendre, P., D. Borcard & D. W. Roberts. 2012. Variation partitioning involving orthogonal spatial eigenfunction submodels. *Ecology* **93**: 1234-1240. [873, 890, 905]

Legendre, P. & A. Chodorowski. 1977. A generalization of Jaccard's association coefficient for *Q* analysis of multi-state ecological data matrices. *Ekol. Pol.* **25**: 297-308. [282, 283, 287, 290, 342, 343, 524, 525]

Legendre, P., M. R. T. Dale, M.-J. Fortin, P. Casgrain & J. Gurevitch. 2004. Effects of spatial structures on the results of field experiments. *Ecology* **85**: 3202-3214. [18, 22]

Legendre, P., M. R. T. Dale, M.-J. Fortin, J. Gurevitch, M. Hohn & D. Myers. 2002. The consequences of spatial structure for the design and analysis of ecological field surveys. *Ecography* **25**: 601-615. [16, 18, 20, 21]

Legendre, P., S. Dallot & L. Legendre. 1985. Succession of species within a community: chronological clustering, with applications to marine and freshwater zooplankton. *Am. Nat.* **125**: 257-288. [321, 768, 774–779, 842]

Legendre, P., M. De Cáceres & D. Borcard. 2010. Community surveys through space and time: testing the space-time interaction in the absence of replication. *Ecology* **91**: 262-272. [900]

Legendre, P. & P. Dutilleul. 1991. Comments on Boyle's "Acidity and organic carbon in lake water: variability and estimation of means". *J. Paleolimnol.* **6**: 103-110. [18]

Legendre, P. & P. Dutilleul. 1992. Introduction to the analysis of periodic phenomena. 11-25 *in*: M. A. Ali [ed.] *Rhythms in fishes*. NATO ASI Series, Vol. A-236. Plenum, New York. [714, 715, 718, 719, 752]

Legendre, P. & M.-J. Fortin. 1989. Spatial pattern and ecological analysis. *Vegetatio* **80**: 107-138. [9, 601, 790, 791, 793, 807, 821, 832, 844]

Legendre, P. & M.-J. Fortin. 2010. Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data. *Mol. Ecol. Resour.* **10**: 831-844. [262, 363, 602–604]

Legendre, P. & E. D. Gallagher. 2001. Ecologically meaningful transformations for ordination of species data. *Oecologia* **129**: 271-280. [328, 330–332, 463, 648]

Legendre, P., R. Galzin & M. Harmelin-Vivien. 1997a. Relating behavior to habitat: solutions to the fourth-corner problem. *Ecology* **78**: 547-562. [613, 614, 618, 621–623]

Legendre, P. & F.-J. Lapointe. 2004. Assessing congruence among distance matrices: single malt Scotch whiskies revisited. *Aust. N. Z. J. Stat.* 46: 615-629. [217]

Legendre, P., F.-J. Lapointe & P. Casgrain. 1994. Modeling brain evolution from behavior: a permutational regression approach. *Evolution* **48**: 1487-1499. [606, 607]

Legendre, P. & L. Legendre. 1982. Échantillonnage et traitement des données. 163-216 *in*: S. Frontier [ed.] *Stratégies d'échantillonnage en écologie*. Collection d'Écologie, No. 17. Masson, Paris, and Les Presses de l'Université Laval, Québec. [199, 200, 202, 737]

Legendre, P. & L. Legendre. 1998. *Numerical ecology, 2nd English edition*. Elsevier Science BV, Amsterdam, The Netherlands. xv + 853 pp. [xiii, xv, 349, 666]

Legendre, P. & V. Legendre. 1984c. Postglacial dispersal of freshwater fishes in the Québec peninsula. *Can. J. Fish. Aquat. Sci.* **41**: 1781-1802. [840, 844, 847, 849]

Legendre, P., F. Long, R. Bergeron & J. M. Levasseur. 1978. Inventaire aérien de la faune dans le Moyen Nord québecois. *Can. J. Zool.* **56**: 451-462. [235, 283]

Legendre, P. & B. H. McArdle. 1997. Comparison of surfaces. *Oceanol. Acta* **20**: 27-41. [790, 826]

Legendre, P., X. Mi, H. Ren, K. Ma, M. Yu, I. F. Sun & F. He. 2009. Partitioning beta diversity in a subtropical broad-leaved forest of China. *Ecology* **90**: 663-674. [402, 874, 875]

Legendre, P., N. L. Oden, R. R. Sokal, A. Vaudor & J. Kim. 1990. Approximate analysis of variance of spatially autocorrelated regional data. *J. Classif.* **7**: 53-75. [18, 20, 21]

Legendre, P., J. Oksanen & C. J. F. ter Braak. 2011. Testing the significance of canonical axes in redundancy analysis. *Methods Ecol. Evol.* **2**: 269-277. [634, 635]

Legendre, P., D. Planas & M.-J. Auclair. 1984a. Succession des communautés de gastéropodes dans deux milieux différant par leur degré d'eutrophisation. *Can. J. Zool.* **62**: 2317-2327. [768]

Legendre, P. & D. J. Rogers. 1972. Characters and clustering in taxonomy: a synthesis of two taximetric procedures. *Taxon* **21**: 567-606. [282, 338, 378]

Legendre, P., S. F. Thrush, V. J. Cummings, P. K. Dayton, J. Grant, J. E. Hewitt, A. H. Hines, B. H. McArdle, R. D. Pridmore, D. C. Schneider, S. J. Turner, R. B. Whitlatch & M. R. Wilkinson. 1997b. Spatial structure of bivalves in a sandflat: scale and generating processes. *J. Exp. Mar. Biol. Ecol.* **216**: 99-128. [828]

Legendre, P. & M. Troussellier. 1988. Aquatic heterotrophic bacteria: modeling in the presence of spatial autocorrelation. *Limnol. Oceanogr.* **33**: 1055-1067. [606]

Legendre, P., M. Troussellier & B. Baleux. 1984b. Indices descriptifs pour l'étude de l'évolution des communautés bactériennes. 71-84 *in*: A. Bianchi [ed.] *Bactériologie marine – Colloque international no 331*. Éditions du CNRS, Paris. [257]

Legendre, P., M. Troussellier, V. Jarry & M.-J. Fortin. 1989. Design for simultaneous sampling of ecological variables: from concepts to numerical solutions. *Oikos* **55**: 30-42. [844]

Lehn, W. H. 1979. Atmospheric refraction and lake monsters. *Science (Wash. D. C.)* **205**: 183-185. [226]

Lehn, W. H. & I. Schroeder. 1981. The Norse merman as an optical phenomenon. *Nature (Lond.)* **289**: 362-366. [226]

Lekan, J. F. & R. E. Wilson. 1978. Spatial variability of phytoplankton biomass in the surface waters of Long Island Sound. *Estuar. Coast. Mar. Sci.* **6**: 239-251. [757]

Léonard, R., P. Legendre, M. Jean & A. Bouchard. 2008. Spatial patterns of submerged macrophyte communities in a fluvio-lacustrine section of the St. Lawrence River: a landscape ecology perspective. *Landscape Ecol.* **23**: 91-105. [877]

Levene, H. 1960. Robust tests for equality of variances. 278–292 *in*: I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow & H. B. Mann [eds.] *Contributions to probability and statistics*. Stanford University Press, Stanford, California. [682]

Levin, S. A. 1992. The problem of pattern and scale in ecology. *Ecology* 73: 1943-1967. [9]

Levin, S. A. 2000. Multiple scales and the maintenance of biodiversity. *Ecosystems* 3: 498-506. [9]

Levings, C. D. 1975. Analyses of temporal variation in the structure of a shallow-water benthic community in Nova Scotia. *Int. Revue ges. Hydrobiol.* **60**: 449-470. [768]

Levins, R. 1968. *Evolution in changing environments: some theoretical explorations*. Princeton University Press, Princeton, New Jersey. [260]

Lilliefors, H. W. 1967. The Kolmogorov-Smirnov test for normality with mean and variance unknown. *J. Amer. Statist. Assoc.* **62**: 399-402. [190, 191]

Lindeman, R. L. 1942. The trophic-dynamic aspect of ecology. *Ecology* **23**: 399-418. [785]

Line, J. M., C. J. F. ter Braak & H. J. B. Birks. 1994. WACALIB version 3.3 – A computer program to reconstruct environmental variables from fossil assemblages by weighted averaging and to derive sample-specific errors of prediction. *J. Paleolimnol.* **10**: 147-152. [673]

Lingoes, J. C. 1971. Some boundary conditions for a monotone analysis of symmetric matrices. *Psychometrika* **36**: 195-203. [502]

Lipschutz, S. & M. L. Lipson. 2009. *Schaum's outline of linear algebra. 4th edition*. McGraw-Hill, New York. viii + 425 pp. [569]

Little, R. J. A. & D. B. Rubin. 1987. *Statistical analysis with missing data*. Wiley, New York. xiv + 278 pp. [54–56]

Lloyd, M. & R. J. Ghelardi. 1964. A table for calculating the «equitability» component of species diversity. *J. Anim. Ecol.* **33**: 217-225. [255, 256]

Lloyd, M., R. F. Inger & F. W. King. 1968. On the diversity of reptile and amphibian species in a Bornean rain forest. *Am. Nat.* **102**: 497-515. [253]

Logerwell, E. A., R. P. Hewitt & D. A. Demer. 1998. Scale-dependent spatial variance patterns and correlations of seabirds and prey in the southeastern Bering Sea as revealed by spectral analysis. *Ecography* **21**: 212-223. [759, 762]

Longhurst, A. 2007. *Ecological geography of the sea. 2nd edition*. Elsevier Academic Press, San Diego. 560 pp. [248]

Loreau, M. 2010. *The challenges of biodiversity science*. Excellence in Ecology, 17 [O. Kinne, Ed.]. International Ecology Institute, Oldendorf/Luhe. xxviii + 120 pp. [247, 250]

Lotwick, H. W. & B. W. Silverman. 1982. Methods for analysing spatial processes of several types of points. *J. Roy. Statist. Soc. B* **44**: 406-413. [652]

Ludwig, A., M. Bigras-Poulin & P. Michel. 2009. The analysis of crow population dynamics as a surveillance tool. *Transbound. Emerg. Dis.* **56**: 337-345. [750]

Lukaszewicz, J. 1951. Sur la liaison et la division des points d'un ensemble fini. *Colloq. Math.* **2**: 282-285. [346, 350]

MacArthur, R. H. 1957. On the relative abundance of bird species. *Proc. Natl. Acad. Sci. USA* **43**: 293-295. [256]

MacArthur, R., H. Recher & M. Cody. 1966. On the relation between habitat selection and species diversity. *Am. Nat.* **100**: 319-332. [260]

Macnaughton-Smith, P., W. T. Williams, M. B. Dale & L. G. Mockett. 1964. Dissimilarity analysis: a new technique of hierarchical sub-division. *Nature (Lond.)* **202**: 1034-1035. [380]

MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. 281-297 *in*: L. M. Le Cam & J. Neyman [eds.] *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. Vol. 1*. University of California Press, Berkeley. [386]

Madansky, A. 1959. The fitting of straight lines when both variables are subject to error. *J. Amer. Statist. Assoc.* **54**: 173-205. [546]

Magnan, P., M. A. Rodriguez, P. Legendre & S. Lacasse. 1994. Dietary variation in a freshwater fish species: Relative contributions of biotic interactions, abiotic factors, and spatial structure. *Can. J. Fish. Aquat. Sci.* **51:** 2856-2865. [54, 892]

Magurran, A. E. 2004. *Measuring biological diversity.* Blackwell Pub., Oxford. viii + 256 pp. [260]

Mahalanobis, P. C. 1936. On the generalized distance in statistics. *Proc. Natl. Inst. Sci. India* **2**: 49-55. [303]

Manly, B. F. J. 1986. Randomization and regression methods for testing for associations with geographical, environmental, and biological distances between populations. *Res. Popul. Ecol. (Kyoto)* **28**: 201-218. [606]

Manly, B. F. J. 1997. *Randomization, bootstrap and Monte Carlo methods in biology. 2nd edition*. Chapman and Hall, London. xix + 399 pp. [25–27, 30, 31]

Mantel, N. 1967. The detection of disease clustering and a generalized regression approach. *Cancer Res.* **27**: 209-220. [412, 598, 600]

Marcotorchino, J. F. & P. Michaud. 1979. *Optimisation en analyse ordinale des données*. Masson, Paris. xii + 211 pp. [269]

Mardia, K.V., J.T. Kent, and J.M. Bibby. 1979. Multivariate analysis. Academic Press, London. xv + 521 pp. [143]

Margalef, R. 1958. Information theory in ecology. *General Systems* **3**: 36-71. [xii, 252, 253, 255, 257]

