



The role of odds ratios in joint species distribution modeling

Alan E. Gelfand^{1,2} · Shinichiro Shirota^{1,2}

Received: 15 July 2020 / Revised: 21 December 2020 / Accepted: 4 January 2021 /
Published online: 9 February 2021

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC part of Springer Nature 2021

Abstract

Joint species distribution modeling is attracting increasing attention these days, acknowledging the fact that individual level modeling fails to take into account expected dependence/interaction between species. These joint models capture species dependence through an associated correlation matrix arising from a set of latent multivariate normal variables. However, these associations offer limited insight into realized dependence behavior between species at sites. We focus on presence/absence data using joint species modeling, which, in addition, incorporates spatial dependence between sites. For pairs of species selected from a collection, we emphasize the induced odds ratios (along with the joint occurrence probabilities); they provide a better appreciation of the practical dependence between species that is implicit in these joint species distribution modeling specifications. For any pair of species, the spatial structure enables a spatial odds ratio surface to illuminate how dependence varies over the region of interest. We illustrate with a dataset from the Cape Floristic Region of South Africa consisting of more than 600 species at more than 600 sites. We present the spatial distribution of odds ratios for pairs of species that are positively correlated and pairs that are negatively correlated under the joint species distribution model.

Keywords Allopatry · Gaussian process · Latent variables · Odds ratio · Spatial dependence · Species richness · Sympatry

Handling Editor: Luiz Duczmal.

✉ Shinichiro Shirota
shinichiro.shirota@gmail.com

¹ Department of Statistical Science, Duke University, Durham, NC, USA

² Department of Commerce, Meiji University, Tokyo, Japan

1 Introduction

Recently, numerous publications on joint species distribution modeling (JSDM) have appeared in the literature (Pollock et al. 2014; Thorson et al. 2015; Ovaskainen et al. 2016; Rota et al. 2016; Clark et al. 2017). A comparison of such modeling has been presented in Wilkinson et al. (2018). Such effort reflects the appreciation that observation of a community at a site anticipates dependence between the species present at that site. That is, stacked species distribution modeling (Guisan and Rahbek 2011; Calabrese et al. 2014), i.e., modeling the species independently but looking at the results jointly, need not perform well. For example, with presence/absence data, such modeling may tend to overestimate the probability of presence for each species at a site. Hence the number of presences, the richness, at a site (Guisan and Rahbek 2011; Clark et al. 2017) may be overestimated. This can be potentially more problematic when a large number of species are examined.

Specifically, Guisan and Rahbek (2011) offer the following criticisms of stacked species distribution modeling: (i) without adding a dispersal filter, it may incorrectly predict species in areas that appear environmentally suitable but that are outside their colonizable or historical range; (ii) it does not consider any constraints based on the carrying capacity of the local environment which determine the maximum number of species that may co-occur; and (iii) it does not explicitly consider any rules based on biotic interactions that control species co-occurrences and can exclude species from a community. As a result, too many species can be predicted to occur in a geographical unit by stacked species distribution models (e.g., Graham and Hijmans 2006; Pineda and Lobo 2009).

Joint species distribution modeling has been developed for presence/absence data, for count (abundance) data, and for composition data (Clark et al. 2017). Here, we focus primarily on presence/absence data, considering a large number, S , of species and a large number, n , of sites. Site i , $i = 1, 2, \dots, n$ provides an $S \times 1$, vector, \mathbf{Y}_i with entries 1 (presence) or 0 (absence). The sites may be viewed, hence modeled, as independent or as spatially dependent, as appropriate. Regardless, with j denoting the species, dependence among species, i.e., among the Y_{ij} in \mathbf{Y}_i , is modeled within a site.

The joint species distribution modeling challenge for presence/absence is the need to model, for each site i , the set of 2^S probabilities associated with the set of possible realizations of \mathbf{Y}_i . This is infeasible even for relatively small S (while we imagine S of order 10^2 or more). Rota et al. (2016) attempt modeling of such probabilities in the context of animal occupancy, employing a multivariate Bernoulli specification with logit links in order to specify regression models. They adopt log odds ratios as second-order interaction parameters under this specification and illustrate with a collection of four species. For plant species, where typically S is large, the solution commonly adopted in the literature is to introduce latent variables, \mathbf{Z}_i which drive the responses \mathbf{Y}_i . The \mathbf{Z}_i are modeled as $S \times 1$ multivariate normal vectors, which enable tractable model specification and fitting.

The contribution here is to note the strong limitation with regard to understanding dependence in response between pairs of species through the correlations between the corresponding pairs of latent normal variables. It has been remarked, e.g., Clark

et al. (2017), that the $S \times S$ correlation matrix associated with the latent multivariate normal vectors is primarily a device for creating dependence among species and that the pairwise correlations are difficult to interpret. We clarify that the dependence structure associated with the latent multivariate normal model used to drive the JSDM offers little useful information regarding whether joint occurrence of species at a site is encouraged or discouraged. We demonstrate that odds ratios offer a quantitatively more clear and useful story in this regard and propose their use for inference. Specifically, in terms of joint occurrence, the odds ratios reflect the contribution of the environmental covariates in addition to the pairwise correlations; as a result, they will vary across sites. Again, Rota et al. (2016), mentioned above, allude to odds ratios for species pairs. Lane et al. (2014) also examine odds ratios (in a non-model based fashion) with regard to pairwise association between species. In fact, they also remark that “the correlations provide little information on species presence.”

Further, we work with a spatial JSDM, following Shirota et al. (2019), which, for each pair of species, enables an odds ratio surface over the study region to illuminate the nature of pairwise species dependence over the region. Thus, we are able to capture presence/absence dependence between species at a site as well as spatial dependence in presence/absence across sites.

Our approach is applicable for JSDM modeling in general; in fact, it can proceed from the output of fitting any of the current crop of such models. However, in order to illustrate, we employ a plant communities dataset from the Cape Floristic Region (CFR) in South Africa (following Shirota et al. 2019). This dataset consists of presence–absence measurements for 639 tree species at 662 locations to which the spatial JSDM is fitted.

The format of the paper is as follows. The second section reminds us of 2×2 tables and associated dependence. Section 3 reviews the basics of JSDMs and also supplies connection to odds ratios. Section 4 brings in formal spatial modeling while Section 5 presents the results for the foregoing CFR dataset. Section 6 offers a brief conclusion with some possible extensions.

2 2 × 2 tables and familiar ecological dependence notions

For a single site and two species, a joint model requires specification of four probabilities, (p_{00} , p_{01} , p_{10} , and p_{11}) which sum to 1. The subscripts indicate absence (0) or presence (1) of the first and second species, respectively. Customarily, the probabilities are presented in a 2×2 table (where the rows are associated with species 1 and the columns are associated with species 2), as in Table 1(a).

