Ovaskainen, Otso; Roy, David B.; Fox, Richard; Anderson, Barbara J. 2016. **Uncovering hidden spatial structure in species communities with spatially explicit joint species distribution models**. *Methods in Ecology and Evolution*, 7 (4). 428-436. 10.1111/2041-210X.12502

Received Date : 04-Jul-2015

Revised Date : 19-Oct-2015

Accepted Date : 25-Oct-2015

Article type : Research Article

Editor: David Orme

Uncovering hidden spatial structure in species communities with spatially explicit joint species distribution models

Otso Ovaskainen[1,2,*], David B. Roy[3], Richard Fox[4] and Barbara J. Anderson[5]

[1] Metapopulation Research Centre, Department of Biosciences, P.O. Box 65, FI-00014, University of Helsinki, Finland.

[2] Centre for Biodiversity Dynamics, Department of Biology, Norwegian University of Science and Technology, N-7491 Trondheim, Norway.

[3] Centre for Ecology and Hydrology, Wallingford, Oxfordshire, OX10 8BB, UK

[4] Butterfly Conservation, East Lulworth, Wareham, Dorset, BH20 5QP, UK.

[5] Landcare Research, Private Bag 1930, Dunedin 1954, New Zealand

* Corresponding author: Metapopulation Research Centre, Department of Biosciences, P.O. Box 65, FI-00014, University of Helsinki, Finland. E-mail: otso.ovaskainen@helsinki.fi

Abstract

Modern species distribution models account for spatial autocorrelation in order to obtain unbiased statistical inference on the effects of covariates, to improve the model's predictive ability through spatial interpolation, and to gain insight in the spatial processes shaping the data. Somewhat analogously, hierarchical approaches to community-level data have been developed to gain insights into community-level processes, and to improve species-level inference by borrowing information from other species that are either ecologically or phylogenetically related to the focal species. We unify spatial and community-level structures

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/2041-210X.12502

by developing spatially explicit joint species distribution models. The models utilize spatially structured latent factors to model missing covariates as well as species-to-species associations in a statistically and computationally effective manner. We illustrate that the inclusion of the spatial latent factors greatly increases the predictive performance of the modelling approach with a case study of 55 species of butterfly recorded on a 10 km x 10 km grid in Great Britain consisting of 2,609 grid cells.

Key words. Joint species distribution models, community models, spatial models, latent factors

Introduction

Conceptual and theoretical research in community ecology has long emphasized that the dynamics and distributions of species communities are shaped by the interplay between i) environmental filtering, ii) species interactions, and iii) spatial and stochastic processes (Leibold et al. 2004). One reason why metacommunity theories are still poorly linked with data is the lack of statistical frameworks that enable these three factors to be integrated and that would be applicable for data typically available in community ecological studies (Logue et al. 2011). As we briefly review below, the last decade has brought major statistical advances in species distribution modelling that helps to bridge this gap between theory and data: joint species distribution modelling facilitates the assessment of environmental filtering and species interactions, whereas spatially and spatio-temporally structured species distributions models enable one to incorporate the effects of spatial and stochastic processes. In this paper, we bring these developments together by developing a statistical framework for spatially explicit joint species distribution modelling.

In their influential review, Ferrier and Guisan (2006) classified strategies for analysing community-level species distribution data into the three categories of ‘assemble first, predict later’ (e.g. modelling species richness as the response variable), ‘predict first, assemble later’ (e.g. summing the predictions of single-species models to predict species richness), and ‘assemble and predict together’. Since their review, a great amount of methodological progress has taken place in the category of ‘assemble and predict together’, i.e. joint species distribution models that include simultaneously both species- and community-level components. Such models have been shown to have better predictive power than single-species models, in particular for rare species for which model parameterization may not be feasible without borrowing information from other species (Ovaskainen and Soininen 2011, Bonthoux et al. 2013, Hui et al. 2013).

Joint species distribution models extend single-species approaches in two principally different ways: by modelling environmental filtering at the community level, and by accounting for statistical co-occurrence among the species. In the context of regression-based models, one approach for seeking community level patterns in environmental filtering is to treat the species-specific regression coefficients (related to occurrence and/or detectability) as random effects, and thus assuming that they follow either univariate or multivariate normal

distributions across the species (Dorazio and Royle 2005, Dorazio et al. 2006, Kery et al. 2009, Russell et al. 2009, Dorazio et al. 2010, Zipkin et al. 2010, Ovaskainen and Soininen 2011, Jackson et al. 2012, Olden et al. 2014). Another approach that similarly allows sharing information among species is the use of mixture models, which use model-based grouping of species into 'species archetypes' (Dunstan et al. 2011, Hui et al. 2013). Not only model construction, but also model selection can be conducted either at the species level or at the community level (Madon et al. 2013).

As reviewed by Kissling et al. (2012) and Wisz et al. (2013), statistical co-occurrence among species (generated by species interactions or missing covariates) can be incorporated into joint species distribution models in several ways. The most straightforward alternative is to use some species as predictors for others. In communities with a large number of species, this however leads to the problem of multiple testing, which can be counteracted by including as predictors only the most abundant species (le Roux et al. 2014), or only those species that are part of the food web of the focal species (Pellissier et al. 2013). Another alternative is the use of multivariate regression models (Ovaskainen et al. 2010, Sebastian-Gonzalez et al. 2010, Clark et al. 2014, Pollock et al. 2014) or neural network models (Harris 2015) in which the response variable is the vector of occurrences (Ovaskainen et al. 2010, Sebastian-Gonzalez et al. 2010, Pollock et al. 2014, Harris 2015) or abundances (Clark et al. 2014) of all species. In this context, neural network models can be used to identify non-linear relationships between species. With rich enough data, one may attempt to infer more refined aspects of species associations, e.g. the presence of so called competitive intransitivity (Ulrich et al. 2014). But as statistical co-occurrence patterns can be created either by missing environmental covariates or by biotic interactions (Morales-Castilla et al.), the results of such multivariate regression models need to be interpreted with caution (Ovaskainen et al. 2010, Pollock et al. 2014).

