



ОБЗОР ПРОГРАММНЫХ СРЕДСТВ, ИСПОЛЬЗУЕМЫХ ДЛЯ АНАЛИЗА ГЕОБОТАНИЧЕСКИХ ДАННЫХ

А. Новаковский

м.н.с. отдела компьютерных систем, технологий и моделирования
E-mail: novakovsky@ib.komisc.ru, тел.: (8212) 21 67 52

Научные интересы: *теория графов, кластерный анализ, программирование, классификация растительности*

Введение

На сегодняшний день существует множество программных средств, разработанных для автоматизации обработки геоботанических данных. Однако специалисты-геоботаники зачастую слабо информированы о них. Трудности вызывают как понимание используемых алгоритмов, так и интерпретация полученных результатов. Такое положение объясняется рядом причин. Первое – языковой барьер, большинство программ и сопроводительной документации разработано в странах Европейского союза и США. Второе – недостаточная математическая подготовка, которая затрудняет понимание используемых алгоритмов. Третье – большинство программ является платными.

Цель настоящего обзора – познакомить читателя с основными математическими подходами, используемыми в геоботанических исследованиях; дать краткое описание возможностей наиболее распространенных программных средств, реализующих эти подходы; привести интернет-ссылки, по которым можно было бы найти демонстрационные версии предлагаемых к обсуждению программ. В работе рассматриваются следующие программные средства: TURBOVEG, TWINSPAN, CONOCO, SYNTAXON, CAP, JUICE, BIOTAS и GRAPHS.

Основные принципы обработки данных

При обработке геоботанической информации используются два основных подхода – ординация и классификация (кластеризация).

Принцип *ординации* заключается в анализе изучаемых показателей в качестве точек многомерного пространства, где свойства выступают как координаты, задающие местоположение этих точек. Использование данного принципа позволяет уменьшить размерность пространства (до 2- или 3-мерного) и имеет своей целью графически показать взаимное расположение исследуемых объектов (рис. 1а). Одним из основных условий при уменьшении размерности является сохранение расстояний между объектами (или, если точнее, минимально возможное их изменение). Такое представление дает возможность визуального выделения существующих трендов в наборах данных и облегчает поиск факторов, наиболее сильно влияющих на исследуемые объекты. К методам ординации относятся: многомерное шкалирование, принцип главных компонент, анализ соответствий, канонический анализ соответствий и т.д.

Под *классификацией* (кластеризацией) понимается задача разбиения всей совокупности рассматриваемых

объектов на отдельные группы (классы) со сходными характеристиками (рис. 1б). Причем зачастую, между такими группами выстраиваются иерархические отношения, т.е. классы могут включать в себя друг друга. Существует несколько подходов в классификации. Первый, наиболее распространенный и математически формализованный, основывается на использовании мер сходства/расстояния. При таком подходе для всех пар рассматриваемых объектов рассчитывается степень их сходства между собой, и на основе этих данных выделяют классы. В геоботанике обычно используют коэффициенты сходства Жаккара, Сьеренсена, Стургена-Радулеску, Шимкевича-Симпсона и др., а также Евклидово расстояние, расстояние χ^2 и т.д. [2, 7, 15]. Кроме матриц сходств, в литературе описывается множество других методов кластеризации, основанных на оценивании функций плотности статистического распределения, эвристических алгоритмах перебора, идеях математического программирования [4, 13, 14, 21]. Подробное описание методов ординации можно найти и в Internet на сайте «Oklahoma State University»: <http://www.okstate.edu/artsci/botany/ordinate>.

В геоботанических исследованиях для классификации растительности часто используется довольно специфический метод, который основан на обработке валовых таблиц. Суть данного подхода заключается в последовательной перестановке строк и столбцов таблицы с целью нахождения блочно-диагонального вида (рис. 1в). Такое представление позволяет выделить группы растений, четко характеризующие геоботанические таксоны, другими словами, выделяется набор наиболее информативных признаков и одновременно производится разбиение множества объектов по этим признакам [8, 9]. В сравнении с использованием мер сходства/различия геоботанический подход к классификации менее формализован и допускает большую свободу его применения.

