Ecology, 92(2), 2011, pp. 289–295

Making more out of sparse data: hierarchical modeling of species communities

Otso Ovaskainen[1,3] and Janne Soininen[2]

[1]*Department of Biosciences, University of Helsinki, Viikinkaari 1, FI-00014 University of Helsinki, Finland*
[2]*Department of Environmental Sciences, University of Helsinki, Viikinkaari 1, FI-00014 University of Helsinki, Finland*

Abstract. Community ecologists and conservation biologists often work with data that are too sparse for achieving reliable inference with species-specific approaches. Here we explore the idea of combining species-specific models into a single hierarchical model. The community component of the model seeks for shared patterns in how the species respond to environmental covariates. We illustrate the modeling framework in the context of logistic regression and presence–absence data, but a similar hierarchical structure could also be used in many other types of applications. We first use simulated data to illustrate that the community component can improve parameterization of species-specific models especially for rare species, for which the data would be too sparse to be informative alone. We then apply the community model to real data on 500 diatom species to show that it has much greater predictive power than a collection of independent species-specific models. We use the modeling approach to show that roughly one-third of distance decay in community similarity can be explained by two variables characterizing water quality, rare species typically preferring nutrient-poor waters with high pH, and common species showing a more general pattern of resource use.

Key words: Bayesian inference; diatoms; hierarchical modeling; predictive power; species community; species distribution model; statistical modeling.

Introduction

One central question in basic and applied ecology is how the abundance and distribution of species depend on environmental covariates. This question can be approached by species distribution modeling, including a wide variety of statistical techniques for correlating species occurrences against environmental covariates (e.g., Guisan and Zimmermann 2000, Pearce and Ferrier 2000, Thuiller et al. 2003, Guisan and Thuiller 2005, Elith et al. 2006, Latimer et al. 2006, Phillips et al. 2006). Species distribution models (or bioclimate envelope models) are increasingly used to predict how species may respond to changing environmental conditions, such as habitat loss or climate change (e.g., Warren et al. 2001, Pearson and Dawson 2003). Species distribution models are central also in conservation planning, as they can be used to rank the outcomes of alternative conservation scenarios (Moilanen et al. 2005, Peralvo et al. 2007, Kremen et al. 2008).

Often the interest is not on a single species but on an entire community of species. In this case, the data may be too sparse for reliable inference in the case of rare species, which are often excluded from the analyses. This is problematic especially in conservation applications, where the interest is especially on the rare species. One approach that avoids this problem is to formulate a model directly for a summary statistic, such as species richness or community similarity (e.g., Currie 1991, Nekola and White 1999, Green et al. 2004, Kreft et al. 2008, Morlon et al. 2008). Other widely used approaches include a variety of ordination techniques, such as canonical correspondence analysis (ter Braak 1986) and nonmetric multidimensional scaling (Kruskal 1964), and the use of multivariate regression trees (De'ath 2002). These approaches make it possible to utilize data on all species, and they help to get the "big picture" out of the data, but they fail to show in detail how the relationship between environmental covariates and the species community builds up from species-specific responses (Gelfand et al. 2005, 2006). Moreover, these approaches can be difficult to use for predicting, e.g., how the community would respond to changing environmental conditions.

In this paper, we propose a hierarchical approach that combines a set of species-specific models by a community-level model. Our method builds from multivariate adaptive regression splines (e.g., Leathwick et al. 2006), which are of multivariate nature in the sense that the selection of environmental covariates is based on data on all species. However, in this modeling approach each species is eventually modeled independently, leading to limited statistical power especially for the case of rare species. The novelty of our approach is that species-specific models are linked to each other by a higher-level structure, and they are thus fitted simultaneously to the data. Informally, the model seeks for general patterns in

Manuscript received 22 June 2010; revised 22 September 2010; accepted 30 September 2010. Corresponding Editor: J. Elith.
[3] E-mail: otso.ovaskainen@helsinki.fi

how the individual species respond to the environmental covariates. The hierarchical structure makes it possible to include also species with very limited data, and it thus facilitates the analysis of sparse data sets with large numbers of very rare species. On top of the species-specific inference, the approach provides a compact summary of the entire community, which can be used to assess, e.g., how community similarity depends on variation in environmental covariates and other factors.

We first present the general modeling approach and illustrate its performance with simulated data. We then apply the community model to presence–absence data on diatoms in Finnish streams, and compare its predictive power against a collection of species-specific models.