Joint species distribution models can also be effective tools for bringing functional and phylogenetic perspectives to the analysis of species distribution data. Species traits can be used to model the responses of the species to environmental covariates (Pollock et al. 2012, Brown et al. 2014) and to facilitate the estimation of the species-to-species correlation matrices by considering them as functions of trait dissimilarity (Dorazio and Connor 2014). Accounting for phylogenetic constrains is necessary for obtaining unbiased inference in analyses that consider each species as a data point, and it can also be helpful for disentangling the effects of environmental filtering from those of biotic interactions (Helmus et al. 2007, Ives and Helmus 2011). Further, bringing the phylogenetic perspective to joint species distribution models shifts the emphasis from measures of community similarity based on species identity to corresponding measures based on phylogenetic similarity (Ives and Helmus 2010).

In parallel to the developments aiming to move from single-species perspectives to multi-species perspectives, the need for using spatially explicit species distribution models has become increasingly acknowledged in ecological research, both due to interest on spatial processes per se, and due to the need to account for non-independent data points (Dormann et

al. 2007). The estimation of spatially structured residuals has been facilitated by computational advances in Bayesian inference, both on Markov Chain Monte Carlo (MCMC) sampling methods (Latimer et al. 2009, Chakraborty et al. 2010) and on methods based upon the integrated nested Laplace approximation (Blangiardo et al. 2013). Another increasingly popular approach for bringing spatial structure into species distribution models is the use of spatial eigenvectors derived from the distance matrix among the sampling sites (Borcard and Legendre 2002, Dray et al. 2006, Dray et al. 2012).

The aim of this paper is to integrate joint species distribution modelling and spatially explicit species distribution modelling. Such developments were pioneered by Latimer et al. (2009), who incorporated for each species a spatially structured residual, and estimated species-to-species correlation structure among the spatial effects. As the approach of Latimer et al. (2009) requires the estimation of species-specific spatial effects, it is not suited for large species communities that are often dominated by rare species. Latimer et al. (2009) parameterized their model for four common species only. Here we overcome this limitation by modelling spatial effects at the community level. To do so, we utilize latent factor models, which have recently emerged in the ecological literature (Walker and Jackson 2011, Hui 2015), and for which computationally efficient sampling algorithms are available (Bhattacharya and Dunson 2011). The use of spatial latent factors was recently introduced in the community context by Thorson et al. (2015). While our work is closely related to that developed independently by Thorson et al. (2015), it has the following differences: (i) our modelling approach is developed in the Bayesian framework, and it thus provides the full posterior distribution of parameter uncertainty, (ii) we combine spatial factors with fixed effects and thus partition variation between measured and unmeasured covariates, (iii) we apply the model to all species that make up the community, instead of restricting the analyses to common species only, and (iv) we demonstrate how the approach can be used to assess the geographic scaling of spatial covariance patterns. We demonstrate the predictive power of our modelling approach with data consisting of the occurrences of 55 species of butterflies sampled in Great Britain during 1995-1999 on 2,841 grid cells at the resolution of 10 km x 10 km (Asher et al. 2001).

Joint species distribution modelling with spatially structured latent factors

We model the presence-absences or abundances of a set of species using the statistical framework of spatially explicit joint species distribution models. The main advantage of this modelling framework is the use of spatially structured latent factors, which makes it possible to capture the effects of missing covariates, the effects of biotic interactions, or the combination of these two. The computational and statistical efficiency of the approach arises from there generally being far fewer latent factors than there are species. This is because all species are modelled with the help of a shared set of latent factors, each species having its own loading for each latent factor. If the latent factors were known covariates, these loadings would simply correspond to regression coefficients which could be estimated using standard techniques. However, it is often the case that species distributions are partly determined by unknown or unmeasurable covariates, or by biotic interactions. These ‘hidden covariates’ are

here accounted for by the latent factors, and as they are not known *a priori*, they must be estimated. During the model fitting process, also the spatial scale at which each latent factor varies is estimated. For instance, if the latent factor corresponds to a large-scale macro-climatic gradient, the corresponding spatial scale will be much larger than if the latent factor corresponds to a small-scale micro-climate gradient or small-scale biotic interactions. Informally, the latent factors, and their spatial scales, are estimated so that they explain as much of the variation in the distributions of all the species simultaneously. Also the number of latent factors is estimated, with the aim of including a sufficient number of latent factors to allow the model to capture as much of the biologically relevant variation as possible, but to avoid over-fitting and thus the inclusion of latent factors that model noise rather than signal.

Before turning to the formal description of the model, we illustrate its main idea in Fig. 1. For simplicity, we don't include any measured covariates. Thus the predictors of the model consist only of two latent factors η_1 and η_2, shown respectively in panels **a** and **b** of Fig. 1. The example is constructed to mimic the case of two competing species with overlapping resource use, e.g. two birds which are both restricted to coniferous forest but that compete for nesting locations within each stand. In such a case, one would expect to see negative co-occurrence over short spatial scales, but positive co-occurrence over large spatial scales (Araujo and Rozenfeld 2014). In Fig. 1, the latent factor η_1 represents the shared resource, and it varies at the large characteristic spatial scale $\alpha_1 = 10$ spatial units (in Fig. 1, the spatial unit corresponds to the grid cell size). The latent factor η_2 represents the influence of competition, and it varies at the smaller spatial scale of $\alpha_2 = 2$ spatial units. Each species j has its own loading λ_{jh} for each latent factor h, so that species-specific occurrence probabilities are modelled as linear combinations of the latent factors. In the example of Fig. 1, the loadings of species 1 are $\lambda_{11} = 1, \lambda_{12} = 1$, so that the linear predictor for species 1 (illustrated in panels **ce**) is $L_1 = \eta_1 + \eta_2$. The loadings of species 2 are $\lambda_{21} = 1, \lambda_{12} = -2$, so that the linear predictor for species 2 (illustrated in panels **df**) is $L_2 = \eta_1 - 2\eta_2$. As both species have a positive loading to the latent factor η_1, they show positive co-occurrence over large spatial scales. But as their loadings have opposite signs to the latent factor η_2, their co-occurrence pattern is negative over short spatial scales (Figs. 1**gh**).