The probabilities in Table 1(b) provide an example of strong sympatry, i.e., when the species are present, there is a strong chance that they are present together. Formally, sympatry is characterized by encouraging joint occurrence or joint absence, e.g., $P(\text{species 2 present}|\text{species 1 present}) > P(\text{species 2 present})$. With the notation above, we have $\frac{p_{11}}{p_1} > p_{.1}$, a departure from independence. Expressed through odds, we have $\frac{p_{11}}{p_{10}} > \frac{p_{01}}{p_{00}}$, the odds for the presence of species 2 when species 1 is present are greater than the odds when species 1 is not present. The odds ratio,

Table 1 2×2 tables : (a) A generic 2×2 table, (b) A table showing strong sympatry, (c) A table showing strong allopatry

		0		1	
(a)					
0		p_{00}	p_{01}	$p_{0.}$	
1		p_{10}	p_{11}	$p_{1.}$	
		$p_{.0}$	$p_{.1}$		1
(b)					
0		.14	.02	.16	
1		.04	.80	.84	
		.18	.82		1
(c)					
0		.03	.49	.52	
1		.47	.01	.48	
		.5	.5		1

$\theta \equiv \frac{p_{11}p_{00}}{p_{10}p_{01}} > 1$, is capturing positive dependence. For Table 1(b), $\theta = 140$, far above 1, very strong positive dependence between the species. Odds ratios are often presented on the ln scale, placing them on the whole real line, “symmetrizing” departure from independence around 0. Here, $\ln\theta = 4.94$.

Table 1(c) captures very strong allopatry. Formally, allopatry is characterized by discouraging co-occurrence, e.g., $P(\text{species 2 present}|\text{species 1 present}) < P(\text{species 2 present})$. Now, we have $\frac{p_{11}}{p_{1.}} < \frac{p_{01}}{p_{.1}}$, again a departure from independence. Expressed through odds, we have $\frac{p_{11}}{p_{10}} < \frac{p_{01}}{p_{00}}$. The odds for the presence of species 2 when species 1 is present are less than the odds when species 1 is not present; the odds ratio, $\theta = \frac{p_{11}p_{00}}{p_{10}p_{01}} < 1$, capturing negative dependence. For Table 1(c), $\theta = .0013$, well below 1; we have very strong negative dependence between the species. Also, $\ln\theta = -6.64$.

In summary, the odds ratio provides a useful tool for learning about species dependence with regard to presence/absence. Below, we employ odds ratios to assess to departure from independence, with interpretation in terms of encouraging or discouraging joint occurrence or joint absence for pairs of species. This is analogous to specifying correlation pairwise in providing multivariate normal distributions. Importantly, since independence modeling underlies stacked species distribution models, such models will not be able to capture sympatric or allopatric behavior for pairs of species.

In the context of a JSDM, for a pair of species, if both are somewhat prevalent, we can adopt the interpretation above for θ . Suppose one species, say species 1, is prevalent, i.e. $p_{10} + p_{11} \equiv q$ with q not very small and species 2 is rare, i.e., $p_{01} + p_{11} = \epsilon$ for ϵ very small, with say, $p_{11} = k\epsilon$, $0 < k < 1$. Then, the four cell probabilities are $p_{00} = (1 - q) - (1 - k)\epsilon$, $p_{01} = (1 - k)\epsilon$, $p_{10} = q - k\epsilon$, and $p_{11} = k\epsilon$. So, $\theta = \frac{((1-q)-(1-k)\epsilon)(k\epsilon)}{((1-k)\epsilon)(q-k\epsilon)} \approx k(1 - q)/q(1 - k)$. With regard to the interpretation of θ , it only depends upon the size of k relative to the size of q . The fact that the second species

is rare does not inhibit the interpretation of θ . However, suppose joint absence occurs nearly all of the time, i.e., $p_{00} = 1 - \epsilon$. As a result, all three remaining probabilities are very small, say $p_{01} = a\epsilon$, $p_{10} = b\epsilon$, and $p_{11} = c\epsilon$ with $a + b + c = 1$. Then, $\theta = \frac{(1-\epsilon)c\epsilon}{a\epsilon b\epsilon} = \frac{(1-\epsilon)c}{ab\epsilon} \approx c/ab\epsilon$. So, regardless of a , b , or c , θ will be very large, and this is driven by the very large probability of joint absence. Strong sympatry results from the high probability of joint absence and θ is less meaningful.

Under a JSMDM we typically consider that many possible species can potentially occur at a site, e.g., $O(10^2)$, as in our dataset. However, typically we find just a few species at a given site, perhaps at most 10. As a result, at site \mathbf{s} , with the indicator function, $1_j(\mathbf{s})$, indicating presence or absence of species j at location (\mathbf{s}) , most of the $P(1_j(\mathbf{s}) = 1) \equiv p_j(\mathbf{s}) \approx 0$. Therefore, for many species pairs, we will be in the last case above. So, we will obtain more useful insight from odds ratios when at least one species is somewhat prevalent. Arguably, species interaction is of lesser interest for pairs of rare species.

2.1 Species richness

We offer a brief remark on species richness which records the number of distinct species present at the site and is commonly used to characterize species distributions at sites. With the notation above, the observed richness at site \mathbf{s} is $\text{Rich}(\mathbf{s}) = \sum_{j=1}^S 1_j(\mathbf{s})$. (To be clear, $\{1_j(\mathbf{s}), j = 1, 2, \dots, S\}$ does not constitute a multinomial trial but, rather, a set of dependent Bernoulli trials.) Whether we model species independently, using stacked species distribution models or dependently, using JSMDM's, $E(\text{Rich}(\mathbf{s})) = \sum_{j=1}^S E(1_j(\mathbf{s})) = \sum_{j=1}^S P(1_j(\mathbf{s}) = 1) = \sum_{j=1}^S p_j(\mathbf{s})$.

The practical issue is whether the marginal expectations across the species under a joint model will tend to agree with the corresponding expectations across species under independent models. Because the joint model considers the data for all of the species at a site while the individual models consider the data only for the individual species at the site, unconstrained by the overall presence/absence at the site, intuitively, we might anticipate the latter expectations to be larger, suggesting prediction of higher richness using a stacked species distribution model. Regardless, we expect to incorrectly estimate uncertainty in richness when the indicator variables in the sum are not independent.

3 Joint species distribution modeling

We offer a brief review of joint species distribution modeling in the nonspatial case. We take our data to consist of observations Y_{ij} , a binary variable indicating presence ($Y_{ij} = 1$) or absence ($Y_{ij} = 0$) of species j at site i . Following either Clark et al. (2017) or Ovaskainen et al. (2016), we associate with each Y_{ij} a latent normal variable Z_{ij} . In fact, we have a latent multivariate normal vector \mathbf{Z}_i which yields the observed \mathbf{Y}_i .

