Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2016, **7**, 549–555 doi: 10.1111/2041-210X.12501

Using latent variable models to identify large networks of species-to-species associations at different spatial scales

Otso Ovaskainen[1,2]*, Nerea Abrego[2,3,4], Panu Halme[3,5] and David Dunson[6]

[1]*Department of Biosciences, University of Helsinki, PO Box 65, FI-00014 Helsinki, Finland;* [2]*Centre for Biodiversity Dynamics, Department of Biology, Norwegian University of Science and Technology, N-7491 Trondheim, Norway;* [3]*Department of Biological and Environmental Science, University of Jyväskylä, PO Box 35, FI-40014 Jyväskylä, Finland;* [4]*Plant Biology and Ecology (Botany), Faculty of Science and Technology, University of the Basque Country (UPV/EHU), PO Box 644, E-48080 Bilbao, Spain;* [5]*Natural History Museum, University of Jyväskylä, PO Box 35, 40014 Jyväskylä, Finland; and* [6]*Department of Statistical Science, Duke University, PO Box 90251, Durham, NC, USA*

Summary

1. We present a hierarchical latent variable model that partitions variation in species occurrences and co-occurrences simultaneously at multiple spatial scales. We illustrate how the parameterized model can be used to predict the occurrences of a species by using as predictors not only the environmental covariates, but also the occurrences of all other species, at all spatial scales.

2. We leverage recent progress in Bayesian latent variable models to implement a computationally effective algorithm that enables one to consider large communities and extensive sampling schemes.

3. We exemplify the framework with a community of 98 fungal species sampled in *c*. 22 500 dead wood units in 230 plots in 29 beech forests.

4. The networks identified by correlations and partial correlations were consistent, as were networks for natural and managed forests, but networks at different spatial scales were dissimilar.

5. Accounting for the occurrences of the other species roughly doubled the predictive powers of the models compared to accounting for environmental covariates only.

Key-words: biotic interaction, co-occurrence, correlation, hierarchical model, joint species distribution model, partial correlation

Introduction

Biotic interactions play a key role in shaping the assembly and dynamics of species communities at different spatiotemporal scales (e.g. Morin 2011). However, it is often challenging to identify the positions of species within interaction networks and to infer the directions and strengths of the interactions. Thus, one of the major challenges in community ecology is to develop empirical and statistical methods to detect biotic interactions and assess their influence in structuring species communities (Morales-Castilla *et al*. 2015).

In this paper, we focus on the classical problem of inferring biotic interactions from snapshot data on species occurrence or abundance and data on environmental covariates. In theory, this aim is not feasible, as co-occurrence patterns generated by biotic interactions or shared responses to unmeasured environmental covariates are statistically indistinguishable (Kissling *et al*. 2012). However, species-to-species associations revealed by such analyses can be considered as data-driven hypotheses about biotic interactions (e.g. Ovaskainen, Hottola & Siitonen 2010; Morales-Castilla *et al*. 2015), which can be tested (e.g. with experiments). The early methods simply compared the numbers of co-occurrences to those based on a random distribution (e.g. Connor & Simberloff 1983; Hastings 1987; Stone & Roberts 1990; Gotelli 2000) and thus ignored the fact that co-occurrences can result also from shared niche. To correct for this, joint species distribution models emerged in the ecological literature, facilitating the assessment of environmental filtering and species interactions in an integrated framework (Ovaskainen, Hottola & Siitonen 2010; Clark *et al*. 2014; Pollock *et al*. 2014; Warton *et al*. In Press). One approach for inferring biotic interactions is to regress the occurrences of each species against both environmental variables and the occurrences of the other species (e.g. Dorazio *et al*. 2006; Kissling *et al*. 2012; Dorazio & Connor 2014). Another approach consists of modelling the residual co-occurrence after accounting for the influences of measured covariates (Peres-Neto, Olden & Jackson 2001; Ovaskainen, Hottola & Siitonen 2010; Kissling *et al*. 2012; Pollock *et al*. 2014). However, both of these approaches are feasible only for small species communities, as they estimate species-by-species effects without any structural assumptions.

*Correspondence author. E-mail: otso.ovaskainen@helsinki.fi.

Latent variable models have recently emerged in the ecological literature (Walker & Jackson 2011; Hui *et al*. 2015; Letten *et al*. 2015; Thorson *et al*. 2015; Warton *et al*. In Press) and comprise an efficient method for finding structure in multivariate data (Walker & Jackson 2011; Hui *et al*. 2015). In particular, latent variable models make the estimation of large co-occurrence matrices feasible because they can greatly reduce the numbers of parameters to be estimated. Latent variables can be used in a flexible way within hierarchical generalized models, and they can have, for example, spatial or temporal structures (Thorson *et al*. 2015).

In this paper, we present a hierarchical joint species distribution modelling framework that attributes variation in species occurrence and co-occurrence to the influences of environmental variables and species-to-species associations. We extend earlier approaches in four ways. (i) We include latent variables simultaneously at multiple spatial scales and thus present an alternative to the spatially explicit approach of Thorson *et al*. (2015). (ii) We show how the latent variable model can be used to predict the occurrences of a particular species by using as predictors not only the environmental covariates, but also the occurrences of all other species, at all spatial scales. (iii) We assess both correlations and partial correlations, the former being more informative about realized co-occurrence patterns and the latter being more likely to reveal causal relationships. (iv) We utilize recent progress in Bayesian latent variable models (Bhattacharya & Dunson 2011) to greatly improve the computational efficiency of the methods, thus enabling one to consider large communities (here 98 species) and extensive sampling designs (here *c*. 22 500 sampling units). We illustrate the use of the method with fungal occurrence data from both managed and natural beech forests.