Обзор программ

TURBOVEG. Все описываемые ниже программы работают с массивами данных, которые требуется готовить специальным образом. Поэтому сначала остановимся на специализированной программе, предназначенной для ввода и хранения больших объемов фитоценологической информации – TURBOVEG (рис. 2а). Причем количество хранимых элементов практически не ограничено. Основным предназначением TURBOVEG является функция преобразования накопленных данных в форматы, понятные для других программ

Community Ordinations – каноническая ординация растительных сообществ). Программа используется как для выявления особенностей распределения растительных сообществ в соответствии с ведущими экологическими факторами, так и для подтверждения или опровержения результатов, полученных другими путями. В программе реализованы следующие методы:

- канонический анализ соответствий;
- принцип главных компонент;
- анализ избыточности.

В качестве результата работы этих алгоритмов строятся ординационные диаграммы и биплоты. Кроме графического результата программа выдает корреляционные матрицы внешних факторов с осями биплотов и внешних факторов друг с другом, что также немаловажно при интерпретации результатов. Подробности о возможностях программы и принципах ее работы можно найти в работе [21] и на авторской странице в интернет <http://www.microcomputerpower.com/catalog/canoco.html>.

Достоинства и недостатки. Версия CONOCO для DOS доступна в интернете и не требует регистрации. К недостаткам относится сложность работы с DOS-версией программы, поскольку это требует использования разных запускаемых модулей для обработки данных и графического отображения результатов.

SYNTAXON. Одной из очень часто упоминаемых программ, предназначенных для проведения математической обработки фитоценологического материала, является SYN-TAX (SYNTAXON). К сожалению, поскольку программа является платной, мы не смогли проверить ее в действии и руководствовались только содержанием сайта группы разработчиков <http://ramet.elte.hu/~podani/>.

SYN-TAX предназначена для обработки геоботанических данных методами ординации и кластеризации. Программа состоит из трех модулей:

- ординации, включающий в себя основные методы ординации (метрическая и неметрическая ординация, анализ соответствий, анализ избыточности, канонический корреляционный анализ и т. д.) и позволяющий графически представлять результаты анализа;
- иерархической кластеризации (включает методы ближнего и дальнего соседа, кластеризации, основанной на энтропии и т. д.);
- неиерархической кластеризации, в котором реализованы такие алгоритмы как: метод k -средних, метод нечеткой кластеризации и т. д.

Для определения мер сходства в программе предлагается к использованию 33 коэффициента: качественные, количественные, ранговой и простой корреляции и т. д.

К достоинствам SYNTAXON можно отнести возможность графического представления получаемых результатов и использование большого числа коэффициентов сходства и различия, к недостаткам – плату за ее использование.

CAP. Одной из самых современных и обладающих простым интерфейсом пользователя является программа CAP – «Community Analysis Package» (рис. 26). Программа предназначена для анализа растительных данных методами ординации, кластеризации и класси-

фикации. Ее отличительной особенностью является возможность работы с трансформированными данными, что позволяет уменьшить влияние субъективного фактора. Причем легкость преобразованию дает возможность простого подбора наиболее подходящей формы трансформации. В программе реализованы следующие способы преобразования данных: логарифмический, экспоненциальный, синусоидальный, извлечения корня, изменение на определенную константу и т. д. В CAP реализованы следующие методы обработки данных:

- методы ординации: анализ главных компонент, анализ соответствий и т. д.;
- методы кластеризации: единичное присоединение, полное присоединения, присоединение по среднему и др.;
- отдельным пунктом меню вынесено использование алгоритма TWINSpan, предназначенного для проведения классификации растительных сообществ;
- существует возможность строить таблицы сходства с использованием качественных, количественных коэффициентов сходства и мер расстояния (всего 27 коэффициентов).

К положительным моментам можно отнести возможность графического отображения результатов работы алгоритмов, и множества вспомогательных данных, таких как матрицы корреляций, статистические параметры для результатов и т. д.