Following Ovaskainen et al. (2016), in the specification for Z_{ij} , we let L_{ij}^F and L_{ij}^R denote the fixed and random effects contributions, respectively, which are included

additively in the modeling of Z_{ij} . The L notation is intended to suggest a linear form. We assume the $S \times 1$ vector $\mathbf{L}_i^R \sim N(\mathbf{0}, H)$ with H an $S \times S$ correlation matrix capturing dependence between species. So, marginally, $L_{ij}^R \sim N(0, 1)$ and $H_{jj'} \equiv \rho^{(j,j')}$ is the correlation between species j and j' .

We adopt a *functional* relationship between Y_{ij} and Z_{ij} , that is, $Y_{ij} = 1(Z_{ij} > 0)$ and $P(Y_{ij} = 1) = P(Z_{ij} > 0)$ under the specification that $Z_{ij} = L_{ij}^F + L_{ij}^R + \epsilon_{ij}$ where the ϵ_{ij} are pure error terms, i.e., $\epsilon \sim N(0, 1)$. Dependence is introduced through the specification for L_{ij}^R , making Z_{ij} and $Z_{ij'}$ dependent ($\text{corr}(Z_{ij}, Z_{ij'}) = \rho^{(j,j')}$) and therefore, Y_{ij} and $Y_{ij'}$ dependent.

Given L_{ij}^F and L_{ij}^R , the Z_{ij} are conditionally independent and $P(Y_{ij} = 1) = P(Z_{ij} > 0) = \Phi(L_{ij}^F + L_{ij}^R)$. Now, it is the Φ 's that are dependent, yielding dependence between species at the so-called second stage of a hierarchical model.

Ovaskainen et al. (2016) employ a *conditional* specification, $[Y_{ij}|Z_{ij}]$. Under a probit link function, they specify $P(Y_{ij} = 1) = \Phi(Z_{ij})$. So, if $Z_{ij} = L_{ij}^F + L_{ij}^R$, they obtain $\Phi(L_{ij}^F + L_{ij}^R)$. The distinction between functional and conditional specification seems muddled in, e.g., the review paper of Wilkinson et al. (2018). In the sequel, we work exclusively with the functional specification, as above.

The correlations arising from the latent Z 's ignore the issue that, while many species can potentially occur at a site, typically we find just a few. While we might be considering say 600+ species (as in our example) that could be present at a site, we will rarely see more than say 10. In fact, these correlations do not vary with site and have little to do with the actual realization of \mathbf{Y}_i at site i . Those species we see at the site i will typically be in response to the available environmental predictors at the site. Furthermore, a positive association may be suggestive of co-occurrence or of a potential substitution effect, i.e., a particular species is present but another, say similar one, could equally well have been successful there. That is, the joint distribution at a site is not a multinomial trial where the occurrence of one species precludes the occurrence of the other, yielding negative correlations. The joint distribution allows presence of the pair of species and, if they respond similarly, adjusted for the environment, we can have positive associations. This leads to discussion presented in Zobel and Anton (1997) and in Ovaskainen et al. (2016) regarding the nature of the species pool.

As a last comment here, why not specify the ϵ_{ij} to be standard logistic random errors rather than standard normal random errors to attempt to directly connect probabilities to logits needed for log odds ratios, as in Sect. 2? That is, we could follow the path in Ovaskainen et al. (2010), which employs a multivariate logistic link. However, the incompatibility between the logistic distribution for the ϵ 's and the normal distribution for the L^R 's muddies the computational waters. For instance, $p_{i,11}^{(j,j')} = P(Z_{ij} \geq 0, Z_{ij'} \geq 0) = \int a(L_{ij}^R)a(L_{ij'}^R)f(L_{ij}^R, L_{ij'}^R)dL_{ij}^RdL_{ij'}^R$ where $a(L_{ij}^R) = \frac{e^{L_{ij}^F + L_{ij}^R}}{1 + e^{L_{ij}^F + L_{ij}^R}}$ and $f(\cdot, \cdot)$ is a bivariate normal density. Such integrals are awkward to work with for no benefit.

3.1 Connection to odds ratios

For the JSDMs above, again, dependence across species is captured through the pairwise correlation between species in the latent bivariate normal distribution. We do not model the $p_{a,b}^{(j,j')}$, $a, b = 0, 1$ directly but, rather, we model the parameters in the latent multivariate normal distribution and, as a result, each of these probabilities is a smooth function of these parameters.

However, there is no direct connection between say, $\rho^{(j,j')}$ and the odds ratio associated with the induced 2×2 table of joint probabilities for the species pair, (j, j') at site i . Specifically, suppose the latent bivariate normal distribution for $\begin{pmatrix} Z_{ij} \\ Z_{ij'} \end{pmatrix}$ has mean $\begin{pmatrix} \mu_i^{(j)} \\ \mu_i^{(j')} \end{pmatrix}$ and correlation matrix $\begin{pmatrix} 1 & \rho^{(j,j')} \\ \rho^{(j,j')} & 1 \end{pmatrix}$. Then,

$$\theta_i^{(j,j')} = \frac{p_{i,00}^{(j,j')} p_{i,11}^{(j,j')}}{p_{i,10}^{(j,j')} p_{i,01}^{(j,j')}} = \frac{P(Z_{ij} < 0, Z_{ij'} < 0)P(Z_{ij} \geq 0, Z_{ij'} \geq 0)}{P(Z_{ij} \geq 0, Z_{ij'} < 0)P(Z_{ij} < 0, Z_{ij'} \geq 0)}. \tag{1}$$

The expressions for the double integrals in (1) show that each probability is a function of $\mu_i^{(j)}$, $\mu_i^{(j')}$, and $\rho^{(j,j')}$. The Appendix provides the attractive result that $\theta_i^{(j,j')}$ is non-decreasing in $\rho^{(j,j')}$ for fixed $\mu_i^{(j)}$ and $\mu_i^{(j')}$. However, in the presence of $\mu_i^{(j)}$, $\mu_i^{(j')}$, the latent correlations do not determine the strength/magnitude of the odds ratios.

In more detail, the result in the Appendix should be applied to $W_{ij} = Z_{ij} - \mu_i^{(j)}$ where say, $\mu_i^{(j)} = \mathbf{X}_i^T \boldsymbol{\beta}_j$ and $W_{ij'} = Z_{ij'} - \mu_i^{(j')}$ where say, $\mu_i^{(j')} = \mathbf{X}_i^T \boldsymbol{\beta}_{j'}$. As a result, $P(Z_{ij} < 0, Z_{ij'} < 0) = P(W_{ij} < c_{ij}, W_{ij'} < c_{ij'})$, where $c_{ij} = -\mathbf{X}_i^T \boldsymbol{\beta}_j$ and $c_{ij'} = -\mathbf{X}_i^T \boldsymbol{\beta}_{j'}$, is non-decreasing in $\rho^{(j,j')}$ for any \mathbf{X}_i , $\boldsymbol{\beta}_j$, and $\boldsymbol{\beta}_{j'}$ and therefore so is the associated odds ratio, $\theta(\mathbf{X}_i, \boldsymbol{\beta}_j, \boldsymbol{\beta}_{j'})$. Then, for a given $\rho^{(j,j')}$, we can see the response of $\theta(\mathbf{X}_i, \boldsymbol{\beta}_j, \boldsymbol{\beta}_{j'})$ to changes in \mathbf{X}_i for given coefficient vectors; we can understand how the odds ratio varies across environmental niches. In different words, JSDMs *disentangle* the role of the environment from the role of biotic interactions in the model specification. With these models, odds ratios provide a measure of association that *unifies* the effects of the biotic and abiotic conditions while enabling assessment of the effect of each on the association.

