ISSN 2079-0864, Biology Bulletin Reviews, 2022, Vol. 12, No. 1, pp. 84–93. © Pleiades Publishing, Ltd., 2022.
Russian Text © The Author(s), 2021, published in Zhurnal Obshchei Biologii, 2021, Vol. 82, No. 2, pp. 143–154.

Models of Joint Distribution of Species on the Example of Benthic Communities from Small Rivers of the Volga Basin

V. K. Shitikov[a], T. D. Zinchenko[a], *, and L. V. Golovatyuk[a]

[a] *Institute of Ecology of the Volga Basin, Russian Academy of Sciences, Togliatti, 445003 Russia*

**e-mail: zinchenko.tdz@yandex.ru*

Received October 20, 2020; revised November 5, 2020; accepted December 23, 2020

Abstract—The theoretical and practical aspects of building of joint species-distribution models, a modern tool for the analysis of ecological communities, are considered. It is shown that it is inappropriate to use the MaxEnt method or other methods based on the concept of "pseudo-absence" points when the observational data are quantitative indicators of the population density (in particular, the abundance of species in hydrobiological studies). Contemporary multidimensional models of the joint distribution of communities should include a set of parameters that assess the impact of the following groups of fixed and random factors on the species occurrence: (a) the covariates and categorical variables describing the environmental conditions and characteristics of biotopes, (b) the main indicators characterizing each species and the phylogenetic structure of communities, (c) the functions of spatial autocorrelation of the data at observation points, (d) the residual (i.e., not caused by external factors) associativity of species. Analysis of the published data and practical examples of implementation showed that the mentioned requirements, in general, are satisfied by the methodological platform and the R package Hierarchical Modeling of Species Communities (HMSC). They form the basis for the construction of multidimensional hierarchical generalized linear models with mixed parameters estimated by Bayesian procedure. The main concepts and blocks of the HMSC platform are described, and the results of models based on the authors' data and long-term hydrobiological studies of benthic communities in 132 small and medium-sized rivers in the middle and lower Volga basin are discussed. The parameters of a set of one-dimensional candidate models for the abundance distribution of the subfamily Prodiamesinae (Diptera, Chironomidae) are analyzed, and a map of its forecast range within the region is constructed. To illustrate the multidimensional case, a model of the joint spatial distribution of 31 species of chironomids is constructed and its coefficients are analyzed. A residual correlation graph of statistically significant interspecies interactions has been built. It is concluded that the HMSC method and software package can be effectively used to solve fundamental problems of communities' ecology: the ways in which the areas of individual populations, the structure of their communities, and the nature of interspecific interactions depend the on environmental conditions and methods to predict future trends of these processes in response to global changes.

DOI: 10.1134/S2079086422010078

INTRODUCTION

The structure of the spatial distribution of communities and its relationship with the living conditions of populations are the most important directions in ecological studies. After the development of the BIOCLIM package in the 1980s (Busby, 1991), modeling of the species distribution (Species Distribution Models, SDM) and environmental niches (Environmental Niche Models, ENM) became a powerful tool for (macro)ecological and biogeographic studies and the assessment of the role of factors affecting the species distribution (Peterson et al., 2011). These methods were also very effective in paleoecology, phylogenetics, bioresource management, and wildlife conservation (Araújo et al., 2019). There is a vast body of literature on various SDM/ENM methods, the use of which has received a great deal of attention in the works of foreign ecologists (Franklin, 2009; Guisan et al., 2017) and in a detailed review of Russian colleagues (Lisovsky et al., 2020).

Analysis of the species spatial distribution is based on two different conceptual approaches. The process-based SDMs (also known as rank models of population dynamics; Zurell et al., 2016) explicitly include model structures and parameters describing the mechanisms of the main ecological processes in communities. The need to estimate the coefficients of the rates of reproduction, mortality, dispersal and demographic stochasticity (Vellend, 2016; Rosenberg et al., 2020), as well as their dependence on selective data acquisition processes, make such an approach still difficult to use, although an accounting of the basic processes in communities should be welcomed in all cases (D'Amen et al., 2017).

Another approach can be termed correlative, in that it is based on the identification of statistical dependences between environmental factors and the data on the species occurrence.

Dozens of methods of SDM construction have been described (Norberg et al., 2019). They differ in many aspects, including the composition of the source data ("only presence" of species at sampling points, "presence–absence," or quantitative assessment of abundance), structural assumptions of models (a generalized linear model, support vectors, or random forest), algorithms to obtain solutions (with the maximum-likelihood estimation or a Bayesian approach) and technical implementation (whether the method is available as an R package or as an independent software product). Works ranking the totality of the constructed models according to the degree of their competence and the construction of their ensembles (collectives) are also under way; the predictions of several models are weighted and averaged in them (Breiner et al., 2018).

Of the many applied algorithms, we should note the most commonly used method of maximum entropy, which is implemented in the MaxEnt program (Phillips et al., 2006; Lisovsky and Dudov, 2020). The algorithm predicts the probability of species presence at an arbitrary point of the geographical space based only on the points where it has been already recorded (*PO*, *presence-only*). The use of the MaxEnt software results in the calculation of an exponential function, the arguments of which are partial functions of particular predictors (linear, quadratic, multiple, etc.) with λ coefficients that estimate the contribution of the corresponding environmental factor. A step-by-step selection of the optimal model and the adjustment of λ coefficients is carried out with allowance for the minimization of the prediction error, both for the initial sample of *PO* and for the set of randomly selected points at which, as it is assumed, the species is absent (*pseudo-absence*, *PA*, or "*background*" *points*). The success of the algorithm largely depends on the selection of the form of particular functions, the sample size of *PA*, the prefiltering of the source data, the use of a correction layer, etc. (Lisovsky and Dudov, 2020).

The use of random background points is a classical approach that is known as the Resource Selection Function (Johnson, 1980), which presupposes a comparison of current habitat conditions with estimates of the availability of necessary resources for the community. In fact, it is often very difficult to confirm a species absence, and it has therefore been shown that this approach evaluates not so much the desired probability of the presence of a species as the heterogeneity of the empirical data used. In particular, indicators of success in the prediction of absence are often determined by "capricious zeroes," i.e., those points where the species simply cannot occur (Hastie and Fithian, 2013; Guisande et al., 2017). Therefore, if the presence/absence data are available or if there are especially quantitative estimates of the population abundance, it is appropriate to apply the adequate statistical methods.