Margalef, R. 1968. *Perspectives in ecological theory*. Univ. Chicago Press, Chicago. viii + 111 pp. [2, 768]

Margalef, R. 1974. *Ecologia*. Ediciones Omega, Barcelona. xv + 951 pp. [37, 249, 250, 255]

Margalef, R. & F. González Bernáldez. 1969. Grupos de especies asociadas en el fitoplancton del mar Caribe (NE de Venezuela). *Invest. Pesq.* **33**: 287-312. [452]

Margalef, R. & E. Gutiérrez. 1983. How to introduce connectance in the frame of an expression for diversity. *Am. Nat.* **121**: 601-607. [254]

Marquardt, D. W. & R. D. Snee. 1975. Ridge regression in practice. *Am. Statist.* **29**: 3-20. [564]

Matheron, G. 1962. *Traité de géostatistique appliquée. Tomes 1 et 2*. Éditions Technip, Paris. 334 pp., 172 pp. [831]

Matheron, G. 1965. *Les variables régionalisées et leur estimation – Une application de la théorie des fonctions aléatoires aux sciences de la nature*. Masson, Paris. 305 pp. [790, 831]

Matheron, G. 1970. *La théorie des variables régionalisées, et ses applications*. Les Cahiers du Centre de Morphologie Mathématique. Fontainebleau, fascicule 5. 212 pp. [831]

Matheron, G. 1971. *The theory of regionalised variables and its applications*. Les Cahiers du Centre de Morphologie Mathématique, Fasc. 5, ENSMP, Paris. 211 pp. [831]

Matheron, G. 1973. The intrinsic random functions and their applications. *Adv. Appl. Prob.* **5**: 439-468. [831]

Matula, D. W. & R. R. Sokal. 1980. Properties of Gabriel graphs relevant to geographic variation research and the clustering of points in the plane. *Geogr. Anal.* **12**: 205-222. [835]

Maxwell, J. C. 1871. Remarks on the mathematical classification of physical quantities. *Proc. Lond. Math. Soc.* **3**: 224-232. [109, 111]

McArdle, B. 1988. The structural relationship: regression in biology. *Can. J. Zool.* **66**: 2329-2339. [546, 547, 550, 552, 554, 556]

McArdle, B. H. & M. J. Anderson. 2001. Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology* **82**: 290-297. [263]

McBratney, A. B. & R. Webster. 1981. The design of optimal sampling schemes for local estimation and mapping of regionalized variables – II. *Comput. Geosci.* **7**: 335-365. [21]

McBratney, A. B., R. Webster & T. M. Burgess. 1981. The design of optimal sampling schemes for local estimation and mapping of regionalized variables – I. *Comput. Geosci.* **7**: 331-334. [21]

McCoy, E. D., S. S. Bell & K. Walters. 1986. Identifying biotic boundaries along environmental gradients. *Ecology* **67**: 749-759. [294, 772]

McCullagh, P. & J. A. Nelder. 1983. *Generalized linear models*. Chapman and Hall, London. 261 pp. [586]

McCune, B. 1994. Improving community analysis with the Beals smoothing function. *Écoscience* **1**: 82-86. [334]

McCune, B. 1997. Influence of noisy environmental data on canonical correspondence analysis. *Ecology* **78**: 2617-2623. [673]

McCune, B. & T. F. H. Allen. 1985. Will similar forest develop on similar sites? *Can. J. Bot.* **63**: 367-376. [601]

McGeoch, M. A. & S. L. Chown. 1998. Scaling up the value of bioindicators. *Trends Ecol. Evol.* **13**: 46-47. [401]

McIntyre, R. M. & R. K. Blashfield. 1980. A nearest-centroid technique for evaluating the minimum-variance clustering procedure. *Multivar. Behav. Res.* **15**: 225-238. [417]

Mead, R. 1988. *The design of experiments – Statistical principles for practical applications*. Cambridge University Press, Cambridge. xiv + 620 pp. [535]

Mendelssohn, R. & P. Cury. 1987. Fluctuations of a fortnightly abundance index of the Ivoirian coastal pelagic species and associated environmental conditions. *Can. J. Fish. Aquat. Sci.* **44**: 408-421. [56]

Méot, A., P. Legendre & D. Borcard. 1998. Partialling out the spatial component of ecological variation: questions and propositions in the linear modeling framework. *Environ. Ecol. Stat.* **5**: 1-26. [854]

Mesplé, F., M. Troussellier, C. Casellas & P. Legendre. 1996. Evaluation of simple statistical criteria to qualify a simulation. *Ecol. Model.* **88**: 9-18. [544, 555]

Meulman, J. 1982. *Homogeneity analysis of incomplete data*. DSWO Press, Leiden. 168 pp. [465]

Miller, J. K. 1975. The sampling distribution and a test for the significance of the bimultivariate redundancy statistic: a Monte Carlo study. *Multivar. Behav. Res.* **10**: 233-244. [633]

Miller, J. K. & S. D. Farr. 1971. Bimultivariate redundancy: a comprehensive measure of interbattery relationship. *Multivar. Behav. Res.* **6**: 313-324. [566, 632]

Milligan, G. W. 1979. Ultrametric hierarchical clustering algorithms. *Psychometrika* **44:** 343-346. [377]

Milligan, G. W. 1980. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika* **45**: 325-342. [415]

Milligan, G. W. 1981. A Monte Carlo study of thirty internal criterion measures for cluster analysis. *Psychometrika* **46**: 187-199. [416]

Milligan, G. W. 1996. Clustering validation – Results and implications for applied analyses. 341-375 *in*: P. Arabie, L. J. Hubert & G. De Soete [eds.] *Clustering and Classification*. World Scientific Publ. Co., River Edge, New Jersey. [350, 389, 415]

Milligan, G. W. & M. C. Cooper. 1985. An examination of procedures for determining the number of clusters in a data set. *Psychometrika* **50**: 159-179. [389]

Milligan, G. W. & M. C. Cooper. 1987. Methodological review: clustering methods. *Appl. Psychol. Meas.* **11**: 329-354. [386]

Milligan, G. W. & M. C. Cooper. 1988. A study of standardization of variables in cluster analysis. *J. Classif.* **5**: 181-204. [44]

Minchin, P. R. 1987. An evaluation of the relative robustness of techniques for ecological ordination. *Vegetatio* **69**: 89-107. [487, 512, 517]

Monestiez, P. 1978. Méthodes de classification automatique sous contraintes spatiales. 367-379 *in*: J. M. Legay & R. Tomassone [eds.] *Biométrie et écologie*. Inst. nat. Rech. agronomique, Jouy-en-Josas. [840]

Montgomery, D. C. & E. A. Peck. 1982. *Introduction to linear regression analysis*. Wiley, New York. 504 pp. [589, 590]

Mood, A. M. 1969. Macro-analysis of the American educational system. *Oper. Res*. **17**: 770-784. [571]

Mood, A. M. 1971. Partitioning variance in multiple regression — Analyses as a tool for developing learning models. *Am. Educ. Res. J*. **8**: 191-202. [571]

Moran, P. A. P. 1950. Notes on continuous stochastic phenomena. *Biometrika* **37**: 17-23. [793]

Moreau, G. & L. Legendre. 1979. Relation entre habitat et peuplements de poissons: essai de définition d'une méthode numérique pour des rivières nordiques. *Hydrobiologia* **67**: 81-87. [306]

Moretti, M., M. De Cáceres, C. Pradella, M. K. Obrist, B. Wermelinger, P. Legendre & P. Duelli. 2010. Fire-induced taxonomic and functional changes in saproxylic beetle communities in fire sensitive regions. *Ecography* **33**: 760-771. [399]

Moretti, M. & C. Legg. 2009. Combining plant and animal traits to assess community functional responses to disturbance. *Ecography* **32**: 299-309. [702]

Morice, E. 1968. *Dictionnaire de statistique*. Dunod, Paris. ix + 196 pp. [234]

Morrall, R. A. A. 1974. Soil microfungi associated with aspen in Saskatchewan: synecology and quantitative analysis. *Can. J. Bot*. **52**: 1803-1817. [455]

Morrison, D. F. 1990. *Multivariate statistical methods. 3rd edition*. McGraw-Hill, New York. xvii + 495. [2, 10, 146, 429]

Motyka, J. 1947. O zadaniach i metodach badan geobotanicznych. Sur les buts et les méthodes des recherches géobotaniques. *Annales Universitatis Mariae Curie-Sklodowska (Lublin, Polonia), Sectio C, Supplementum I*. viii + 168 pp. [285, 311]

Motyka, J., B. Dobrzanski & S. Zawadzki. 1950. Wstepne badania nad lakami poludniowo-wschodniej Lubelszczyzny. Preliminary studies on meadows in the south-east of the province Lublin. *Annales Universitatis Mariae Curie-Sklodowska (Lublin, Polonia), Sectio E*, **13**: 367-447. [285, 311]

Muirhead, R. J. 1982. *Aspects of multivariate statistical theory*. Wiley, New York. xix + 673 pp. [143]

Muller, R. A. & G. J. MacDonald. 2002. *Ice ages and astronomical causes: data, spectral analysis and mechanisms*. Springer-Verlag, Berlin. xvii + 323 pp. [754]

Murray, J. L. S. & P. A. Jumars. 2002. Clonal fitness of attached bacteria predicted by analog modeling. *BioScience* **52**: 343-355. [124]

Murtagh, F. 1985. A survey of algorithms for contiguity-constrained clustering and related problems. *Comput. J*. **28**: 82-88. [843]

Myers, D. E. 1982. Matrix formulation of co-kriging. *Math. Geol*. **14**: 249-257. [57]

Myers, D. E. 1983. Estimation of linear combinations and co-kriging. *Math. Geol*. **15**: 633-637. [57]

Myers, D. E. 1984. Co-kriging — New developments. 295-305 *in*: G. Verly, M. David, A. G. Journel & A. Marechal [eds.] *Geostatistics for natural resources characterization, Part 1. Vol. C 122*. NATO ASI Series. D. Reidel Publ. Co., Dordrecht. [57]

Myers, D. E. 1997. Statistical models for multiple-scaled analysis. 273-293 *in*: D. A. Quattrochi & M. F. Goodchild [eds.] *Scale in remote sending and GIS*. Lewis, New York. [813]

Myster, W. & S. T. A. Pickett. 1992. Dynamics of associations between plants in ten old fields during 31 years of succession. *J. Ecol.* **80**: 291-302. [319, 390]

Nardi, M., E. Morgan & F. Scapini. 2003. Seasonal variation in the free-running period in two *Talitrus saltator* populations from Italian beaches differing in morphodynamics and human disturbance. *Estuar. Coast. Shelf Sci.* **58**, Supplement: 199-206. [744]

Nekola, J. C. & P. S. White. 1999. The distance decay of similarity in biogeography and ecology. *J. Biogeogr.* **26**: 867–878. [598]

Nemec, A. F. L. & R. O. Brinkhurst. 1988. Using the bootstrap to assess statistical significance in the cluster analysis of species abundance data. *Can. J. Fish. Aquat. Sci.* **45**: 965-970. [416]

Neter, J., M. H. Kutner, C. J. Nachtsheim & W. Wasserman. 1996. *Applied linear statistical models*. Irwin, Chicago. xv + 1408 pp. [538, 558]

Neu, C. W., C. R. Byers & J. M. Peek. 1974. A technique for analysis of utilization-availability data. *J. Wildl. Manag.* **38**: 541-545. [244]

Neyman, J. & E. S. Pearson. 1966. *Joint statistical papers of J. Neyman and E. S. Pearson*. University of California Press, Berkeley. vii + 299 pp. [28]

Nie, N. H., C. H. Hull, J. G. Jenkins, K. Steinbrenner & D. H. Bent. 1975. *SPSS - Statistical package for the social sciences. 2 edition*. McGraw-Hill, New York. xxiv + 675 p. pp. [593, 596]

Nishisato, S. 1980. *Analysis of categorical data – Dual scaling and its applications*. Mathematical expositions No. 24. University of Toronto Press, Toronto. xiii + 276 pp. [464]

Noorduijn, S. L., A. Ghadouani, R. Vogwill, K. R. J. Smettem & P. Legendre. 2010. Water table response to an experimental alley farming trial: dissecting the spatial and temporal structure of the data. *Ecol. Appl.* **20**: 1704-1720. [877]

North, P. M. 1977. A novel clustering method for estimating numbers of bird territories. *Appl. Stat. 26*: 149-155. [773]

Norusis, M. J. 1990. *SPSS advanced statistics user's guide*. SPSS Inc., Chicago. 285 pp. [587]

Noy-Meir, I. & D. Anderson. 1971. Multiple pattern analysis or multiscale ordination: towards a vegetation hologram. 207-232 *in*: G. P. Patil, E. C. Pielou & E. W. Water [eds.] *Statistical ecology: populations, ecosystems, and systems analysis*. Pennsylvania State University Press, University Park, Pennsylvania. [894]