Figure 1 illustrates the foregoing in the lower panels by showing the log odds ratio ($\ln \theta$) and the joint occurrence probability (p_{11}) as a function of ρ for given pairs of μ 's. The monotonicity is observed in both $\ln \theta$ and p_{11} with respect to ρ . In fact, according to the μ 's, we can have p_{11} arbitrarily large with $\rho < 0$. That is, ρ supplies essentially no information regarding the probability of joint occurrence of the species. (A similar argument can be made in terms of joint absence.) The Appendix further argues that we can not have contradictory signs for ρ and $\ln \theta$. If $\rho > 0 (< 0)$, $\ln \theta \geq (\leq) 0$. However, again according to the μ 's, for a given $\rho > 0$, $\ln \theta$ can take arbitrary positive values; for a given $\rho < 0$, $\ln \theta$ can take arbitrary negative values. Again, ρ provides little indication of the strength of dependence between species.

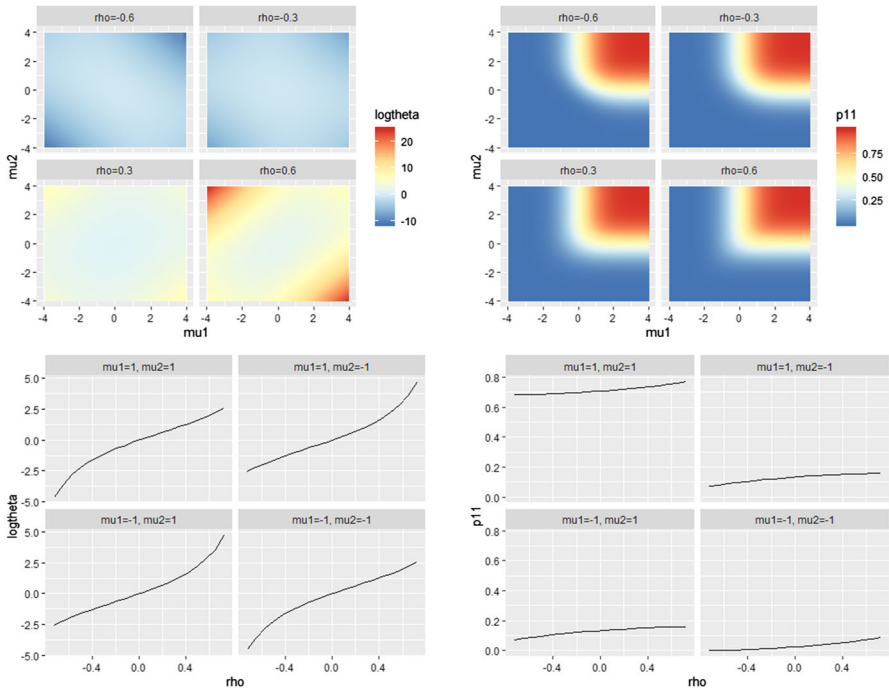


Fig. 1 $\ln \theta$ (left) and p_{11} (right) as a function of ρ (bottom) and μ 's (top)

We can ask whether dependence using either θ ($\ln \theta$) or p_{11} can demonstrate monotonicity in the level of a predictor, i.e., in a component of \mathbf{X} ? For p_{11} (and p_{00}) the answer is yes; this is clear since $\mu^{(j)}$ and $\mu^{(j')}$ will be monotone in the predictor. The upper right panel of Fig. 1 reveals this. For θ the answer is no. Technical analysis is messy, but the upper left panel of Fig. 1 demonstrates this by following straight lines from left to right or top to bottom. Therefore, while there is a relationship between the odds ratio and the latent means, hence, regressors, it is not monotone. As a result, spatial variation in the odds ratio surface need not align with spatial variation in a regressor.

Lastly, working within a Bayesian framework, analysis of odds ratios and joint occurrence probabilities is a post-model fitting activity. We can examine any pair of species at any location since these probabilistic objects are functions of the model parameters (along with the environmental predictors); posterior samples of model parameters provide posterior samples of these objects. As a result, we can obtain the posterior distribution for, say, $\theta_i^{(j,j')}$. We can summarize this distribution with a posterior mean but, we can also calculate probabilities such as $P(\theta_i^{(j,j')} > 1 | \text{data})$, the probability of positive association. The Appendix comments on such calculation.

4 Spatial dependence considerations

Bringing in spatial dependence, first-order spatial explanation is captured through spatially referenced predictors (covariates) as a spatial regression. Second-order explanation brings in spatial dependence in the sense that, for a species, locations closer to each other are anticipated to provide a similar probability of presence (adjusted for covariates). Spatial random effects, customarily assumed to come from a Gaussian process using so-called geostatistical models (Banerjee et al. 2014; Cressie and Wikle 2011), are employed. With binary responses, they are introduced through a latent Gaussian process as follows.

We modify the notation above by attaching the location \mathbf{s}_i to site i and writing $Y_{ij} \equiv Y_j(\mathbf{s}_i)$. Now we envision point level modeling with a conceptual presence/absence variable, $Y_j(\mathbf{s})$ for species j at every location, \mathbf{s} , in a study region, say D . For species j , we have a realization of a presence/absence surface, $\{Y_j(\mathbf{s}) : \mathbf{s} \in D\}$ which is observed at $\{\mathbf{s}_i, i = 1, 2, \dots, n\}$.

With regard to the Z 's, extending Sect. 4, at location \mathbf{s} , we have $Z_j(\mathbf{s}) = L_j^F(\mathbf{s}) + L_j^R(\mathbf{s}) + \epsilon_j(\mathbf{s})$. Association between the Z 's is identical to that between the L^F 's. With the functional specification, now $Y_j(\mathbf{s}) = 1(Z_j(\mathbf{s}) > 0)$, which is referred to as a clipped Gaussian field in the literature (e.g., De Oliveira 2000).

There are two types of dependence in play here. The first is with regard to presence/absence among species at a site, while the second is with regard to presence/absence across sites. With multiple species, the latter requires the cross-covariance for the latent Gaussian variables, the covariance between, say $Z_j(\mathbf{s})$ and $Z_{j'}(\mathbf{s}')$ and supplies dependence between the presence of species j at location \mathbf{s} and species j' at location \mathbf{s}' . See Banerjee et al. (2014) for a full discussion of cross covariance specification and see Shirota et al. (2019) for joint species distribution modeling, taking into account spatial dependence between locations.