Material and methods

THE EXAMPLE DATA

We illustrate the methods with data on wood-inhabiting fungi acquired in Navarre (northern Spain) with a hierarchical design: the inventories were conducted in 29 beech forests, out of which 8 were natural and 21 managed. Within each forest, 5–10 randomly located 10 × 10 m sample plots (in total 230 plots) were visited once, from late September to early November in 2011 and 2012. In each plot, all dead wood pieces (called henceforth resource units) larger than 0·2 cm in diameter were checked (3809 in natural and 18 651 in managed forests) and the occurrences of all wood-inhabiting macromycetes were recorded. Data on the diameter, length and decay stage of all resource units were recorded, whether there were fungal fruit bodies present. The spatial levels of resource units and forests are biologically motivated, whereas the level of plot is biologically arbitrary and motivated by the sampling design. For more details on the sampling design, see Abrego & Salcedo (2013, 2014) and Appendices S1 and S2 (Supporting information). The data are included in File S3.

STATISTICAL ANALYSIS

To examine the stability of species-to-species associations under different environmental conditions, we analysed the data separately for natural forests (which may be considered as the reference level) and for managed forests (which may be considered as a perturbation). As our main interest is in the estimation of species-to-species associations, for which there is no statistical power for species with very few occurrences, we included in the analyses only those 98 species (out of the 326 species) that occurred at least five times in the reference situation, that is in natural forests.

We model the occurrences of the species using a joint species distribution model, which accounts for the influences of the measured covariates, for random variation in species occurrence, as well as for species-to-species associations. To define the model, we denote the resource units by the index $i = 1,\ldots,n$ and the species by the index $j = 1,\ldots,m$, where n is the number of resource units and m is the number of species. We denote the presence–absence data by y_{ij}, so that $y_{ij} = 1$ if the species j was found in resource unit i and otherwise $y_{ij} = 0$. We model species occurrence with probit regression as $y_{ij} = 1_{z_{ij} > 0}$, where the latent occurrence score z_{ij} is defined as

$$z_{ij} = L_{ij} + \varepsilon^R_{ij} + \varepsilon^P_{P(i)j} + \varepsilon^F_{F(i)j} + \epsilon_{ij}. \qquad \text{eqn 1}$$

In Eq. 1, linear predictor L_{ij} includes the effects of the measured covariates, whereas the random effects ε^R_{ij}, $\varepsilon^P_{P(i)j}$ and $\varepsilon^F_{F(i)j}$ model variation in species occurrence and co-occurrence at the resource unit (R), plot (P) and forest (F) levels. The indices $P(i)$ and $F(i)$ denote the plot and the forest to which the resource unit i belongs, respectively. The residual terms are distributed independently as $\epsilon_{ij} \sim N(0, 1)$, and they correspond to the probit link function.

We modelled the linear predictor as $L_{ij} = \sum_{k=1}^{q} x_{ik}\beta_{jk}$, where the $q = 4$ covariates x_{ik} are the intercept ($x_{i1} = 1$), the log-transformed volume of the resource unit (x_{i2}), the decay class of the resource unit (x_{i3}) and the squared decay class of the resource unit (x_{i4}). The β_{jk} are the regression coefficients to be estimated. They model the overall prevalence of each species (through the intercept) as well as species niches, that is the influences of the sizes and the decay classes of the resource units on occurrence probability.

To improve statistical efficiency for data with many rare species, we follow Ovaskainen & Soininen (2011) in borrowing information across species. To do so, we assume that species-specific regression coefficients follow, independently among the species, the multivariate normal distribution $\boldsymbol{\beta}_j \sim N(\boldsymbol{\mu}, \varphi_j^{-1}\boldsymbol{\Sigma})$. Here, the $q \times 1$ vector $\boldsymbol{\beta}_j$ includes the regression coefficients of the species j. The $q \times 1$ vector $\boldsymbol{\mu}$ measures the response of an average species to the environmental covariates, and the $q \times q$ matrix $\boldsymbol{\Sigma}$ measures the amount of variation among the species-specific regression coefficients (the diagonal elements of $\boldsymbol{\Sigma}$) as well as the amount of covariation among the responses to the different covariates (the off-diagonal elements of $\boldsymbol{\Sigma}$). As an extension of the model of Ovaskainen & Soininen (2011), we applied a scale mixture approach (Andrews & Mallows 1974) by including the term φ_j, assumed to be distributed as $\varphi_j \sim \Gamma(\upsilon/2, \upsilon/2)$. The inclusion of this term allows for 'outlier species' by widening the tails of the multivariate normal distribution.

We assumed that the random effects ε^R_{ij}, $\varepsilon^P_{P(i)j}$ and $\varepsilon^F_{F(i)j}$ are distributed according to the multivariate normal distributions $\boldsymbol{\varepsilon}^R_i \sim N(0, \boldsymbol{\Omega}^R)$, $\boldsymbol{\varepsilon}^P_P \sim N(0, \boldsymbol{\Omega}^P)$ and $\boldsymbol{\varepsilon}^F_F \sim N(0, \boldsymbol{\Omega}^F)$, where $\boldsymbol{\Omega}^R$, $\boldsymbol{\Omega}^P$, and $\boldsymbol{\Omega}^F$ are species-to-species variance–covariance matrices. From the viewpoint of a single species, this corresponds to an ANOVA-type structure. The diagonal elements Ω^R_{jj}, Ω^P_{jj} and Ω^F_{jj} of the variance–covariance matrices quantify the amount of random variation for species j at each spatial level. As repeated measurements from the same plots and forests are not statistically independent, the random effects at these levels are needed to control for the dependency structure in the data. Our main interest is, however, in the non-diagonal elements of the variance–co-

variance matrices, that is $\Omega^R_{jj'}$, $\Omega^P_{jj'}$ and $\Omega^F_{jj'}$ with $j \neq j'$. These measure the covariance in the occurrence between the species j and j', and they thus assess whether the two species occur more or less often than by expected by chance, after accounting for the influences of their niche.