Material and Methods

A hierarchical model of species community

While almost any kind of species-specific models can be connected by a hierarchical structure, we assume here for simplicity logistic regression applied to presence–absence data. We thus consider a community of n species inhabiting a set of m sites. We denote the data by the $n \times m$ dimensional matrix $\mathbf{y}$, with $y_{ij} = 1$ if species i is present in site j and $y_{ij} = 0$ if the species is absent from the site. We assume that the presence of the species is influenced by k covariates, which are organized into the $m \times k$ matrix $\mathbf{X}$. For each species i, the logistic regression models reads as

$$\mathrm{logit}\Big(\mathrm{P}(y_{ij} = 1)\Big) = \sum_{l=1}^{k} X_{jl}\beta_{il} \tag{1}$$

where the β_{il} are the regression coefficients to be estimated.

To combine the species-specific models into a model of the entire species community, we organize the regression coefficients β_{il} into a $n \times k$ matrix $\boldsymbol{\beta}$. We use a dot to single out a row or a column from a matrix, so that, e.g., $\boldsymbol{\beta}_{i\cdot} = (\beta_{il})_{l=1}^{k}$ denotes the vector of k regression coefficients for species i. We then assume that the responses of the species to the environmental covariates, measured by the regression coefficients, stem from a common distribution. As the baseline model, we assume that the $\boldsymbol{\beta}_i$ are distributed (independently among the species) multinormally as

$$\boldsymbol{\beta}_{i\cdot} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V}). \tag{2}$$

Here $\boldsymbol{\mu}$ is a vector of length k measuring the response of a typical species to the covariates. The diagonal elements of the $k \times k$ variance-covariance matrix $\mathbf{V}$ measure how much the species vary in their responses to the environmental covariates, and the off-diagonal terms measure the covariances in responses to pairs of environmental covariates.

Simulated data

To illustrate how the hierarchical approach works in practice, we first consider two examples with simulated data. In both examples, we generated a community of $n = 100$ species inhabiting a set of $m = 100$ sites. Each species was assumed to follow Eq. 1 with $k = 2$ parameters. These are the intercept, which relates to the overall prevalence of the species, and a regression coefficient measuring the species response to a single environmental covariate. In the first hypothetical community (H1), we set the community-level parameter values $\boldsymbol{\mu}$ and $\mathbf{V}$ so that the occupancy probability of a typical species increases with an increasing value of the covariate, and that this is especially the case for the rare species (Fig. 1A). In the second hypothetical community (H2), we assumed that the species may be specialized either to a low or a high value of the covariate. We generated this community by assuming a mixture model, in which a species belongs with probability p_l to group $l = 1, 2$, with parameters

$$\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\mu}_l, \mathbf{V}_l). \tag{3}$$

The community H2 was parameterized so that the rare species are specialized to either positive or negative values of the environmental covariate, whereas the common species are generalists (Fig. 1D).

We generated simulated presence–absence data for both of the hypothetical species communities (H1 and H2; see Appendix A for parameter values). For the sake of comparison, we fitted the species-specific models (Eq. 1) either independently for each species or by combining them with a community component (Eqs. 2 and 3).

Diatoms in Finnish streams

We next consider a case study on stream diatoms to test how the community model performs with real data compared to a set of independent species-specific models. Diatoms are unicellular microscopic algae that live in running waters either attached to benthic surfaces, or freely on various substrata. Diatoms disperse efficiently and are highly diverse both locally and regionally (Stevenson et al. 1996).

The sampling design is hierarchical, consisting of seven stream systems (called regions), within each of which there are 15 sampling sites. These seven drainage systems were chosen because they cover a large geographical extent and as long a nutrient concentration gradient as possible for streams in Finland. The sampled areas within the regions ranged between 1000 km^2 and 10 094 km^2. The presence–absence of diatoms was sampled in these 105 sites, with observations on 365 species. As is usual with presence–absence data, the absences are not as reliable as the presences, but for simplicity we omit here any observation error. For details on the sampling scheme, see Soininen (2008) and Appendix C. We assumed the following model:

$$\mathrm{logit}\Big(P(y_{ij} = 1)\Big) = \sum_{l} X_{jl}\beta_{il} + \varepsilon_j \tag{4}$$

where the random effect ε_j accounts for residual variation among the sampling sites. As the random