The SDM models were mainly developed to model the range of only one species, while it is often necessary to estimate the joint distribution of many species that form communities (Clark et al., 2014; Warton et al., 2015). One possible approach is the use of aggregated models of distribution (*stacked* SDM, SSDM), in which a set of models for particular species is built in the first stage and their results are then combined (Calabrese et al., 2014). In contrast, another generalized method of analysis (*joint* SDM, JSDM) combines the species level of the model data into one model, which is simultaneously adjusted to the structure of the entire community. This makes it possible not only to identify interspecific associations but also to correlate the patterns obtained with the characteristics of species (Abrego et al., 2017) and their phylogenetic features or patterns of coexistence (Pollock et al., 2014). Lastly, the SDFA class of models (Spatial Dynamic Factor Analysis; Thorson et al., 2016) considers the distribution of the community structure under the effect of environmental factors not only in space, but also in time.

Changes in the patterns of interspecific interactions associated with differences in environmental conditions were found for a wide range of taxonomic groups (e.g., Brooker, 2006). Conclusions about the presence and strength of interspecific interactions are traditionally made based data from observations of the species occurrence with different statistical methods: multidimensional ordination, pair correlation, models of aggregation and segregation of species, etc. (Legendre, P. and Legendre, L., 2012). An important problem here is that conclusions about coexistence determined by interspecific interactions are mixed with effects generated by joint variation in the species response to abiotic changes. Since JSDM explicitly includes the measured ecological covariates, the estimates of species associativity found with their help are more adequate for identifying true interactions than "raw" indices of co-occurrence (Warton et al., 2015).

Hereinafter, the JSDM construction technique is considered based on a version of the Generalized Linear Mixed Models (GLMMs). According to statistical terminology, it is interpreted as a multidimensional, hierarchical, generalized linear model with mixed parameters based on the Bayesian procedure for their evaluation. As a working example, we used the results of long-term hydrobiological studies of benthic communities in small and medium rivers in the middle and lower Volga River regions (Zinchenko, 2009, 2011; Golovatyuk et al., 2018). The results of calculations were obtained with the statistical environment R ver. 3.6 and the HMSC (Hierarchical Modelling of Species

Communities) package developed by Ovaskainen et al. In this regard, the methodological material is subsequently presented based on a book (Ovaskainen and Abrego, 2020) and previous articles by this research group (Ovaskainen et al., 2016a, 2016b, 2017; Tikhonov et al., 2017, 2020).

MATERIALS AND METHODS

Description of the HMSC Statistical Model

A typical dataset obtained during ecological studies of the communities includes a set of species $j = 1... n_s$ identified at multiple n_y biotopes (strictly speaking, at *sampling units*) $i = 1... n_y$. The generalized linear mixed GLMM model used can be applied to various indicators of the species abundance y_{ij} (presence/absence, abundance, biomass, coverage, etc.) by including various communication functions and postulating error-distribution laws. In the context of HMSC, the selected data are adjusted with a multidimensional model, i.e., the number of response variables coincides with the number of species n_s. For each species, a statistical distribution $y_{ij} \sim D\{L_{ij}, \sigma_j^2\})$ is given, where L_{ij} is the mathematical expectation of the species density j at point i and σ_j^2 is the variance parameter (not used in the case of a Poisson or Bernoulli distribution). In the case of a normal distribution, the value of L_{ij} is modeled as a linear function of two groups of predictors representing fixed and random factors:

$$L_{ij} = \sum_{k=1}^{n_c} x_{ij}\beta_{jk} + \varepsilon_{ij}, \quad \text{where} \quad \varepsilon_{ij} = \sum_{h=1}^{n_f} \eta_{ih}\lambda_{jh}(z_i). \quad (1)$$

The first term of expression (1), which models the effect of fixed factors, is a common form of linear regression, where x_{ik} is the value of the kth environmental variable at point i, $k = 1... n_c$ and β_{jk} is the regression coefficient representing the proportion of the linear response of the species j to this covariate. To enable parametrization of the model with sparse data or rare species, the distribution of regression coefficients is assumed to be $\beta_{j\cdot} \sim N(\mu, \mathbf{V})$, where vector μ is an estimate of the average response of the species to the measured covariates, and the variance-covariance matrix $\mathbf{V}$ corresponds to the variation of particular species relative to the mathematical expectation. Hereinafter, the dot in the expression $\beta_{j\cdot}$ means that the index k runs through all values from 1 to n_c for each fixed j.

The contribution of a set of random factors, including spatial autocorrelation and interspecific interactions, is modeled by the second term ε_{ij}, which is the sum of the products of n_f latent factors and their loads. Here, η_{ih}, $h = 1... n_f$ is the factor value for a selected point i, and $\lambda_{jh}(z_{i\cdot})$ is the factor loading on the species j from the latent factor h, which generalizes an arbitrary set of predictors $z_{i\cdot}$. If, in particular, we accept that $\lambda_{jh}(z_i) = \sum_{k=1}^{n_c} x_{ik}\lambda_{jhk}$, then the structure of covariances between species ε_{ij} becomes a function of the state of the environment determined by the initial set of variables, the covariates x. Some random factors may be related to the nested structure of the research plan (e.g., reservoir basin → river → sampling point), therefore, the model under consideration is interpreted as hierarchical.

The coefficients of model (1) are calculated based on data from observations x with the Bayesian methodology, which is based on an iterative process of adjusting the initial (a priori) estimates of the model parameters $\boldsymbol{\theta}$ and obtaining their resulting (a posteriori) distribution. This process is realized via the construction of long, iterative sequences of several Markov chain Monte Carlo (MCMC) algorithms, for which the transition distribution is determined by the function $P(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{X})$. The process of modeling is often quite long and continues until the distribution of the current values of the process approaches some stationary distribution. Visual and formal diagnostic techniques are used to check the convergence of chains. The cross-validation algorithm is used to check the adequacy of the model and to compare its different variants.