To improve statistical efficiency, we parameterized the variance–covariance matrices $\mathbf{\Omega}^R$, $\mathbf{\Omega}^P$ and $\mathbf{\Omega}^F$ using a latent variable approach. We thus modelled, for example, the resource unit-level random effects as $\varepsilon^R_{ij} = \sum_k \eta^R_{ik}\lambda^R_{kj}$, where the index k runs over the latent factors. The latent variables η^R_{ik} are assumed to be distributed as $\eta^R_{ik} \sim N(0, 1)$, and consequently, the variance–covariance matrix can be computed from the factor loadings as $\mathbf{\Omega}^R = (\boldsymbol{\lambda}^R)^T\boldsymbol{\lambda}^R$. The latent variables can represent either missing environmental covariates or the outcome of ecological interactions (Warton *et al.* In Press). In the case of a missing covariate, η^R_{ik} models the value of unmeasured environmental covariate in resource unit i, for example its humidity. The factor loading λ^R_{kj} then represents the response of species j to that unmeasured variable. In the case of ecological interactions, the latent variables η^R_{ik} do not have a straightforward interpretation, but they generate patterns where two species either co-occur more often than by random (if the factor loadings have the same sign) or where the two species co-occur less often by random (if the factor loadings have opposite signs). The reason why the latent variable approach makes it feasible to estimate large variance–covariance matrices is that the number of latent factors is typically much smaller than the number of species, and thus, the model has much fewer parameters to be estimated (Warton *et al.* In Press). To keep the model structure simple, we assume that the random effects at the three hierarchical spatial scales are independent of each other, and thus, we do not consider interactions among the spatial scales.

We note that as we model occurrences at the resource unit level, the model assumes that the number of occurrences is directly proportional to the number of resource units, and thus, it accounts for the direct effect of variation in resource availability. However, the model does not account for possible nonlinear effects between species occurrence and resource availability. For example, in a forest with limited amount of dead wood, a species may be below its extinction threshold (e.g. Hanski & Ovaskainen 2000) and thus absent even if there would be suitable resources for it. In theory, such nonlinear effects could be included in the fixed effects, for example, by adding the density of resources as a plot- or forest-level predictor. However, as we have not done so here due to the lack of data, in the present model the influences of population dynamical feedbacks are included in the random effects.

We fitted the model using Bayesian inference. We assumed an inverse Wishart prior for $\mathbf{\Sigma}$ with $q + 1$ degrees of freedom and the variance–covariance matrix set to identity matrix. We assume for each component of μ a normal prior with zero mean and unit variance. The parameter υ that relates to the inflation of variance (to allow for 'outlier species') was set to $\upsilon = 4$. To facilitate the estimation of high-dimensional variance–covariance matrices, we assumed sparse infinite factor priors for the latent loadings (Bhattacharya & Dunson 2011). We parameterized the prior (for all three spatial levels) as

$$\lambda_{kj} \sim N\left(0, \phi_{kj}^{-1}\tau_k^{-1}\right), \phi_{kj} \sim \text{Gamma}(3/2, 3/2),$$
$$\tau_k = \prod_{l=1}^{k} \delta_l, \delta_l \sim \text{Gamma}(50, 1).$$

As described in more detail in Bhattacharya & Dunson (2011), this prior allows for uncertainty in the number of factors and sparsity structure. The amount of shrinkage increases rapidly with factor index k, and thus, the model assigns non-negligible factor loadings only for those factors that are strongly supported by the data. We sampled the posterior distribution using the Gibbs sampler developed by Bhattacharya & Dunson (2011) with small modifications to fit the present model structure. The Gibbs sampler was implemented in Matlab, the code being available as File S1 and a tutorial as Appendix S4. We run the MCMC chain to 100 000 iterations, out of which the initial half was discarded. Using a desktop PC, the computational time was *c.* 6 h for the natural forest data and *c.* 20 h for the managed forest data.

At the level of the latent occurrence scores, the covariance structure of the latent variable part of the model is $\mathbf{\Omega} + \mathbf{I}$, where $\mathbf{\Omega} = \mathbf{\Omega}^R + \mathbf{\Omega}^P + \mathbf{\Omega}^F$, and the identity matrix arises from the residual term $\boldsymbol{\epsilon}_{ij}$. We measure the strengths of species-to-species associations by correlation matrices corresponding to each of the spatial scales ($\mathbf{R}^R$, $\mathbf{R}^P$, $\mathbf{R}^F$) as well as their sum ($\mathbf{R}$). We define these by adding to the corresponding variance–covariance matrices $\mathbf{\Omega}^R$, $\mathbf{\Omega}^P$, $\mathbf{\Omega}^F$ and $\mathbf{\Omega}$ the identity matrix $\mathbf{I}$ and then scaling by the standard deviations. We also derived partial correlation matrices by inverting the correlation matrices. Partial correlations are expected to be more informative about causal links among the species than the raw correlations, as a measure for the influence of one species that controls for the effects of all others.