For the joint modeling, we now envision site-specific two-way tables at each \mathbf{s} , comprised of $p_{ab}^{(j,j')}(\mathbf{s})$ ($a, b, = 0, 1$), resulting in a “surface” of such tables, which can be summarized by a log odds ratio surface, $\ln\theta(\mathbf{s})^{(j,j')}$. The spatial dependence modeling provides smoothing for these surfaces as well as interpolation of these surfaces to unobserved locations. The key point is extension of earlier marginal spatial work which presented individual probability of presence surfaces for each of a set of species (Latimer et al. 2006; Gelfand et al. 2005). Now, we examine, pairwise, probability of joint occurrence, of joint absence, and log odds ratios as surfaces over the region of interest.

It is worth remarking that, here and in our application, we are viewing the sampling sites as points within a fairly large study region. For species j , we envision a binary variable (presence/absence) at each point. We are not considering presence/absence associated with areal units, i.e., whether there is at least one individual of species j in the areal unit. This leads to very different modeling for presence/absence (and is beyond the scope here). See Gelfand and Shirota (2019) for careful discussion and development of the differences.

4.1 Model and inference details

We briefly present the spatial JSDM for presence/absence data from Shirota et al. (2019). Let $D \subset \mathbb{R}^2$ be a bounded study region with $\mathcal{S} = \{s_1, \dots, s_n\} \in D$ a set of plot locations, and $\mathbf{Z}(s_i) \in \mathbb{R}^S$ be an $S \times 1$ latent vector of continuous variables at location s_i . Under independence for the locations, the model for \mathbf{Z}_i is specified as

$$\mathbf{Z}(s_i) = \mathbf{B}\mathbf{X}(s_i) + \boldsymbol{\epsilon}(s_i), \quad \boldsymbol{\epsilon}(s_i) \stackrel{iid}{\sim} \mathcal{N}_S(\mathbf{0}, \mathbf{6}), \quad \text{for } i = 1, \dots, n \quad (2)$$

where \mathbf{B} is an $S \times p$ coefficient matrix, $\mathbf{X}(s_i)$ is a $p \times 1$ covariate vector at location s_i and $\mathbf{6}$ is a $S \times S$ covariance matrix for species. This model has order $\mathcal{O}(S^2)$ parameters, $S(S+1)/2$ parameters from $\mathbf{6}$ and pS parameters from \mathbf{B} . For example, for $S = 300$ species and $p = 3$ covariates, the model contains 46,050 parameters.

Taylor-Rodríguez et al. (2017) propose a dimension reduction approximation to $\mathbf{6}$ which allows the number of parameters to grow linearly in S . They approximate $\mathbf{6}$ with $\mathbf{6}^* = \mathbf{3}\mathbf{3}^T + \sigma_\epsilon^2 \mathbf{I}_S$ and replace the above model with

$$\mathbf{Z}(s_i) = \mathbf{B}\mathbf{X}(s_i) + \mathbf{3}\mathbf{w}(s_i) + \boldsymbol{\epsilon}(s_i), \quad \boldsymbol{\epsilon}(s_i) \sim \mathcal{N}_S(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_S), \quad \text{for } i = 1, \dots, n \quad (3)$$

where the random vectors $\mathbf{w}(s_i)$ are i.i.d, distributed as $\mathcal{N}_r(\mathbf{0}, \mathbf{I}_r)$, and $\mathbf{3}$ is an $S \times r$ matrix with $r \ll S$. Now, $\mathbf{6}^*$ has only $Sr + 1$ parameters, the estimation problem of $\mathcal{O}(S^2)$ parameters is reduced to that of $\mathcal{O}(S)$ parameters. This specification is referred to as the dimension reduced nonspatial model. That is, we have spatially referenced predictors but no spatial dependence.

Spatial dependence is introduced through the $\mathbf{w}(s_i)$. We suppose the $r \times 1$ vector $\mathbf{w}(s_i)$ is a realization of an r dimensional Gaussian process at s_i . In fact, we let the components of the vector be associated with independent and identically distributed Gaussian processes with common (decay parameter) exponential covariance function.¹ This model is referred to as the dimension reduced spatial model. The full specification is more complex than we have presented here, including the use of Dirichlet process specifications as part of a dimension reduction to cluster species. That hierarchical model, specified within a Bayesian framework, with prior distributions, model fitting, and prediction is exactly what we used in the example of the next section. In the interest of flow and space for the manuscript, we omit details; the interested reader is encouraged to consult Shirota et al. (2019) for explicit development.

For binary response data in the form of presence-absence, again we employ the functional relationship form, $Y^{(j)}(s_i) = 1$ if $Z^{(j)}(s_i) \geq 0$, $= 0$ if $Z^{(j)}(s_i) < 0$ for $j = 1, \dots, S$, $i = 1, \dots, n$. The data-augmentation algorithm proposed by Chib (1998) for multivariate probit regression, which improves the mixing of the Markov chain Monte Carlo (MCMC) algorithm, is implemented for model fitting.

¹ Under the dimension reduction, we can include at most $r \ll S$ decay parameters where r is say 3 to 5. The effect of adopting a common decay parameter for the latent GP's is expected to be negligible.

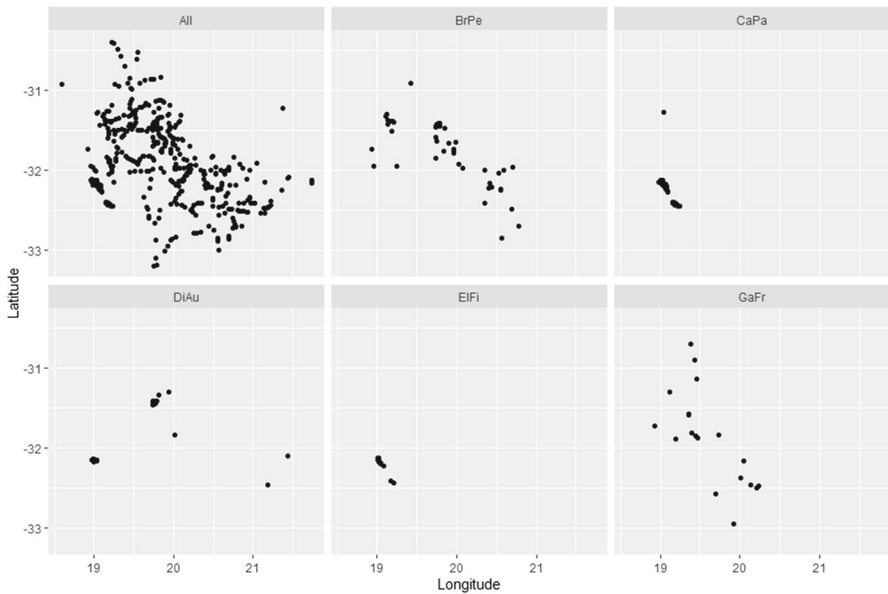


Fig. 2 All 662 locations in CFR and the distribution of the presence of the 5 selected species

5 The Cape Floristic Region data

Our data is extracted from a large database studying the distribution of plants in the Cape Floristic Region (CFR) of South Africa (Takhtajan 1986). The CFR is one of the six floral kingdoms in the world and is located in the southwestern part of South Africa. Though, geographically it is relatively small, it is extremely diverse (9, 000+ species) and highly endemic (70% occur only in the CFR (Rebello 2001)). There are more than 40, 000 geo-coded sites within the CFR where sampling has been recorded. The database from which our dataset was extracted consists of more than 1,400 sites with more than 2,800 species, spanning six regions. The subset we use comes from one of these regions with $n = 662$ sites and $S = 639$ species. The response is binary, presence-absence for each species and site.