Relationship between the Model and the Main Theoretical Constructions of Community Ecology

After construction and diagnostics, the parameterized HMSC model (like any JSDM) can be used to explain the environmental processes in communities and/or for predictions. Figure 1 shows the relationships between the informational structure of the HMSC platform and the main tasks of the ecology of communities. The rectangles include the designations of the source data matrices, and the ellipses show the calculated parameters of the model (1), which can be used in the analysis of the structure of ecological niches and interspecific interactions in the community.

Some of the coefficients $\beta_{j\cdot} \sim N(\mu, \mathbf{V})$ of the HMSC model describing fixed effects determine the extent to which the variability of environmental factors $\mathbf{X}$ affects the occurrence and/or abundance of species. Each species has its own vector of β parameters that limits a certain volume of hyperspace and, hence, its ecological niche. However, the niche boundaries are determined not only by the external parameters but also by the variability of intrapopulation characteristics $\boldsymbol{\Gamma}$ (species-specific traits), such as the body size, morphological features, or mode of nutrition in animals, the seed size or life form in plants, etc.

The phylogenetic relationship between species is another important, fixed effect determining the division of a community into ecological niches. In order to

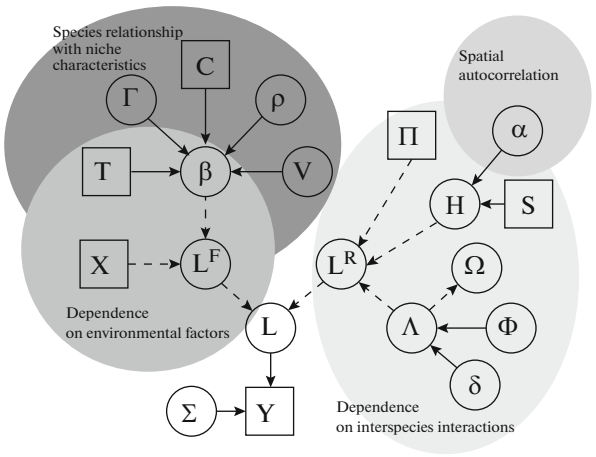

Fig. 1. Relationships between theoretical constructions of the community ecology and THE statistical structure of the HMSC platform (Ovaskainen and Abrego, 2020). Matrices of initial data: **Y**, species abundance; **X**, environmental factors; **T**, species properties; **C**, phylogenetic correlations; **Π**, research plan; **S**, geographical coordinates. Model variables and parameters: **L**, linear predictors; $\mathbf{L}^F$, fixed effects; $\mathbf{L}^R$, random effects; β, species niches; **Γ**, effect of species characteristics in niche; ρ, phylogenetic signal in niche; **V**, residual covariance in niche; **H**, factor loadings of biotopes; α, spatial scale of biotopes; **Λ**, factor loadings of species; **Ω**, species-to-species association matrix; **Φ**, local losses of species loadings; δ, global losses of species loadings; **Σ**, matrix of residual variance.

structure niches based on this feature, the phylogenetic tree is transformed into a matrix $\mathbf{C}$ $n_s \times n_s$, the elements of which (c_{ij} = 0–1) evaluate phylogenetic correlations defined as the proportion of the total evolutionary time for each pair of species i and j. HMSC realizes a phylogenetic correlation model as $\beta_f \sim N(\mu_f, \mathbf{W})$, where $\mathbf{W} = \rho\mathbf{C} + (1 - \rho)\mathbf{I}$, and the calculated parameter ρ = 0–1 measures the strength of the phylogenetic signal. If we assume that the niches are completely phylogenetically structured, then ρ = 1 and the coefficients of the model have a multidimensional normal distribution $\beta_f \sim N(\mu_f, \mathbf{C})$. This model has the same expectation μ_f for all species, but it predicts that phylogenetically close species will, on average, have a smaller statistical dispersion than phylogenetically distant species.

The totality of random effects of HMSC (shown on the right in Fig. 1) simulates the effect of different biotic or abiotic factors on the variability of the response **Y** (without changing its mathematical expectation). In most cases, in the realization of the research plan, the sample points are associated with spatial coordinates, and the dependence between the residues of ε_{ij} is then due to a phenomenon called spatial autocorrelation (observations at points located close to each other will be probably more similar than for sample units located far from each other). The HMSC models any user-defined autocovariance structure dependent on the distance d_{ij} between the selected points i–j. The exponential function $f(d_{ij}) = \sigma_S^2 \exp(-d_{ij}/\alpha)$ is most frequently used where the spatial variance (σ_S^2) and the scale vector (α) are positive parameters of the spatial random effect, which is estimated via construction of the model.

If the hypothesis that all species in a community function independently is rejected, then the totality of statistically significant positive or negative interactions between species may eventually affect the individual abundance **Y** of each of them. For this reason, it is advisable to include a random effect in the multidimensional analysis that takes into account additional information about the species that co-occur "more often than by accident." The last phrase indicates the simultaneous presence of a pair of species at the ith site with a probability exceeding that expected from the similarity of the $\beta_{i.}$ parameters of their niches. The associativity effect of species in a matrix form is written as $L_i^R \sim N(0, \mathbf{\Omega})$, where $\mathbf{\Omega} = \mathbf{\Lambda}^T\mathbf{\Lambda}$ is an ecologically limited correlation matrix of species. Thus, this group of random effects generates residual covariances over and above those taken into account by fixed effects, i.e., it distinguishes only those associations that cannot be explained by ecological covariates x_{ik} that are already included in the model.

Composition of the Initial Data

The construction of HMSC models is considered on the example of an analysis of hydrobiological survey data on bottom communities of the middle and lower Volga basin (Zinchenko, 2011) in different months of the growing season 1990–2019. The hydrobiological survey of macrozoobenthos was carried out in 90 small and 12 medium plain rivers that are tributaries of the Kuibyshev, Saratov, and Volgograd reservoirs, including six rivers of the arid region of the Elton Lake basin (Fig. 2). The medium rivers were divided into approximately homogeneous sections: upper, middle, and lower reaches and the mouth. Each of small rivers was taken as an integral object. Thus, 132 local communities were studied, in each of which up to 40 species of macrozoobenthos were identified with common methods. A total of 1400 samples were analyzed, and 740 species and taxa above the rank of species were identified. The specific abundance (ind./m^2) was used to form the response matrix **Y**.