We compared the relative roles of species-to-species associations and the match between species niche and resource availability in two different ways. First, we performed a variance partitioning of the latent score z_{ij}, thus examining for each species j how much of the variation in the latent scores over the resource units can be attributed to the fixed effects and how much to the random effects at the levels of resource units, plots and forests. We measured variation due to fixed effects as the variance of L_{ij} over the resource units i, whereas we measured variation due to random effects by the diagonal elements of the matrices $\mathbf{\Omega}^R$, $\mathbf{\Omega}^P$ and $\mathbf{\Omega}^F$. Secondly, we constructed posterior predictive data in two ways. In Prediction 1, we accounted only for the environmental covariates (i.e. the match between niche and resource availability), whereas in Prediction 2, we accounted additionally for the information on the occurrences of all the other species at all spatial scales. We computed for each species the Tjur R^2 values (Tjur 2009) for both Predictions 1 and 2 and assessed the importance of species-to-species associations by examining if and how much more accurate Prediction 2 was than Prediction 1.

When generating the Predictions 1 and 2, we sampled the regression parameters β as well as the species scores (i.e. the factor loadings) λ^R, λ^P and λ^F from the posterior distribution and the environmental scores (i.e. the latent variables) η^R, η^P and η^F as described in more detail below. We then computed the linear predictor $L_{ij} = \sum_{k=1}^{q} x_{ik}\beta_{jk}$, the latent occurrence score as $z_{ij} = L_{ij} + \varepsilon^R_{ij} + \varepsilon^P_{P(i)j} + \varepsilon^F_{F(i)j} + \boldsymbol{\epsilon}_{ij}$ and finally simulated species occurrence as $y_{ij} = 1_{z_{ij}>0}$. When generating Prediction 1, we sampled the environmental scores η^R, η^P and η^F from their priors. When generating Prediction 2 for species j, we sampled the environmental scores η^R, η^P and η^F conditionally on the occurrences of all other species j' except the species j. To do so, we utilized the same Gibbs sampler for η^R, η^P and η^F which we used for model parameterization. We then computed the effects $\varepsilon^R_{ij} + \varepsilon^P_{P(i)j} + \varepsilon^F_{F(i)j}$ for species j and generated the predictions as for Prediction 1.

Results

The proportions of resource units with 0, 1 and 2 or more fungal species were 35%, 39% and 26% in natural forests, whereas the corresponding proportions were 41%, 41% and 18% in managed forests. The estimates of the community-level response μ showed that the occurrence probabilities of most species increased with increasing volume of the resource unit and peaked at an intermediate decay class, both in natural and in managed forests (File S2), as expected from previous studies

(e.g. Nordén *et al.* 2013). Counting of cases for which the 95% credibility intervals of the species-specific estimates (β) did not cross zero (File S2), in the natural forests (respectively, managed forests) 70 species (respectively, 93 species) showed a positive response, whereas one species (respectively, none of the species) showed a negative response to resource unit volume. The estimate of β_{j4} was negative, meaning that the species achieve their maximal occurrence probability at an intermediate decay class, for 66 species in the natural forests and for 98 species in the managed forests, whereas for none of the species β_{j4} was positive.

For both the natural and the managed forests, on average more than half of the explained variance in species occurrence (at the level of the latent occurrence scores) was attributed to the fixed effects, whereas the remaining variance was attributed relatively evenly among the resource unit, plot and forest levels (Fig. 1). The largest numbers of species-to-species associations were recorded at the forest level, at which there were almost twice as many associations in natural forests compared to managed forests (Fig. 2a, Table 1). The numbers of positive and negative associations were roughly equal at all levels (Fig. 2a, Table 1). The numbers of associations were consistently smaller but not very much smaller if measured by partial correlations rather than by correlations (Table 1).

The correlations and partial correlations were highly consistent (Fig. 2b). In other words, if a species pair showed a positive (or respectively, a negative) correlation at a given spatial level, the same species pair typically showed also a positive (or respectively, a negative) partial correlation. Importantly, the results were consistent between natural and managed forests (Fig. 2c). Thus, if a species pair showed a positive or a negative correlation at a given level in the natural forest data, it was likely to do so also in the managed forest data. As we parameterized the model independently for these two data sets, this result suggests that the estimated associations are robust. Species pairs that co-occurred positively at the resource unit level tended to co-occur also at plot and forest levels, but the results were not consistent between plot and forest levels (Fig. 2d).

As expected for species with very low prevalence, the species-specific Tjur R^2 values were generally low, and they increased with prevalence (Fig. 3). The Prediction 2 that accounted for the occurrences of the other species was consistently better than Prediction 1 that accounted for the effects of environmental covariates only (Fig. 3). On average, the Tjur R^2 values were 2·6 times (respectively, 1·7 times) higher for Prediction 2 than for Prediction 1 for natural (respectively, managed) forests.

Discussion

Quantifying networks of ecological interactions and assessing their changes due to global environmental change has been the focus of much recent research (Fortuna & Bascompte 2006; Tylianakis *et al.* 2008; Araújo *et al.* 2011). Many of the terrestrial and marine ecosystems involve latent networks that cannot be measured directly. For example, while fungal interactions have been tested experimentally in the small scale (e.g. Boddy 2000; Heilmann-Clausen & Boddy 2005), a direct and robust measurement of large fungal interactive networks is presently not feasible. This is not only due to the high number of species involved, but also due to the differences between laboratory and natural conditions.

