Community Confounding In Joint Species Distribution Models

Justin J. Van Ee (vanee002@colostate.edu)
Colorado State University
Jacob S. Ivan
Colorado Parks and Wildlife
Mevin B. Hooten
The University of Texas at Austin

Research Article

Keywords: models, relationships, epidemic, community

Posted Date: December 6th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1111644/v1

Version of Record: A version of this preprint was published at Scientific Reports on July 18th, 2022. See the published version at https://doi.org/10.1038/s41598-022-15694-6.

Community Confounding in Joint Species Distribution Models

Justin J. Van Ee[1,*], Jacob S. Ivan[2], and Mevin B. Hooten[3]

[1]Colorado State University, Department of Statistics, Fort Collins, 80523, USA

[2]Colorado Parks and Wildlife, Fort Collins, 80526, USA

[3]University of Texas at Austin, Department of Statistics and Data Sciences, Austin, 78712, USA

[*]vanee002@colostate.edu

ABSTRACT

Joint species distribution models have become ubiquitous for studying species-habitat relationships and dependence among species. Accounting for community structure often improves predictive power, but can also alter inference on species-habitat relationships. Modulated species-habitat relationships are indicative of community confounding: The situation in which interspecies dependence and habitat effects compete to explain species distributions. We discuss community confounding in a case study of mammalian responses to the Colorado bark beetle epidemic in the subalpine forest by comparing the inference from independent single species distribution models and a joint species distribution model. We present a method for measuring community confounding and develop a restricted version of our hierarchical model that orthogonalizes the habitat and species random effects. Our results indicate that variables associated with the severity and duration of the bark beetle epidemic suffer from community confounding. This implies that mammalian responses to the bark beetle epidemic are governed by interconnected habitat and community effects. Disentangling habitat and community effects can improve our understanding of the ecological system and possible management strategies. We evaluate restricted regression as a method for alleviating community confounding and distinguish it from other inferential methods for confounded models.

Introduction

Ecological datasets that provide insights about collections of organisms have become prevalent over the last decade thanks to efforts like Long Term Ecological Research Network (LTER), National Ecological Observatory Network (NEON), citizen science surveys, etc.[1]. In addition, technology has improved our ability to fit modern statistical models to these datasets jointly. As a consequence, joint species distribution models (JSDM)[2–4] have become popular for modeling dependence among species simultaneously with environmental drivers of occurrence and/or abundance. JSDMs provide inference for species-habitat relationship conditional on the community. Species habitat preferences are modulated by interspecies dynamics[5–7]. Therefore, we should consider the relationships between community and environmental effects in JSDMs.

We describe a statistical model that we used to improve our understanding of montaine mammal communities in what follows. Critically, we consider potential confounding among interspecies relationships and environmental relationships in our model and demonstrate how to orthogonalize these mechanisms in the model and the resulting inference compared to models where species are treated independently. Unlike previous approaches that have applied restricted regression techniques similar to ours, we use it in the context of well-known ecological models for species occupancy and abundance. While such approaches have become well-known in spatial statistics and environmental science, they have not been adopted in settings involving the multivariate analysis of community data.

Joint Species Distribution Model

We present a JSDM based on a multispecies extension to the Royle-Nichols model[8]. The Royle-Nichols model accounts for heterogeneity in detection induced by the species' latent intensity, a surrogate related to true species abundance. Abundance estimation often requires an explicit spatial region that is closed to emmigration and immigration. Intensity is a measure of how frequently members of the species use a region and does not require a population closure assumption. In the *Model* section, we further discuss the distinctions between abundance and intensity in the Royle-Nichols model.

The Royle-Nichols model utilizes occupancy survey data but provides inference distinct from the basic occupancy model[9]. In the Royle-Nichols model, we estimate individual detection probability for homogeneous members of the population, whereas in an occupancy model, we estimate probability of observing at least one member of the population given that the site is occupied. Furthermore, the Royle-Nichols model allows us to relate habitat covariates to the latent intensity associated with a species at a site, while in an occupancy model, habitat covariates are associated with the species latent probability of occupancy at a site. Species intensity and occupancy may be governed by different mechanisms, and inference from an intensity model can be distinct from that provided by an occupancy model[10–12]. Cingolani et al.[11] proposed that, in plant communities, certain environmental filters preclude species from occupying a site and an additional set of filters may regulate if a species can flourish. Hence, certain covariates that were unimportant in an occupancy model may improve predictive power in an intensity model.