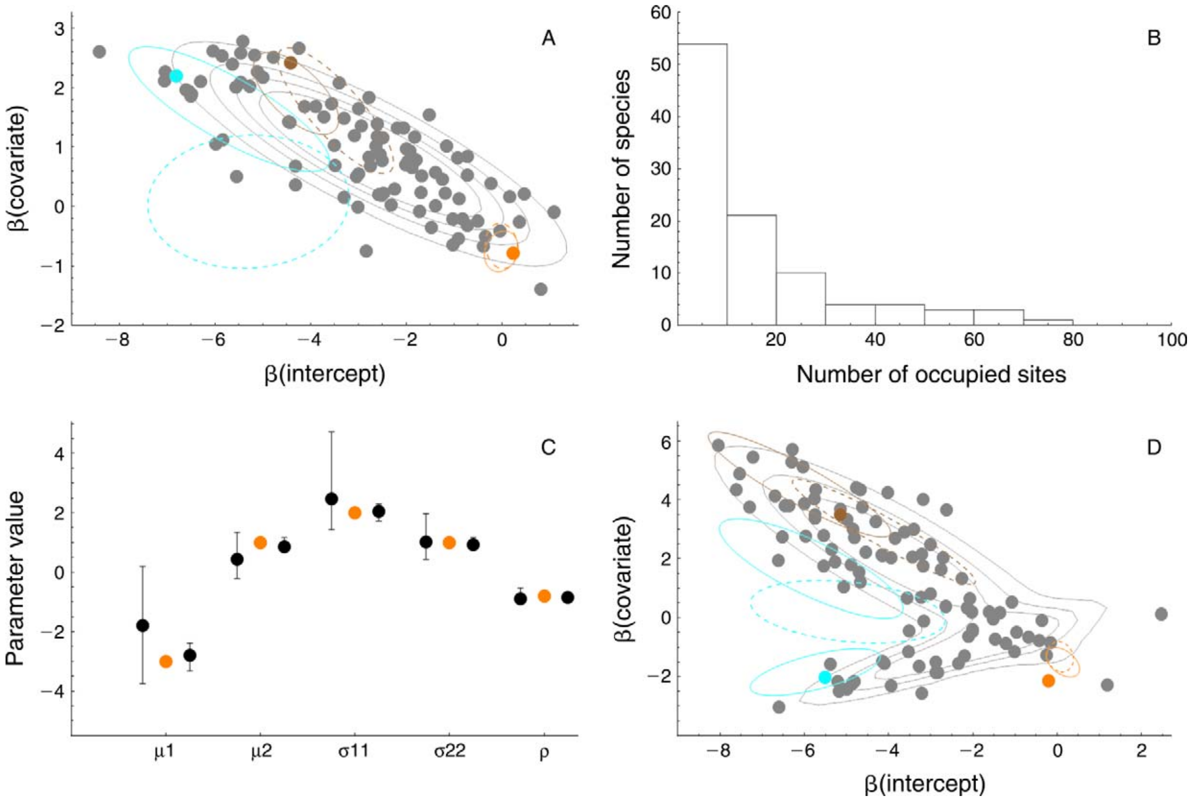

FIG. 1. Community models fitted to simulated data. The dots in panels A and D depict species-specific parameter values in the hypothetical communities H1 and H2, respectively. The gray contour lines show the estimated probability density of the community-level structure based on fitting the model to presence–absence data. The cyan, brown, and orange dots show a very rare, a relatively rare, and a common species, respectively, and the colored ellipses indicate the 75% posterior quantiles for these species, based on fitting the community model to data on 100 species (continuous lines) or the focal species only (dashed lines). Panel B shows the species abundance distribution of the community H1. In panel C the orange dots show the true community-level parameters for H1, and the black dots show the mean parameter estimates (error bars indicate the central 95% posterior credibility interval) if including five (left) or 100 (right) species. For parameter values, see Appendix A.

effect depends only on the site but not on the species, it influences species richness but not community composition. The motivation for this choice is that species richness often varies more among sites than would be expected from independent variation in species occurrences. Each ε_j was assumed to be distributed (independently among the sites) as $\varepsilon_j \sim \mathbf{N}(0, \sigma^2)$. We included in the model $k = 11$ covariates, consisting of seven region-specific intercepts (modeling the overall prevalence of each species in each region), and linear and quadratic effects of two environmental covariates describing water quality, obtained as the first two principal components of eight measured environmental parameters. The first component PC1 (30% of the total variance) increases with decreasing water conductivity, decreasing amount of total phosphorus and decreasing current velocity. This main gradient in the data thus reflects variation in ionic concentration and primary productivity, both of which have been identified as key variables associated with the distribution of diatoms in streams (Soininen et al. 2004). The second component PC2 (25% of the total variance) increases with decreasing water color and with increasing water pH. Thus PC2 reflects the variation in humic content in the water, also known to be important for determining diatom distributions across geographic regions in freshwater systems (Fallu et al. 2002, Soininen et al. 2004).