In this paper, we have illustrated how latent variables models can be used to infer associative networks at various spatial scales, as well as to use the inferred networks for generating improved predictions on species occurrence. We have illustrated the methods in the context of fungal communities, for which few studies have thus far attempted to assess the direction and strength of interactions under natural conditions (Ovaskainen, Hottola & Siitonen 2010; Ottosson *et al.* 2014). Wood-inhabiting fungal communities are known to be highly interactive in various facilitative and antagonistic ways

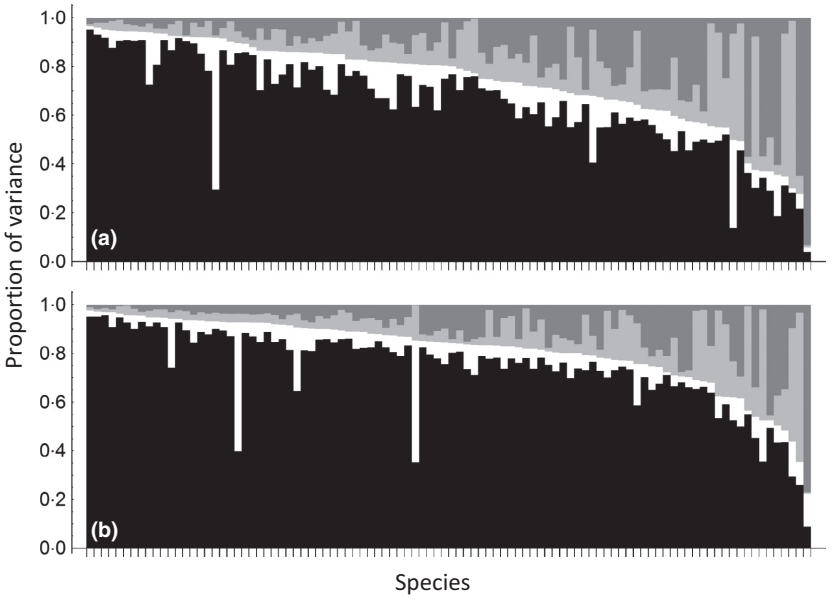

Fig. 1. The relative proportions of variance in species occurrence (at the level of the latent occurrence scores) attributed to the fixed effects and to the random effects at the three spatial scales. The variance attributed to the fixed effects (resource unit-level covariates) is shown by black, whereas the random effects are shown by white (resource unit level), light grey (plot level) and dark grey (forest level). The panels a and b show the data for the natural and managed forests, respectively. The species have been ordered according to the proportion of variance attributed to the fixed effects. The average proportions of variance for the natural forests are 0·65 (fixed effects), 0·09 (resource unit level), 0·12 (plot level) and 0·14 (forest level), whereas the corresponding numbers for the managed forests are 0·74, 0·07, 0·09 and 0·10, respectively.

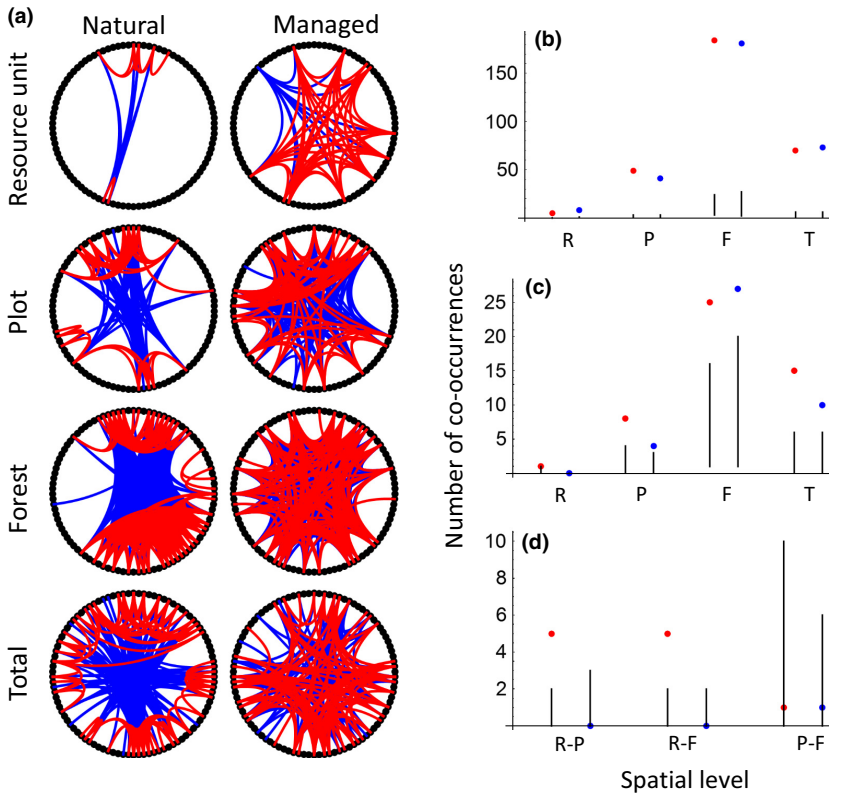

Fig. 2. Species-to-species associations detected at different spatial levels in natural and managed forests. In panel a, the species are shown by the black dots, and species pairs are connected by a red (respectively, blue) line if they showed a positive (respectively, negative) association with at least 95% posterior probability. In order to visualize network structure, the species were ordered circularly based on the angle derived from two leading eigenvectors of the correlation matrix. The ordering was based on the results for natural forests and done independently for each spatial level. For the same figure with species names as well as same figure for partial correlations, see Appendix S3. Panels b, c, d assess the consistency of the results between correlations and partial correlations (b; data shown for natural forests), between natural and managed forests (c; data shown for correlations) and among different spatial levels (d; data shown for correlations in natural forests). The red (respectively, blue) dots show the number of species pairs that showed a positive (respectively, negative) co-occurrence with at least 95% posterior probability in both of the two models that are compared. The lines show the 95% coverage of null expectations obtained by permuting the species identities 100 times. The letters R, P, F and T refer to the resource unit, plot, forest and total levels.