Thirty environmental factors were monitored simultaneously at the same sampling points: the hydrological parameters of the watercourses, water-quality indicators, the content of the main chemical components (composition of bottom sediments, oxygen saturation of water, mineralization, etc.), and others. Raster tables with a resolution of 2.5' containing the main meteorological and geomorphological parameters for the study region were downloaded from

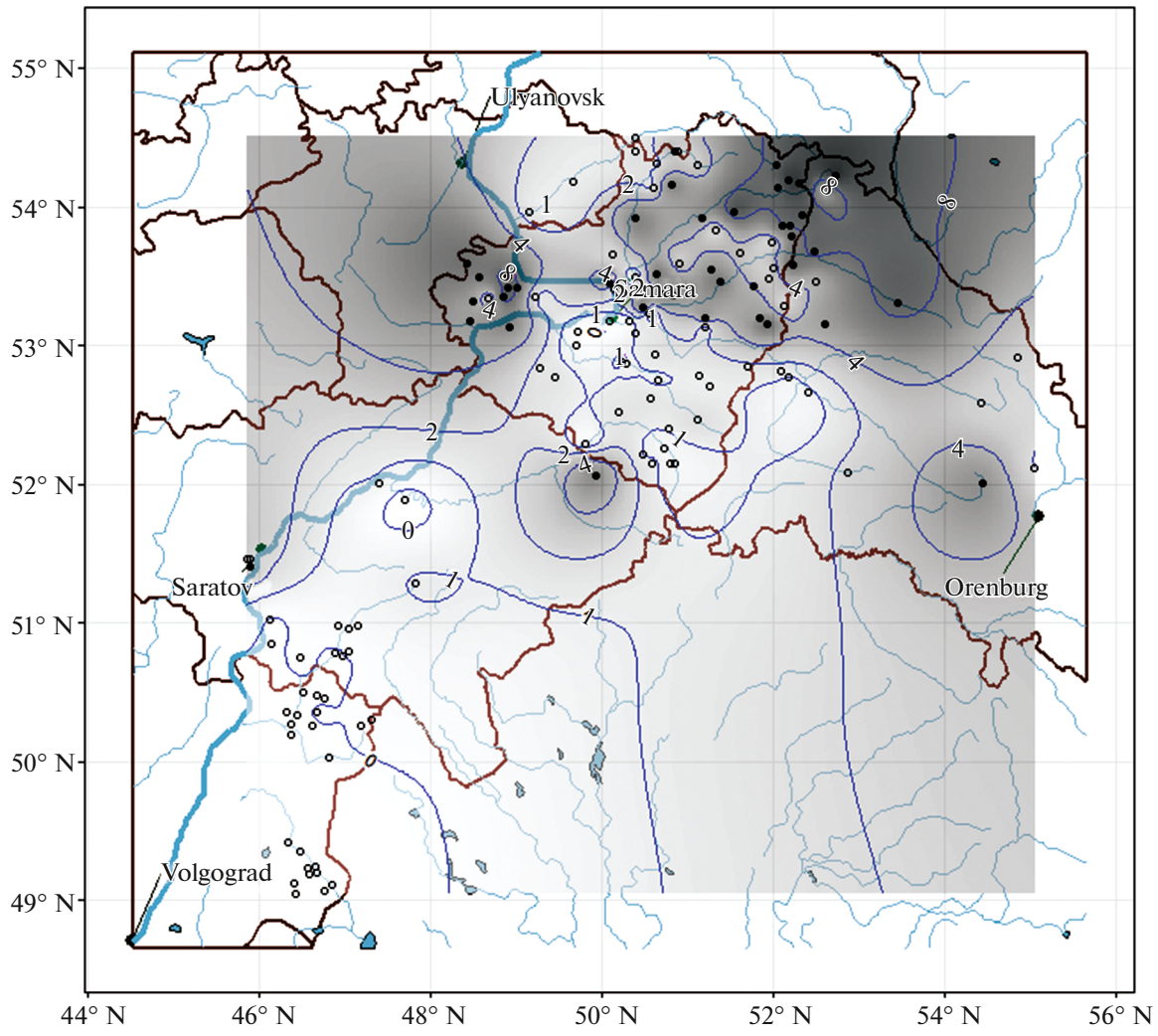

Fig. 2. Map of the study area, regions of hydrobiological survey (• indicates points where Prodiamesinae larvae were found, and **o** indicates their absence in a sample), and the spatial distribution of the subfamily abundance, ln(ind.)/m^2, predicted with the HMSC model.

the open-access datasets WorldClim and Environmental Rasters for Ecological Modeling (ENVIREM). These data were used to model fixed $\mathbf{L}^F$ and random $\mathbf{L}^R$ effects.

RESULTS

Models of the Spatial Distribution of One Species

A set of one-dimensional HMSC models of the distribution of the abundance of the most important species and taxonomic groups was constructed based on data from a study of benthic communities. This made it possible to draw certain conclusions about their relationship with environmental factors and environmental preferences within the studied region. We consider the analysis method on the example of the subfamily Prodiamesinae (Diptera, Chironomidae), all species of which were conditionally assumed to be ecologically identical, and their abundances were summed up and logarithmized. In total, the species of this taxon were detected in 41 of the surveyed 132 river sites.