Скачать демонстрационную версию программы и документацию можно по интернет-адресу: <http://www.pisces-conservation.com/indexsoftprog.html>.

JUICE. Одной из наиболее интересных программ, является JUICE (рис. 2в). Программа предназначена для анализа и классификации больших массивов геоботанических данных, как стандартными средствами, так и с помощью оригинального метода классификации «COCTAIL» и оптимизирована для совместной работы с TWINSpan и TURBOVEG [23, 24]. К достоинствам можно отнести возможность импорта данных из других форматов: текстовых файлов, формата RTF, Excel таблиц и баз данных Access. Обработка в JUICE может осуществляться как вручную (перестановкой строк и столбцов валовой таблицы), так и при помощи встроенных алгоритмов:

- сортировка строк и колонок валовой фитоценологической таблицы по алфавиту, среднему проективному покрытию, выбранным жизненным формам или другим характеристикам;
- автоматическое построение синоптических таблиц;
- выявление индикаторных, доминантных и константных видов;
- возможность вычисления коэффициентов β -разнообразия;
- использование подходов других программных продуктов – проведение классификации методами TWINSpan и COCTAIL.

Резюмируя, можно сказать, что программа JUICE позволяет строить синоптические таблицы, проводить сортировку по описаниям и видам, а также классифицировать растительность различными методами (TWINSpan, COCTAIL). Все алгоритмы могут работать с большими объемами информации (до 65 тысяч