To examine the predictive power of the community model, we mimicked the situation in which species data would be available only for some of the sites. These training data originate from 35 sites out of the 105 sites, with five sites selected randomly from each region. We assumed that the diatom researcher would have the prior information that the entire community of stream diatoms includes ~500 species (Soininen et al. 2009, Heino et al. 2010). The training data had a presence of 280 species, and so we included 220 additional species with no occurrences, so that the 500×35 data matrix (called training data A) describes the presence–absences of all species in the target community. Alternatively, we assumed that the diatom researcher would account only for the observed species, and thus the training data B are organized into a 280×35 matrix. We fitted for each of the 500 species an independent species-specific model

(Eq. 4), and for the two sets of training data (A and B) the community model with the unimodal structure of Eq. 2. We then asked how well the models predicted the occurrence of all species in the full set of 105 sites, for which we assumed the environmental covariates PC1 and PC2 to be known.

To achieve as accurate parameter estimates as possible, we finally fitted the community model to the full diatom data on 500 species on 105 sites. Based on Eq. 2, we used the parameterized model to measure the covariance between the communities j and j' by $\mathrm{Cov}_{jj'} = \mathbf{x}_j^\top \mathbf{V}\mathbf{x}_{j'} + \sigma^2\delta_{jj'}$, where $\mathbf{x}_j$ is the vector of environmental and spatial covariates (row of the design matrix $\mathbf{X}$) and $\delta_{jj'}$ is Kronecker's delta ($\delta_{jj} = 1$ and $\delta_{jj'} = 0$ if $j \neq j'$). We measured correlation in community similarity as

$$\rho_{jj'} = \frac{\mathrm{Cov}_{jj'}}{\sqrt{\mathrm{Var}_j \times \mathrm{Var}_{j'}}}. \tag{5}$$

To examine how much of the community dissimilarity between two diatom communities can be attributed to dissimilarity in environmental conditions, we set PC1 and PC2 to their mean values for all sites, and recomputed the correlation in community similarity, denoted by $\rho^*_{jj'}$.

Parameter estimation

We fitted the models to data using a Bayesian approach. To do so, prior distributions need to be defined for the community-level parameters values $\boldsymbol{\mu}$ and $\mathbf{V}$. For the ease of posterior sampling, we assumed the conjugate normal-inverse-Wishart prior (Gelman et al. 2004) for ($\boldsymbol{\mu}$, $\mathbf{V}$), i.e., $\mathbf{V} \sim \text{Inv-Wishart}_{\upsilon_0}(\Lambda_0^{-1})$and $\boldsymbol{\mu}|\mathbf{V} \sim \mathbf{N}(\boldsymbol{\mu}_0, \mathbf{V}/\kappa_0)$, where we set $\boldsymbol{\Lambda}_0$ to identity matrix and $\boldsymbol{\mu}_0$ to zero vector. For the simulated data, we set $\kappa_0 = 5$ and $\upsilon_0 = k + 5$. In case of the mixture model, we assumed this prior for both components of the mixture, and a uniform prior distribution in (0, 1) for the probability p_1 (this defines a joint prior for [p_1, p_2], as $p_2 = 1 - p_1$).

In case of the real data, we tested the sensitivity of the results to the prior distribution by either setting $\kappa_0 = 0$ and $\upsilon_0 = k$ (prior 1) or $\kappa_0 = 5$ and $\upsilon_0 = k + 5$ (prior 2). Here prior 1 is less informative than prior 2. For the variance component σ^2 we assumed the Inverse-χ^2 prior with parameter 1. Independent species models were fitted with exactly the same procedure by including in the data matrix only one species at a time. Details of the Bayesian MCMC scheme are given in Appendix B, and the Mathematica 7.0 source code is given in the Supplement.