The full model (*m*1) was built based on four geophysical and climatic parameters, three water-quality parameters (fixed factors), the geographical coordinates of the sites, and the categories of ground types of river bottoms (random factors *Rivers* and *Ground*, respectively). An a posteriori distribution of the model coefficients was obtained based on 30000 iterations of four Markov chain Monte Carlo algorithms. The degree of confidence for the coefficients was estimated with 2.5–97.5% quantiles (Table 1) and additional statistics, such as the effective chain length and the scaling factor. The relative importance of each parameter used to predict the magnitude of the response was estimated by their proportion in the decomposition of the

Table 1. Posterior distribution of HMSC-model coefficients for the prediction of the spatial distribution of Prodiamesinae taxa abundance

Name and designation of factors		Average	Standard deviation	Quantiles of distribution		Proportion of response variance explained	
				2.5%	97.5%	factor	group
Average annual temperature	β[*MTemp*]	–0.0532	0.0685	–0.182	0.086	12.16%	43.84%
Precipitation in the driest quarter	β[*PrecDQ*]	–0.0148	0.0514	–0.117	0.082	4.36%	
Altitude	β[*Alt*]	0.0246	0.0098	0.00553	0.044	22.14%	
Roughness index of relief	β[*TRI*]	0.0116	0.0096	–0.0071	0.031	5.18%	
Water mineralization	β[*Miner*]	1.5E-05	6.1E-05	–0.0001	0.00014	2.24%	6.64%
Ammonium nitrogen	β[NH_4]	0.0287	0.116	–0.202	0.251	1.99%	
Oxygen saturation	β[O_2]	–0.0089	0.0143	–0.0382	0.0193	2.41%	
Spatial scale	α[*Rivers*]	2.861	2.98	0	9.83	48.68%	49.52%
Category of grounds	λ[*Ground*]	–0.618	0.69	–2.44	–0.01	0.84%	

total explained variance for all fixed and random factors.

The quality of the resulting model was evaluated based on the residual standard deviation (*RMSE* = 3.11), the coefficient of determination, which determines the proportion of the total variance of the response variable **Y** (which is explained by the structure of the model ($R^2 = 0.613$)), as well as the widely applicable information criterion (*WAIC* = 9634).

As follows from an analysis of the model coefficients (Table 1), the biotope characteristics, the hydrochemical parameters of the water quality, and the sediment composition account for only 7.5% of the explained variance. These factors can be considered statistically insignificant, since the 95% confidence interval of their coefficients includes zero. In this regard, three more candidate models with fewer variables were considered:

—(*m*2) with only seven fixed factors: *RMSE* = 3.99, $R^2 = 0.287$;

—(*m*3) based on the factors characterizing the environmental conditions in the biotope (*Miner*, NH_4, O_2, and *Ground*): *RMSE* = 4.48, $R^2 = 0.116$;

—(*m*4) with climatic (*MTemp*, *PrecDQ*) and geophysical (*Alt*, *TRI*) fixed factors and a random factor *Rivers*, which determines the spatial autocorrelation dependence: *RMSE* = 2.184, $R^2 = 0.854$, *WAIC* = 9262.

Figure 2 shows a schematic map of the study area. The sampling sites are marked with circles (the points where Prodiamesinae were recorded are filled in black). The distribution of the population density of this subfamily predicted with the *m*4 HMSC model is shown in grays of different intensities. The contour lines mark the isolines of the abundance logarithmized (ind./m^2).

Joint Distribution of the Species Ensemble

The multidimensional HMSC model was constructed to estimate the spatial distribution of a community of 31 species of chironomid larvae. Table 2 presents their names, the frequency of occurrence, and the phylogenetic tree. A transformation leading to the χ^2-distance was previously performed for the species abundance. This is probably the most reasonable compromise when one takes into account both the role of the leading components and the contribution of rare or not numerous taxa (Legendre and Gallagher, 2001).

The variables used as model predictors were the same as those for the *m*1 model (Table 1) with the addition of matrix **C** of phylogenetic correlations. The biotope characteristics expressed by the categorical variable *Ground* (from 1, pure sand or pebbles, to 6, black silt and plant residues) were interpreted this time as a fixed factor. In Table 2, cells for species whose a posteriori distribution of coefficients is statistically significantly shifted to the positive region, i.e., in the direction of increasing corresponding predictor, are marked in black. The reverse situation, in which a decrease in the independent variable leads to an increase in the number of species, is marked in gray. The proportions of the VR variance explained by the constructed model and of the overall variance of the response **Y** are presented for each group of factors (geoclimatic and hydrochemical parameters, as well as the spatial autocovariance *Rivers*).

The phylogenetic signal ρ has a posteriori distribution with an average of 0.991 ± 0.00024, which provides a strong evidence of a very significant effect of the taxonomic hierarchy in the determination of ecological niches.

The vector of the spatial scaling factor α has a characteristic pulsating sequence of values with averages $\alpha_1 = 2.97$, $\alpha_2 = 0.006$, $\alpha_3 = 3.34$, $\alpha_4 = 2.35$, $\alpha_5 = 0.22$,

Table 2. Results of construction of the HMSC model for chironomid community

Phylogenetic tree of chironomid community, species names and codes		W, samples	Geoclimatic parameters					Characteristics of biotopes					VR, % *Rivers*	R^2 determinations
			MTemp	*Prec DQ*	*Alt*	*TRI*	VR, %	*Miner*	NH_4	O_2	*Ground*	VR, %		
Tanytarsus pallidicornis	ChTar.p.	121					45.8					28.7	25.6	0.043
Tanytarsus kharaensis	ChTar.kr	69					14.0			■		26.6	59.3	0.330
Tanytarsus sp.	ChTar.sp	333	▒			■	46.8			■		20.2	33.0	0.154
Paratanytarsus confusus	ChPtt.co	123		■			60.0					20.9	19.0	0.074
Chironomus obtusidens	ChChi.o.	46					44.4					27.5	28.1	0.033
Chironomus aprilinus	ChChi.ap	64	■				34.5				■	23.7	41.9	0.243
Chironomus salinarius	ChChi.sr	148					5.0	■	■			71.5	23.5	0.985
Dicrotendipes nervosus	ChDic.n.	90	▒		▒		65.0					14.4	20.6	0.074
Endochironomus albipennis	ChEch.a.	53					51.6	▒				16.7	31.7	0.148
Glyptotendipes salinus	ChGly.sl	75					5.7			■	■	8.1	86.2	0.834
Glyptotendipes gripekoveni	ChGly.g.	49				▒	48.0					26.8	25.2	0.034
Microchironomus tener	ChMch.t.	96			▒	▒	63.3					17.2	19.5	0.097
Microchironomus deribae	ChMch.d.	72					4.8		▒	■	■	10.7	84.5	0.971
Microtendipes pedellus	ChMit.p.	55			▒	■	57.2				▒	23.6	19.2	0.103
Paratendipes albimanus	ChPat.a.	61	▒				56.5					20.3	23.2	0.056
Parachironomus varus	ChPchvr.	50					48.9					24.2	26.9	0.098
Polypedilum nubeculosum	ChPol.n.	448	■				30.8	▒				12.3	56.9	0.182
Polypedilum bicrenatum	ChPol.b.	68		■		▒	68.4					15.4	16.2	0.145
Polypedilum scalaenum	ChPol.s.	118					49.5					27.8	22.7	0.041
Stictochironomus crassiforceps	ChSchcs.	112					32.1				■	31.3	36.6	0.097
Procladius ferrugineus	ChPrc.f.	395					42.8			▒		40.0	17.2	0.120
Ablabesmyia monilis	ChAbl.m.	78					42.5					20.8	36.7	0.056
Tanypus punctipennis	ChTan.p.	112	▒	■	■	■	73.3			▒		3.9	22.8	0.287
Cricotopus salinophilus	ChCri.sf	211					5.6	■	■			45.0	49.4	0.958
Cricotopus bicinctus	ChCri.b.	177			■		54.9				▒	31.4	13.6	0.127
Corynoneura scutellata	ChCor.s.	19				▒	49.7					26.5	23.8	0.031
Paracladius conversus	ChPlc.co	98			■	■	64.1					22.1	13.9	0.148
Psectrocladius sordidellus	ChPse.s.	75	■				56.6			▒		19.5	23.8	0.224
Rheocricotopus fuscipes	ChRhe.f.	49			■		63.7					22.1	14.2	0.087
Prodiamesa olivacea	ChPro.o.	126			■		56.4				▒	29.2	14.4	0.152
Monodiamesa bathyphila	ChMnd.ba	58			■		60.0				■	25.1	15.0	0.101