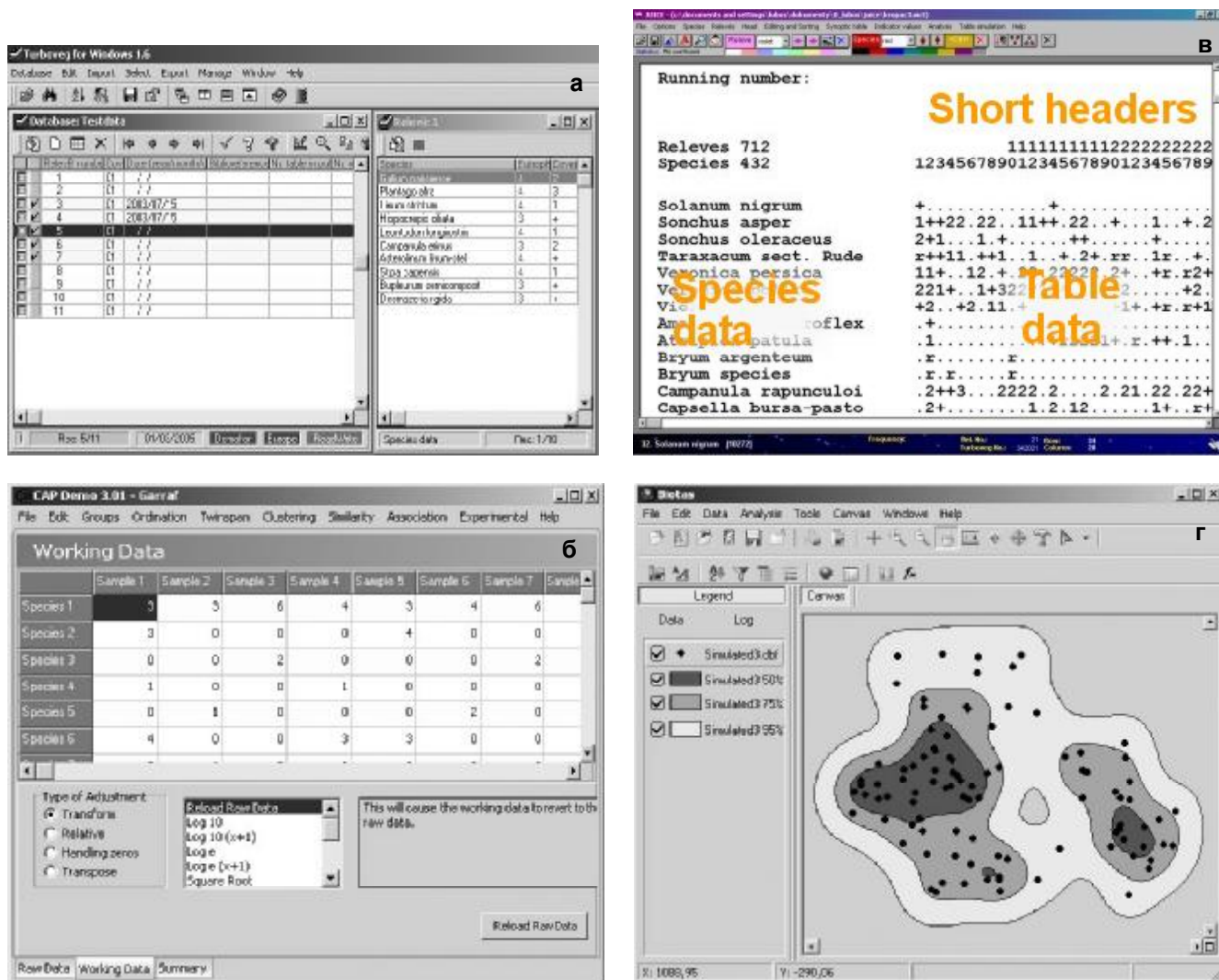


Рис. 2. Внешний вид программ: TURBOVEG (а) – слева: список геоботанических описаний, справа: список видов для выделенного описания; CAP (б) – закладка преобразования; JUICE (в) – валовая таблица геоботанических описаний; BIOTAS (г) – справа: построенные контуры встречаемости видов.

описаний). Программа является бесплатной. Кроме того, хорошо проработана система помощи пользователю, находящаяся на авторском сайте: <http://www.sci.muni.cz/botany/juice.htm>. К недостаткам можно отнести отсутствие графического представления данных.

BIOTAS. Рассмотрим программу, не предназначенную для проведения классификации растительности, которая, однако, может представлять определенный интерес для геоботаников. Речь идет о программе BIOTAS, разработанной для визуализации и анализа пространственных данных. Другими словами, программа позволяет визуальным образом отображать информацию о местах встреч тех или иных видов и накладывать карты соответствующей местности с целью нахождения взаимосвязей между рельефом, климатическими и другими условиями с встречаемостью тех или иных видов (рис. 2г). Кроме простого наложения, в BIOTAS реализованы и другие алгоритмы анализа пространственных данных:

- реализована возможность использования разнообразных фильтров (в том числе и случайного) для отбора информации для дальнейшего анализа;

- существует возможность анализа плотности пространственного распределения исследуемых объектов с помощью сеток плотности;

- введены алгоритмы проверки данных на возможность их кластеризации;

- в программе реализованы несколько алгоритмов, определяющих контуры распространения исследуемого вида с разными степенями достоверности (рис. 2г);

- реализован алгоритм анализа пространственных данных в зависимости от временных параметров, например его можно использовать при изучении миграции животных;

- введены функции определения линейной или полиномиальной регрессии;

- реализованы дополнительные функции, позволяющие рассчитывать статистические параметры (минимальное, максимальное и среднее значения, дисперсию, стандартное отклонение, медиану и др.), строить различные по форме гистограммы (обычные, круговые и т.п.).

Демонстрационную версию программы и инструкцию пользователя можно найти на сайте разработчи-

	A	B	C	D	E	F	G	H	I
1	Picea	Pinus	Abies	Betula	Populus	Salix	Alnus	Sorbus	
2			-36	-32	-34	-50	-53	-54	-62
3				-62	-42	-74	-74	-75	-80
4					-42	-53	-47	-48	-49
5						-31	-25	-48	-43
6							5	-50	-38
7								-42	-30
8									-30
9									30