Community Confounding

Species distributions are shaped by habitat as well as competition and mutualism within the community[7,13,14]. Community confounding occurs when species distributions are explained by a convolution of habitat and interspecies effects and can lead to inferential discrepancies between a joint and single species distribution model. Former studies have incorporated interspecies dependence into an occupancy model[15–18], and others have addressed spatial confounding[1,19–21], but none of these considered community confounding.

We address community confounding by formulating a version of our model that orthogonalizes the habitat effects and species random effects. Orthogonalizing the fixed and random effects is common practice in spatial statistics and often referred to as restricted spatial regression[22–26]. Restricted regression has been applied to spatial generalized linear mixed models (SGLMM) for observations $\boldsymbol{y}$, which can be expressed as

$$\boldsymbol{y} \sim [\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\psi}], \tag{1}$$

$$g(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\eta}, \tag{2}$$

$$\boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{H}(\boldsymbol{\phi})), \tag{3}$$

where $g(\cdot)$ is a link function, $\boldsymbol{\psi}$ are additional parameters for the data model, and $\boldsymbol{\phi}$ parameterizes the random effect. In the SGLMM, prior information facilitates the estimation of $\boldsymbol{\eta}$, which would not be estimable otherwise due to its shared column space with $\boldsymbol{\beta}$[24]. This is analogous to applying a ridge penalty to $\boldsymbol{\eta}$, which stabilizes the likelihood. Another method for fitting the confounded SGLMM is to specify a restricted version:

$$\boldsymbol{y} \sim [\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\psi}], \tag{4}$$

$$g(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\theta} + (\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{\eta}, \tag{5}$$

$$\boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{H}(\boldsymbol{\phi})), \tag{6}$$

where $\boldsymbol{P}_X$ is the projection matrix onto the column space of $\boldsymbol{X}$. In the unrestricted SGLMM, variability in the latent mean $\boldsymbol{\mu}$ in the direction of $\boldsymbol{X}$ is explained by both the regression coefficients $\boldsymbol{\beta}$ in (2) and random effect $\boldsymbol{\eta}$. In the restricted model, however, all variability in the direction of $\boldsymbol{X}$ is explained solely by the regression coefficients $\boldsymbol{\theta}$ in (5)[25]. We refer to $\boldsymbol{\beta}$ as the conditional effects because they depend on $\boldsymbol{\eta}$, and $\boldsymbol{\theta}$ as the unconditional effects.

Restricted regression, as specified (4)-(6), was proposed by Reich et al.[22]. Reich et al.[22] described a disease-mapping example in which the inclusion of a spatial random effect rendered one covariate effect unimportant that was important in the non-spatial model. Spatial maps indicated an association between the covariate and response, making inference from the spatial model appear untenable. Reich et al.[22] proposed restricted spatial regression as a method for recovering the posterior expectations of the non-spatial model and shrinking the posterior variances which tend to be inflated for the unrestricted SGLMM.

Several modifications of restricted spatial regression have been proposed[24,27–30]. All restricted spatial regression methods seek to provide posterior means $\mathrm{E}(\theta_j|\boldsymbol{Y})$ and marginal posterior variances $\mathrm{Var}(\theta_j|\boldsymbol{Y})$, $j = 1, \ldots, p$ that satisfy the following two conditions[31]:

1. $\mathrm{E}(\boldsymbol{\theta}|\boldsymbol{Y}) = \mathrm{E}(\boldsymbol{\beta}_{\mathrm{NS}}|\boldsymbol{Y})$ and,
2. $\mathrm{Var}\left(\beta_{\mathrm{NS},j}|\boldsymbol{Y}\right) \leq \mathrm{Var}(\theta_j|\boldsymbol{Y}) \leq \mathrm{Var}\left(\beta_{\mathrm{Spatial},j}|\boldsymbol{Y}\right)$ for $j = 1, \ldots, p$,

where $\boldsymbol{\beta}_{NS}$ and $\boldsymbol{\beta}_{Spatial}$ are the regression coefficients corresponding to the non-spatial and unrestricted spatial models, respectively.