W, species occurrence in samples (n = 1400); VR, the proportion of response variance explained by a group of factors. Statistically significant factors (designations given in Table 1) positively affecting the species abundance are marked in black, and factors with a negative effect are in gray.

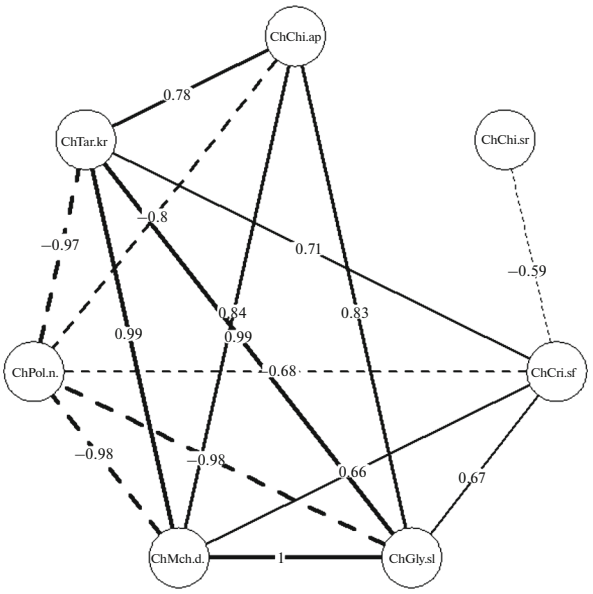

Fig. 3. Graph of correlations between the abundance of chironomid species after the elimination of environmental factors.

and we have not yet found a reasonable explanation for this phenomenon.

From the matrix **Ω**, which determines the residual covariance relationships between the considered species, only those negative or positive associations for which a posteriori probability is at least 0.95 were selected. They comprised seven of the 31 species, and their correlation graph is presented in Fig. 3.

DISCUSSION

Analysis of the coefficients of the HMSC model makes it possible to determine the priorities of external factors according to the degree of their effect on the spatial distribution of the population density of species. In particular, the presented results indicate a strong dependence of the abundance of Prodiamesinae on the cartographic coordinates and the altitude above sea level of the surveyed area. This is usually typical for ranges that limit a clearly defined geographical cluster. Actually, the larvae of this subfamily are expressed rheo- and oxybionts and inhabit stony–sandy biotopes of flowing rivers of the Arctic-alpine type (Makarchenko, E. and Makarchenko, M., 1999). In our study, they mainly occurred in the rivers of the Bugulma–Belebey Upland of the forest–steppe province of the high Transvolga region of Samara oblast (Fig. 2).

Some other species are characterized by a pronounced dependence on the hydrochemical conditions of the aquatic environment and the biotope type. For example, *Chironomus salinarius* and *Cricotopus salinophilus* are typical halophiles and inhabit watercourses of the Elton region with high mineralization and ammonium-ion content. The proportion of variance in the abundance of these species, which is explained by hydrochemical parameters, is high and constitutes up to 71.5%, while the coefficient of determination R^2 reaches 0.985. Similarly, euryoxybiont species such as *Prodiamesa olivacea* occur mainly in rivers on sandy–silty sediments, and their occurrence therefore depends very little on the effect of geoclimatic parameters. The proportion of variance explained by spatial autocorrelation is very high in most cases, which probably follows from the mosaic nature of the environment (Hutchinson, 1959), where particular internal patterns exist.

The low coefficient of determination in many species from Table 2 is explained not by the weak possibilities for the construction of HMSC models but by a number of objective reasons. First, many eurybiont species relatively evenly occur throughout the territory; the boundaries of their ranges are unclear, and there are no distinctly expressed geographical clusters. Second, we used a very limited list of environmental covariates $x_{.k}$, and it is quite possible to suggest that the leading factor determining the population density of a species was just not included in this list (this may be an unaccounted-for hydrochemical parameter, the flow velocity, or some hard-to-formalize landscape feature).

In this regard, it is extremely important to select the composition of environmental predictors that are used in the model construction. In this case, there is no general theory, and the selection of potentially important factors is usually based on the experience and intuition of the researcher. However, it is not always reasonable to seek to use as many source variables as possible just in case. This causes an unjustified increase in the complexity of the model and, consequently, the risk of retraining, which often leads to a decrease in the predictive power of the model rather than an increase. This is confirmed by a significant increase in the coefficient of determination R^2 of the model ($m4$) as compared to the full model ($m1$).