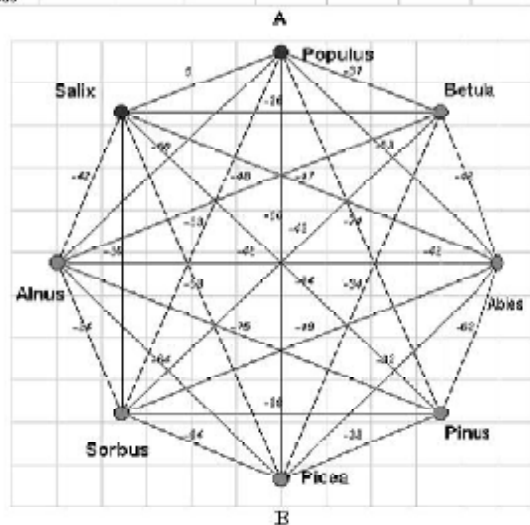


Рис. 3. Табличное представление коэффициентов сходства (а) и представление этих же коэффициентов в виде графа (б).

ков: <http://www.ecostats.com/purchase/order.htm>.

GRAPHS. Рассмотрим более подробно разрабатываемый нами на базе Института биологии Коми НЦ УрО РАН программный модуль GRAPHS. В качестве основы для модуля был взят программный продукт Microsoft Excel, входящий в состав пакета Microsoft Office. Эта программа широко используется геоботаниками для хранения данных и обладает гибким средством программирования VBA [10]. Модуль предназначен для проведения полуавтоматической классификации геоботанических описаний на основе использования коэффициентов сходства и графов, построенных с их использованием. Всего реализован расчет 12 коэффициентов, среди них качественные (Жаккара, Сьеренсена-Чекановского, Стургена-Радулеску и др.) и количественные (Сьеренсена-Чекановского с учетом обилий, Шимкевича-Симпсона, ранговой корреляции Кэнделла и корреляции Пирсона и т.д.), а также коэффициенты сопряженности видов.

Для визуального отображения полученных результатов используются графы, вершинами которых являются рассматриваемые объекты (геоботанические описания или виды), а ребрами – коэффициенты сходства между ними (рис. 3). Такой подход разрабатывался многими исследователями [1, 5, 12, 15]. С целью облегчения восприятия полученных графов в модуле используются различные визуальные эффекты (изменение стиля, цвета, размеров вершин, длины ребер) и алгоритмы теории графов (разбиение на компоненты связности, построение дендритов и дендрограмм, представление в виде звезды). Описание используемых алгоритмов можно найти в работах посвященных теории графов [6, 11].

Для проверки работоспособности модуля, нами проводилась его апробация на различных данных, собранных сотрудниками Института биологии Коми НЦ УрО

РАН, Ботанического института РАН им. В.Л. Комарова и Санкт-Петербургского университета. Апробация показала хорошее совпадение результатов ручной классификации (экспертной оценки) с результатами работы модуля, причем это относится как к методу Браун-Бланке, так и эколого-фитоценологическому подходу. В качестве иллюстрации остановимся на результатах обработки данных д.б.н. С.В. Дегтевой (Институт биологии Коми НЦ УрО РАН), проводившейся с помощью модуля GRAPHS.

Для анализа использованы около 100 геоботанических описаний, выполненных в десяти административных районах Республики Коми. На основе этих данных, с использованием программы Microsoft Excel, была составлена сводная таблица геоботанических описаний для формации сероольховых лесов. При классификации массива геоботанических описаний ассоциации выделялись на основании критерия участия в сложении нижних ярусов сообществ наиболее обильных видов (доминантов) и эколого-ценотических групп сопряженных видов. Всего было выделено 10 ассоциаций, на каждую из которых приходилось от 3 до 18 описаний. Подробнее с методикой классификации и характеристиками выделенных ассоциаций можно познакомиться в работе С.В. Дегтевой [3]. Использование модуля GRAPHS дало возможность проверить насколько оценка дифференциации растительных сообществ и их объединение в группы, основанные на большом опыте работы в поле и при обработке описаний, совпадает с результатами, полученными при независимой обработке с использованием количественных коэффициентов.