Results

Model performance against simulated data

The hypothetical species community H1 is dominated by rare species, but contains also few very common species (Fig. 1B). In case of presence–absence, the data are maximally informative if the species occurs in half of the sites, i.e., when the model intercept is close to zero. For such a species (shown by orange color in Fig. 1A) the parameter estimates are almost identical whether the species is treated independently or as part of the community. In contrast, the cyan species of Fig. 1A is so rare that actually the data do not contain a single observation of this species. Thus, when the species is treated independently, there is no information on how this species might respond to the environmental covariate (dashed line in Fig. 1A). However, the community model has learned from the other species that especially the rare species respond positively to the environmental covariate, and correctly assumes that this is the case also for the focal species (continuous line in Fig. 1A). The brown color in Fig. 1A represents an intermediate case of a relatively rare species, for which the parameter estimate is influenced by but not dominated by the community-level model. The accuracy in the estimated community-level parameters increases with increasing number of species included in the data analysis (Fig. 1C), feeding back to the accuracy of the parameter estimates for the individual species.

Fig. 1D shows that if one fits a mixture model to data generated by a mixture model, the properties of the original community can be recovered (contour lines in Fig. 1D). Now for the rare species (shown by cyan color) there remain uncertainty on which kind of environmental conditions that species is specialized to. The bimodal structure of species-specific responses can however be identified only with data on large enough number of species (not shown). Thus, more data are needed to obtain a reliable characterization of a complexly structured community (here, the bimodal model) than of a simply structured community (here, the unimodal model).

Model performance against the diatom data

As shown above, community modeling can lead to improved inference compared to species-specific modeling if the community follows a joint structure that is reflected by the model assumptions. However, it is not a priori clear whether real communities can be described by simple structures such as Eq. 2. Thus, we next turn to a more critical test, and evaluate the predictive power of the community model in case of real data on diatoms in Finnish streams.

Fig. 2A, B evaluates the performances of the community model, and the collection of independently parameterized single-species models, in using training data to predict the community structure in all sites. The independent species models (orange lines in Fig. 2A, B) lead to biased estimates, and their predictions are very sensitive on the assumed prior distribution. This is not surprising, given that almost half of the species are completely missing in the training data from 35 sites. In contrast, the community model leads to an accurate prediction which is insensitive to the assumed prior distribution (black lines in Fig. 2A, B). However, this is the case only for training data A, where also the