Computer-intensive methods for the selection of informative variables and constructing models of optimal complexity have recently been greatly developed (genetic algorithm, resampling, cross-checking (Shitikov and Rosenberg, 2014; Shitikov and Mastitsky, 2017)). The use of these algorithms in HMSC is often problematic due to the high resource intensity of the construction of MCMC chains of sufficient efficiency. The recommendation to use R^2 or informational criteria to compare models is not, in the full sense, a test of scientific hypotheses, since the differences in these values are not statistically interpretable. For example, can the decrease in the Widely criterion from 9634 to 9262 be considered essential for the selection of the candidate model, or is it due to occasional circumstances?

When modeling the relationship between the environmental conditions and the species occurrence, the authors of HMSC closely link their approach with the concept of an ecological niche. However, the use of the term "niche" in the context of correlative SDMs is subject to criticism, because, "in order to model a niche, it is necessary to understand how the morphology, physiology, and especially the behavior of organisms are determined by environmental factors and to assess how habitat conditions affect the adaptability of a species (growth, survival, and reproduction)" (Kearney, 2006, p. 186). In the considered example concerning the abundance of Prodiamesinae, the leading factors, the altitude above sea level and average annual temperature, are, of course, not directly parameters of the fundamental niche and are far from being limited to them, although they indirectly determine the mechanisms of ecological processes and characteristics of biotopes.

Ovaskainen et al. broaden the scope of the fundamental niche by including a matrix of interspecific associations $\mathbf{\Omega}$ in the analysis that is independent of environmental factors, thereby restoring the concept of competitive exclusion and symbiotic relations, which is important for niche theory. Nevertheless, the very concept of an "ecological niche" continues to be rather vague. It is possible to set the hyperspace dimension and estimate the centers of statistical distributions of all independent variables for each of the species, but this does not yet make it possible to operate mathematically with many theoretical constructions of the niche. Such important concepts as "hyperspace volume" and "packing density" require the preliminary determination of not only average, but also boundary (i.e. normative), values of niche parameters, and a quantitative assessment of the differences between the types of niches and the degree of their overlap is associated with the justification of a multidimensional distance metric.

CONCLUSIONS

In conclusion, it should be noted that the developed methodological platform and the package of functions of HMSC community modeling, in our opinion, is a universal and complex computing environment that allows the integration of many data sets and provides answers to a wide range of issues. Specialists in the field of community ecology may be interested in the following attractive opportunities of the validated method.

The models of the distribution and subsequent prediction of the species population density are built not only on the basis of the set of x_k parameters determining environmental conditions; they also take into account the variability of characteristic features of the species γ, the phylogenetic relationship ρ, and the function of spatial autocovariance α.

The use of HMSC enables the construction of SDM models, both at the level of particular species and collectively for arbitrary communities; in the latter case, the response prediction is made with consideration of the contribution of interspecific relationships calculated from the matrix $\mathbf{\Omega}$ of residual associations.

The method can be applied to different variants of research schemes (including hierarchical, temporal or spatial plans) and to many types of empirical data distributions (presence/absence, counting parameters and continuous values).

ACKNOWLEDGMENTS

We are grateful to to O . Ovaskainen (University of Helsinki, Finland) and G. Tikhonov (Aalto University, Finland) for their methodological support and valuable comments on the manuscript.

COMPLIANCE WITH ETHICAL STANDARDS

Conflict of interests. The authors declare that they have no conflicts of interest.

Statement on animal welfare. All applicable international, national, and/or institutional guidelines for the care and use of animals were followed.

REFERENCES

Abrego, N., Norberg, A., and Ovaskainen, O., Measuring and predicting the influence of traits on the assembly processes of wood-inhabiting fungi, *J. Ecol.*, 2017, vol. 105, no. 4, pp. 1070–1081.

Araújo, M.B., Anderson, R.P., Barbosa, A.M., Beale, C.M., Dormann, C.F., et al., Standards for distribution models in biodiversity assessments, *Sci. Adv.*, 2019, vol. 5, no. 1, p. eaat4858.

Breiner, F.T., Nobis, M.P., Bergamini, A., and Guisan, A., Optimizing ensembles of small models for predicting the distribution of species with few occurrences, *Methods Ecol. Evol.*, 2018, vol. 9, pp. 802–808.

Brooker, R.W., Plant-plant interactions and environmental change, *New Phytol.*, 2006, vol. 171, pp. 271–284.

Busby, J.R., BIOCLIM—a bioclimate analysis and prediction system, *Plant Prot. Q.*, 1991, vol. 6, pp. 8–9.

Calabrese, J.M., Certain, G., Kraan, C., and Dormann, C.F., Stacking species distribution models and adjusting bias by linking them to macroecological models, *Global Ecol. Biogeogr.*, 2014, vol. 23, pp. 99–112.

Clark, J.S., Gelfand, A.E., Woodall, C.W., and Zhu, K., More than the sum of the parts: forest climate response from joint species distribution models, *Ecol. Appl.*, 2014, vol. 24, pp. 990–999.

D'Amen, M., Rahbek, C., Zimmermann, N.E., and Guisan, A., Spatial predictions at the community level: from current approaches to future frameworks, *Biol. Rev.*, 2017, vol. 92, pp. 169–187.

Franklin, J., *Mapping Species Distributions: Spatial Inference and Prediction*, Cambridge: Cambridge Univ. Press, 2009.

Golovatyuk, L.V., Shitikov, V.K., and Zinchenko, T.D., Estimation of the zonal distribution of species of bot-

tom communities in lowland rivers of the Middle and Lower Volga basin, *Biol. Bull.* (Moscow), 2018, vol. 45, no. 10, pp. 1262–1268.

Guisan, A., Thuiller, W., and Zimmermann, N.E., *Habitat Suitability and Distribution Models: With Applications in R*, Cambridge: Cambridge Univ. Press, 2017.

Guisande, C., Garcia-Rosello, E., Heine, J., Gonzalez-Dacosta, J., Gonzalez-Vilas, L., et al., SPEDInstabR: an algorithm based on a fuctuation index for selecting predictors in species distribution modeling, *Ecol. Inf.*, 2017, vol. 37, pp. 18–23.

Hastie, T. and Fithian, W., Inference from controversy, *Ecography*, 2013, vol. 36, pp. 864–867.

Hutchinson, G.E., Homage to Santa Rosalia or Why are there so many kinds of animals? *Am. Nat.*, 1959, vol. 43, no. 870, pp. 145–159.