The inferential impacts of spatial confounding on the regression coefficients has been debated. Hodges and Reich[23] outlined five viewpoints on spatial confounding and restricted regression in the literature and refuted the two following views:

1. Adding the random effect $\boldsymbol{\eta}$ corrects for bias in $\boldsymbol{\beta}$ resulting from missing covariates.

2. Estimates of $\boldsymbol{\beta}$ in a SGLMM are shrunk by the random effect and hence conservative.

The random effect $\boldsymbol{\eta}$ can increase or decrease the magnitude of $\boldsymbol{\beta}$, and the change may be galvanized by mechanisms not related to missing covariates. Therefore, we cannot assume the regression coefficients in the SGLMM will exceed those of the restricted model, nor should we regard the estimates in either model as biased due to misspecification. Confounding in the SGLMM causes $\text{Var}(\beta_j|\boldsymbol{Y}) \geq \text{Var}(\theta_j|\boldsymbol{Y})$, $j = 1, ..., p$, because of the shared column space of the fixed and random effects. Thus, we refer to the conditional coefficients as conservative with regards to their credible intervals, not their posterior expectations.

Reich et al.[22] argued that restricted spatial regression should always be applied because the spatial random effect is generally added to improve predictions and/or correct the fixed effect variance estimate. While it may be inappropriate to orthogonalize a set of fixed effects in an ordinary linear model, orthogonalizing the fixed and random effect in a spatial model is permissible because the random effect is generally not of inferential interest. Paciorek[32] provided the alternative perspective that, if confounding exists, it is inappropriate to attribute all contested variability in $\boldsymbol{y}$ to the fixed effects. Hanks et al.[25] discussed factors for deciding between the unrestricted and restricted SGLMM on a continuous spatial support. The restricted SGLMM leads to improved computational efficiency, but the unconditional effects are less conservative under model misspecification and more prone to type-S errors: The Bayesian analogue of Type I error. Fitting the unrestricted SGLMM when the fixed and random effect are truly orthogonal does not introduce bias, but it will increase the fixed effect variance. Given these considerations, Hanks et al.[25] suggested a hybrid approach where the conditional effects, $\boldsymbol{\beta}$, are extracted from the resticted SGLMM. This is possible because the restricted SGLMM is a reparameterization of the unrestricted SGLMM. This hybrid approach leads to improved computational efficiency but yields the more conservative parameter estimates.

Restricted regression has also been applied in time series applications. Dominici et al.[33] debiased estimates of fixed effects confounded by time using restricted smoothing splines. Without the temporal random effect, Dominici et al.[33] asserted all temporal variation in the response would be wrongly attributed to temporally correlated fixed effects. Houseman et al.[34] used restricted regression to ensure identifiability of a nonparametric temporal effect and highlighted certain covariate effects that were more evident in the restricted model (i.e., the unconditional effects' magnitude was greater). Furthermore, restricted regression is implicit in restricted maximum likelihood estimation (REML). REML is often employed for debiasing the estimate of the variance of $\boldsymbol{Y}$ in linear regression and fitting linear mixed models that are not estimable in their unrestricted format[35]. Because REML is generally applied in the context of variance and covariance estimation, considerations regarding the effects of REML on inference for the fixed effects are lacking in the literature.

In ecological science, multispecies models often include random effects to account for interspecies dependence. Interspecies dependence in the random effects can be characterized by a covariance matrix. Unlike a spatial or temporal random effect, we consider species random effects to be inferentially important, rather than a tool for improving predictions or catch-all for missing covariates. A restriction approach in a multispecies model attributes contested variation between the fixed effects (habitat information) and random effect (community information) to the fixed effect.

We demonstrate restricted regression in multispecies models by comparing inference between a restricted and unrestricted

model for the camera trap data. Furthermore, we discuss community confounding and its relevance to restricted regression in the context of multispecies models. We present a method for measuring confounding and highlight its inferential utility. We also discuss other inferential methods for confounded models and consider their appropriateness in the multispecies context.