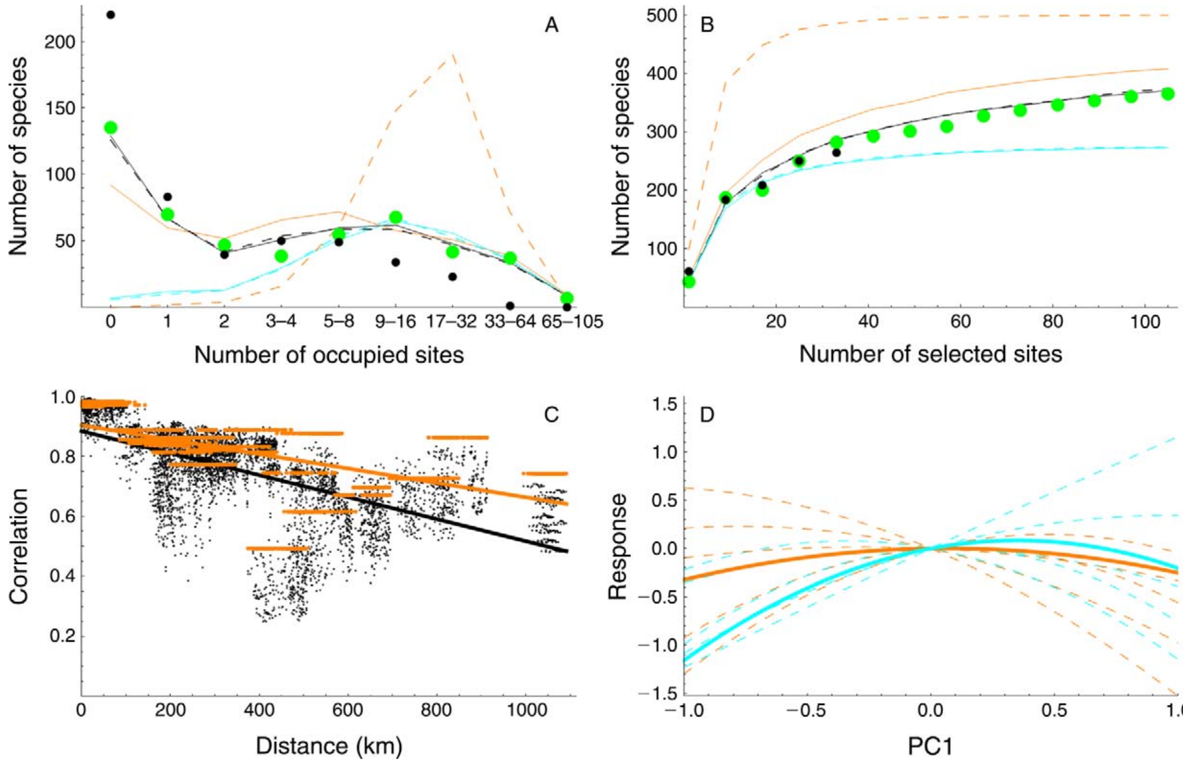

FIG. 2. Community model applied to diatom data. Panels A (species abundance distribution) and B (expected number of species in a random selection of sites) compare model predictions (based on training data on 35 sites, black dots) to the full data on 105 sites (green dots). The lines show the median model prediction based on independent species models on all of the 500 species (orange) or the community model (black for training data A, cyan for training data B), assuming prior 1 (continuous lines) or prior 2 (dashed lines). Panels C and D show model predictions based on the full data on 500 species on 105 sites. In panel C, the black and orange dots show the correlation coefficients $\rho_{jj'}$ and $\rho^*_{jj'}$, respectively (see Eq. 5), for all pairs of sites. The lines show linear regressions to these data against the Euclidian distance between the sites, and they thus partition the pattern of distance decay into environmental and spatial covariates. Panel D shows the predicted responses of a rare (intercept 2 standard deviations below the mean, cyan color) and a common (intercept 2 standard deviations above the mean, orange color) species to the environmental covariate PC1. The continuous lines show the expected (among-species) response; the dashed lines show five individual species randomized from the community. The response shows the value of the linear predictor (normalized to zero for $PC1 = 0$), which for logistic regression can be viewed as the change in odds ratio. Panel D shown for region 1 (other regions show a similar pattern), corresponding Fig. on PC2 is given in Appendix C.

unobserved species were included in the analyses. In case of training data B, where the missing species were excluded, the community model greatly underestimates the number of rare species in the community (cyan lines in Fig. 2A, B).

Patterns of community structure

We next analyze the diatom community structure based on fitting the community model to the full data from 105 sites, including the unobserved species. Community similarity, measured by the correlation $\rho_{jj'}$ (Eq. 5) is positive between all pairs of study sites (Fig. 2C). Thus, the local communities in any two sites are more similar than expected by random, reflecting the fact that the same species tend to be common (or rare) across the whole study area. Measuring by the slope of distance decay in Fig. 2C (−0.00037 for $\rho_{jj'}$ and −0.00024 for $\rho^*_{jj'}$), 35% of the similarity in community composition can be attributed to the measured environmental conditions (PC1 and PC2), the rest representing unexplained spatial variation. The community-level parameters $\boldsymbol{\mu}$ and $\mathbf{V}$ (Appendix C) indicate that rare species tend to prefer high values of PC1 (nutrient poor waters) and PC2 (high pH), whereas a typical common species is a generalist with respect to the environmental conditions. There is however much variation among the species, both rare and common species including generalists and specialists with respect to PC1 and PC2 (Fig. 2D and Appendix C).

DISCUSSION

Advances in computational methods have made it increasingly feasible to fit complex hierarchical models to ecological data structured by space, sampling design and other such factors (Diez and Pulliam 2007, Cressie et al. 2009, Latimer et al. 2009). In this study, we have used a hierarchical structure to combine a set of species-specific models into a model of the species community.

We have demonstrated both with simulated data and with real data that the community approach can lead to improved inference compared to the application of species-specific models. A major strength of the community approach is that sparse data on rare species need not be excluded, but become valuable, as they provide information on community-level characteristics. The community-level model (Eq. 2) provides a compact summary which can be used to analyze higher-level patterns such as species richness and community turnover (Fig. 2C). Unlike in models of distance-decay in community similarity (e.g., Nekola and White 1999, Green et al. 2004), we do not model these patterns directly, but they emerge from the species-based description of the community.

Our case study with the diatom community showed that even the simple multivariate normal structure of Eq. 2 can be a good description of a species community. To which extent this holds true among different types of species communities is an open empirical question. However, our modeling approach is not restricted to the multivariate normal model, as one can assume an arbitrary community-level model, such as the mixture model. Given a set of alternative models, standard tools of model selection and validation can be applied. As community modeling can help to get more out of sparse data, we believe that it will facilitate the analysis of many basic and applied questions in ecology. We have illustrated community modeling specifically for logistic regression, but it would be straightforward to apply this approach to any generalized linear (or additive) models, and to append the model with additional components. For example, spatial autocorrelation could be readily incorporated by assuming an appropriate covariance structure (within and among species) for the random effect of ε (Eq. 4).