Let us then turn to a more formal definition of the model. We index by $i = 1, \ldots, n_y$ the sampling units and by $j = 1, \ldots, n_s$ the species. Whilst we exemplified the modelling framework with two species, the model is equally well suitable for communities consisting of a large number of species.
In case of presence-absence data, we model the presence ($y_{ij} = 1$) or absence ($y_{ij} = 0$, including the possibility of non-detection) of species j on sampling unit i by probit regression, implemented as $y_{ij} = 1_{z_{ij} > 0}$, where the latent liability $z_{ij} = L_{ij} + e_{ij}$ includes the linear predictor L_{ij} and the residual which models the probit link-function and is distributed

$\varepsilon_{ij} \sim N(0,1)$ independently among the species and the sampling units. The linear predictors are further modelled as

$$L_{ij} = \sum_{k=1}^{n_c} x_{ik}\beta_{kj} + z_{ij}, \qquad \text{where} \quad z_{ij} = \sum_{h=1}^{n_f} \eta_{ih}\lambda_{jh}. \quad \text{(Eq. 1)}$$

Here x_{ik} is the measured covariate $k = 1, \ldots, n_c$ for sampling unit i, β_{kj} is the regression coefficient measuring how species j responds to the covariate k, η_{ih} is the (unmeasured) latent factor $h = 1, \ldots, n_f$, and the factor loading λ_{jh} measures how species j responds to the latent factor h. We included the intercept in the model by setting $x_{i1} = 1$.

In the standard non-spatial latent factor model (Bhattacharya and Dunson 2011), the latent factors are assumed to be normally distributed with zero mean and unit variance, $\eta_{ih} \sim N(0,1)$. To bring spatial structure for the latent factors, we assumed a spatially homogeneous Gaussian process with $\mathrm{Cov}(\eta_{ih}, \eta_{i'h'}) = \delta_{hh'} f_h(d_{ii'})$, where $d_{ii'}$ is the spatial distance between the sampling units i and i', and $\delta_{hh'}$ is Kronecker delta with value 1 for $h = h'$ and with value 0 for $h \neq h'$. The function $f_h(d)$ is a spatial covariance function, normalized to $f_h(0) = 1$ so that its unit is correlation and that the marginal distributions of the latent factors have zero mean and unit variance, similar to non-spatial latent factor models. Here we assume the exponential function $f_h(d) = \exp(-d/\alpha_h)$, where α_h is the spatial scale of the latent factor h. We note that the standard latent factor model with spatially independent factors induces the covariance structure $\varepsilon_{..} \sim N(0, \Psi)$ with $\Psi = (\Lambda^{\mathrm{T}}\Lambda) \otimes \mathrm{I}_{n_y}$ where Λ is a matrix of the factor loadings (Bhattacharya and Dunson 2011). In addition to determining covariance at zero distance (Thorson et al. 2015), the spatial structure of the latent factors induces a spatial covariance $\rho_{jj'}(d)$ in the latent factors influencing the occurrences of species j and j', given at distance d by

$$\rho_{jj'}(d) = \left[\mathrm{Cov}(z_{ij}, z_{i'j'}) | d_{ii'} = d\right] = \sum_{h=1}^{n_f} \lambda_{jh}\lambda_{j'h} f_h(d). \quad \text{(Eq. 2)}$$

This characterizes species-to-species associations not only at the local level ($d = 0$) but also their spatial decay, similar to Latimer et al. (2009). As illustrated by Fig 1**h**, spatial covariance between species can be positive or negative, corresponding to positive or negative co-occurrence.

To ensure that not more factors are selected than is necessary to explain the data, we follow Bhattacharya and Dunson (2011) by defining a multiplicative gamma process shrinkage prior on the factor loadings. This variant of the Bayesian approach also avoids the need to use a pre-specified structure for the loading matrix as assumed by Thorson et al. (2015). We extended the Gibbs sampling algorithm of Bhattacharya and Dunson (2011) to the present

case by utilizing Bayesian multivariate regression for the fixed effects, and by incorporating a discrete grid sampler for the spatial scale parameters α_h. The technical details of the sampling algorithm are presented in Supplementary material, as well as a Matlab code for model parameterization.

A case study with butterfly data for Great Britain

To illustrate the modelling approach, we consider a case study of 55 species of butterflies. We use the 1995-1999 atlas data (Asher et al. 2001) as presence-absence for each of the 10 x 10 km Ordnance Survey grid cells for Great Britain (n=2,609 cells). Based on previous studies on butterflies in Great Britain (Hodgson et al. 2011, Bennie et al. 2013), we included as measured covariates (1) the number of growing degree days above 5 degrees Celsius, and the percentage of the grid cell cover that consists of (2) broadleaved woodland, (3) coniferous woodland, and (4) calcareous substrates (Fig. 2; see Supplementary material for details on the data). To make the test case challenging, we randomly selected only 300 cells that were used as training data to parameterize the model (Fig. 2), and thus used the remaining 2,309 cells for model validation. To assess the influence of spatially structured latent factors on model performance, we first fitted the model with just the four covariates. We then fitted the model with the four covariates and the spatially structured latent factors.

While the focus in this paper is on the spatial part of the model, we note that the model belongs to the standard framework of generalized linear mixed models, allowing one to incorporate various hierarchical layers and covariance structures. As an example, we model here the responses of the species (β_{kj}) to the measure covariates (x_{ik}) as a function of their functional group. To do so, we classified the species as wider countryside species, specialist species, and migratory species. Similarly to Brown et al. (2014) and Ovaskainen et al. (2010), we model the vector of regression coefficients for species j with the multivariate normal distribution $\beta_{.j} \sim N(\mu_{g(j)}, V)$, where the expected response $\mu_{g(j)}$ is assumed to be specific to the functional group (g) to which the species belongs to.

Failing to account for spatial autocorrelation (i.e., assuming independence among the data points) is expected to lead to biased estimates of fixed effects (known covariates) and overestimation of their statistical significance (Legendre et al. 2002). With the butterfly data, the estimates for the fixed effects were more pronounced and had tighter credibility intervals in the model without latent factors than in the model that also includes latent factors. In the model without the latent factors, the 95% credibility interval for the effect of covariates 1, 2, 3 and 4 did not cross zero respectively for 50, 42, 11 and 40 species. In contrast, when latent factors were included, 95% credibility interval for the effect of covariates 1, 2, 3 and 4 did not cross zero for 28, 6, 8 and 10 species (see Supplementary material for the species-specific results). The likely overestimation of fixed effects is visible both in species- and community-level predictions, which reflect the covariate layers more pronouncedly than the data. For example, areas with calcareous substrate (Fig. 2d) differ in their species richness from the

surrounding areas in a more pronounced way in the model prediction (Fig. 3e) than in the data (Fig. 3d). With the inclusion of the latent factors, this mismatch between the data and the model prediction (Fig. 3f) disappears.