Table 1. The numbers of species pairs with positive (respectively, negative) associations with at least 95% posterior support. The numbers of latent variables identified for natural forests were 4, 4 and 5 for the resource unit, plot and forest levels, respectively. The corresponding numbers for managed forests were 4, 5 and 4

Forest type	Measure of association	Resource unit	Plot	Forest	Total
Natural	Correlation	7 (8)	54 (47)	250 (242)	105 (115)
	Partial correlation	5 (8)	51 (41)	218 (210)	82 (80)
Managed	Correlation	46 (10)	81 (82)	137 (151)	98 (72)
	Partial correlation	37 (10)	75 (69)	95 (123)	84 (51)

(Boddy, Frankland & van West 2008), and our results suggest that they are indeed highly non-randomly arranged. After controlling for log size and decay class, the numbers of positive and negative associations were roughly equal at all levels considered, reflecting no overall variation in species richness, but marked variation in community composition (Halme *et al.* 2013).

In wood-inhabiting fungi, direct interactions are necessarily restricted to the scale of an individual resource unit, within which the mycelia grow and compete for nutrients. In spite of this, we identified the smallest number of associations exactly at this scale. The explanation for the apparent contradiction is solely the question of statistical power. As two or more species were found to fruit only on a minority of the resource units, the data contain only weak signals of associations at this level. This is especially the case for negative associations, as it is difficult to tell whether a species is missing due to a negative association or merely by chance. To obtain more statistical power at the resource unit level, for example, DNA-based data on mycelial abundances would be required (e.g. Ovaskainen *et al.* 2013).

In contrast, we identified a large number of associations at the plot and forest levels. An interesting question is what generates the associations at these larger scales, as fungi do not interact directly at these levels. While variation caused by unmeasured environmental covariates is one straightforward explanation, we argue that also species interactions are likely to play a key role. As one example, fungi follow predictable patterns of community succession during the decay process of

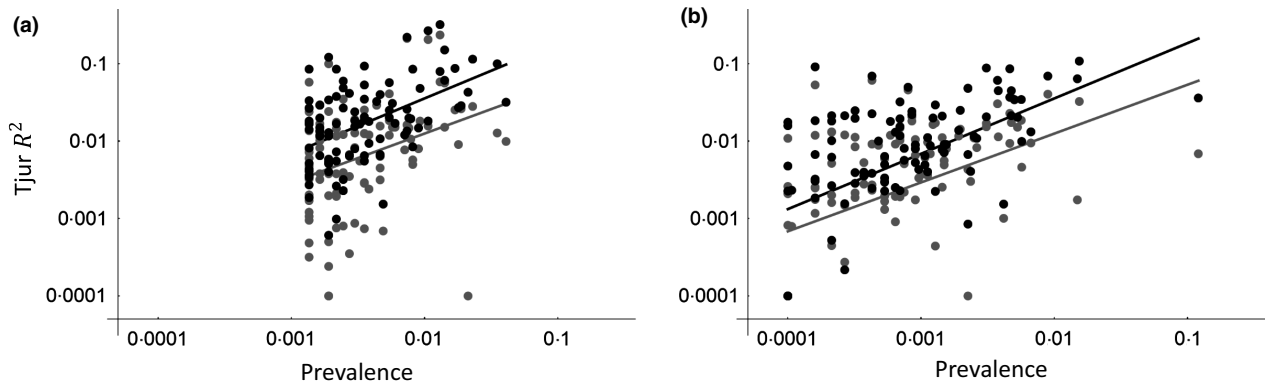

Fig. 3. The predictive powers of the models based on environmental covariates and occurrences of the other species. The dots show the species-specific Tjur R^2values and the lines the linear regressions. The colours indicate whether the predictions are based on environmental variables only (Prediction 1; grey) or on environmental variables and the occurrences of all other species (Prediction 2; black). Panel a corresponds to natural and panel b to managed forests.

the resource units. The probability of a particular species fruiting later during the succession can strongly depend on which predecessor species have taken part in the decay process of the resource unit earlier, for example, because different predecessor species modify the substrate in different ways (Fukami *et al*. 2010; Ottosson *et al*. 2014; Hiscox *et al*. 2015). In snapshot data, such associations are not visible at the resource unit scale, because the species do not fruit on the same resource unit at the same time. However, such associations can be seen at the plot or forest scales. For example, assume that a particular predecessor species is currently abundant in a certain forest on the early decay stage resource units. As the current occurrences are due to colonizations by members of the earlier generation, it is likely that the same species was abundant also some time ago and thus that those resource units that are currently in advanced decay stages were earlier decayed by the same predecessor species. Consequently, those species that are favoured by the initial decay caused by this predecessor species are likely to be currently abundant in the same forest, creating a positive association at the forest level.

More generally, at larger spatial scales, the effects of direct interactions become merged with the effects of dispersal, population dynamic feedbacks and patterns generated by variation in the abiotic environment (Götzenberger *et al*. 2012). A more mechanistic understanding of such feedbacks would call for studies of species occurrences and co-occurrences that would combine fruit body data with other kinds of data, for example, data on the mycelial stage (e.g. Ovaskainen *et al*. 2013), phenological data (Halme & Kotiaho 2012) and dispersal data (e.g. Norros *et al*. 2014). Examining the relationship between functional traits and species occurrence and co-occurrence at different spatial and temporal scales provides an important challenge for future work on fungal communities.