Noy-Meir, I., D. Walker & W. T. Williams. 1975. Data transformation in ecological ordination - II. On the meaning of data standardization. *J. Ecol.* **63**: 779-800. [328, 332]

Obenchain, R. L. 1977. Classical F-tests and confidence regions for ridge regression. *Technometrics* **19**: 429-439. [564]

Ochiai, A. 1957. Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions. *Bull. Jpn. Soc. Sci. Fish.* **22**: 526-530. [277]

O'Connor, I. & L. W. Aarssen. 1987. Species association patterns in abandoned sand quarries. *Vegetatio* **73**: 101-109. [319, 390]

Oden, N. L. 1984. Assessing the significance of a spatial correlogram. *Geogr. Anal.* **16**: 1-16. [745, 799]

Oden, N. L. & R. R. Sokal. 1986. Directional autocorrelation: an extension of spatial correlograms to two dimensions. *Syst. Zool.* **35**: 608-617. [807, 819]

Oden, N. L., R. R. Sokal, M.-J. Fortin & H. Goebl. 1993. Categorical wombling: detecting regions of significant change in spatially located categorical variables. *Geogr. Anal.* **25**: 315-336. [790, 844, 846]

Odum, E. P. 1950. Bird populations of the Highlands (North Carolina) plateau in relation to plant succession and avian invasion. *Ecology* **31**: 587-605. [xii, 311]

Ohtani, K. 2000. Bootstrapping R^2 and adjusted R^2 in regression analysis. *Ecol. Model.* **17**: 473-483. [565]

Okubo, A. 1987. Fantastic voyage into the deep – Marine biofluids mechanics. 32-47 *in*: E. Teramoto & M. Yamaguti [eds.] *Mathematical topics in population biology, morphogenesis and neurosciences*. Lecture Notes in Biomathematics. Springer-Verlag, Berlin. [129]

Olea, R. A. 1991. *Geostatistical glossary and multilingual dictionary*. Oxford University Press. New York. [37]

Oliver, M. A. & R. Webster. 1989. A geostatistical basis for spatial weighting in multivariate classification. *Math. Geol.* **21**: 15-35. [843]

O'Neill, R. V., R. H. Gardner, B. T. Milne, M. G. Turner & B. Jackson. 1991. Heterogeneity and spatial hierarchies. 85-96 *in*: J. Kolasa & S. T. A. Pickett [eds.] *Ecological heterogeneity*. Springer-Verlag, New York. [9]

Orlóci, L. 1966. Geometric models in ecology. I. The theory and application of some ordination methods. *J. Ecol.* **54**: 193-215. [509]

Orlóci, L. 1967a. Data centering: a review and evaluation with reference to component analysis. *Syst. Zool.* **16**: 208-212. [451]

Orlóci, L. 1967b. An agglomerative method for classification of plant communities. *J. Ecol.* **55**: 193-206. [301]

Orlóci, L. 1975. *Multivariate analysis in vegetation research*. Dr. W. Junk B. V., The Hague. ix + 276 pp. [270, 464]

Orlóci, L. 1978. *Multivariate analysis in vegetation research. 2nd edition*. Dr. W. Junk B. V., The Hague. ix + 451 pp. [270, 289, 300, 312, 329, 403, 476]

Orlóci, L. 1981. Probing time series vegetation data for evidence of succession. *Vegetatio* **46**: 31-35. [768]

Ouellette, M.-H., P. Legendre & D. Borcard. 2012. Cascade multivariate regression tree: a novel approach for modelling nested explanatory sets. *Methods Ecol. Evol.* **3**: 234-244. [409]

Paloheimo, J. E. & L. M. Dickie. 1965. Food and growth of fishes. I. A growth curve derived from experimental data. *J. Fish. Res. Board. Can.* **22**: 521-542. [127, 137]

Passy, S. 2007. Community analysis in stream biomonitoring: what we measure and what we don't. *Environ. Monit. Assess.* **127**: 409-417. [855]

Patrick, R. 1949. A proposed biological measure of stream conditions, based on a survey of the Conestoga basin, Lancaster County, Pennsylvania. *Proc. Acad. Nat. Sci. Phila.* **101**: 277-341. [200, 251]

Patten, B. C. 1962. Species diversity in net phytoplankton of Raritan Bay. *J. Mar. Res.* **20**: 57-75. [256]

Pawitan, Y. & J. Huang. 2003. Constrained clustering of irregularly sampled spatial data. *J. Stat. Comput. Sim.* **73**: 853-865. [841]

Pearson, K. 1897. On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proc. R. Soc. London* 60: 487-498. [43]

Pearson, K. 1900. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philos. Mag., Ser. 5* **50**: 157-172. [230]

Pearson, K. 1901. On lines and planes of closest fit to systems of points in space. *Philos. Mag.* **2**: 559-572. [430, 546]

Pearson, K. 1926. On the coefficient of racial likeness. *Biometrika* **18**: 105-117. [306]

Peet, R. K. 1974. The measurement of species diversity. *Annu. Rev. Ecol. Syst.* **5**: 285-307. [250, 251, 255]

Peet, R. K., R. G. Knox, J. S. Case & R. B. Allen. 1988. Putting things in order: the advantages of detrended correspondence analysis. *Am. Nat.* **131**: 924-934. [486]

Peitgen, H.-O., H. Jürgens & D. Saupe. 2004. *Chaos and fractals: new frontiers of science, 2nd ed.* Springer-Verlag, New York. xiii + 864 pp. [2]

Peli, T. & D. Malah. 1982. A study of edge detection algorithms. *Computer Graphics Image Process.* **20**: 1-21. [844]

Pélissier, R., P. Couteron & S. Dray. 2008. Analyzing or explaining beta diversity: Comment. *Ecology* **89**: 3227-3232. [607]

Percival, D. B. & A. T. Walden. 2000. *Wavelet methods for time series analysis*. Cambridge University Press, New York. xxv + 594 pp. [767]

Peres-Neto, P. R. & D. A. Jackson. 2001. How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test. *Oecologia* **129**: 169-178. [612, 704]

Peres-Neto, P. R. & P. Legendre. 2010. Estimating and controlling for spatial structure in the study of ecological communities. *Global Ecol. Biogeogr.* **19**: 174-184. [20, 902]

Peres-Neto, P. R., P. Legendre, S. Dray & D. Borcard. 2006. Variation partitioning of species data matrices: estimation and comparison of fractions. *Ecology* **87**: 2614-2625. [566, 571, 573, 578, 633, 667]

Perruchet, C. 1981. Classification sous contrainte de contiguïté continue. 71-92 *in*: *Classification automatique et perception par ordinateur.* Séminaires de l'Institut national de Recherche en Informatique et en Automatique (C 118), Rocquencourt. [840]

Perruchet, C. 1983a. Significance tests for clusters*:* overview and comments. 199-208 *in*: J. Felsenstein [ed.] *Numerical taxonomy.* NATO Advanced Study Institute Series G (Ecological Sciences), No. 1. Springer-Verlag, Berlin. [415]

Perruchet, C. 1983b. Une analyse bibliographique des épreuves de classifiabilité en analyse des données. *Statist. Anal. Données* **8**: 18-41. [415]

Petchey, O. L. & K. J. Gaston. 2002. Functional diversity (FD), species richness and community composition. *Ecol. Lett.* **5**: 402-411. [255]

Peters, R. H. 1983. *The ecological implications of body size*. Cambridge Univ. Press, Cambridge. xii + 329 pp. [537]

Petrie, W. M. F. 1899. Sequences in prehistoric remains. *J. Anthropol. Inst.* **29**: 295-301. [403]

Piazza, A. & L. L. Cavalli-Sforza. 1975. Spectral analysis of patterned covariance matrices and evolutionary relationships. 76-105 *in*: G. F. Estabrook [ed.] *Proceedings of the eight international conference on numerical taxonomy.* W. H. Freeman, San Francisco. [381]

Pickett, S. T. A. & P. S. White [eds.] 1985. *The ecology of natural disturbance and patch dynamics*. Academic Press, New York. xiv + 472 pp. [785]

Pielou, E. C. 1966. The measurement of diversity in different types of biological collections. *J. Theor. Biol.* **13**: 131-144. [xii, 253, 255]

Pielou, E. C. 1969. *An introduction to mathematical ecology*. John Wiley & Sons, New York. viii + 286 pp. [xii, 250, 254]

Pielou, E. C. 1975. *Ecological diversity*. John Wiley & Sons, New York. viii + 165 pp. [221, 227, 249, 250, 256, 257]

Pielou, E. C. 1977. *Mathematical ecology. 2nd edition*. Wiley, New York. x + 385 pp. [789]

Pillai, K. C. S. & Y.-S. Hsu. 1979. Exact robustness studies of the test of independence based on four multivariate criteria and their distribution problems under violations. *Ann. I. Stat. Math.* **31**: 85-101. [694]

Pinel-Alloul, B. 1995. Spatial heterogeneity as a multiscale characteristic of zooplankton community. *Hydrobiologia* **300/301**: 17-42. [855]

Pinel-Alloul, B., E. Magnin & G. Codin-Blumer. 1982. Effet de la mise en eau du réservoir Desaulniers (Territoire de la Baie de James) sur le zooplancton d'une rivière et d'une tourbière réticulée. *Hydrobiologia* **86**: 271-296. [778]

Pinel-Alloul, B., G. Méthot, L. Lapierre & A. Willsie. 1996. Macroinvertebrate community as a biological indicator of ecological and toxicological factors in Lake Saint-François (Québec). *Environ. Pollut.* **91**: 65-87. [855]

Pinel-Alloul, B., G. Méthot, G. Verreault & Y. Vigneault. 1990. Phytoplankton in Quebec lakes: variation with lake morphometry, and with natural and anthropogenic acidification. *Can. J. Fish. Aquat. Sci.* **47**: 1047-1057. [370]

Pinel-Alloul, B., T. Niyonsenga & P. Legendre. 1995. Spatial and environmental components of freshwater zooplankton structure. *Écoscience* **2**: 1-19. [855]

Pinty, J. J. & C. Gaultier. 1971. *Dictionnaire pratique de mathématiques et statistiques en sciences humaines*. Editions universitaires, Paris. 298 pp. [227]

Pitard, F. F. 1992. *Pierre Gy's sampling theory and sampling practice. Volume I: Heterogeneity and sampling*. CRC Press Inc., Boca Raton, Florida. 214 pp. [788]

Plackett, R. L. 1974. *The analysis of categorical data*. Griffin's statistical monographs and courses, no. 35. Griffin, London. viii + 159 pp. [235]

Planes, S., A. Lefèvre, P. Legendre & R. Galzin. 1993. Spatio-temporal variability in fish recruitment to a coral reef (Moorea, French Polynesia). *Coral Reefs* **12**: 105-113. [842]

Platt, T. 1969. The concept of energy efficiency in primary production. *Limnol. Oceanogr.* **14**: 653-659. [114]

Platt, T. 1972. Local phytoplankton abundance and turbulence. *Deep-Sea Res.* **19**: 183-187. [757]

Platt, T. 1978. Spectral analysis of spatial structure in phytoplankton populations. 73-84 *in*: J. H. Steele [ed.] *Spatial pattern in plankton communities*. Plenum Press, New York. [757]

Platt, T. 1981. Thinking in term of scale: introduction to dimensional analysis. 112-121 *in*: T. Platt, K. H. Mann & R. E. Ulanowicz [eds.] *Mathematical models in biological oceanography*. The UNESCO Press, Paris. [125, 127, 129]

Platt, T. & K. L. Denman. 1975. Spectral analysis in ecology. *Annu. Rev. Ecol. Syst.* **6**: 189-210. [754, 757]

Platt, T., L. M. Dickie & R. W. Trites. 1970. Spatial heterogeneity of phytoplankton in a near-shore environment. *J. Fish. Res. Board Can.* **27**: 1453-1473. [732, 734, 760, 761]

Platt, T. & D. V. Subba Rao. 1970. Energy flow and species diversity in a marine phytoplankton bloom. *Nature (Lond.)*. **227**: 1059-1060. [114, 732, 734, 760, 761]

Platt, T. & D. V. Subba Rao. 1973. Some current problems in marine phytoplankton productivity. *Fish. Res. Board Can. Tech. Rep.* **307**: 1-90. [123, 124]

Podani, J. 1999. Extending Gower's general coefficient of similarity to ordinal characters. *Taxon* **48**: 331-340. [279]

Podani, J. & B. Csányi. 2010. Detecting indicator species: some extensions of the IndVal measure. *Ecol. Indic.* **10**: 1119-1124. [400]

Powell, T. M., P. J. Richerson, T. M. Dillon, B. A. Agee, B. J. Dozier, D. A. Godden & L. O. Myrup. 1975. Spatial scales of current speed and phytoplankton biomass fluctuations in Lake Tahoe. *Science (Wash. D. C.)* **189**: 1088-1090. [757]