The posterior median estimates (95% credibility intervals) for spatial scale parameters of the two most dominant spatial latent factors are $\alpha_1 = 170\ (120 - 260)$ km and $\alpha_2 = 170\ (120 - 250)$ km. The first latent factor identifies essentially a north-south gradient, whereas the second one recognises that the south-eastern part of Great Britain differs in terms of its butterfly community composition from the rest of the country and especially from the north-western part (Fig. 2). The model with spatially structured latent factors appears to better predict the data than the model without the latent factors (Figs. 3 and 4). As expected, the predictive power is generally poorest for species with very low or very high prevalence, as the occurrence of these species varies little. The mean R^2 value is 30% for the model without latent factors, whereas it is 42% for the model with latent factors. Across the 55 species, the model with latent factors had higher R^2 and AUC (Fielding and Bell 1997) values for 54 and 51 species, respectively, when evaluated against the validation data. In addition to improving the average AUC for occurrence of individual species from 0.86 to 0.91, including the latent factors improved the root mean squared error of species richness, reducing it from 4.9 species to 3.2 species

Taking the average over the species, in the model with latent factors the proportions of variance (at the level of the linear predictor) attributed to covariates 1-4 were 28%, 3%, 1% and 3%, whereas they were 54% and 11% for the latent factors 1-2 (Fig. 4). Thus, the covariates contributed 35% to the explained variation and the latent factors the remaining 65%, reflecting the increase in the model's predictive power achieved by adding the latent factors. Taking the average over the species, the amount of variance (at the level of the linear predictor) attributed to covariates 1-4 in the model with latent factors were reduced to 71%, 51%, 102% and 55% of the corresponding values in the model without latent factors. Thus, the latent factors absorbed some of the variation attributed to fixed effects in the model without the latent factors.

Discussion

Statistical methods for joint species distribution modelling have become well established, but thus far they have lacked a spatially explicit perspective (with the exceptions of Latimer et al. 2009, Thorson et al. 2015). In this paper we have utilized recent progress in latent factor modelling to develop a general statistical framework for spatially explicit joint species distribution modelling. As illustrated by our results, the inclusion of spatial latent factors improves statistical inference of joint species distribution models in three ways. First, failing to account for spatial structure corresponds to the assumption of independent data points, which leads to biased estimates for the effects of measured covariates. The inclusion of spatially structured latent factors is analogous to the inclusion of a spatially structured residual in single species models, and thus it corrects the inference on the effects of measured

covariates. Second, the incorporation of spatial latent factors enables spatial interpolation which can greatly improve the predictive power of the model. This was indeed the case in the butterfly case study, where most of the explained variation was attributed to the spatial latent factors. Third, the spatial latent factors identified by the models can be informative, as they can be interpreted as the covariates that influence the species community but are missing from the model, or as the end results of biotic interactions. For example, whilst we included the growing degree days above 5 degrees as a covariate, the first and most important latent factor identified by the model corresponded to a North-South gradient. This suggests that the North-South gradient correlates with relevant covariates other than the number of growing degree days, or that there is random species turnover along this gradient.

As community-level data on species occurrence or abundance typically come from a spatial setting, the possibility to account for spatiality in the analysis phase enables many kinds of applications. Thus we expect the method presented here to be generally useful for data typically collected in community ecological studies: presence-absence or abundance data acquired for a set of sampling sites, some environmental covariates describing the properties of those sites, and the spatial coordinates of those sites. With such data, the modelling approach presented here can be used for assessing the geographic scaling of covariance patterns (Eq. 2; illustrated in Fig. 1**h**), which can provide information on the type of species interactions (Araujo and Rozenfeld 2014). More generally, the generalized linear modelling framework allows one to partition variation in any community metric (e.g., species richness, evenness, or community dissimilarity) to the influences of measured covariates, to the influences of spatially structured latent factors, and to unexplained residual variation. We note that while we have utilized here atlas data that form a regular grid, the method applies directly also to any spatially irregular sampling design. Furthermore, by replacing the distance in two-dimensional space to the one-dimensional distance over time, the method applies as such also for time-series data. In this case, the exponential correlation structure assumed here corresponds to the widely applied AR(1) autoregressive model.

The model presented here adds the influence of measured covariates compared to the approach presented by (Thorson et al. 2015), but is still ignores many aspects that other approaches developed in community ecology account for. However, as our modelling approach is based on the standard framework of hierarchical generalized linear mixed models, it is of a very general nature and easily extendable to components implemented in previous research to joint species distribution modelling. These include the influence of species traits (Pollock et al. 2012, Brown et al. 2014) and phylogenetic constraints (Helmus et al. 2007, Ives and Helmus 2011), the use of abundance data instead of presence-absence data (Clark et al. 2014), and accounting for detectability (Dorazio et al. 2006).

As illustrated here, spatially explicit community modelling is expected to be useful especially for problems that involve spatial interpolation, e.g. for predicting species distribution maps from a sparse set of observations. But much interest in community ecology relates also to extrapolation, i.e. predicting the occurrences of species under environmental conditions not present in the training data, e.g. after climate change, after habitat loss, or in an unexplored

region. While it is not expected that spatially explicit community modelling will be able to provide improved mean predictions for extrapolation, it can improve the assessment of uncertainty in such predictions. This is because the modelling framework is able to identify how much of the current species occurrences are influenced by such unknown variation that is structured by space and thus not likely to be just noise. Assuming that the same proportion of the variance will be attributed to un-modelled but spatially structured variation also in the extrapolated situation will enable constructing more realistic confidence intervals than just ignoring such variation.

Acknowledgements

We thank David Dunson, Chaozhi Zheng, Nerea Abrego, William Lee, two anonymous reviewers and members of the UKPopNet (NERC R8-H12-012 and English Nature) for insightful discussions and valuable comments. We thank the large number of volunteer recorders contributing GB butterfly occurrence records. These datasets are operated by Butterfly Conservation and the NERC Centre for Ecology & Hydrology, and financially supported by a consortium of government agencies. OO was supported by the Academy of Finland (grant nr 250444) and BJA by NERC grant NE/FO18606/1 and a RSNZ Rutherford Discovery Fellowship.