The key idea behind the community-level model is that the responses of the individual species to the environmental covariates follow a joint structure. We note that we however have assumed that the species occur statistically independently of each other. Other approaches to community modeling (Latimer et al. 2009, Ovaskainen et al. 2010) relax this assumption by examining if some species pairs occur more or less often together than expected from their ecological niches. As these approaches fit a species-to-species correlation matrix, they are limited for cases with a large amount of data on a small number of species, in contrast to the present model which is best suited for sparse data for a large number of species.

Our results on factors behind community turnover in diatoms (Fig. 2C) suggest that environmental covariates play a substantial role but they still fail to explain much of the variation. Diatoms are thought as efficient indicators of environmental conditions, each species having quite distinct environmental preferences (e.g., Stoermer and Smol 1999). For example, we found that rare species tend to prefer nutrient poor waters, and thus streams with high water quality are important for conservation of rare species. We emphasize, though, that there was much variation among species (Fig. 2D), and many species showed relatively generalistic responses (see also Pither and Aarssen 2005). In line with our results, recent papers have suggested that a major part of the community patterns in diatoms may be in fact generated by other than local factors, such as historical events or dispersal limitation (Vyverman et al. 2007, Heino et al. 2010). These results give thus further evidence that distributional patterns of microorganisms may not be fundamentally different from those observed for macroorganisms, and that also microorganisms can show strong spatial structure free of environmental constraints (see, e.g., Green et al. 2004).

Acknowledgments

This work was initiated through stimulating discussions during a series of UKPopNet (NERC and Natural English) working group meetings hosted by Barbara J. Anderson (University of York). We thank Atte Moilanen, Chris Thomas, Chaozhi Zheng, and two anonymous reviewers for valuable comments, and Sami Ojanen for help with preparing the supplementary material. The study was supported by the Academy of Finland (Grant no. 124242 to O. Ovaskainen and 126718 to J. Soininen) and the European Research Council (ERC Starting Grant no. 205905 to O. Ovaskainen).

Literature Cited

Cressie, N., C. A. Calder, J. S. Clark, J. M. V. Hoef, and C. K. Wikle. 2009. Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. Ecological Applications 19:553–570.

Currie, D. J. 1991. Energy and large-scale patterns of animal-species and plant-species richness. American Naturalist 137:27–49.

De'ath, G. 2002. Multivariate regression trees: a new technique for modeling species-environment relationships. Ecology 83:1105–1117.

Diez, J. M., and H. R. Pulliam. 2007. Hierarchical analysis of species distributions and abundance across environmental gradients. Ecology 88:3144–3152.

Elith, J., et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29:129–151.

Fallu, M. A., N. Allaire, and R. Pienitz. 2002. Distribution of freshwater diatoms in 64 Labrador (Canada) lakes: species–environment relationships along latitudinal gradients and reconstruction models for water colour and alkalinity. Canadian Journal of Fisheries and Aquatic Sciences 59:329–349.

Gelfand, A. E., A. M. Schmidt, S. Wu, J. A. Silander, A. Latimer, and A. G. Rebelo. 2005. Modelling species diversity through species level hierarchical modelling. Journal of the Royal Statistical Society Series C—Applied Statistics 54:1–20.

Gelfand, A. E., J. A. Silander, S. S. Wu, A. Latimer, P. O. Lewis, A. G. Rebelo, and M. Holder. 2006. Explaining species distribution patterns through hierarchical modeling. Bayesian Analysis 1:41–91.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2004. Bayesian data analysis. Second edition. Chapman and Hall/CRC, Boca Raton, Florida, USA.

Green, J. L., A. J. Holmes, M. Westoby, I. Oliver, D. Briscoe, M. Dangerfield, M. Gillings, and A. J. Beattie. 2004. Spatial scaling of microbial eukaryote diversity. Nature 432:747–750.

Guisan, A., and W. Thuiller. 2005. Predicting species distribution: offering more than simple habitat models. Ecology Letters 8:993–1009.

Guisan, A., and N. E. Zimmermann. 2000. Predictive habitat distribution models in ecology. Ecological Modelling 135:147–186.