Prentice, I. C. 1980. Multidimensional scaling as a research tool in quaternary palynology: a review of theory and methods. *Rev. Palaeobot. Palyno.* **31**: 71-104. [270, 310, 328, 332]

Press, W. H., S. A. Teukolsky, W. T. Vetterling & B. P. Flannery. 2007. *Numerical recipes – The art of scientific computing. 3rd edition*. Cambridge Univ. Press, Cambridge. xxi + 1235 pp. [85, 503, 516, 718]

Priestley, M. B. 1964. The analysis of two dimensional stationary processes with discontinuous spectra. *Biometrika* **51**: 195-217. [793, 807]

Priestley, M. B. 1981a. *Spectral analysis and time series. 1. Univariate series*. Academic Press, London. xviii + 653 pp. [714]

Priestley, M. B. 1981b. *Spectral analysis and time series. 2. Multivariate series, prediction and control*. Academic Press, London. xviii + 237 pp. [714]

Prim, R. C. 1957. Shortest connection networks and some generalizations. *Bell Syst. Tech. J.* **36**: 1389-1401. [346]

Pugesek, B. H., A. Tomer & A. von Eye. 2003. *Structural equation modeling: applications in ecological and evolutionary biology*. Cambridge University Press, Cambridge. xiii + 409 pp. [593]

Quenouille, M. H. 1950. *Introductory statistics*. London. [46]

Quinghong, L. & S. Bråkenhielm. 1995. A statistical approach to decompose ecological variation. *Water Air Soil Pollut.* **85**: 1587-1592. [856]

Ragan, M. A. 1992. Phylogenetic inference based on matrix representation of trees. *Mol. Phylogenet. Evol.* **1**: 53-58. [417]

Rajski, C. 1961. Entropy and metric spaces. 44-45 *in*: C. Cherry [ed.] *Information theory*. Butterworths, London. [233]

Rand, W. M. 1971. Objective criteria for the evaluation of clustering methods. *J. Amer. Statist. Assoc.* **66**: 846-850. [413]

Rao, C. R. 1948. The utilization of multiple measurements in problems of biological classification. *J. Roy. Statist. Soc. B* **10**: 159-203. [674]

Rao, C. R. 1951. An asymptotic expansion of the distribution of Wilks' criterion. *Bull. Internat. Stat. Inst.* **33**: 177-181. [682]

Rao, C. R. 1952. *Advanced statistical methods in biometric research*. John Wiley & Sons, New York. xvii + 390 pp. [674]

Rao, C. R. 1964. The use and interpretation of principal component analysis in applied research. *Sankhyaá, Ser. A* **26**: 329-358. [381, 430, 451, 492, 630, 635, 637]

Rao, C. R. 1973. *Linear statistical inference and its applications. 2nd edition*. Wiley, New York. 625 pp. [635, 637]

Rao, C. R. 1982. Diversity and dissimilarity coefficients: a unified approach. *Theor. Popul. Biol.* **21**: 24-43. [255, 264]

Rao, C. R. 1995. A review of canonical coordinates and an alternative to correspondence analysis using Hellinger distance. *Qüestiió* (*Quaderns d'Estadística i Investigació Operativa*) **19**: 23-63. [310]

Ratkowsky, D. A. 1983. *Nonlinear regression modeling – A unified practical approach*. Marcel Dekker Inc., New York. viii + 276 pp. [538, 583]

Raup, D. M. & R. E. Crick. 1979. Measurement of faunal similarity in paleontology. *J. Paleontol.* **53**: 1213-1227. [293, 295]

Redford, A. J., R. M. Bowers, R. Knight, Y. Linhard & N. Fierer. 2010. The ecology of the phyllosphere: geographic and phylogenetic variability in the distribution of bacteria on tree leaves. *Environ. Microbiol.* **12**: 2885-2893. [518]

Rejmánek, M. & J. Leps. 1996. Negative associations can reveal interspecific competition and reversal of competitive hierarchies during succession. *Oikos* **76**: 161-168. [390]

Rendón, E., I. Abudez, A. Arizmendi & E. M. Quiroz. 2011. Internal versus external cluster validation indexes. *International Journal of Computers and Communications* **5**: 27-34. [418]

Rendu, J.-M. 1981. *An introduction to geostatistical methods of mineral evaluation*. South African Institute of Mining and Metallurgy, Johannesburg. 84 pp. [831]

Renfrew, C. & P. G. Bahn. 2008. *Archaeology: theories, methods and practice. 5th edition*. Thames & Hudson, London. 656 pp. [403]

Renshaw, E. & E. D. Ford. 1984. The description of spatial pattern using two-dimensional spectral analysis. *Vegetatio* **56**: 75-85. [793, 807]

Rényi, A. 1961. On measures of entropy and information. 547-561 *in*: J. Neyman [ed.] *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability.* University of California Press, Berkeley. [250]

Reynolds, C. S. 1987. Community organization in the freshwater plankton. Ch. 14. *in*: J. H. R. Gee & P. S. Giller [eds.] *Organization of communities: past and present*. Blackwell Scient. Publ., Oxford. [785]

Reyssac, J. & M. Roux. 1972. Communautés planctoniques dans les eaux de Côte d'Ivoire. Groupes d'espèces associées. *Mar. Biol. (Berl.)* **13**: 14-33. [317, 452, 479]

Ricker, W. E. 1973. Linear regression in fishery research. *J. Fish. Res. Board Can.* **30**: 409-434. [547]

Rigler, F. H. 1982. The relation between fisheries management and limnology. *Trans. Am. Fish. Soc.* **111**: 121-132. [226]

Ripley, B. D. 1981. *Spatial statistics*. Wiley, New York. x + 252 pp. [790, 793, 807]

Ripley, B. D. 1987. Spatial point pattern analysis in ecology. 407-429 *in*: P. Legendre, & L. Legendre [eds.] *Developments in numerical ecology.* NATO ASI Series, Vol. G 14. Springer-Verlag, Berlin. [790]

Robert, P. & Y. Escoufier. 1976. A unifying tool for linear multivariate statistical methods: the RV-coefficient. *Appl. Stat.–J. Roy. St. C* **25**: 257-265. [699]

Robertson, G. P. 1987. Geostatistics in ecology: interpolating with known variance. *Ecology* **68**: 744-748. [832]

Robinson, G. K. 1982. Behrens-Fisher problem. 205-209 *in*: S. Kotz & N. L. Johnson [eds.] *Encyclopedia of statistical sciences. Vol. 1*. Wiley, New York. [25]

Roche, C. 1978. Exemple de classification hiérarchique avec contrainte de contiguïté. Le partage d'Aix-en-Provence en quartiers homogènes. *C. Anal. Données* **3**: 289-305. [840]

Rodríguez, M. A. & P. Magnan. 1995. Application of multivariate analyses in studies of the organization and structure of fish and invertebrate communities. *Aquat. Sci.* **57**: 199-216. [855]

Rogers, D. J. & T. T. Tanimoto. 1960. A computer program for classifying plants. *Science (Wash. D. C.)* **132**: 1115-1118. [274]

Rohlf, F. J. 1963. Classification of *Aedes* by numerical taxonomic methods (Diptera: Culicidae). *Ann. Entomol. Soc. Am.* **56**: 798-804. [353]

Rohlf, F. J. 1970. Adaptive hierarchical clustering schemes. *Syst. Zool.* **19**: 58-82. [523]

Rohlf, F. J. 1972. An empirical comparison of three ordination techniques in numerical taxonomy. *Syst. Zool.* **21**: 271-280. [428]

Rohlf, F. J. 1974. Methods for comparing classifications. *Annu. Rev. Ecol. Syst.* **5**: 101-113. [413]

Rohlf, F. J. 1978. A probabilistic minimum spanning tree algorithm. *Information Processing Letters* **7**: 44-48. [349]

Rohlf, F. J. 1982a. Single linkage clustering algorithms. 267-284 *in*: P. R. Krishnaih [ed.] *Handbook of Statistics*. North-Holland, Amsterdam. [349]

Rohlf, F. J. 1982b. Consensus indices for comparing classifications. *Math. Biosci.* **59:** 131-144. [413, 418]

Rolecek, J., L. Tichy, D. Zeleny & M. Chytry. 2009. Modified TWINSPAN classification in which the hierarchy respects cluster heterogeneity. *J. Veg. Sci.* **20**: 596-602. [383]

Ross, G. J. S. 1990. *Nonlinear estimation*. Springer-Verlag, New York. viii + 189 pp. [538, 583, 716]

Rossi, R. E., D. J. Mulla, A. G. Journel & E. H. Franz. 1992. Geostatistical tools for modeling and interpreting ecological spatial dependence. *Ecol. Monogr.* **62**: 277-314. [790, 807, 811, 817, 832]

Roux, G. & M. Roux. 1967. À propos de quelques méthodes de classification en phytosociologie. *Rev. Stat. Appl.* **15**: 59-72. [464]

Roux, M. & J. Reyssac. 1975. Essai d'application au phytoplancton marin de méthodes statistiques utilisées en phytosociologie terrestre. *Ann. Inst. Océanogr. (Paris)* **51**: 89-97. [309]

Roxburgh, S. H. & P. Chesson. 1998. A new method for detecting species associations with spatially autocorrelated data. *Ecology* **79**: 2180-2192. [390]

Roy, M. L., A. G. Roy & P. Legendre. 2010. The relations between 'standard' fluvial habitat variables and turbulent flow at multiple scales in morphological units of a gravel-bed river. *River Res. Appl.* **26**: 439-455. [877]

Royston, J. P. 1982a. An extension of Shapiro and Wilk's *W* test for normality to large samples. *Appl. Statist.* **31**: 115-124. [191]

Royston, J. P. 1982b. Algorithm AS 177. Expected normal order statistics (exact and approximate). *Appl. Statist.* **31**: 161-165. [191]

Royston, J. P. 1982c. Algorithm AS 181. The *W* test for normality. *Appl. Statist.* **31**: 176-177. [191]

Rubenstein, D. I. & M. A. R. Koehl. 1977. The mechanisms of filter feeding: some theoretical considerations. *Am. Nat.* **111**: 981-994. [129]

Russell, P. F. & T. R. Rao. 1940. On habitat and association of species of anopheline larvae in south-eastern Madras. *J. Malar. Inst. India* **3**: 153-178. [277]

Sakai, A. K. & N. L. Oden. 1983. Spatial pattern of sex expression in silver maple (*Acer saccharinum* L.): Morisita's index and spatial autocorrelation. *Am. Nat.* 122: 489-508. [800]

Sale, P. F. 1978. Coexistence of coral reef fishes – A lottery for living space. *Environ. Biol. Fishes* **3**: 85-102. [619]

Sanders, H. L. 1960. Benthic studies in Buzzards Bay. III. The structure of the soft-bottom community. *Limnol. Oceanogr.* **5**: 138-153. [406]

Sanders, H. L. 1968. Marine benthic diversity: a comparative study. *Am. Nat.* **102**: 243-282. [251]

Särndal, C.-E. 1978. Design-based and model-based inference in survey sampling. *Scand. J. Stat.* **5**: 27-52. [6]

SAS Institute Inc. 2011. SAS/STAT 9.3 User's Guide. SAS Institute Inc., Cary, North Carolina. xii + 8621 pp. [386]

Scheider, W. & P. Wallis. 1973. An alternate method of calculating the population density of monsters in Loch Ness. *Limnol. Oceanogr.* **18**: 343. [226]

Scherrer, B. 1982. Techniques de sondage en écologie. 63-162 *in*: S. Frontier [Ed.] *Stratégies d'échantillonnage en écologie*. Collection d'Écologie, No. 17. Masson, Paris et les Presses de l'Université Laval, Québec. [21]

Schmidt, E. 1907. Zur Theorie der linearen und nichtlinearen Integralgleichungen. I Teil. Entwicklung willkürlichen Funktionen nach System vorgeschriebener. *Math. Ann.* 63: 433-476. [103]

Schneider, D. C. 1994. *Quantitative ecology – Spatial and temporal scaling*. Academic Press, San Diego. xv + 395 pp. [109, 124, 786]

Schnell, G. D. 1970. A phenetic study of the suborder Lari (Aves). I. Methods and results of principal components analyses. *Syst. Zool.* **19**: 35-57. [523]

Schoener, T. W. 1970. Nonsynchronous spatial overlap of lizards in patchy habitats. *Ecology* **51**: 408-418. [243]

Schoener, T. W. & G. H. Adler. 1991. Greater resolution of distributional complementarities by controlling for habitat affinities: A study with Bahamian lizards and birds. *Am. Nat.* **137**: 669-692. [243]