Data Accessibility

The data used in the case study are provided in the supplementary material. These data were obtained from the following sources:

- British butterfly distribution data for the 1995-1999 atlas period gathered by the Butterflies for the New Millennium recording scheme (Asher et al. 2001) are provided in the supplementary material. These data are held by Butterfly Conservation and the Centre for Ecology & Hydrology, and are available through http://butterfly-conservation.org/111/butterflies-for-the-new-millennium.html and http://data.nbn.org.uk (contact: Richard Fox, rfox@butterfly-conservation.org). These data maybe used, with appropriate acknowledgement, under a creative commons license (https://creativecommons.org/licenses/by/4.0/).

- UK climate data: are provided in the supplementary material. We used mean annual number of growing degree days above 5 degrees Celsius for 1995-1999 for Britain at a 10 km Ordnance Survey grid resolution were derived from CRU ts2.1 and CRU 61-90 climate datasets (Barrow, Hulme & Jiang 1993). This involved the anomalies at

0.5 deg grid resolution being interpolated onto the UK Ordnance Survey 10 km grid and combined with the TIGER climate data (Hill 1995) from mean elevations within grid cells. These data may be used, with appropriate acknowledgement, under a creative commons license (https://creativecommons.org/licenses/by/4.0/). Original data are data are available from http://www.alarmproject.net/climate/climate/

- UK Land cover data (LCM2000): are provided in the supplementary material. We used percent cover broadleaved woodland, and percent coniferous woodland (Fuller et al. 2002). Percent cover was calculated as a percentage of the land area within UK Ordnance Survey 10 km grid cell. These data are licensed and and must be used in accordance with the open government License (OGL; http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/). Original data are available through http://www.ceh.ac.uk/landcovermap2000.html.

- UK geology data: are provided in the supplementary material. We used the 1 kilometre resolution Soil Parent Material model detailing 6 basic parent-material parameters (derived from the 1:50 000 scale version). We calculated the sum of 1km squares within each UK Ordnance Survey 10 km grid cell with a calcareous content value of "HIGH". Original data were downloaded on 2015-10-01, licensing restrictions, terms and conditions and original data are available (with appropriate acknowledgement) from http://www.bgs.ac.uk/downloads/start.cfm?id=2899.

Figure legends

Figure 1. Illustration of the joint species distribution modelling with spatially structured latent factors. The panels **a** and **b** show latent factors η_1 and η_2 which have exponentially decaying correlation structure at spatial scales $\alpha_1 = 10$ and $\alpha_2 = 2$. The panels **c** and **d** show respectively the linear predictors for species 1 ($L_1 = \eta_1 + \eta_2$) and species 2 ($L_2 = \eta_1 - 2\eta_2$), which combine the latent factors with the loading matrix $\Lambda = (1 \quad 1; 1 \quad -2)$. The panels **e** and **f** show the occurrence patterns of species 1 and 2 and panel **g** their co-occurrence pattern, with red and blue denoting occurrences of species 1 and 2, and black denoting the co-occurrence of both species. Panel **g** shows the spatial covariance functions (Eq. 2), with red and blue depicting the within species covariances $\rho_{11}(d)$ and $\rho_{11}(d)$, and black depicting the between species covariance $\rho_{12}(d)$.

Figure 2. Measured environmental covariates and model-identified latent factor used to model the butterfly community. The upper panels show the measured covariates 1-4: the number of growing degree days above 5 degrees (**a**), and the fraction of each grid cell consisting of broadleaved woodland (**b**), coniferous woodland (**c**), and calcareous substrates (**d**). The lower line of panels show the two most dominant latent factors (i.e., hidden environmental covariates) identified by the model: $\eta_{\cdot 1}$ (**e**) and $\eta_{\cdot 2}$ (**f**). The black squares in panel **g** show the 300 randomly selected 10 km x 10 km grid cells that were used to parameterize the model. The remaining 2,309 (shown by grey) were used to test the predictive performance of the model.

Figure 3. Visual comparison of model predictions and data. Panel **a** shows data (black corresponding to presence and grey to absence) for one of the 55 butterfly species (Green hairstreak*; Callophrys rubi*) and panel **d** the observed species richness. Panels **bc** show the model predictions for the occurrence probability of Green hairstreak based on the model without (**b**) and with (**c**) spatial latent factors. Similarly, panels **ef** show the model predictions for species richness based on the model without (**e**) and with (**f**) latent factors. All predictions are based on fitting the models to data on the 300 training sites shown in Fig. 2**g**.

Figure 4. The predictive performance of the community model. Panel **a** shows species-specific Tjur (2009) R^2 values as a function of the species prevalence, and panel **b** compares predicted species richness to observed species richness. Both panels are based on fitting the models without spatial latent factors (grey dots) and with spatial latent factors (black dots) to data on 300 training sites (Fig. 2**g**), and comparing model predictions to the data for the 2,309 validation sites. Panel **c** shows the relative proportions of variance attributed to the measured covariates and to the spatial latent factors. The measured covariates 1-4 are ordered from bottom to top, and coloured grey (covariate 1, i.e. growing degree days), and 3 levels of red (covariates 2-4). The latent factors 1-2 are shown on top of the measured covariates, and are coloured as light blue (latent factor 1) or dark blue (latent factor 2).

Supplementary material

Supplementary methods. Technical details of the Monte Carlo Markov Chain algorithm used to sample the posterior distribution.

Matlab implementation of the statistical model. The files include a source code, tutorial for its use, a simulated example, and the butterfly example of this paper.

Supplementary information for the case study.

Supplementary Table S1. Posterior estimates for the effects of the measured covariates derived from the model without latent factors. The sheets show the posterior mean estimate, the posterior median estimate, and the 0.025 and 0.975 quantiles.

Supplementary Table S2. Posterior estimates for the effects of the measured covariates derived from the model with latent factors. The sheets show the posterior mean estimate, the posterior median estimate, and the 0.025 and 0.975 quantiles.

References

Araujo, M. B. and A. Rozenfeld. 2014. The geographic scaling of biotic interactions. Ecography 37:406-415.

Asher, J., M. Warren, R. Fox, P. Harding, G. Jeffcoate, and S. Jeffcoate. 2001. The millennium atlas of butterflies in Britain and Ireland. Oxford University Press.

Bennie, J., J. A. Hodgson, C. R. Lawson, C. T. R. Holloway, D. B. Roy, T. Brereton, C. D. Thomas, and R. J. Wilson. 2013. Range expansion through fragmented landscapes under a variable climate. Ecology Letters 16:921-929.