We found that the directions and strengths of associative interactions were consistent for natural and managed forests, suggesting the robustness of our method, as well as at least partial conservativeness of the associations under environmental perturbations. The associations were also highly consistent whether measured by correlations or partial correlations, possibly due to the generally low prevalence of the species, and thus only limited potential for generating indirect associations. In contrast, the estimated associations were not consistent at the plot and forest levels. This result is in concordance with earlier results suggesting that the mechanisms structuring assembly patterns differ at different spatial scales (Götzenberger *et al*. 2012).

We have demonstrated how the statistical model of species co-occurrence can be used to ask how the occurrence of a particular species is influenced by the occurrences of the other species. While we generated here the conditional predictions (Prediction 2) for each species conditional on the occurrences of all other species in all resource units, model predictions can more generally be conditioned to any subset of other species and any subset of the sampling units. Thus, one can, for example, ask how knowledge about the occurrences of a particular species group A improves the predictions for another species group B. This enables the systematic detection of indicator species, whose presence reflects the presences of other species in the community (Caro 2010). Good indicator species (species in group A) are such that they are easy to detect and that knowledge about their presence–absence much improves predictions for species of interest (species in group B), for example species of conservation concern that are difficult to observe directly.

Acknowledgements

We thank Izaro Errasti for help in digitalizing the data, Isabel Salcedo for help in the design of the data collection and the identification of the specimens and Iñaki Odriozola and anonymous reviewers for providing helpful comments. We acknowledge funding from the Academy of Finland (Grant No. 250444 to OO), the Research Council of Norway (CoE Grant No. 223257), Maj and Tor Nessling Foundation (Grant to PH) and the University of the Basque Country (UPV/EHU PhD fellowship PIF10/2010/PIF10008 to NA).

References

Abrego, N. & Salcedo, I. (2013) Variety of woody debris as the factor influencing wood-inhabiting fungal richness and assemblages: is it a question of quantity or quality? *Forest Ecology and Management*, **291**, 377–385.

Abrego, N. & Salcedo, I. (2014) Response of wood-inhabiting fungal community to fragmentation in a beech forest landscape. *Fungal Ecology*, **8**, 18–27.