Schönemann, P. H. 1966. A generalized solution of the orthogonal Procrustes problem. *Psychometrika* **31**: 1-10. [703]

Schönemann, P. H. & R. M. Carroll. 1970. Fitting one matrix to another under choice of a central dilation and a rigid motion. *Psychometrika* **35**: 245-256. [703]

Schuster, A. 1898. On the investigation of hidden periodocities with application to a supposed 26 day period of meteorological phenomena. *Terrestrial Magnetism* **3**: 13-41. [748, 793]

Searle, S. R. 1987. *Linear models for unbalanced data*. Wiley. xxiv + 536 pp. [568]

Shannon, C. E. 1948. A mathematical theory of communications. *Bell System Technical Journal* **27**: 379-423. [221, 225]

Shao, K. & F. J. Rohlf. 1983. Sampling distributions of consensus indices when all bifurcating trees are equally likely. 132-137 *in*: J. Felsenstein [ed.] *Numerical taxonomy*. NATO ASI Series, Vol. G-1. Springer-Verlag, Berlin. [529]

Shao, K. & R. R. Sokal. 1986. Significance tests of consensus indices. *Syst. Zool.* **35**: 582-590. [529]

Shapiro, S. S. & M. B. Wilk. 1965. An analysis of variance test for normality (complete samples). *Biometrika* **52**: 591-611. [190, 191]

Sharma, S., P. Legendre, M. De Cáceres & D. Boisclair. 2011. The role of environmental and spatial processes in structuring native and non-native fish communities across thousands of lakes. *Ecography* **34**: 762-771. [893]

Sharpe, D. M., G. R. Guntenspergen, C. P. Dunn, L. A. Leitner & F. Stearns. 1987. Vegetation dynamics in a southern Wisconsin agricultural landscape. 137-155 *in*: M. G. Turner [ed.] *Landscape heterogeneity and disturbance*. Ecological Studies 64. Springer-Verlag, New York. [489, 490]

Sheldon, R. W. & S. R. Kerr. 1972. The population density of monsters in Loch Ness. *Limnol. Oceanogr.* **17**: 796-798. [226]

Shepard, R. N. 1962. The analysis of proximities: multidimensional scaling with an unknown distance function. *Psychometrika* **27**: 125-139. [414, 427, 512]

Shepard, R. N. 1966. Metric structures in ordinal data. *J. Math. Psychol.* **3**: 287-315. [512]

Shipley, B. 2002. *Cause and correlation in biology: a user's guide to path analysis, structural equations and causal inference. 1st paperback edition*. Cambridge University Press, Cambridge. xii + 317 pp. [593]

Shumway, R. H. & D. S. Stoffer. 1982. An approach to time series smoothing and forecasting using the EM algorithm. *J. Time Ser. Anal.* **3**: 253-264. [56]

Shumway, R. H. & D. S. Stoffer. 2011. *Time series analysis and its applications – With R examples. 3rd edition*. Springer Texts in Statistics. Springer, New York. viii + 596 pp. [714, 722]

Sidák, Z. 1967. Rectangular confidence regions for the means of multivariate normal distributions. *J. Amer. Statist. Assoc.* 62 : 626-633. [23]

Siegel, S. 1956. *Nonparametric statistics for the behavioral sciences*. McGraw-Hill Series in Psychology, McGraw-Hill, New York. xvii + 312 pp. [25, 31, 201, 202, 213, 215, 230, 396]

Siegel, S. and N. J. Castellan, Jr. 1988. *Nonparametric statistics for the behavioral sciences. 2nd edition*. McGraw-Hill, New York. xxiii + 399 pp. [25, 201, 213, 215]

Simon, H. A. 1962. The architecture of complexity. *Proc. Am. Philos. Soc.* **106**: 467-482. [9]

Simpson, E. H. 1949. Measurement of diversity. *Nature (Lond.)* **163**: 688. [253]

Slutzky, E. 1927. The summation of random causes as the source of cyclic processes. In Russian. Translation revised by the author in 1937. *Econometrica* **5**: 105-146. [725]

Smol, J. P. & E. F. Stoermer [eds.]. 2010. *The diatoms – Applications for the environmental and earth sciences. 2nd edition*. Cambridge University Press, New York. xviii + 667 pp. [672]

Smouse, P. E., J. C. Long & R. R. Sokal. 1986. Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *Syst. Zool.* **35**: 627-632. [604, 606]

Sneath, P. H. A. 1957. The application of computers to taxonomy. *J. Gen. Microbiol.* **17**: 201-226. [341]

Sneath, P. H. A. 1966. A comparison of different clustering methods as applied to randomly-spaced points. *Classification Soc. Bull.* **1**: 2-18. [352]

Sneath, P. H. A. & R. R. Sokal. 1973. *Numerical taxonomy – The principles and practice of numerical classification*. W. H. Freeman, San Francisco. xv + 573 pp. [35, 36, 44, 270, 289, 346–349, 352, 353, 355, 357, 360, 369, 373, 412, 418, 523]

Snedecor, G. W. & W. G. Cochran. 1967. *Statistical methods*. 6th Iowa State Univ. Press, Ames. xiv + 593 pp. [25]

Soares, A., J. Távora, L. Pinheiro, C. Freitas & J. Almeida. 1992. *Predicting probability maps of air pollution concentration – A case study on Barreiro/Seixal industrial area*. Fourth international geostatistics congress, 13-18 september 1992, Troya, Portugal. [832]

Sokal, R. R. 1979. Ecological parameters inferred from spatial correlograms. 167-196 *in*: G. P. Patil & M. L. Rosenzweig [eds.] *Contemporary quantitative ecology and related ecometrics*. *Vol. 12*. Statistical Ecology Series. International Co-operative Publ. House, Fairland, Maryland. [804]

Sokal, R. R. 1986. Spatial data analysis and historical processes. 29-43 *in*: E. Diday *et al*. [eds.] *Data analysis and informatics, IV*. North-Holland, Amsterdam. [529, 819]

Sokal, R. R., I. A. Lengyel, P. A. Derish, M. C. Wooten & N. L. Oden. 1987. Spatial autocorrelation of ABO serotypes in mediaeval cemeteries as an indicator of ethnic and familial structure. *J. Archaeol. Sci*. **14**: 615-633. [31]

Sokal, R. R. & C. D. Michener. 1958. A statistical method for evaluating systematic relationships. *Univ. Kans. Sci. Bull*. **38**: 1409-1438. [274, 340, 355]

Sokal, R. R., N. L. Oden & B. A. Thomson. 1997a. A simulation study of microevolutionary inferences by spatial autocorrelation analysis. *Biol. J. Linn. Soc*. **60**: 73-93. [805]

Sokal, R. R., N. L. Oden, J. Walker & D. M. Waddle. 1997b. Using distance matrices to choose between competing theories and an application to the origin of modern humans. *J. Hum. Evol*. **32**: 501-522. [805]

Sokal, R. R. & F. J. Rohlf. 1962. The comparison of dendrograms by objective methods. *Taxon* **11**: 33-40. [346, 412]

Sokal, R. R. & F. J. Rohlf. 1981. Taxonomic congruence in Leptopodomorpha re-examined. *Syst. Zool*. **30**: 309-325. [48]

Sokal, R. R. & F. J. Rohlf. 1995. *Biometry – The principles and practice of statistics in biological research. 3rd edition*. W. H. Freeman, New York. xix + 887 pp. [23, 26, 31, 48–50, 190, 191, 201, 230, 231, 235, 244, 538, 543, 546, 548, 550, 554, 562, 592, 593, 595, 601, 610, 720, 848]

Sokal, R. R. & P. H. A. Sneath. 1963. *Principles of numerical taxonomy*. W. H. Freeman, San Francisco. xvi + 359 pp. [xv, 270, 274–277]

Sokal, R. R. & J. D. Thomson. 1987. Applications of spatial autocorrelation in ecology. 431-466 *in*: P. Legendre & L. Legendre [eds.] *Developments in numerical ecology*. NATO ASI Series, Vol. G-14. Springer-Verlag, Berlin. [800]

Sokolove, P. G. & W. N. Bushell. 1978. The chi square periodogram: its utility for analysis of circadian rhythms. *J. Theor. Biol*. **72**: 131-160. [741]

Somers, K. M. & R. H. Green. 1993. Seasonal patterns in trap catches of the crayfish Cambarus bartoni and *Orconectes virilis* in six south-central Ontario lakes. *Can. J. Zool*. **71**: 1136-1145. [601]

Song, C. Q., B. Y. Wang & X. J. Sun. 1996. Implication of paleovegetational changes in Diaojiao Lake, Inner Mongolia. (In Chinese). *Acta Bot. Sin*. **38**: 568-575. [773]

Sørensen, T. 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analysis of the vegetation on Danish commons. *Biol. Skr*. **5**: 1-34. [350, 351]

Sousa, W. P. 1979. Experimental investigations of disturbance and ecological succession in a rocky intertidal algal community. *Ecol. Monogr*. **49**: 227-254. [785]

Southwood, T. R. E. 1966. *Ecological methods with particular reference to the study of insect populations*. Chapman and Hall, London. xviii + 391 pp. [50]

Southwood, T. R. E. 1987. The concept and nature of the community. 3-27 *in*: J. H. R. Gee & P. S. Giller [eds.] *Organization of communities: past and present*. Blackwell Scientific Publ., Oxford. [785]

Soyer, J. 1970. Bionomie benthique du plateau continental de la côte catalane française. III. Les peuplements de copépodes harpacticoïdes (Crustacea). *Vie Milieu **21**: 337-511*. [406]

Späth, H. 1975. *Cluster-Analyse-Algorithmen zur Objektklassifizierung und Datenreduktion*. R. Oldenbourg Verlag, München. 217 pp. [386]

Späth, H. 1980. *Cluster analysis algorithms*. Ellis Horwood, Chichester. 226 pp. [386]

Spearman, C. 1904. "General intelligence," objectively determined and measured. *Am. J. Psychol*. **15**: 201-292. [xv, 340]

Sprules, W. G. 1980. Nonmetric multidimensional scaling analyses of temporal variation in the structure of limnetic zooplankton communities. *Hydrobiologia* **69**: 139-146. [518, 768]

Steel, M. A. & D. Penny. 1993. Distributions of tree comparison metrics. Some new results. *Syst. Biol*. **42**: 126-141. [529]

Steiner, D., K. Baumberger & H. Maurer. 1969. Computer-processing and classification of multi-variate information from remote sensing imagery – A review of the methodology as applied to a sample of agricultural crops. 895-907 *in*: *Proc. sixth int. Symp. on Remote Sensing of Environment. Vol. II*. Willow Run Laboratories, Inst. of Science and Technology, Univ. of Michigan. [688]

Stephens, M. A. 1974. EDF statistics for goodness of fit and some comparisons. *J. Amer. Statist. Assoc*. **69**: 730-737. [188, 190, 191]

Stephenson, W., W. T. Williams & S. D. Cook. 1972. Computer analyses of Petersen's original data on bottom communities. *Ecol. Monogr.* **42**: 387-415. [306]

Steven, D. M. & R. Glombitza. 1972. Oscillatory variation of a phytoplankton population in a tropical ocean. *Nature (Lond.)* **237**: 105-107. [733, 735, 737, 738, 750, 751]

Stewart, D. & W. Love. 1968. A general canonical correlation index. *Psychol. Bull*. **70**: 160-163. [630, 695]

Stewart, G. W. 1993. On the early history of the singular value decomposition. *SIAM Review* 35: 551-566. [103]

Stewart-Oaten, A., W. M. Murdoch & K. R. Parker. 1986. Environmental impact assessment: 'pseudoreplication' in time? *Ecology* **67**: 929-940. [267]

St-Louis, N. & P. Legendre. 1982. A water quality index for lake beaches. *Water Res*. **16**: 945-948. [723]

Student [W. S. Gosset]. 1914. The elimination of spurious correlation due to position in time or space. *Biometrika* **10**: 179-180. [822]

Sturges, H. A. 1926. The choice of a class interval. *J. Amer. Statist. Assoc*. **21**: 65-66. [796]

Swan, J. M. A. 1970. An examination of some ordination problems by use of simulated vegetational data. *Ecology* **51**: 89-102. [483, 487]

Taguchi, S. 1976. Short-term variability of photosynthesis in natural marine phytoplankton populations. *Mar. Biol. (Berl.)* **37**: 197-207. [754]

Tardif, J., P. Dutilleul & Y. Bergeron. 1998. Variations in periodicities of the ring width of black ash (*Fraxinus nigra* Marsh.) in relation to flooding and ecological site factors at Lake Duparquet in northwestern Quebec. *Biol. Rhythm Res.* **29**: 1-29. [753]

Taylor, L. R. 1961. Aggregation, variance, and the mean. *Nature* **189**: 732-735. [50]