Bhattacharya, A. and D. B. Dunson. 2011. Sparse Bayesian infinite factor models. Biometrika 98:291-306.

Blangiardo, M., M. Cameletti, G. Baio, and H. Rue. 2013. Spatial and spatio-temporal models with R-INLA. Spatial and spatio-temporal epidemiology 4:33-49.

Bonthoux, S., A. Baselga, and G. Balent. 2013. Assessing Community-Level and Single-Species Models Predictions of Species Distributions and Assemblage Composition after 25 Years of Land Cover Change. Plos One 8.

Borcard, D. and P. Legendre. 2002. All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. Ecological Modelling 153:51-68.

Brown, A. M., D. I. Warton, N. R. Andrew, M. Binns, G. Cassis, and H. Gibb. 2014. The fourth-corner solution - using predictive models to understand how species traits interact with the environment. Methods in Ecology and Evolution 5:344-352.

Chakraborty, A., A. E. Gelfand, A. M. Wilson, A. M. Latimer, and J. A. Silander, Jr. 2010. Modeling large scale species abundance with latent spatial processes. Annals of Applied Statistics 4:1403-1429.

Clark, J. S., A. E. Gelfand, C. W. Woodall, and K. Zhu. 2014. More than the sum of the parts: forest climate response from joint species distribution models. Ecological Applications 24:990-999.

Dorazio, R. M. and E. F. Connor. 2014. Estimating Abundances of Interacting Species Using Morphological Traits, Foraging Guilds, and Habitat. Plos One 9.

Dorazio, R. M., M. Kery, J. A. Royle, and M. Plattner. 2010. Models for inference in dynamic metacommunity systems. Ecology 91:2466-2475.

Dorazio, R. M. and J. A. Royle. 2005. Estimating size and composition of biological communities by modeling the occurrence of species. Journal of the American Statistical Association 100:389-398.

Dorazio, R. M., J. A. Royle, B. Soderstrom, and A. Glimskar. 2006. Estimating species richness and accumulation by modeling species occurrence and detectability. Ecology 87:842-854.

Dormann, C. F., J. M. McPherson, M. B. Araujo, R. Bivand, J. Bolliger, G. Carl, R. G. Davies, A. Hirzel, W. Jetz, W. D. Kissling, I. Kuehn, R. Ohlemueller, P. R. Peres-Neto, B. Reineking, B. Schroeder, F. M. Schurr, and R. Wilson. 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography 30:609-628.

Dray, S., P. Legendre, and P. R. Peres-Neto. 2006. Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). Ecological Modelling 196:483-493.

Dray, S., R. Pelissier, P. Couteron, M. J. Fortin, P. Legendre, P. R. Peres-Neto, E. Bellier, R. Bivand, F. G. Blanchet, M. De Caceres, A. B. Dufour, E. Heegaard, T. Jombart, F. Munoz, J. Oksanen, J.

Thioulouse, and H. H. Wagner. 2012. Community ecology in the age of multivariate multiscale spatial analysis. Ecological Monographs 82:257-275.
Dunstan, P. K., S. D. Foster, and R. Darnell. 2011. Model based grouping of species across environmental gradients. Ecological Modelling 222:955-963.
Ferrier, S. and A. Guisan. 2006. Spatial modelling of biodiversity at the community level. Journal of Applied Ecology 43:393-404.
Fielding, A. H. and J. F. Bell. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation 24:38-49.
Fuller, R. M., G. M. Smith, J. M. Sanderson, R. A. Hill, and A. G. Thomson. 2002. The UK Land Cover Map 2000: Construction of a parcel-based vector map from satellite images. Cartographic Journal 39:15-25.
Harris, D. J. 2015. Generating realistic assemblages with a joint species distribution model. Methods in Ecology and Evolution 6:465-473.
Helmus, M. R., K. Savage, M. W. Diebel, J. T. Maxted, and A. R. Ives. 2007. Separating the determinants of phylogenetic community structure. Ecology Letters 10:917-925.
Hodgson, J. A., C. D. Thomas, T. H. Oliver, B. J. Anderson, T. M. Brereton, and E. E. Crone. 2011. Predicting insect phenology across space and time. Global Change Biology 17:1289-1300.
Hui, F. K. C., Taskinen, S., Pledger, S., Foster, S. D., and Warton, D. I. 2015. Model-Based Approaches to Unconstrained Ordination. Methods in Ecology and Evolution:in press.
Hui, F. K. C., D. I. Warton, S. D. Foster, and P. K. Dunstan. 2013. To mix or not to mix: comparing the predictive performance of mixture models vs. separate species distribution models. Ecology 94:1913-1919.
Ives, A. R. and M. R. Helmus. 2010. Phylogenetic Metrics of Community Similarity. American Naturalist 176:E128-E142.
Ives, A. R. and M. R. Helmus. 2011. Generalized linear mixed models for phylogenetic analyses of community structure. Ecological Monographs 81:511-525.
Jackson, M. M., M. G. Turner, S. M. Pearson, and A. R. Ives. 2012. Seeing the forest and the trees: multilevel models reveal both species and community patterns. Ecosphere 3.
Kery, M., J. A. Royle, M. Plattner, and R. M. Dorazio. 2009. Species richness and occupancy estimation in communities subject to temporary emigration. Ecology 90:1279-1290.
Kissling, W. D., C. F. Dormann, J. Groeneveld, T. Hickler, I. Kuehn, G. J. McInerny, J. M. Montoya, C. Roemermann, K. Schiffers, F. M. Schurr, A. Singer, J.-C. Svenning, N. E. Zimmermann, and R. B. O'Hara. 2012. Towards novel approaches to modelling biotic interactions in multispecies assemblages at large spatial extents. Journal of Biogeography 39:2163-2178.
Latimer, A. M., S. Banerjee, H. Sang, Jr., E. S. Mosher, and J. A. Silander, Jr. 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.
le Roux, P. C., L. Pellissier, M. S. Wisz, and M. Luoto. 2014. Incorporating dominant species as proxies for biotic interactions strengthens plant community models. Journal of Ecology 102:767-775.
Legendre, P., M. R. T. Dale, M. J. Fortin, J. Gurevitch, M. Hohn, and D. Myers. 2002. The consequences of spatial structure for the design and analysis of ecological field surveys. Ecography 25:601-615.
Leibold, M. A., M. Holyoak, N. Mouquet, P. Amarasekare, J. M. Chase, M. F. Hoopes, R. D. Holt, J. B. Shurin, R. Law, D. Tilman, M. Loreau, and A. Gonzalez. 2004. The metacommunity concept: a framework for multi-scale community ecology. Ecology Letters 7:601-613.
Logue, J. B., N. Mouquet, H. Peter, H. Hillebrand, and G. Metacommunity Working. 2011. Empirical approaches to metacommunities: a review and comparison with theory. Trends in Ecology & Evolution 26:482-491.
Madon, B., D. I. Warton, and M. B. Araujo. 2013. Community-level vs species-specific approaches to model selection. Ecography 36:1291-1298.