Andrews, D.F. & Mallows, C.L. (1974) Scale mixtures of normal distributions. *Journal of the Royal Statistical Society Series B (Methodological)*, **36**, 99–102.
Araújo, M.B., Rozenfeld, A., Rahbek, C. & Marquet, P.A. (2011) Using species co-occurrence networks to assess the impacts of climate change. *Ecography*, **34**, 897–908.
Bhattacharya, A. & Dunson, D.B. (2011) Sparse Bayesian infinite factor models. *Biometrika*, **98**, 291–306.
Boddy, L. (2000) Interspecific combative interactions between wood-decaying basidiomycetes. *FEMS Microbiology Ecology*, **31**, 185–194.
Boddy, L., Frankland, J.C. & van West, P. (2008) *Ecology of Saprotrophic Basidiomycetes*. Elsevier Academic Press, London, UK.
Caro, T.M. (2010) *Conservation by Proxy: Indicator, Umbrella, Keystone, Flagship, and Other Surrogate Species*, 2nd edn. Island Press, Washington, Covelo, London.
Clark, J.S., Gelfand, A.E., Woodall, C.W. & Zhu, K. (2014) More than the sum of the parts: forest climate response from joint species distribution models. *Ecological Applications*, **24**, 990–999.
Connor, E.F. & Simberloff, D. (1983) Interspecific competition and species co-occurrence patterns on islands: null models and the evaluation of evidence. *Oikos*, **41**, 455–465.
Dorazio, R.M. & Connor, E.F. (2014) Estimating abundances of interacting species using morphological traits, foraging guilds, and habitat. *PLoS ONE*, **9**, e94323.
Dorazio, R.M., Royle, J.A., Söderström, B. & Glimskär, A. (2006) Estimating species richness and accumulation by modeling species occurrence and detectability. *Ecology*, **87**, 842–854.
Fortuna, M.A. & Bascompte, J. (2006) Habitat loss and the structure of plant-animal mutualistic networks. *Ecology Letters*, **9**, 281–286.
Fukami, T., Dickie, I.A., Wilkie, J.P., Paulus, B.C., Park, D., Roberts, A., Buchanan, P.K. & Allen, R.B. (2010) Assembly history dictates ecosystem functioning: evidence from wood decomposer communities. *Ecology Letters*, **13**, 675–684.
Gotelli, N.J. (2000) Null model analysis of species co-occurrence patterns. *Ecology*, **81**, 2606–2621.
Götzenberger, L., de Bello, F., Bråthen, K.A., Davison, J., Dubuis, A., Guisan, A. *et al.* (2012) Ecological assembly rules in plant communities-approaches, patterns and prospects. *Biological Reviews*, **87**, 111–127.
Halme, P. & Kotiaho, J.S. (2012) The importance of timing and number of surveys in fungal biodiversity research. *Biodiversity and Conservation*, **21**, 205–219.
Halme, P., Ódor, P., Christensen, M., Piltaver, A., Veerkamp, M., Walleyn, R., Siller, I. & Heilmann-Clausen, J. (2013) The effects of habitat degradation on metacommunity structure of wood-inhabiting fungi in European beech forests. *Biological Conservation*, **168**, 24–30.
Hanski, I. & Ovaskainen, O. (2000) The metapopulation capacity of a fragmented landscape. *Nature*, **404**, 755–758.
Hastings, A. (1987) Can competition be detected using species co-occurrence data? *Ecology*, **68**, 117–123.
Heilmann-Clausen, J. & Boddy, L. (2005) Inhibition and stimulation effects in communities of wood decay fungi: exudates from colonized wood influence growth by other species. *Microbial Ecology*, **49**, 399–406.
Hiscox, J., Savoury, M., Muller, C.T., Lindahl, B.D., Rogers, H.J. & Boddy, L. (2015) Priority effects during fungal community establishment in beech wood. *ISME Journal*, **9**, 2246–2260.
Hui, F.K.C., Taskinen, S., Pledger, S., Foster, S.D. & Warton, D.I. (2015) Model-based approaches to unconstrained ordination. *Methods in Ecology and Evolution*, **6**, 399–411.
Kissling, W.D., Dormann, C.F., Groeneveld, J., Hickler, T., Kühn, I., McInerny, G.J. *et al.* (2012) Towards novel approaches to modelling biotic interactions in multispecies assemblages at large spatial extents. *Journal of Biogeography*, **39**, 2163–2178.
Letten, A.D., Keith, D.A., Tozer, M.G. & Hui, F.K.C. (2015) Fine-scale hydrological niche differentiation through the lens of multi-species co-occurrence models. *Journal of Ecology*, **103**, 1264–1275.
Morales-Castilla, I., Matias, M.G., Gravel, D. & Araújo, M.B. (2015) Inferring biotic interactions from proxies. *Trends in Ecology & Evolution*, **30**, 347–356.
Morin, P.J. (2011) *Community Ecology*, 2nd edn. Wiley-Blackwell, Oxford, UK.
Nordén, J., Penttilä, R., Siitonen, J., Tomppo, E. & Ovaskainen, O. (2013) Specialist species of wood-inhabiting fungi struggle while generalists thrive in fragmented boreal forests. *Journal of Ecology*, **101**, 701–712.
Norros, V., Rannik, Ü., Hussein, T., Petäjä, T., Vesala, T. & Ovaskainen, O. (2014) Do small spores disperse further than large spores? *Ecology*, **95**, 1612–1621.
Ottosson, E., Nordén, J., Dahlberg, A., Edman, M., Jönsson, M., Larsson, K. *et al.* (2014) Species associations during the succession of wood-inhabiting fungal communities. *Fungal Ecology*, **11**, 17–28.
Ovaskainen, O., Hottola, J. & Siitonen, J. (2010) Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions. *Ecology*, **9**, 2514–2521.
Ovaskainen, O. & Soininen, J. (2011) Making more out of sparse data: hierarchical modeling of species communities. *Ecology*, **92**, 289–295.
Ovaskainen, O., Schigel, D., Ali-Kovero, H., Auvinen, P., Paulin, L., Norden, B. & Norden, J. (2013) Combining high-throughput sequencing with fruit body surveys reveals contrasting life-history strategies in fungi. *ISME Journal*, **7**, 1696–1709.
Peres-Neto, P.R., Olden, J.D. & Jackson, D.A. (2001) Environmentally constrained null models: site suitability as occupancy criterion. *Oikos*, **93**, 110–120.
Pollock, L.J., Tingley, R., Morris, W.K., Golding, N., O'Hara, R.B., Parris, K.M., Vesk, P.A. & McCarthy, M.A. (2014) Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). *Methods in Ecology and Evolution*, **5**, 397–406.
Stone, L. & Roberts, A. (1990) The checkerboard score and species distributions. *Oecologia*, **85**, 74–79.
Thorson, J.T., Scheuerell, M.D., Shelton, A.O., See, K.E., Skaug, H.J. & Kristensen, K. (2015) Spatial factor analysis: a new tool for estimating joint species distributions and correlations in species range. *Methods in Ecology and Evolution*, **6**, 627–637.
Tjur, T. (2009) Coefficients of determination in logistic regression models—A new proposal: the coefficient of discrimination. *American Statistician*, **63**, 366–372.
Tylianakis, J.M., Didham, R.K., Bascompte, J. & Wardle, D.A. (2008) Global change and species interactions in terrestrial ecosystems. *Ecology Letters*, **11**, 1351–1363.
Walker, S.C. & Jackson, D.A. (2011) Random-effects ordination: describing and predicting multivariate correlations and co-occurrences. *Ecological Monographs*, **81**, 635–663.
Warton, D., Blanchet, G., O'Hara, R., Ovaskainen, O., Taskinen, S., Walker, S. & Hui, F.K.C. (In Press) So many variables: joint modelling in community ecology. *Trends in Ecology & Evolution*. doi: 10.1016/j.tree.2015.09.007.

Received 24 July 2015; accepted 25 October 2015
Handling Editor: David Warton

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Appendix S1. Information on the sampled beech forests, their management history, their spatial locations and details on the sampling scheme.

Appendix S2. Literature consulted for the identification of fungal species.

Appendix S3. Supporting figures.

Appendix S4. Tutorial for the Matlab code.

File S1. Matlab code for model parameterization.

File S2. Community-level estimates for natural and managed forests, respectively. The sheets named 'mu' include parameter estimates for μ, the sheets named 'beta' include the parameter estimates for β, and the sheets named 'corr(Sigma)' for the off-diagonal elements of Σ, scaled by standard deviations. The subscripts '*_n*' and '*_m*' in the sheet names refer to natural and managed forests, respectively. For each parameter, the posterior mean estimate and the 0·025 and 0·975 posterior quantiles are shown.

File S3. The data used in the paper.

Data accessibility

All data used in the case study are provided in the supplementary material.