Teissier, G. 1948. La relation d'allométrie: sa signification statistique et biologique. *Biometrics* **4**: 14-53. [547, 549]

ter Braak, C. J. F. 1983. Principal components biplots and alpha and beta diversity. *Ecology* **64**: 454-462. [443]

ter Braak, C. J. F. 1985. Correspondence analysis of incidence and abundance data: properties in terms of a unimodal response model. *Biometrics* **41**: 859-873. [478]

ter Braak, C. J. F. 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology* **67**: 1167-1179. [410, 660, 661, 666]

ter Braak, C. J. F. 1987a. The analysis of vegetation-environment relationships by canonical correspondence analysis. *Vegetatio* **69**: 69-77. [661, 666]

ter Braak, C. J. F. 1987b. Calibration. 78-90 *in*: R. H. G. Jongman, C. J. F. ter Braak & O. F. R. van Tongeren [eds.] *Data analysis in community and landscape ecology*. Pudoc, Wageningen, The Netherlands. Reissued in 1995 by Cambridge Univ. Press, Cambridge. [672, 674]

ter Braak, C. J. F. 1987c. Ordination. 91-173 *in*: R. H. G. Jongman, C. J. F. ter Braak & O. F. R. van Tongeren [eds.] *Data analysis in community and landscape ecology*. Pudoc, Wageningen, The Netherlands. Reissued in 1995 by Cambridge Univ. Press, Cambridge. [271, 428, 456, 457, 470, 471, 477, 485, 486, 490, 491, 526, 632, 635, 642, 661, 708]

ter Braak, C. J. F. 1988a. Partial canonical correspondence analysis. 551-558 *in*: H.-H. Bock [ed.] *Classification and related methods of data analysis*. North-Holland, Amsterdam. [667]

ter Braak, C. J. F. 1988b. CANOCO – an extension of DECORANA to analyze species-environment relationships. *Vegetatio* **75**: 159-160. [490, 629, 661]

ter Braak, C. J. F. 1988c. *CANOCO – a FORTRAN program for canonical community ordination by [partial] [detrended] [canonical] correspondence analysis, principal component analysis and redundancy analysis (version 2.1)*. Agricultural Mathematics Group, Ministry of Agriculture and Fisheries, Wageningen. ii + 95 pp. [490, 661]

ter Braak, C. J. F. 1990. *Update notes: CANOCO version 3.10*. Agricultural Mathematics Group, Wageningen. [21, 470, 471, 490, 634, 651, 658, 661]

ter Braak, C. J. F. 1992. Permutation versus bootstrap significance tests in multiple regression and ANOVA. 79-86 *in*: K.-H. Jöckel, G. Rothe & W. Sendler [eds.] *Bootstrapping and related techniques*. Springer-Verlag, Berlin. [651]

ter Braak, C. J. F. 1994. Canonical community ordination. Part I: Basic theory and linear methods. *Écoscience* **1**: 127-140. [443, 639]

ter Braak, C. J. F. 1995. Non-linear methods for multivariate statistical calibration and their use in palaeoecology: a comparison of inverse (*k*-nearest neighbours, partial least squares and weighted averaging partial least squares) and classical approaches. *Chemometrics Intelligent Lab. Syst.* **28**: 165-180. [672, 709]

ter Braak, C. J. F., A. Cormont & S. Dray. 2012. Improved testing of species trait-environment relationships in the fourth-corner problem. *Ecology* **93**: 1525-1526. [613, 620]

ter Braak, C. J. F. & S. Juggins. 1993. Weighted averaging partial least squares regression (WA-PLS): an improved method for reconstructing environmental variables from species assemblages. *Hydrobiologia* **269**: 485-502. [672, 673]

ter Braak, C. J. F. & C. W. N. Looman. 1987. Regression. 29-77 *in*: R. H. G. Jongman, C. J. F. ter Braak & O. F. R. van Tongeren [eds.] *Data analysis in community and landscape ecology*. Pudoc, Wageningen, The Netherlands. Reissued in 1995 by Cambridge Univ. Press, Cambridge. [568, 588]

ter Braak, C. J. F. & I. C. Prentice. 1988. A theory of gradient analysis. *Adv. Ecol. Res.* **18**: 271-317. [271, 327, 632, 672]

ter Braak, C. J. F. & A. P. Schaffers. 2004. Co-correspondence analysis: a new ordination method to relate two community compositions. *Ecology* **85**: 834-846. [699]

ter Braak, C. J. F. & P. Smilauer. 1998. *CANOCO reference manual and user's guide to Canoco for Windows – Software for canonical community ordination (version 4)*. Microcomputer Power, Ithaca, New York. [490, 519, 661]

ter Braak, C. J. F. & P. Smilauer. 2002. CANOCO reference manual and CanoDraw for Windows user's guide – Software for canonical community ordination (version 4.5). Microcomputer Power, Ithaca, New York. 500 pp. [558, 651]

ter Braak, C. J. F. & H. van Dam. 1989. Inferring pH from diatoms: a comparison of old and new calibration methods. *Hydrobiologia* **178**: 209-223. [672]

ter Braak, C. J. F. & P. F. M. Verdonschot. 1995. Canonical correspondence analysis and related multivariate methods in aquatic ecology. *Aquat. Sci.* **57**: 255-289. [666]

Thiessen, A. W. 1911. Precipitation averages for large areas. *Monthly Weather Review* **39**: 1082-1084. [839]

Thioulouse, J., D. Chessel & S. Champely. 1995. Multivariate analysis of spatial patterns: a unified approach to local and global structures. *Environ. Ecol. Stat.* **2**: 1-14. [852]

Thompson, S. K. 1992. *Sampling*. Wiley, New York. 368 pp. [241]

Thorrington-Smith, M. 1971. West Indian Ocean phytoplankton: a numerical investigation of phytohydrographic regions and their characteristic phytoplankton associations. *Mar. Biol. (Berl.)* **9**: 115-137. [317, 391]

Thrush, S. F., D. C. Schneider, P. Legendre, R. B. Whitlatch, P. K. Dayton, J. E. Hewitt, A. H. Hines, V. J. Cummings, S. M. Lawrie, J. Grant, R. D. Pridmore, S. J. Turner & B. H. McArdle. 1997. Scaling-up from experiments to complex ecological systems: where to next? *J. Exp. Mar. Biol. Ecol.* 216: 243-254. [788]

Tinkler, K. J. 1972. The physical interpretation of eigenfunctions of dichotomous matrices. *T. I. Brit. Geogr.* **55**: 17-46. [859, 884]

Tobler, W. 1970. A computer movie simulating urban growth in the Detroit region. *Econ. Geogr.* **46**: 234-240. [859]

Torgerson, W. S. 1958. *Theory and methods of scaling*. Wiley, New York. 460 pp. [492]

Toussaint, G. 1980. The relative neighbourhood graph of a finite planar set. *Pattern Recogn.* **12**: 261-268. [835, 838]

Tranter, D. J. & P. E. Smith. 1968. Filtration performance. 27-56 *in*: Zooplankton sampling. Monographs on Oceanographic Methodology, 2. UNESCO, Paris. [129]

Trexler, J. C. & J. Travis. 1993. Nontraditional regression analyses. *Ecology* **74**: 1629-1637. [588, 591]

Troussellier, M. & P. Legendre. 1981. A functional evenness index for microbial ecology. *Microb. Ecol.* **7**: 283-296. [257]

Troussellier, M. & P. Legendre. 1989. Dynamics of fecal coliform and culturable heterotroph densities in an eutrophic ecosystem: stability of models and evolution of these bacterial groups. *Microb. Ecol.* **17**: 227-235. [564]

Troussellier, M., P. Legendre & B. Baleux. 1986. Modeling of the evolution of bacterial densities in an eutrophic ecosystem (sewage lagoons). *Microb. Ecol.* **12**: 355-379. [564, 596]

Tukey, J. W. 1958. Bias and confidence in not-quite large samples. (Abstract). *Ann. Math. Statist.* **29**: 614. [31]

Tukey, J. W. 1977. *Exploratory data analysis*. Addison-Wesley, Reading, Mass. xvi + 688 pp. [722]

Tuomisto, H. 2010. A diversity of beta diversities: straightening up a concept gone awry. Part 2. Quantifying beta diversity and related phenomena. *Ecography* **33**: 23-45. [260]

Tuomisto, H. & A. D. Poulsen. 2000. Pteridophyte diversity and species composition in four Amazonian rain forests. *J. Veg. Sci.* **11**: 383-396. [870]

Tuomisto, H. & K. Ruokolainen. 1994. Distribution of *Pteridophyta* and Melastomataceae along an edaphic gradient in an Amazonian rain forest. *J. Veg. Sci.* **5**: 25-34. [774]

Tuomisto, H. & K. Ruokolainen. 2006. Analyzing or explaining beta diversity? Understanding the targets of different methods of analysis. *Ecology* **87**: 2697–2708. [607]

Tuomisto, H. & K. Ruokolainen. 2008. Analyzing or explaining beta diversity: Reply. *Ecology* **89**: 3244-3256. [607]

Tuomisto, H., K. Ruokolainen, M. Aguilar & A. Sarmiento. 2003. Floristic patterns along a 43-km long transect in an Amazonian rain forest. *J. Ecol.* **91**: 743-756. [402, 774, 779]

Ulrych, T. J. & T. N. Bishop. 1975. Maximum entropy spectral analysis and autoregressive decomposition. *Rev. Geophys. Space Phys.* **13**: 183-200. [765]

Ulrych, T. J. & R. W. Clayton. 1976. Time series modelling and maximum entropy. *Phys. Earth Planet. Inter.* **12**: 188-200. [765]

Ulrych, T. J. & O. Jensen. 1974. Cross-spectral analysis using maximum entropy. *Geophysics* **39**: 353-354. [766]

Underwood, A. J. 1991. Beyond BACI: experimental designs for detecting human environmental impacts on temporal variations in natural populations. *Aust. J. Mar. Freshwater Res.* **42**: 569-587. [267]

Underwood, A. J. 1992. Beyond BACI: the detection of environmental impacts on populations in the real, but variable, world. *J. Exp. Mar. Biol. Ecol.* **161**: 145-178. [267]

Underwood, A. J. 1994. On beyond BACI: sampling designs that might reliably detect environmental disturbances. *Ecol. Appl.* **4**: 3-15. [267]

Underwood, A. J. 1997. *Experiments in ecology – Their logical design and interpretation using analysis of variance*. Cambridge University Press, Cambridge. xviii + 504 pp. [535]

Upton, G. J. G. 1978. *The analysis of cross-tabulated data*. John Wiley & Sons, New York. xii + 148 pp. [235]

Upton, G. J. G. & B. Fingleton. 1985. *Spatial data analysis by example. Vol. 1: Point pattern and quantitative data*. Wiley, Chichester. xi + 410 pp. [601, 790, 835, 839]

Upton, G. J. G. & B. Fingleton. 1989. *Spatial data analysis by example. Vol. 2: categorical and directional data*. Wiley, Chichester. xi + 416 pp. [790]

Urban, D. L., E. S. Minor, E. A. Treml & R. S. Schick. 2009. Graph models of habitat mosaics. *Ecol. Lett.* **12**: 260-273. [884]

van den Brink, P. J., P. J. den Besten, A. bij de Vaate & C. J. F. ter Braak. 2009. Principal response curves technique for the analysis of multivariate biomonitoring time series. *Environ. Monit. Assess.* **152**: 271-281. [657]

van den Brink, P. J. & C. J. F. ter Braak. 1998. Multivariate analysis of stress in experimental ecosystems by Principal Response Curves and similarity analysis. *Aquat. Ecol.* **32**: 163-178. [657]

van den Brink, P. J. & C. J. F. ter Braak. 1999. Principal Response Curves: Analysis of time-dependent multivariate responses of a biological community to stress. *Environ. Toxicol. Chem.* **18**: 138-148. [657]

van den Brink, P. J., N. W. van den Brink & C. J. F. ter Braak. 2003. Multivariate analysis of ecotoxicological data using ordination: Demonstrations of utility on the basis of various examples. *Australas. J. Ecotox.* **9**: 141-156. [657]

van Rijckevorsel, J. L. A. & J. de Leeuw. 1988. *Component and correspondence analysis – Dimension, reduction by functional approximation*. John Wiley & Sons, Chichester. xiii + 146 pp. [464]

Veech, J. A., K. S. Summerville, T. O. Crist & J. C. Gering. 2002. The additive partitioning of species diversity: recent revival of an old idea. *Oikos* **99**: 3-9. [260]

Vellend, M. 2004. Parallel effects of land-use history on species diversity and genetic diversity of forest herbs. *Ecology* **85**: 3043-3055. [295]

Vellend, M., K. Verheyen, K. M. Flinn, H. Jacquemyn, A. Kolb, H. van Calster, G. Peterken, B. J. Graae, J. Bellemare, O. Honnay, J. Brunet, M. Wulf, F. Gerhardt & M. Hermy. 2007. Homogenization of forest plant communities and weakening of species-environment relationships via agricultural land use. *J. Ecol.* **95**: 565-573. [295]