Morales-Castilla, I., M. G. Matias, D. Gravel, and M. B. Araújo. 2015. Inferring biotic interactions from proxies. Trends in Ecology & Evolution 30:347-356.

Olden, A., O. Ovaskainen, J. S. Kotiaho, S. Laaka-Lindberg, and P. Halme. 2014. Bryophyte Species Richness on Retention Aspens Recovers in Time but Community Structure Does Not. Plos One 9.

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.

Ovaskainen, O. and J. Soininen. 2011. Making more out of sparse data: hierarchical modeling of species communities. Ecology 92:289-295.

Pellissier, L., R. P. Rohr, C. Ndiribe, J.-N. Pradervand, N. Salamin, A. Guisan, and M. Wisz. 2013. Combining food web and species distribution models for improved community projections. Ecology and Evolution 3:4572-4583.

Pollock, L. J., W. K. Morris, and P. A. Vesk. 2012. The role of functional traits in species distributions revealed through a hierarchical model. Ecography 35:716-725.

Pollock, L. J., R. Tingley, W. K. Morris, N. Golding, R. B. O'Hara, K. M. Parris, P. A. Vesk, and M. A. McCarthy. 2014. Understanding co-occurrence by modelling species simultaneously with a Joint Species DistributionModel (JSDM). Methods in Ecology and Evolution 5:397-406.

Russell, R. E., J. A. Royle, V. A. Saab, J. F. Lehmkuhl, W. M. Block, and J. R. Sauer. 2009. Modeling the effects of environmental disturbance on wildlife communities: avian responses to prescribed fire. Ecological Applications 19:1253-1263.

Sebastian-Gonzalez, E., J. Antonio Sanchez-Zapata, F. Botella, and O. Ovaskainen. 2010. Testing the heterospecific attraction hypothesis with time-series data on species co-occurrence. Proceedings of the Royal Society B-Biological Sciences 277:2983-2990.

Thorson, J. T., M. D. Scheuerell, A. O. Shelton, See, Kevin E., H. J. Skaug, and K. Kristensen. 2015. Spatial factor analysis: a new tool for estimating joint species distributions and correlations in species range. Methods in Ecology and Evolution:in press.

Tjur, T. 2009. Coefficients of Determination in Logistic Regression Models-A New Proposal: The Coefficient of Discrimination. American Statistician 63:366-372.

Ulrich, W., S. Soliveres, W. Kryszewski, F. T. Maestre, and N. J. Gotelli. 2014. Matrix models for quantifying competitive intransitivity from species abundance data. Oikos 123:1057-1070.

Walker, S. C. and D. A. Jackson. 2011. Random-effects ordination: describing and predicting multivariate correlations and co-occurrences. Ecological Monographs 81:635-663.

Wisz, M. S., J. Pottier, W. D. Kissling, L. Pellissier, J. Lenoir, C. F. Damgaard, C. F. Dormann, M. C. Forchhammer, J.-A. Grytnes, A. Guisan, R. K. Heikkinen, T. T. Hoye, I. Kuehn, M. Luoto, L. Maiorano, M.-C. Nilsson, S. Normand, E. Ockinger, N. M. Schmidt, M. Termansen, A. Timmermann, D. A. Wardle, P. Aastrup, and J.-C. Svenning. 2013. The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. Biological Reviews 88:15-30.

Zipkin, E. F., J. A. Royle, D. K. Dawson, and S. Bates. 2010. Multi-species occurrence models to evaluate the effects of conservation and management actions. Biological Conservation 143:479-484.