Under restriction to a latitude-longitude rectangle, the upper left panel of Fig. 2 shows the 662 locations of the CFR data and the remaining panels show the distribution of 5 selected species: 1) *Bromus pectinatus* (BrPe); 2) *Cannomois parviflora* (CaPa); 3) *Diospyros austro-africana* (DiAu); 4) *Elegia filacea* (ElFi); and 5) *Galenia fruticosa* (GaFr). For covariate information, we include: (1) elevation, (2) mean annual precipitation, and (3) mean annual temperature; these values are standardized, following Shirota et al. (2019). Figure 3 shows the levels of these three associated standardized covariates at the locations. Altogether, the total number of binary responses is $n \times S = 662 \times 639 = 423, 018$. The overall number of presences is 6,980, 1.65% of the total number of binary responses. Again, though we have many species in our dataset, only a few are present on a given plot.

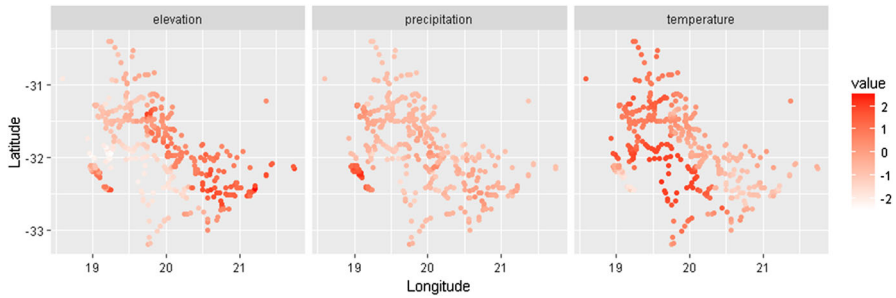


Fig. 3 Standardized covariate surface: elevation (left), mean annual precipitation (middle) and mean annual temperature (right)

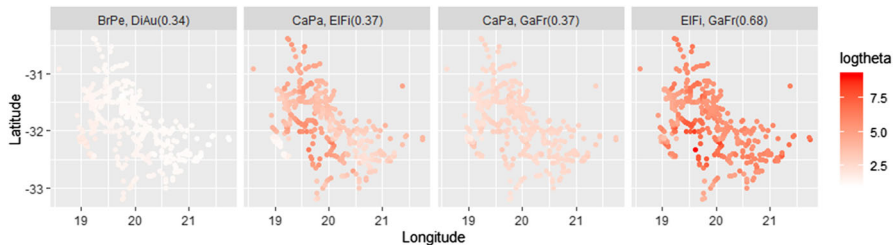


Fig. 4 Posterior mean surfaces of log odds ratio for all pairs of 5 species with positive correlation. The values in parenthesis are the posterior correlations for each pair

It is difficult to grasp a 639×639 correlation matrix and investigation of the resulting potential 20, 416 odds ratio surfaces resulting from all possible pairs is infeasible. So, we focus on the foregoing 5 species which occur in at least 20 of the locations and provide a range of posterior correlations from -0.7 to 0.7 under the model fitting. The 10 posterior mean log odds ratio ($\ln \theta(\mathbf{s})$) surfaces are presented in Figs. 4 (positively correlated pairs) and 5 (negatively correlated pairs).

The surfaces are above 0 for positively correlated pairs and below 0 for negatively correlated pairs, as argued above, but show spatial variation attributable to the covariates. Specifically, the correlations of EIFi with the other species are quite different: -0.30 (BrPe), 0.37 (CaPa), -0.70 (DiAu) and 0.68 (GaFr). For the strongly positively correlated pair, EIFi and GaFr, we see much spatial variation in association with many large positive $\ln \theta(\mathbf{s})$. For the strongly negatively correlated pair EIFi and DiAu, again we see much variation in association with many large negative $\ln \theta(\mathbf{s})$.

The probability of joint occurrence (p_{11}) is very small (very close to 0) across all pairs and most locations (though p_{11} reaches 0.15 for positively correlated pairs at some locations); in the interest of space, we do not show them. In summary, the probabilities of joint absence, the $p_{00}^{(j,j')}(\mathbf{s})$, are almost always very high. Under individual species modeling, $P(Y_j(\mathbf{s}) = 1)$ need not be very small. Thus, while a marginal model may suggest that the chance of absence is not high, the joint chance of absence will almost always be high.

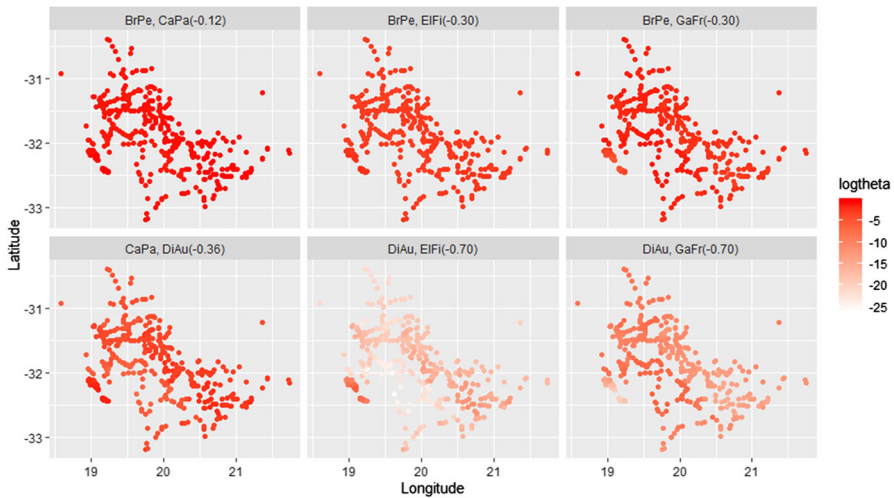


Fig. 5 Posterior mean surfaces of log odds ratio for all pairs of 5 species with negative correlation. The values in parenthesis are the posterior correlations for each pair

6 Conclusion

For the current collection of high-dimensional joint species distribution models, focusing on presence/absence data, we have argued that the correlations which are incorporated in the latent multivariate normal models that drive these JSDMs offer limited inference regarding species interaction. We have shown that odds ratios as well as joint occurrence probabilities (or joint absence probabilities), all induced by the JSDM specification, enable a more clear interpretation. Further, using a JSDM with spatial modeling of presence/absence, we developed odds ratio surfaces as well as joint occurrence surfaces over a study region. In this way, we can assess how joint species dependence varies over the region.