Venables, W. N. & B. D. Ripley. 2002. *Modern applied statistics with S. 4th edition*. Springer-Verlag, New York. xi + 495 pp. [33, 714, 722, 729, 731]

Venrick, E. L. 1971. Recurrent groups of diatoms in the North Pacific. *Ecology* **52**: 614-625. [393]

Ver Hoef, J. M. & D. C. Glenn-Lewin. 1989. Multiscale ordination: a method for detecting pattern at several scales. *Vegetatio* **82**: 59-67. [894]

Verly, G., M. David, A. G. Journel, & A. Marechal. 1984. *Geostatistics for natural resources characterization. Parts 1 and 2*. Reidel, Dordrecht. xxi + 585 pp., xii + 506 pp. [832]

Verneaux, J. 1973. Cours d'eau de Franche-Comté (Massif du Jura). Recherches écologiques sur le réseau hydrographique du Doubs – Essai de biotypologie. *Annales Scientifiques de l'Université de Franche-Comté, Biologie Animale* **3**: 1-260. [659, 695]

Villéger, S., N. W. H. Mason & D. Mouillot. 2008. New multidimensional functional diversity indices for a multifaceted framework in functional ecology. *Ecology* **89**: 2290-2301. [255]

Vinod, H. D. 2011. *Hands-on matrix algebra using R: active and motivated learning with applications*. World Scientific Publishing, Hackensack, New Jersey. 348 pp. [60]

Visscher, J. P. 1928. Reactions of the cyprid larvae of barnacles at the time of attachment. *Biol. Bull. (Woods Hole)* **54**: 327-335. [805]

Voronoï, G. F. 1909. Recherches sur les paralléloèdres primitifs. *Journal für die reine und angewandte Mathematik* **136**: 67-179. [839]

Wagner, H. H. 2003. Spatial covariance in plant communities: integrating ordination, variogram modeling, and variance testing. *Ecology* **84**: 1045-1057. [813, 814, 894]

Wagner, H. H. 2004. Direct multi-scale ordination with canonical correspondence analysis. *Ecology* **85**: 342-351. [814, 894]

Wagner, H. H. & M.-J. Fortin. 2005. Spatial analysis of landscapes: concepts and statistics. *Ecology* **86**: 1975-1987. [896]

Walliser, B. 1977. *Systèmes et modèles*. Seuil, Paris. [38]

Ward, J. H. 1963. Hierarchical grouping to optimize an objective function. *J. Amer. Statist. Assoc.* **58**: 236-244. [360, 365, 419]

Wartenberg, D., S. Ferson & F. J. Rohlf. 1987. Putting things in order: a critique of detrended correspondence analysis. *Am. Nat.* **129**: 434-448. [483, 486]

Warton, D. I., I. J. Wright, D. S. Falster & M. Westoby. 2006. Bivariate line-fitting methods for allometry. *Biol. Rev.* **81**: 259-291. [550, 552, 554, 555]

Watson, L., W. T. Williams & G. N. Lance. 1966. Angiosperm taxonomy: a comparative study of some novel numerical techniques. *J. Linn. Soc. Lond. Bot.* **59**: 491-501. [310]

Webster, R. 1973. Automatic soil-boundary location from transect data. *Math. Geol.* **5**: 27-37. [770]

Webster, R. & T. M. Burgess. 1984. Sampling and bulking strategies for estimating soil properties in small regions. *J. Soil Sci.* **35**: 127-140. [21]

Wegman, E. J. & I. W. Wright. 1983. Splines in statistics. *J. Amer. Statist. Assoc.* **78**: 351-365. [589]

Whittaker, E. T. & G. Robinson. 1924. *The calculus of observations – A treatise on numerical mathematics*. Blackie & Son, London. xvi + 396 pp. [740]

Whittaker, J. 1984. Model interpretation from the additive elements of the likelihood function. *Appl. Statist.* **33**: 52-64. [580, 582]

Whittaker, R. H. 1952. A study of summer foliage insect communities in the Great Smoky Mountains. *Ecol. Monogr.* **22**: 1-44. [305]

Whittaker, R. H. 1956. Vegetation of the Great Smoky Mountains. *Ecol. Monogr.* **26**: 1-80. [12, 582, 669, 707, 785]

Whittaker, R. H. 1960. Vegetation of the Siskiyou Mountains, Oregon and California. *Ecol. Monogr.* **30**: 279-338. [260, 478]

Whittaker, R. H. 1962. Classification of natural communities. *Bot. Rev.* **28**: 1-239. [337, 389]

Whittaker, R. H. 1967. Gradient analysis of vegetation. *Biol. Rev. (Camb.)* **42**: 207-264. [271, 478, 482]

Whittaker, R. H. 1972. Evolution and measurement of species diversity. *Taxon* **21**: 213-251. [249, 260, 261, 271, 319]

Whittaker, R. H. & H. G. Gauch. 1973. Evaluation of ordination techniques. 287-321 *in*: R. H. Whittaker [ed.] *Handbook of vegetation science. Part V.* Dr. W. Junk, The Hague. [452]

Whittington, H. B. & C. P. Hughes. 1972. Ordovician geography and faunal provinces deduced from trilobite distribution. *Phil. Trans. R. Soc. Lond. B* **263**: 235-278. [519]

Whittle, P. 1963. On the fitting of multivariate autoregressions and the approximate canonical factorization of a spectral density matrix. *Biometrika* **50**: 129-154. [782]

Wiens, J. A. 1989. Spatial scaling in ecology. *Funct. Ecol.* **3**: 385-397. [786–788]

Wieser, W. 1960. Benthic studies in Buzzards Bay. II. The meiofauna. *Limnol. Oceanogr.* **5**: 121-137. [404, 406]

Wilhm, J. L. 1968. Use of biomass units in Shannon's formula. *Ecology* **49**: 153-156. [250]

Wilkinson, G. N., S. R. Eckert, T. W. Hancock & O. Mayo. 1983. Nearest neighbour (NN) analysis of field experiments. *J. R. Stat. Soc. Ser. B* **45**: 151-211. [20]

Wilks, S. S. 1932. Certain generalizations in the analysis of variance. *Biometrika* **24**: 471-494. [682]

Wilks, S. S. 1935. The likelihood test of independence in contingency tables. *Ann. Math. Statist.* **6**: 190-196. [230]

Williams, B. K. 1983. Some observations on the use of discriminant analysis in ecology. *Ecology* **64**: 1283-1291. [588]

Williams, B. K. & K. Titus. 1988. Assessment of sampling stability in ecological applications of discriminant analysis. *Ecology* **69**: 1275-1285. [708]

Williams, D. A. 1976a. Improved likelihood ratio tests for complete contingency tables. *Biometrika* **63**: 33-37. [230, 238]

Williams, E. J. 1952. Use of scores for the analysis of association in contingency tables. *Biometrika* **39**: 274-289. [476]

Williams, W. T. 1976b. Hierarchical divisive strategies. *In*: W. T. Williams [ed.] *Pattern analysis in agricultural science*. CSIRO, Melbourne, Australia. [381]

Williams, W. T., H. J. Clay & J. S. Bunt. 1982. The analysis, in marine ecology, of three-dimensional data matrices with one dimension of variable length. *J. Exp. Mar. Biol. Ecol.* **60**: 189-196. [269]

Williams, W. T. & M. B. Dale. 1965. Fundamental problems in numerical taxonomy. *Adv. Bot. Res.* **2**: 35-68. [270, 380]

Williams, W. T. & J. M. Lambert. 1959. Multivariate methods in plant ecology. I. Association-analysis in plant communities. *J. Ecol.* **47**: 83-101. [xii, 377]

Williams, W. T. & J. M. Lambert. 1961. Multivariate methods in plant ecology. III. Inverse association-analysis. *J. Ecol.* **49**: 717-729. [379]

Williams, W. T., J. M. Lambert & G. N. Lance. 1966. Multivariate methods in plant ecology. V. Similarity analyses and information-analysis. *J. Ecol.* **54**: 427-445. [372, 375]

Williams, W. T., G. N. Lance, M. B. Dale & H. T. Clifford. 1971. Controversy concerning the criteria for taxonometric strategies. *Computer J.* **14**: 162-165. [341]

Williams, W. T., G. N. Lance, L. J. Webb, J. G. Tracey & M. B. Dale. 1969. Studies in the numerical analysis of complex rain-forest communities. III. The analysis of successional data. *J. Ecol.* **57**: 515-535. [768]

Williams, W. T. & W. Stephenson. 1973. The analysis of three-dimensional data (sites x species x times) in marine ecology. *J. Exp. Mar. Biol. Ecol.* **11**: 207-227. [269]

Wilson, B. & R. A. Dawe. 2006. Detecting seasonality using time series analysis: comparing foraminiferal population dynamics with rainfall data. *J. Foramin. Res.* **36**: 108-115. [734, 739]

Wirth, M., G. F. Estabrook & D. J. Rogers. 1966. A graph theory model for systematic biology, with an example for the Oncidiinae (Orchidaceae). *Syst. Zool.* **I5**: 59-69. [412, 419]

Wishart, D. 1969. An algorithm for hierarchical classifications. *Biometrics* **25**: 165-170. [365]

Wold, S. 1974. Spline functions in data analysis. *Technometrics* **16**: 1-11. [589]

Wolda, H. 1981. Similarity indices, sample size and diversity. *Oecologia (Berl.)* **50**: 296-302. [323]

Wollenberg, A. L. van den. 1977. Redundancy analysis. An alternative for canonical correlation analysis. *Psychometrika* **42**: 207-219. [630, 635, 637]

Womble, W. H. 1951. Differential systematics. *Science* **114**: 315-322. [844]

Wright, S. 1921. Correlation and causation. *J. Agric. Res.* **20**: 557-585. [592]

Wright, S. 1943. Isolation by distance. *Genetics* **28**: 114-138. [598]

Wright, S. 1960. Path coefficients and path regressions: alternative or complementary concepts? *Biometrics* **16**: 189-202. [592]

Wright, S. P. 1992. Adjusted P-values for simultaneous inference. *Biometrics* **48**: 1005-1013. [23]

Young, F. W. 1985. Multidimensional scaling. 649-659 *in*: S. Kotz & N. L. Johnson [eds.] *Encyclopedia of statistical sciences. Vol. 5*. Wiley, New York. [512]

Yule, G. U. 1927. On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers. *Phil. Trans. R. Soc. Lond. A* **226**: 267-298. [725]

Zar, J. H. 1999. *Biostatistical analysis. 4th edition*. Prentice Hall, Upper Saddle River, New Jersey. xii + 663 pp. + appendices. [215, 566]

Zeeb, B. A., C. E. Christie, J. P. Smol, D. L. Findlay, H. J. Kling & H. J. B. Birks. 1994. Responses of diatom and chrysophyte assemblages in Lake 227 sediments to experimental eutrophication. *Can. J. Fish. Aquat. Sci.* **51**: 2300-2311. [855]

References to R packages

The list provides, in alphabetic order, references for the R packages cited in the Software *sections of the chapters, and listed under the* R packages *entry of the* Subject index. *The references in the list are those provided by the author(s) in the package documentation files. The current package version numbers are also shown.*

ade4, version 1.4-17

Chessel, D., A.-B. Dufour & S. Dray [with contributions of T. Jombart, J. R. Lobry, S. Ollier, S. Pavoine & J. Thioulouse]. 2011. ade4: Analysis of ecological data: exploratory and Euclidean methods in environmental sciences. R package version 1.4-17. http://cran.r-project.org/web/packages/ade4/.