Для сравнения был построен граф в виде дендрограммы, которая строилась методом среднего. В качестве меры сходства использован коэффициент Сьеренсена-Чекановского. После построения дендрограммы вершины графа были окрашены в разные цвета в соответствии с ассоциациями выделенными автором описаний (рис. 4). Это наглядно показало, что выделенные автором группы описаний практически совпадают с результатами, получаемыми с помощью предложенного модуля. На основании полученных результатов можно сделать вывод, что использование программного модуля «GRAPHS» значительно ускоряет процесс обработки описаний и дает вполне адекватную картину дифференциации описаний и их интеграции в синтаксоны. Особенно полезной она может быть для первичного разбиения материала на гомогенные группы, объем которых соответствует ассоциациям при доминантном подходе. Для системы классификации Браун-Бланке группы имеют объем примерно сравнимый с ассоциациями и субассоциациями [10].

К недостаткам модуля можно отнести невозможность отображения большого числа объектов, например – для представления данных в виде окружности не более 30 элементов, для дендрита и звезды до 100, а для дендрограммы до 200.

Выводы

На основе проведенного анализа существующих программных продуктов, отметим следующие моменты. Единственной, найденной нами, специализирован-

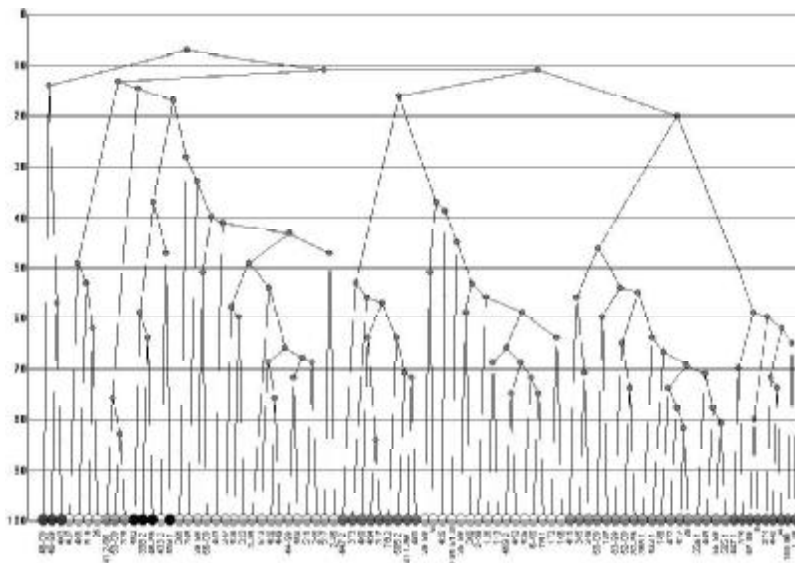


Рис. 4. Дендрограмма сходств по коэффициенту Сьеренсена-Чекановского. По оси абсцисс: названия геоботанических описаний; по оси ординат: коэффициент сходства.

ной программой предназначенной для ввода, хранения и преобразования фитоценологической информации к другим форматам является TURBOVEG. Программа совместима с большинством наиболее распространенных продуктов обработки геоботанических данных. Поскольку программа является платной, то вместо TURBOVEG можно использовать Microsoft Excel.

Для целей ординации данных существует большое количество специализированных программ. Наиболее удобной в использовании, на наш взгляд, является CAP. Другие же программы либо являются платными, либо морально устарели. Кроме CAP можно порекомендовать использовать другие, не рассчитанные на работу с ботаническими данными, программы статистической обработки, например, SPSS, Statistica и др.

Среди программ, предназначенных для классификации растительности, на наш взгляд, наиболее простой в обучении и использовании показала себя программа JUICE. Являясь бесплатной, она в то же время позволяет проводить разнообразную обработку геоботанического материала в виде валовых таблиц, либо в виде синоптических таблиц.

Для графического представления данных и взаимосвязей между ними и при небольшом количестве исследуемых объектов можно использовать модуль GRAPHS.