For a dataset with spatially varying regressors, whether one fits a packaged JSDM model (which assumes independence across plots) or a spatial JSDM, we obtain spatially varying odds ratios (and joint presence probabilities), extending the interpretation beyond pairwise correlations. We unify the effects of the biotic and abiotic conditions while still enabling assessment of the effect of each on the association.

While we have focused on odds ratio and joint presence surfaces, we could consider other summaries in this spirit. For instance, we could be interested in conditional probability surfaces, i.e., the probability of the presence of one species given another one is there and vice versa. Unlike the odds ratio which is symmetric, focusing on whether both species are there or not, these features are asymmetric, accounting for an ordering of the species. In fact, such conditioning can be extended to more than two species to assess the probability of presence given say two other species are there.

A different extension would turn to abundance, often of more interest than presence/absence. When abundances are recorded as counts, frequently they are converted to ordinal classifications, e.g., (i) absent, (ii) 1 to 5, (iii) 6 to 20, (iv) 21 to 50, and

(v) more than 50. In some cases, abundances are presented as *relative* abundances, normalizing individual species abundance by total abundance across all species. Now, we can specify ordinal classifications through proportions. In other cases, abundances are viewed as cover-abundance, i.e., a measure of plant cover (often used in vegetation science). It is based on percentages with several scales of cover-abundance in the literature, e.g. the 5-point cover scale of Braun-Blanquet or the Domin scale (Mueller-Dombois and Ellenberg 1974).

For a pair of species, if there are K ordinal classifications, we replace the foregoing 2×2 table with a $K \times K$ table, letting $Y_{i,j}$ denote the categorical classification for species j at site i , taking ordered values $k = 1, 2, \dots, K$. Then, we can consider local, global, and cumulative odds ratios (Agresti 2012) as measures of species dependence. They are defined as:

- (i) **Local:** $\theta_{i,kk'}^{L(j,j')} = \frac{P(Y_{i,j}=k, Y_{i,j'}=k')P(Y_{i,j}=k+1, Y_{i,j'}=k'+1)}{P(Y_{i,j}=k, Y_{i,j'}=k'+1)P(Y_{i,j}=k+1, Y_{i,j'}=k')}$
- (ii) **Global:** $\theta_{i,kk'}^{G(j,j')} = \frac{P(Y_{i,j} \leq k, Y_{i,j'} \leq k')P(Y_{i,j} > k, Y_{i,j'} > k')}{P(Y_{i,j} \leq k, Y_{i,j'} > k')P(Y_{i,j} > k, Y_{i,j'} \leq k')}$
- (iii) **Cumulative:** $\theta_{i,kk'}^{C(j,j')} = \frac{P(Y_{i,j'} \leq k' | Y_{i,j}=k) / P(Y_{i,j'} > k' | Y_{i,j}=k)}{P(Y_{i,j'} \leq k' | Y_{i,j}=k+1) / P(Y_{i,j'} > k' | Y_{i,j}=k+1)}$

By extending the discussion in Sect. 2, each of these odds ratios has a clear interpretation in terms of a particular type of pairwise species dependence.

Acknowledgements The computational results were obtained using Ox version 6.21 (Doornik 2007).

Appendix

I: We consider, in detail, the connection between the correlation arising under the latent multivariate normal model for species pairs and the associated odds ratio. We draw on some older work relating bivariate normal probabilities to the associated bivariate correlation. There is a substantial literature, with multivariate extensions, and we only note two papers here: Gupta (1963) and Slepian (1962).

The basic result we need is the following:

Theorem: Suppose $\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim \text{BivN}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$. Then, $P(Z_1 \leq c_1, Z_2 \leq c_2)$ is non-decreasing in ρ .

We apply this result to (1). For fixed $\mu_i^{(j)}$ and $\mu_i^{(j')}$, by simple probability calculations, we have $p_{i,00}^{(j,j')}$ non-decreasing in $\rho^{(j,j')}$, we have $p_{i,01}^{(j,j')}$ non-increasing in $\rho^{(j,j')}$, we have $p_{i,10}^{(j,j')}$ non-increasing in $\rho^{(j,j')}$, and we have $p_{i,11}^{(j,j')}$ non-decreasing in $\rho^{(j,j')}$. As a result, the numerator in (1) is non-decreasing in $\rho^{(j,j')}$ while the denominator in (1) is non-increasing in $\rho^{(j,j')}$. So, altogether, we have $\theta_i^{(j,j')}$ non-decreasing in $\rho^{(j,j')}$ for all i and (j, j') pairs.

As a corollary, since $\theta_i^{(j,j')} = 1$ when $\rho^{(j,j')} = 0$, we must have $\theta_i^{(j,j')} \geq 1$ when $\rho^{(j,j')} > 0$ and $\theta_i^{(j,j')} \leq 1$ when $\rho^{(j,j')} < 0$.

II: We offer some brief words regarding how calculation of probabilities is implemented. Under Markov chain Monte Carlo model fitting, we obtain posterior samples, say $\mu_{i,b}^{(j)}, \mu_{i,b}^{(j')}, \rho_b^{(j,j')}, b = 1, 2, \dots, B$. Each term in (1) is a double integral which is a function of $(\mu_i^{(j)}, \mu_i^{(j')}, \rho^{(j,j')})$. So, each sample, $\mu_{i,b}^{(j)}, \mu_{i,b}^{(j')}, \rho_b^{(j,j')}$ produces a posterior realization of each of the four terms on the right side of (1), e.g., $p_{i,00,b}^{(j,j')}$ hence a posterior realization, $\theta_{i,b}^{(j,j')}$. Across $b = 1, 2, \dots, B$, we obtain a posterior sample of size B for each of the terms on the right side of (1) as well as the induced odds ratio. The double integrals that are needed can be computed using approximations or numerically. In fact, if we work with the functional JSDM and specification (ii) above, then Z_{ij} and $Z_{ij'}$ are conditionally independent given $L_{ij}^F, L_{ij'}^F, L_{ij}^R$ and $L_{ij'}^R$. So, we only need to calculate univariate cumulative normal distribution functions.

A fully Monte Carlo alternative is to generate many $(Z_{ij}, Z_{ij'})$ pairs, hence many $(Y_{ij}, Y_{ij'})$ pairs for each $\mu_{i,b}^{(j)}, \mu_{i,b}^{(j')}, \rho_b^{(j,j')}$. The collection can be placed into a 2×2 table from which we can obtain a posterior realization of each of the probabilities on the right side of (1) as well the odds ratio.