AEM, version 0.4-1

Blanchet, F. G. [with contributions from P. Legendre]. 2012. AEM: Tools to construct asymmetric eigenvector maps (AEM) spatial variables. R package version 0.4-1. https://r-forge.r-project.org/R/?group_id=195.

ape, version 3.0-1

Paradis, E., B. Bolker, J. Claude, H. S. Cuong, R. Desper, B. Durand, J. Dutheil, O. Gascuel, C. Heibl, D. Lawson, V. Lefort, P. Legendre, J. Lemon, Y. Noel, J. Nylander, R. Opgen-Rhein, A.-A. Popescu, K. Schliep, K. Strimmer & D. de Vienne. 2012. ape: Analyses of phylogenetics and evolution. R package version 3.0-1. http://cran.r-project.org/web/packages/ape/.

base, **stats** and **splines** in R version 2.14.2

R Development Core Team. 2012. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

BiodiversityR, version 1.6

Kindt, R. 2011. BiodiversityR: GUI for biodiversity and community ecology analysis. R package version 1.6. http://cran.r-project.org/web/packages/BiodiversityR/.

ca, version 0.33

Greenacre, M. & O. Nenadic. 2010. ca: Simple, multiple and joint correspondence analysis. R package version 0.33. http://cran.r-project.org/web/packages/ca/.

car, version 2.0-12

Fox, J., S. Weisberg, D. Bates, D. Firth, M. Friendly, G. Gorjanc, S. Graves, R. Heiberger, R. Laboissiere, G. Monette, H. Nilsson, D. Ogle, B. Ripley & A. Zeileis. 2012. car: Companion to Applied regression. R package version 2.0-12. http://cran.r-project.org/web/packages/car/.

cclust, version 0.6-16

Dimitriadou, E. 2009. cclust: Convex clustering methods and clustering indexes. R package version 0.6-16. http://cran.r-project.org/web/packages/cclust/.

cluster, version 1.14.2

Maechler, M. [based on S original by P. Rousseeuw, A. Struyf & M. Hubert]. 2012. cluster: Cluster analysis extended Rousseeuw *et al*. R package version 1.14.2. http://cran.r-project.org/web/packages/cluster/.

clValid, version 0.6-4

Brock, G., V. Pihur, S. Datta & S. Datta. 2011. clValid: Validation of clustering results. R package version 0.6-4. http://cran.r-project.org/web/packages/clValid/.

cocorresp, version 0.1-9

Simpson, G. L. [from original Matlab routines by C.J.F. ter Braak & A.P. Schaffers]. 2010. cocorresp: Co-correspondence analysis methods. R package version 0.1-9. http://cran.r-project.org/web/packages/cocorresp/.

codep, version 0.1-6

Guénard, G. [with contributions from B. Pagès]. 2010. codep: Multiscale codependence analysis. R package version 0.1-6. http://cran.r-project.org/web/packages/codep/.

const.clust, version 1.2

Legendre, P. 2011. const.clust: Space- and time-constrained clustering package. R package version 1.2. http://numericalecology.com/rcode/.

DAAG, version 1.12

Maindonald , J. & W. J. Braun. 2012. DAAG: Data analysis and graphics data and functions. R package version 1.12. http://cran.r-project.org/web/packages/DAAG/.

DierckxSpline, version 1.1-4

Dorai-Raj, S. & S. Graves. 2009. DierckxSpline: R companion to "Curve and surface fitting with splines". R package version 1.1-4. http://cran.r-project.org/web/packages/DierckxSpline/.

ecodist, version 1.2.7

Goslee, S. & D. Urban. 2012. ecodist: Dissimilarity-based functions for ecological analysis. R package version 1.2.7. http://cran.r-project.org/web/packages/ecodist/.

FactoMineR, version 1.18

Husson, F., J. Josse, S. Le & J. Mazet. 2012. FactoMineR: Multivariate exploratory data analysis and data mining with R. R package version 1.18. http://cran.r-project.org/web/packages/FactoMineR/.

FD, version 1.0-11

Laliberté, E. & B. Shipley. 2011. FD: Measuring functional diversity (FD) from multiple traits, and other tools for functional ecology. R package version 1.0-11. http://cran.r-project.org/web/packages/FD/.

flexclust, version 1.3-2

Leisch, F. [with contributions by E. Dimitriadou]. 2011. flexclust: Flexible cluster algorithms. R package version 1.3-2. http://cran.r-project.org/web/packages/flexclust/.

geoR, version 1.7-2

Ribeiro, P. J. Jr. & P. J. Diggle. 2012. geoR: Analysis of geostatistical data. R package version 1.7-2. http://cran.r-project.org/web/packages/geoR/.

indicspecies, version 1.6.0

De Cáceres, M. & F. Jansen. 2011. indicspecies: Functions to assess the strength and significance of relationship of species site group associations. R package version 1.6.0. http://cran.r-project.org/web/packages/indicspecies/.

kernlab, version 0.9-14

Karatzoglou, A., A. Smola & K. Hornik. 2011. kernlab: Kernel-based machine learning lab. R package version 0.9-14. http://cran.r-project.org/web/packages/kernlab/.

klaR, version 0.6-6

Roever, C., N. Raabe, K. Luebke, U. Ligges, G. Szepannek & M. Zentgraf. 2011. klaR: Classification and visualization. R package version 0.6-6. http://cran.r-project.org/web/packages/klaR/.

labdsv, version 1.4-1

Roberts, D. W. 2010. labdsv: Ordination and multivariate analysis for ecology. R package version 1.4-1. http://cran.r-project.org/web/packages/labdsv/.

lmodel2, version 1.7-0

Legendre, P. 2011. lmodel2: Model II regression. R package version 1.7-0. http://cran.r-project.org/web/packages/lmodel2/.

MASS, version 7.3-17

Ripley, B., K. Hornik, A. Gebhardt & D. Firth. 2012. MASS: Support functions and datasets for Venables and Ripley's MASS. R package version 7.3-17. http://cran.r-project.org/web/packages/MASS/.

Matrix, version 1.0-5

Douglas Bates, D. & M. Maechler. 2012. Matrix: Sparse and dense matrix classes and methods. R package version 1.0-5. http://cran.r-project.org/web/packages/Matrix/.

mice, version 2.11

van Buuren, S. & K. Groothuis-Oudshoorn. 2011. mice: Multivariate imputation by chained equations. R package version 2.11. http://cran.r-project.org/web/packages/mice/.

missMDA, version 1.2

Husson, F. & J. Josse. 2010. missMDA: Handling missing values with/in multivariate data analysis (principal component methods). R package version 1.2. http://cran.r-project.org/web/packages/missMDA/.

mlogit, version 0.2-2

Croissant, Y. 2011. mlogit: Multinomial logit model. R package version 0.2-2. http://cran.r-project.org/web/packages/mlogit/.

mvpart, version 1.6-0

De'ath, G. 2012. mvpart: Multivariate partitioning. R package version 1.6-0. http://cran.r-project.org/web/packages/mvpart/.

MVPARTwrap, version 0.1-8

Ouellette, M.-H. [with contributions from P. Legendre]. 2011. MVPARTwrap: Additional functionalities for package mvpart. R package version 0.1-8. http://cran.r-project.org/web/packages/MVPARTwrap/.

ncf, version 1.1-3

Bjornstad, O. N. 2009. ncf: Spatial nonparametric covariance functions. R package version 1.1–3. http://cran.r-project.org/web/packages/ncf/.

nlme, version 3.1-103

Pinheiro, J., D. Bates, S. DebRoy, D. Sarkar & the R Core team. 2012. nlme: Linear and nonlinear mixed effects models. R package version 3.1-103. http://cran.r-project.org/web/packages/nlme/.

nortest, version 1.0

Gross, J. 2006. nortest: Tests for normality. R package version 1.0. http://cran.r-project.org/web/packages/nortest/.

packfor, version 0.0-8

Dray, S. [with contributions of P. Legendre & G. Blanchet]. 2012. packfor: Forward selection with permutation (Canoco p. 46). R package version 0.0-8. https://r-forge.r-project.org/R/?group_id=195.

pastecs, version 1.3-11

Ibanez, F., P. Grosjean & M. Etienne. 2009. pastecs: Package for analysis of space-time ecological series. R package version 1.3-11. http://cran.r-project.org/web/packages/pastecs/.

PCNM, version 2.1-2

Legendre, P., D. Borcard, F. G. Blanchet & S. Dray. 2012. PCNM: MEM spatial eigenfunction and principal coordinate analyses. R package version 2.1-2. https://r-forge.r-project.org/R/?group_id=195.

penalized, version 0.9-38

Goeman, J. & R. Meijer. 2012. penalized: L1 (lasso) and L2 (ridge) penalized estimation in GLMs and in the Cox model. R package version 0.9-38. http://cran.r-project.org/web/packages/penalized/.

picante, version 1.3-0

Kembel, S. W., D. D. Ackerly, S. P. Blomberg, W. K. Cornwell, P. D. Cowan, M. R. Helmus, Helene Morlon & C. O. Webb. 2011. picante: R tools for integrating phylogenies and ecology. R package version 1.3-0. http://cran.r-project.org/web/packages/picante/.

Rcmdr, version 1.8-3

Fox, J., L. Andronic, M. Ash, M. Bouchet-Valat, T. Boye, S. Calza, A. Chang, P. Grosjean, R. Heiberger, K. K. Pour, G. J. Kerns, R. Lancelot, M. Lesnoff, U. Ligges, S. Messad, M. Maechler, R. Muenchen, D. Murdoch, E. Neuwirth, D. Putler, B. Ripley, M. Ristic & P. Wolf. 2012. Rcmdr: R commander. R package version 1.8-3. http://cran.r-project.org/web/packages/Rcmdr/.

rdaTest, version 1.8

Legendre, P. & S. Durand. 2012. rdaTest: Canonical redundancy analysis. R package version 1.8. http://numericalecology.com/rcode/.

rioja, version 0.7-3

Juggins, S. 2012. rioja: Analysis of Quaternary science data. R package version 0.7-3. http://cran.r-project.org/web/packages/rioja/.

sem, version 2.1-1

Fox, J. & J. Byrnes [with contributions from M. Culbertson, M. Friendly, A. Kramer & G. Monette]. 2011. sem: Structural equation models. R package version 2.1-1. http://cran.r-project.org/web/packages/sem/.

seriation, version 1.0-6

Hahsler, M., C. Buchta & K. Hornik. 2011. seriation: Infrastructure for seriation. R package version 1.0-6. http://cran.r-project.org/web/packages/seriation/.

sgeostat, version 1.0-24

Majure, J. J. & A. Gebhardt. 2012. sgeostat: An object-oriented framework for geostatistical modeling in S+. R package version 1.0-24. http://cran.r-project.org/web/packages/sgeostat/.

smatr, version 3.2.4

Warton, D., R. Duursma, D. Falster & S. Taskinen. 2011. smatr: (Standardised) major axis estimation and testing routines. R package version 3.2.4. http://cran.r-project.org/web/packages/smatr/.

SoDA, version 1.0-4

Chambers, J. M. 2012. SoDA: Functions and examples for "Software for data analysis". R package version 1.0-4. http://cran.r-project.org/web/packages/SoDA/.

spacemakeR, version 0.0-5

Dray, S. 2010. spacemakeR: Spatial modelling. R package version 0.0-5. https://r-forge.r-project.org/R/?group_id=195.

spdep, version 0.5-45

Bivand, R. [with contributions by M. Altman, L. Anselin, R. Assunção, O. Berke, A. Bernat, G. Blanchet, E. Blankmeyer, M. Carvalho, B. Christensen, Y. Chun, C. Dormann, S. Dray, R. Halbersma, E. Krainski, P. Legendre, N. Lewin-Koh, H. Li, J. Ma, G. Millo, W. Mueller, H. Ono, P. Peres-Neto, G. Piras, M. Reder, M. Tiefelsdorf & D. Yu]. 2011. spdep: Spatial dependence: weighting schemes, statistics and models. R package version 0.5-45. http://cran.r-project.org/web/packages/spdep/.

splines: *see entry* **base**, **stats** and **splines**

stats: *see entry* **base**, **stats** and **splines**

STI, version 1.0.2

Legendre, P., D. Borcard & M. De Cáceres. 2010. STI: Space-time ANOVA models without replications. R package version 1.0.2. http://sites.google.com/site/miqueldecaceres/software.

survey, version 3.28

Lumley, T. 2012. survey: Analysis of complex survey samples. R package version 3.28. http://cran.r-project.org/web/packages/survey/.

survival, version 2.36-12

Therneau, T. & T. Lumley2012. survival: Survival analysis, including penalised likelihood. R package version 2.36-12. http://cran.r-project.org/web/packages/survival/.

vegan, version 2.0-3

Oksanen, J., F. G. Blanchet, R. Kindt, P. Legendre, P. R. Minchin, R. B. O'Hara, G. L. Simpson, P. Solymos, M. H. H. Stevens & H. Wagner. 2012. vegan: Community ecology package. R package version 2.0-3. http://cran.r-project.org/web/packages/vegan/.

vegclust, version 1.5.1

De Cáceres, M. 2011. vegclust: Fuzzy clustering of vegetation data. R package version 1.5.1. http://cran.r-project.org/web/packages/vegclust/.

venneuler, version 1.1-0

Wilkinson, L. 2011. venneuler: Venn and Euler Diagrams. R package version 1.1-0. http://cran.r-project.org/web/packages/venneuler/.

waveslim, version 1.7.1

Whitcher, B. 2012. waveslim: Basic wavelet routines for one-, two- and three-dimensional signal processing. R package version 1.7.1. http://cran.r-project.org/web/packages/waveslim/.

wmtsa, version 1.1-1

Constantine, W. & D. Percival. 2011. wmtsa: Wavelet methods for time series analysis. R package version 1.1-1. http://cran.r-project.org/web/packages/wmtsa/.

Subject index

E