ЛИТЕРАТУРА

1. Андреев В.Л. Классификационные построения в экологии и систематике. М.: Наука, 1980. 142 с.
2. Василевич В.И. Статистические методы в геоботанике. Л.: Наука, 1969. 230 с.
3. Дегтева С.В. Сероольшаники Республики Коми // Бот. журн., 2002. Т. 87, № 1. С. 107-121.
4. Ким Дж.-О., Мьюллер Ч.У. Факторный анализ: статистические методы и практические вопросы // Факторный, дискриминантный и кластерный анализ / Под ред. И.С. Енюкова. М., 1989. С. 5-77.
5. Кушнарев Е.Л. Применение графических методов анализа при изучении видовой структуры орнито-сообществ по экологическому профилю // Теоретико-графовые методы в биогеографических исследованиях. Владивосток, 1983. С. 33-48.
6. Липский В. Комбинаторика для программистов. М.: Мир, 1988. 213 с.
7. Лебедева Н.В., Кривошук Д.А. Биологическое разнообразие и методы его оценки. М., 2002. 432 с. – (<http://nature.ok.ru/biodiversity/book3.html>).
8. Миркин Б.М., Наумова Л.Г., Соломещ А.И. Современная наука о растительности. М., 2000. 263 с.
9. Нешатаев Ю.Н. Методы анализа геоботанических материалов. Л.: Изд-во ЛГУ, 1987. 190 с.
10. Новаковский А.Б. Возможности и принципы работы программного модуля «GRAPHS». Сыктывкар, 2004. 28 с.
11. Об использовании мер зависимости для определения степени ассоциации видов растений / Б.И. Семкин и др. // Теоретико-графовые методы в биогеографических исследованиях. Владивосток, 1983. С. 20-32.
12. Свами М., Тхуласираман К. Графы, сети и алгоритмы. М.: Мир, 1984. 454 с.
13. Ту Дж., Гонсалес Р. Принципы распознавания образов. М.: Мир, 1978. 416 с.
14. Шитиков В.К., Розенберг Г.С., Зинченко Т.Д. Количественная гидроэкология: методы системной идентификации. Тольятти, 2003. 463 с. – (<http://www.tolcom.ru/kiril/Library/Book1/content0/content0.htm>).
15. Шмидт В.М. Математические методы в ботанике. Л.: Изд-во ЛГУ, 1984. 287 с.
16. Bruilheide H. Using formal logic to classify vegetation // Folia Geobot. Phytotax, 1997. Vol. 32. P. 41-46.
17. Bruilheide, H. A new measure of fidelity and its application to defining species groups // J. Vegetation Sci., 2000. Vol. 11. P. 295-306.
18. Hennekens S.M. TURBO(VEG). Software package for input, processing, and presentation of phytosociological data. Wageningen (NL)–Landcaster (UK), 1996.
19. Hennekens, S.M., Schaminee J.H.J. TURBOVEG, a comprehensive database management system for vegetation data // J. Vegetation Sci., 2001. Vol. 12. P. 589-591.
20. Hill M.O. DECORANA and TWINSpan, for ordination and classification of multivariate species data: a new edition, together with supporting programs, in FORTRAN 77. Huntingdon (UK), 1994.
21. Jongman R.H.G., Ter Braak C.J.F., Van Tongeren O.F.R. Data analysis in community and landscape ecology. Wageningen (NL), 1987. 299 p.
22. Legendre P. Legendre L. Numerical Ecology. Amsterdam: Elsevier Sci. B.V., 1998. 853 p.
23. Tichy L. JUICE, software for vegetation classification // J. Vegetation Sci., 2002. Vol. 13. P. 451-453. – (<http://www.sci.muni.cz/botany/juice/jvs.pdf>).
24. Tichy L. Formalized reproduction of an expert-based phytosociological classification: A case study of subalpine tall-forb vegetation // J. Vegetation Sci., 2003. Vol. 14. P. 601-610. – (<http://www.sci.muni.cz/botany/juice/JVS2003b.pdf>). ❖