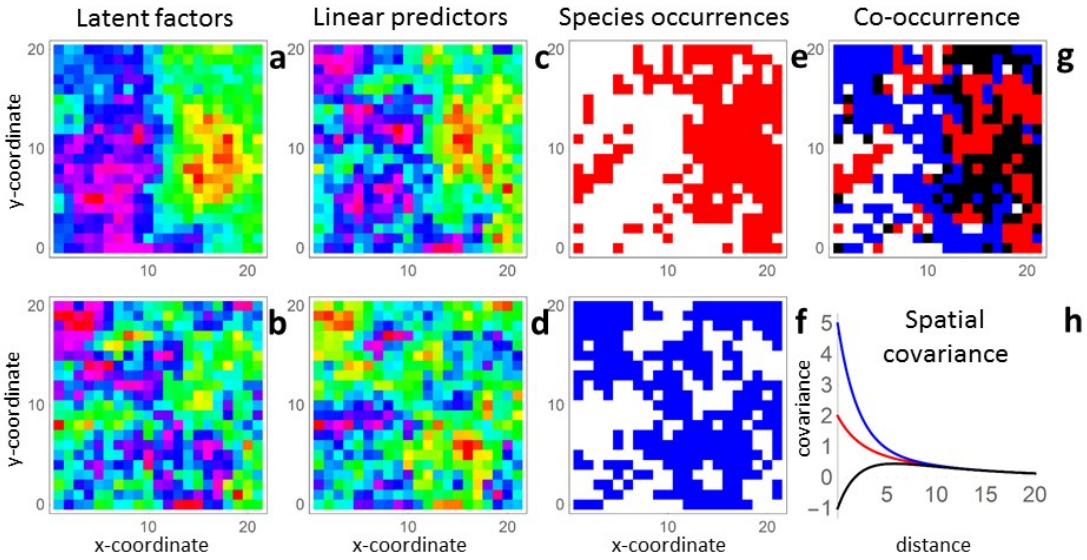

Fig. 1

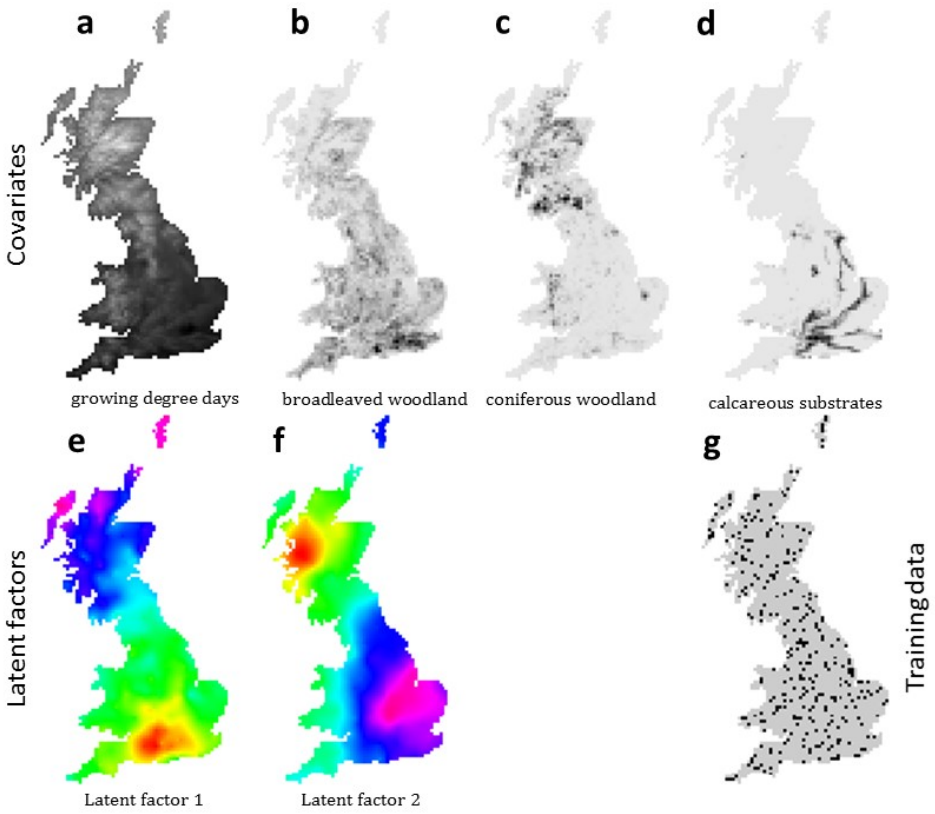

Fig. 2

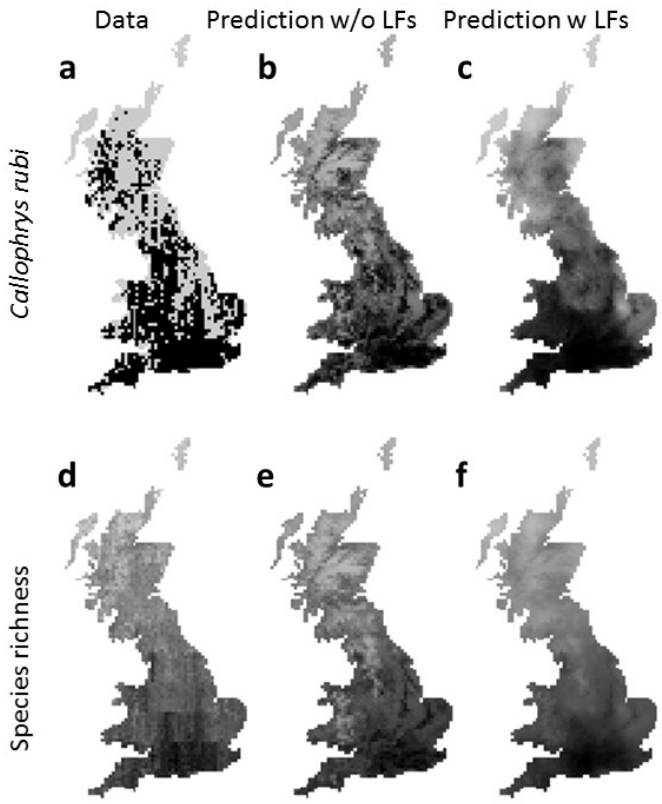

Fig. 3

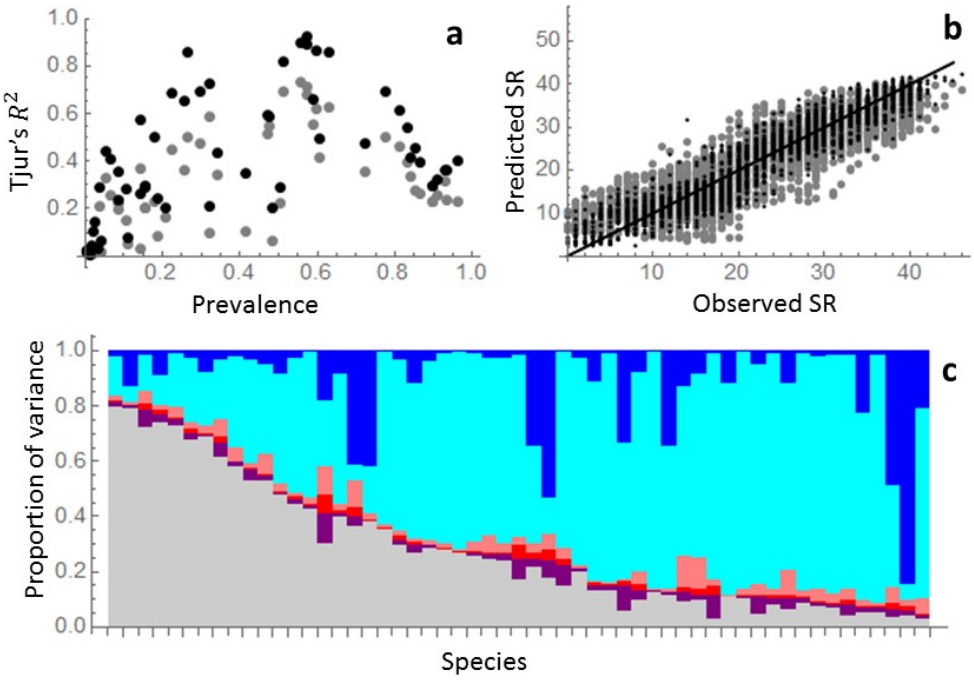

Fig. 4