In what follows, we motivate and formulate a multispecies model. Referencing our multispecies model, we describe community confounding and propose methods to detect and alleviate it. We then apply our model to data collected from 2013-14 during a Colorado Parks and Wildlife camera trap survey. Finally, we discuss our findings and highlight their implications for inference in models that exhibit community confounding.

Model

Royle-Nichols Link Function

We specify a model for analyzing multispecies binary detection data, y_{ijk}, arising from a Bernoulli process with probability of success p_{ijk}, where $i = 1,\ldots,n$, $j = 1,\ldots,J_i$, and $k = 1,\ldots,K$ correspond to sites, occasions, and species, respectively. Occupancy data of this form have traditionally been analyzed in a latent variable framework[9,36,37]. Let $z_{ik} \sim \text{Bern}(\psi_{ik})$ be an indicator on whether species k occupies site i. Given the site is occupied, we detect species k on occasion j with some probability p_{ijk}, such that $(y_{ijk}|z_{ik} = 1) \sim \text{Bern}(p_{ijk})$, but if species k is absent from the site, we have zero probability of detecting it, $P(y_{ijk} = 0|z_{ik} = 0) = 1$.

Royle and Nichols[8] introduced a method for analyzing occupancy data that explicitly modeled the probability of detecting species k at a site as a function of a surrogate related to the true species abundance. Assuming there are N_{ik} individuals of species k in sample region i and that all individuals in species k on the sample unit have identical detection probabilities and are detected independently of other individuals, the probability of detecting at least one of these individuals can be expressed as

$$p_{ijk} = 1 - (1 - r_{jk})^{N_{ik}}, \tag{7}$$

where r_{jk} is a binomial sampling probability that a particular individual of species k is detected on occasion j. While the Royle-Nichols model facilitates inference on number of individuals of species k, N_{ik}, at each site when all the assumptions are met, we do not interpret them as such because sites are not necessarily closed in camera trap studies due to highly mobile species with home ranges larger than the sampling radius of the camera.

The nonlinear function of r_{jk} and N_{ik} in (7) involves more parameters than would be identifiable in a typical occupancy model, especially when the individual detection probability is heterogeneous across occasions (e.g., r_{jk} are heterogeneous). In the heterogeneous case, r_{jk} is connected to covariates with the logit link function:

$$\text{logit}(r_{jk}) = f(\boldsymbol{w}_{jk}, \boldsymbol{\alpha}_k), \tag{8}$$

where $f(\boldsymbol{w}_{jk}, \boldsymbol{\alpha}_k)$ is a linear function of the covariates $\boldsymbol{w}_{jk}$ and regression parameters $\boldsymbol{\alpha}_k$.

Modeling Interspecies Dependence

Following Royle and Nichols[8], we assume $N_{ik} \sim \text{Pois}(\lambda_{ik})$, where λ_{ik} is mean intensity of species k at site i. We let $\boldsymbol{\lambda}$ denote the vectorized intensities of the K species in the community stacked across all n sites. To model interspecies dependence, we specify the conditional multivariate normal distribution:

$$\log(\boldsymbol{\lambda}) \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\eta}, \boldsymbol{H}), \tag{9}$$

$$\boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{spp} \otimes \boldsymbol{I}_n), \tag{10}$$

where $\boldsymbol{X}$ is a block-diagonal matrix of the K species design matrices, $\boldsymbol{\beta} = (\boldsymbol{\beta}'_1, \ldots, \boldsymbol{\beta}'_K)'$ is a stacked vector of species specific regression coefficients, $\boldsymbol{\eta}$ represents the species random effects, and $\boldsymbol{\Sigma}_{spp}$ is a species covariance matrix, and $\boldsymbol{H}$ is a matrix that allows for additional covariance structures such as spatial dependence. For our purpose of comparing the single and joint species distribution models, we marginalized the model over the species random effects and set $\boldsymbol{H} = \boldsymbol{0}$. The resulting distribution for the species latent intensities was $\log(\boldsymbol{\lambda}) \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}_{spp} \otimes \boldsymbol{I}_n)$.