References

- Agresti A (2012) Categorical data analysis, 3rd edn. Wiley, New York
- Banerjee S, Carlin BP, Gelfand AE (2014) Hierarchical modeling and analysis for spatial data, 2nd edn. Chapman and Hall, New York
- Calabrese JM, Certain G, Kraan C, Dormann CF (2014) Stacking species distribution models and adjusting bias by linking them to macroecological models. *Glob Ecol Biogeogr* 23:99–112
- Chib S (1998) Analysis of multivariate probit models. *Biometrika* 85:347–361
- Clark JS, Nemergut D, Seyednasrollah B, Turner PJ, Zhang S (2017) Generalized joint attribute modeling for biodiversity analysis: median-zero, multivariate, multifarious data. *Ecol Monogr* 87:34–56
- Cressie N, Wikle CK (2011) Statistics for spatio-temporal data. Wiley, New York
- De Oliveira V (2000) Bayesian prediction of clipped Gaussian random fields. *Comput Stat Data Anal* 34:299–314
- Doornik J (2007) Ox: object oriented matrix programming. Timberlake Consultants Press, New York
- Gelfand AE, Schmidt AM, Wu S, Silander JA Jr, Latimer A, Rebelo AG (2005) Modelling species diversity through species level hierarchical modelling. *J R Stat Soc Ser C* 54:1–20
- Gelfand AE, Shirota S (2019) Preferential sampling for presence/absence data and for fusion of presence/absence data with presence-only data. *Ecol Monogr* 89:e01372
- Graham CH, Hijmans RJ (2006) A comparison of methods for mapping species ranges and species richness. *Glob Ecol Biogeogr* 15:578–587
- Guisan A, Rahbek C (2011) SESAM - a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages. *J Biogeogr* 38:1433–1444
- Gupta SS (1963) Probability integrals of multivariate normal and multivariate. *Ann Math Stat* 34:792–828
- Lane PW, Lindenmayer DB, Barton PS, Blanchard W, Westgate MJ (2014) Visualization of species pairwise association: a case study of surrogacy in bird assemblages. *Ecol Evol* 4:3279–3289
- Latimer A, Wu S, Gelfand AE, Silander Jr JA (2006) Building statistical models to analyze species distributions. *Ecol Appl* 16:33–50
- Mueller-Dombois D, Ellenberg H (1974) Aims and methods of vegetation ecology. Wiley, New York
- Ovaskainen O, Abrego N, Halme P, Dunson D (2016) Using latent variable models to identify large networks of species-to-species associations at different spatial scales. *Methods Ecol Evo* 7:549–555
- Ovaskainen O, Hottola J, Siitonen J (2010) Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions. *Ecology* 91:2514–2521
- Pineda E, Lobo JM (2009) Assessing the accuracy of species distribution models to predict amphibian species richness patterns. *J Anim Ecol* 78:182–190

- Pollock LJ, Tingley R, Morris WK, Golding N, O'Hara RB, Parris KM, Vesk PA, McCarthy MA (2014) Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). *Methods Ecol Evol* 5:397–406
- Rebello T (2001) SASOL proteas: a field guide to the proteas of South Africa, 2nd edn. Fernwood Press, Halifax
- Rota CT, Ferreira MAR, Kays RW, Forrester TD, Kalies EL, McShea WJ, Parsons AW, Millsbaugh JJ (2016) A multispecies occupancy model for two or more interacting species. *Methods Ecol Evol* 7:1164–1173
- Shirota S, Gelfand AE, Banerjee S (2019) Spatial joint species distribution modeling using Dirichlet processes. *Stat Sin* 29:1127–1154
- Slepian D (1962) The one-sided barrier problem for Gaussian noise. *Bell Syst Techn J* 41:463–501
- Takhtajan A (1986) Floristic regions of the world. University of California Press, California
- Taylor-Rodríguez D, Kaufeld K, Schliep EM, Clark JS, Gelfand AE (2017) Joint species distribution modeling: dimension reduction using Dirichlet processes. *Bayesian Anal* 12:939–967
- Thorson JT, Scheuerell MD, Shelton AO, See KE, Skaug HJ, Kristensen K (2015) Spatial factor analysis: a new tool for estimating joint species distributions and correlations in species range. *Methods Ecol Evol* 6:627–637
- Wilkinson DP, Golding N, Guillera-Aroita G, Tingley R, McCarthy MA (2018) A comparison of joint species distribution models for presence–absence data. *Methods Ecol Evol* 10:198–211
- Zobel DB, Anton JA (1997) A decade of recovery of understory vegetation buried by volcanic tephra from Mount St. Helens. *Ecol Monogr* 67:317–344

Alan E. Gelfand is The James B. Duke Professor of Statistical Science at Duke University. He is former chair of the Department of Statistical Science (DSS) and enjoys a secondary appointment as Professor of Environmental Science and Policy in the Nicholas School. Author of more than 300 papers (more than 240 since 1990), Gelfand is internationally known for his contributions to applied statistics, Bayesian computation and Bayesian inference. (An article in *Science Watch* found him to be the tenth most cited mathematical scientist in the world over the period 1991–2001). Gelfand is an Elected Fellow of the American Statistical Association, the Institute of Mathematical Statistics, and the International Society for Bayesian Analysis. He is an Elected Member of the International Statistical Institute. He is a former President of the International Society for Bayesian Analysis and in 2006 he received the Parzen Prize for a lifetime of research contribution to Statistics. In 2012, he was chosen to give the distinguished Mahalanobis lectures. In 2013, he received a Distinguished Achievement Medal from the ASA Section on Statistics in the Environment. In 2019, he received the S.S. Wilks Memorial Award from the American Statistical Association. Gelfand's primary research focus for the past twenty years has been in the area of statistical modeling for spatial and space-time data. Through a collection of more than 150 papers he has advanced methodology, using the Bayesian paradigm, to associate fully model-based inference with spatial and space-time data displays. His chief areas of application include environmental exposure, spatio-temporal ecological processes, and climate dynamics. He has four books in this area, including the successful "Hierarchical Modeling and Analysis for Spatial Data" with Sudipto Banerjee and Brad Carlin (now second edition), "Hierarchical Modeling for Environmental Data; Some Applications and Perspectives" with James Clark, the "Handbook of Spatial Statistics" with Peter Diggle, Montserrat Fuentes, and Peter Guttorp and the "Handbook of Environmental and Ecological Statistics" with Montserrat Fuentes, Jennifer Hoeting, and Richard Smith. In addition, he has a NSF-CBMS monograph with Erin Schliep entitled, "Bayesian Analysis and Computation for Spatial Point Patterns."

Shinichiro Shirota is Assistant Professor at the Department of Commerce, Meiji University, Japan. He obtained his Ph.D. in Statistics from Duke University in 2017, under the supervision of Professor Alan Gelfand. His primary research focus is in the area of statistical modeling for spatial and space-time data from Bayesian paradigm and statistical computation of Gaussian processes and related probabilistic models. His research areas of application include environmental science, public health and social science.