Heino, J., L. M. Bini, S. M. Karjalainen, H. Mykrä, J. Soininen, L. C. G. Vieira, and J. A. F. Diniz-Filho. 2010. Geographical patterns of micro-organismal community structure: are diatoms ubiquitously distributed across boreal streams? Oikos 119:129–137.

Kreft, H., W. Jetz, J. Mutke, G. Kier, and W. Barthlott. 2008. Global diversity of island floras from a macroecological perspective. Ecology Letters 11:116–127.

Kremen, C., et al. 2008. Aligning conservation priorities across taxa in Madagascar with high-resolution planning tools. Science 320:222–226.

Kruskal, J. B. 1964. Nonmetric multidimensional-scaling: a numerical method. Psychometrika 29:115–129.

Latimer, A. M., S. Banerjee, H. Sang, E. S. Mosher, and J. A. Silander. 2009. Hierarchical models facilitate spatial analysis of large data sets: a case study on invasive plant species in the northeastern United States. Ecology Letters 12:144–154.

Latimer, A. M., S. S. Wu, A. E. Gelfand, and J. A. Silander. 2006. Building statistical models to analyze species distributions. Ecological Applications 16:33–50.

Leathwick, J. R., J. Elith, and T. Hastie. 2006. Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. Ecological Modelling 199:188–196.

Moilanen, A., A. M. A. Franco, R. I. Early, R. Fox, B. Wintle, and C. D. Thomas. 2005. Prioritizing multiple-use landscapes for conservation: methods for large multi-species planning problems. Proceedings of the Royal Society B 272:1885–1891.

Morlon, H., G. Chuyong, R. Condit, S. Hubbell, D. Kenfack, D. Thomas, R. Valencia, and J. L. Green. 2008. A general framework for the distance-decay of similarity in ecological communities. Ecology Letters 11:904–917.

Nekola, J. C., and P. S. White. 1999. The distance decay of similarity in biogeography and ecology. Journal of Biogeography 26:867–878.

Ovaskainen, O., J. Hottola, and J. Siitonen. 2010. Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions. Ecology 91:2514–2521.

Pearce, J., and S. Ferrier. 2000. An evaluation of alternative algorithms for fitting species distribution models using logistic regression. Ecological Modelling 128:127–147.

Pearson, R. G., and T. P. Dawson. 2003. Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? Global Ecology and Biogeography 12:361–371.

Peralvo, M., R. Sierra, K. R. Young, and C. Ulloa-Ulloa. 2007. Identification of biodiversity conservation priorities using predictive modeling: an application for the equatorial pacific region of South America. Biodiversity and Conservation 16:2649–2675.

Phillips, S. J., R. P. Anderson, and R. E. Schapire. 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling 190:231–259.

Pither, J., and L. W. Aarssen. 2005. Environmental specialists: their prevalence and their influence on community-similarity analyses. Ecology Letters 8:261–271.

Soininen, J. 2008. The ecological characteristics of idiosyncratic and nested diatoms. Protist 159:65–72.

Soininen, J., J. Heino, M. Kokocinski, and T. Muotka. 2009. Local-regional diversity relationship varies with spatial scale in lotic diatoms. Journal of Biogeography 36:720–727.

Soininen, J., R. Paavola, and T. Muotka. 2004. Benthic diatom communities in boreal streams: community structure in relation to environmental and spatial gradients. Ecography 27:330–342.

Stevenson, R. J., M. L. Bothwell, and R. Lowe, editors. 1996. Algal ecology. Academic Press, New York, New York, USA.

Stoermer, E. F., and J. P. Smol, editors. 1999. The diatoms: applications for the environmental and earth sciences. Cambridge University Press, Cambridge, UK.

ter Braak, C. J. F. 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology 67:1167–1179.

Thuiller, W., M. B. Araujo, and S. Lavorel. 2003. Generalized models vs. classification tree analysis: Predicting spatial distributions of plant species at different scales. Journal of Vegetation Science 14:669–680.

Vyverman, W., et al. 2007. Historical processes constrain patterns in global diatom diversity. Ecology 88:1924–1931.

Warren, M. S., et al. 2001. Rapid responses of British butterflies to opposing forces of climate and habitat change. Nature 414:65–69.

APPENDIX A

Generation of simulated communities (*Ecological Archives* E092-025-A1).

APPENDIX B

Bayesian estimation schemes (*Ecological Archives* E092-025-A2).

APPENDIX C

Details of the diatom case study (*Ecological Archives* E092-025-A3).

SUPPLEMENT

Source code for parameter estimation (*Ecological Archives* E092-025-S1).