The formulation given in (9) allows for dependence between all K species in the community at each site. Previous joint occupancy models allowed for interspecies dependence in the probability of occupancy[15–17], whereas our model allows for dependence in the species latent intensities. Just as certain habitat features may not preclude species occupancy but can curb intensity, some species may coexist in a region but not be able to jointly flourish[38]. Hence, interspecies dependence that is observed in a joint intensity model may not be present in a joint occupancy model.

Scheffe[39] stipulated that the levels of a random effect are draws from a population, and the draws are not of interest in themselves but only as samples from the larger population, which is of interest. In more recent literature, the term random effect is used more broadly. Hodges and Clayton[40] categorized modern definitions of a random effect into three different varieties. The definition commonly used in spatial statistics is, the levels of the effect arise from a meaningful population, but they are the whole population and these particular levels are of interest. We adopt this definition for the species random effects in (9). In practice, some levels of the population will likely not be included in the species random effects. For example, in Ivan et al.[41], cameras were baited and arranged to capture all members of the mammalian community, but several species were excluded from the species random effects due to a lack of detections.

Priors

We used normal priors for the regression coefficients in both the intensity, $\boldsymbol{\beta}$, and detection, $\boldsymbol{\alpha}$, process and choose the conjugate Inverse-Wishart prior for the species covariance matrix $\boldsymbol{\Sigma}_{spp}$. A more general alternative to the Inverse-Wishart prior is to apply a Cholesky decomposition, $\boldsymbol{\Sigma}_{spp} = \boldsymbol{L}\boldsymbol{D}^{-1}\boldsymbol{L}'$, where $\boldsymbol{L}$ is lower diagonal with ones along the diagonal and $\boldsymbol{D}$ is diagonal with positive diagonal elements, and specify priors for the lower diagonal elements of $\boldsymbol{L}$ and diagonal elements of $\boldsymbol{D}$[42]. We found the Inverse-Wishart prior suitable for our inferential goals, but see Chan and Jeliazkov[42] for alternative covariance matrix priors.

The joint posterior distribution associated with our model is

$$[\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{N},\boldsymbol{\Sigma}_{spp}|\boldsymbol{Y}] \propto \prod_{k=1}^{K}\left(\prod_{i=1}^{n}\left(\prod_{j=1}^{J_i}\left([y_{ijk}|N_{ik},\boldsymbol{\alpha}_k]\right)[N_{ik}|\lambda_{ik}]\right)[\boldsymbol{\alpha}_k][\boldsymbol{\beta}_k]\right)[\boldsymbol{\lambda}|\boldsymbol{\beta}_1,\cdots,\boldsymbol{\beta}_K,\boldsymbol{\Sigma}_{spp}][\boldsymbol{\Sigma}_{spp}]. \tag{11}$$

See Appendix A for the full statement of our Royle-Nichols multispecies model.

Community Confounding

Restricted Regression Approach

We fit a restricted version of the Royle-Nichols multispecies model that orthogonalizes the fixed effects and random effect. We express the species latent intensity process conditionally as

$$\log(\boldsymbol{\lambda}) \sim \mathcal{N}(\boldsymbol{XB}+(\boldsymbol{I}-\boldsymbol{P}_X)\boldsymbol{\eta},\boldsymbol{H}), \tag{12}$$

$$\boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{0},\boldsymbol{\Sigma}_{spp}\otimes\boldsymbol{I}_n), \tag{13}$$

where $\boldsymbol{P}_X$ is the projection matrix onto the column space of $\boldsymbol{X}$. This specification forces the species covariance matrix to explain patterns in the community that are orthogonal to the fixed effects. As in (9), we marginalized over the species random effects but did not set $\boldsymbol{H}=\boldsymbol{0}$ since the resulting normal distribution would be degenerate because $(\boldsymbol{I}-\boldsymbol{P}_X)(\boldsymbol{\Sigma}_{spp}\otimes\boldsymbol{I}_n)(\boldsymbol{I}-\boldsymbol{P}_X)$ is not invertible. To remedy the singularity of the covariance matrix, we applied a ridge penalty by letting $\boldsymbol{H}=\tau^2\boldsymbol{I}$. The resulting distribution was $\log(\boldsymbol{\lambda}) \sim \mathcal{N}\left(\boldsymbol{XB},\tau^2\boldsymbol{I}+(\boldsymbol{I}-\boldsymbol{P}_X)(\boldsymbol{\Sigma}_{spp}\otimes\boldsymbol{I}_n)(\boldsymbol{I}-\boldsymbol{P}_X)\right)$. For implementation, we kept the hierarchical representation in equations (12) and (13) and obtained samples of $(\boldsymbol{I}-\boldsymbol{P}_X)\boldsymbol{\eta}$ in our Markov Chain Monte Carlo (MCMC) through conditioning by Kriging[25,43].