Johnson, D.H., The comparison of usage and availability measurements for evaluating resource preference, *Ecology*, 1980, vol. 61, no. 1, pp. 65–71.

Kearney, M.R., Habitat, environment and niche: What are we modeling? *Oikos*, 2006, vol. 115, no. 1, pp. 186–191.

Legendre, P. and Gallagher, E., Ecologically meaningful transformations for ordination of species data, *Oecologia*, 2001, vol. 129, pp. 271–280.

Legendre, P. and Legendre, L., *Numerical Ecology*, Amsterdam: Elsevier, 2012.

Lissovsky, A.A. and Dudov, S.V., Species-distribution modeling: advantages and limitations of its application. 2. MaxEnt, *Biol. Bull. Rev.*, 2021, vol. 11, no. 3, pp. 265–275.

Lissovsky, A.A., Dudov, S.V., and Obolenskaya, E.V., Species-distribution modeling: advantages and limitations of its application. 1. General approaches, *Biol. Bull. Rev.*, 2021, vol. 11, no. 3, pp. 254–264.

Makarchenko, E.A. and Makarchenko, M.A., Chironomidae—non-biting midges, in *Opredelitel' presnovodnykh bespozvonochnykh Rossii i sopredel'nykh territorii. Tom 4. Vysshie nasekomye. Dvukrylye* (Guide for Identification of Freshwater Invertebrates of Russia and Adjacent Territories, Vol. 4: Higher Insects. Dipterans), St. Petersburg: Zool. Inst., Ross. Akad. Nauk, 1999, pp. 210–296.

Norberg, A., Abrego, N., Blanchet, F.G., Adler, F.R., Anderson, B.J., et al., A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels, *Ecol. Monogr.*, 2019, vol. 89, no. 3, p. e01370.

Ovaskainen, O. and Abrego, N., *Species Distribution Modelling: With Applications in R*, Cambridge: Cambridge Univ. Press, 2020.

Ovaskainen, O., Abrego, N., Halme, P., and Dunson, D., Using latent variable models to identify large networks of species-to-species associations at different spatial scales, *Methods Ecol. Evol.*, 2016a, vol. 7, pp. 549–555.

Ovaskainen, O., Roy, D.B., Fox, R., and Anderson, B.J., Uncovering hidden spatial structure in species communities with spatially explicit joint species distribution models, *Methods Ecol. Evol.*, 2016b, vol. 7, pp. 428–436.

Ovaskainen, O., Tikhonov, G., Norberg, A., Blanchet, F.G., Duan, L., et al., How to make more out of community data? A conceptual framework and its implementation as models and software, *Ecol. Lett.*, 2017, vol. 20, pp. 561–576.

Peterson, A.T., Soberón, J., Pearson, R.G., Anderson, R.P., Martínez-Meyer, E., et al., *Ecological Niches and Geographic Distributions (MPB-49)*, Princeton: Princeton Univ. Press, 2011.

Phillips, S.J., Anderson, R.P., and Schapire, R.E., Maximum entropy modeling of species geographic distributions, *Ecol. Model.*, 2006, vol. 190, nos. 3–4, pp. 231–259.

Pollock, L.J., Tingley, R., Morris, W.K., Golding, N., O'Hara, R.B., et al., Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM), *Methods Ecol. Evol.*, 2014, vol. 5, pp. 397–406.

Rozenberg, G.S., Shitikov, V.K., and Zinchenko, T.D., Mark Vellend. The theory of ecological communities. Princeton; Oxford: Princeton University Press, 2016, *Zh. Obshch. Biol.*, 2020, vol. 81, no. 5, pp. 394–400.

Shitikov, V.K. and Mastitskii, S.E., Classification, regression and other Data Mining algorithms using R, 2017. https://stok1946.blogspot.com. Cited October 10, 2020.

Shitikov, V.K. and Rozenberg, G.S., *Randomizatsiia i butstrep: statisticheskii analiz v biologii i ekologii s ispol'zovaniem R* (Randomization and Bootstrap: Statistical Analysis in Biology and Ecology using R), Tolyatti: Kassandra, 2014.

Thorson, J.T., Ianelli, J.N., Larsen, E.A., Ries, L., Scheuerell, M.D., et al., Joint dynamic species distribution models: a tool for community ordination and spatio-temporal monitoring, *Global Ecol. Biogeogr.*, 2016, vol. 25, pp. 1144–1158.

Tikhonov, G., Abrego, N., Dunson, D., and Ovaskainen, O., Using joint species distribution models for evaluating how species-to-species associations depend on the environmental context, *Methods Ecol. Evol.*, 2017, vol. 8, pp. 443–452.

Tikhonov, G., Opedal, Ø.H., Abrego, N., Lehikoinen, A., de Jonge, M.M.J., et al., Joint species distribution modelling with the R-package H_{MSC}, *Methods Ecol. Evol.*, 2020, vol. 11, pp. 442–447.

Vellend, M., *The Theory of Ecological Communities*, Princeton: Princeton Univ. Press, 2016.

Warton, D.I., Blanchet, F.G., O'Hara, R.B., Ovaskainen, O., Taskinen, S., et al., So many variables: joint modeling in community ecology, *Trends Ecol. Evol.*, 2015, vol. 30, pp. 766–779.

Zinchenko, T.D., Bioindication role of chironomids (Diptera, Chironomidae) in aquatic ecosystems: problems and perspectives, *Usp. Sovrem. Biol.*, 2009, vol. 129, no. 3, pp. 257–270.

Zinchenko, T.D., *Ekologo-faunisticheskaya kharakteristika khironomid (Diptera, Chironomidae) malykh rek basseina Srednei i Nizhnei Volgi (Atlas)* (Ecological-Faunistic Characteristic of Chironomids (Diptera, Chironomidae) from the Small River of Central and Lower Volga: Atlas), Toglyatti: Kassandra, 2011.

Zurell, D., Thuiller, W., Pagel, J., Cabral, J.S., Münkemüller, T., et al., Benchmarking novel approaches for modelling species range dynamics, *Global Change Biol.*, 2016, vol. 22, no. 8, pp. 2651–2664.

Translated by N. Ruban