In the restricted model, we set $\tau^2=2.25$. This choice was supported by the asymptotic equivalence between Poisson and logistic regression. In a generalized occupancy model, the latent probability of occupancy is specified as $\text{logit}(\psi_i) \sim \mathcal{N}(\boldsymbol{x}_i'\boldsymbol{\beta},\tau^2)$.[44] investigated the relation between the prior on $\boldsymbol{\beta}$ and induced prior on the latent probability of success ψ_i in logistic regression; their work showed that specifying an uninformative normal prior on $\boldsymbol{\beta}$, i.e. setting τ^2 large, induces a U-shaped prior for ψ_i with most of the density concentrated near 0 and 1. Broms et al.[16] recommended setting $\tau^2=2.25$ in occupancy models, which results in a relatively flat prior for ψ. For rare species, λ_i in (12) is analogous to ψ_i, and specifying a variance of $\tau^2=2.25$ is minimally informative.

Baddeley[45] motivated the asymptotic equivalence of Poisson and logistic regression in a spatial context where counts of points from a non-homogeneous Poisson process are recorded in a lattice; they showed that, as the grid cells of the lattice become infinitesimally small, the inference yielded from Poisson and logistic regression are equivalent. This result can be applied more generally to any dataset where there is a high proportion of zero counts. We demonstrate the asymptotic equivalence between Poisson and logistic regression in the Royle-Nichols model in Appendix D.

Measuring Confounding

Hefley et al.[26] showed how to assess confounding in SGLMM models by computing the Pearson correlation coefficient between each pair of covariates and eigenvectors from the spectral decomposition of the spatial covariance matrix. We propose another approach relevant to our method that aids in interpretation. We compute the coefficient of determination of each covariate for species k regressed on the estimated latent intensities (no intercept) of the $K-1$ other species in the community. Because the latent intensities are unknown, the coefficents of determination of all covariates are derived quantities and can be computed at each iteration of the MCMC algorithm:

$$R^{2(l)}(\boldsymbol{x}_k) = \frac{SSR^{(l)}(\boldsymbol{x}_k)}{SST(\boldsymbol{x}_k)} = \frac{\left(\boldsymbol{\Lambda}^{(l)}_{-k}\hat{\boldsymbol{\theta}}^{(l)}_{-k} - \bar{\boldsymbol{x}}_k\right)'\left(\boldsymbol{\Lambda}^{(l)}_{-k}\hat{\boldsymbol{\theta}}^{(l)}_{-k} - \bar{\boldsymbol{x}}_k\right)}{(\boldsymbol{x}_k - \bar{\boldsymbol{x}}_k)'(\boldsymbol{x}_k - \bar{\boldsymbol{x}}_k)}, \tag{14}$$

where $\bar{\boldsymbol{x}}_k = (\bar{x}_k, \ldots, \bar{x}_k)$ is the mean of the covariate $\boldsymbol{x}_k$ for species k repeated n times, $\boldsymbol{\Lambda}^{(l)}_{-k} = \left(\boldsymbol{\lambda}^{(l)}_1, \ldots, \boldsymbol{\lambda}^{(l)}_{k-1}, \boldsymbol{\lambda}^{(l)}_{k+1}, \ldots, \boldsymbol{\lambda}^{(l)}_K\right)$ is a matrix of the $K-1$ other species intensities sampled for MCMC iteration l, and $\hat{\boldsymbol{\theta}}^{(l)}_{-k}$ are estimated regression coefficients relating the estimated species intensities at iteration l to $\boldsymbol{x}_k$. The posterior mean $\mathrm{E}\left(R^2(\boldsymbol{x}_k)|\boldsymbol{Y}\right)$ provides a measure of community confounding for the covariate $\boldsymbol{x}_k$ and can help identify which fixed effects will vary between the unrestricted and restricted models. We demonstrate this approach in the following case study.

Camera Trap Survey

Study Area

We analyzed data arising from a study area comprised of subalpine forests in the state of Colorado between 2590 and 3660 m elevation (Figure 1). Sites were restricted to public lands managed by the United States Forest Service, National Park Service, Bureau of Land Management, and Colorado State Forest Service. Forests in our study area were primarily composed of Lodgepole pine (*Pinus contorta*), Engelmann spruce (*Picea engelmannii*), and subalpine fir (*Abies lasiocarpa*). Lodgepole pine was dominant at lower elevations as well as higher elevations that were drier and/or on south-facing slopes; high elevation regions that had cool north-facing slopes were co-dominated by Engelmann spruce and subalpine fir. Lodgepole pine is restricted to the northern two-thirds of Colorado, so all sites in the southern region of the study area were Engelmann spruce, subalpine fir co-dominated. Quaking aspen (*Populus tremuloides*), Douglas-fir (*Pseudotsuga menziesii*), bristlecone pine (*Pinus aristata*), limber pine (*Pinus flexilis*), and blue spruce (*Picea pungens*) were also present at some sites. Mean July and January temperature across the study area were 14°C and -6.1°C respectively. All camera data were collected during summers 2013-2014.

Sampling Design

The primary goal of Ivan et al.[41] was to assess mammalian responses to bark beetle outbreaks, thus sites were randomly selected to facilitate inference on the beetle outbreak covariates. Beetle outbreak covariates included the number of years since the initial outbreak (YSO) and the severity of the outbreak measured by mean overstory mortality (severity). The sample of

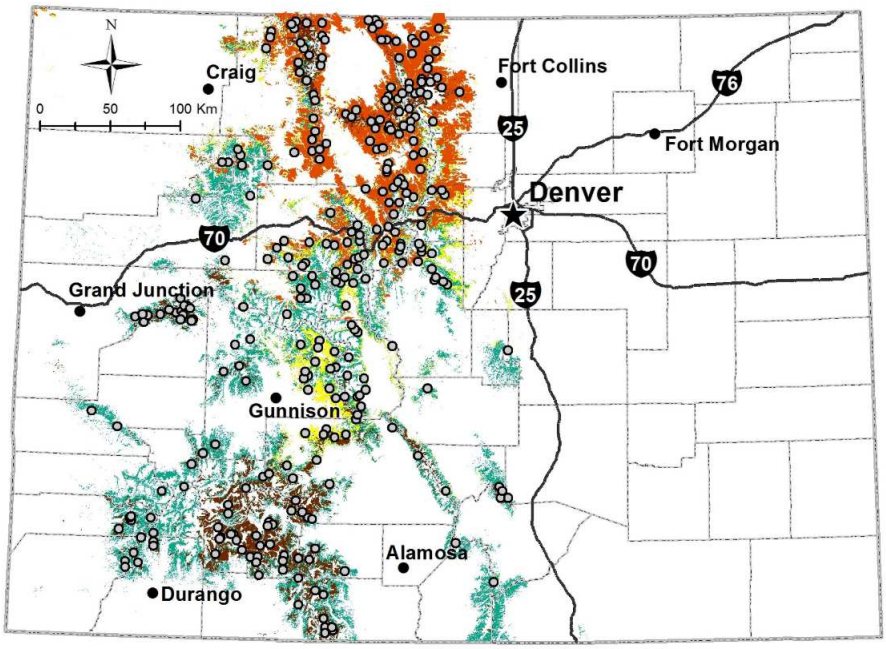

Figure 1. Randomly selected sampling sites (gray circles) where passive infrared game cameras were deployed in spruce-fir (green) and lodgepole pine (yellow) forests in Colorado, USA, 2013–2014. Brown and orange are the approximate extents of spruce beetle and mountain pine beetle impacts in spruce-fir and lodgepole pine forests, respectively, as of 2014. Reprinted from "Mammalian responses to changed forest conditions resulting from bark beetle outbreaks in the southern Rocky Mountains," by J. S. Ivan, 2018, Ecosphere, 9(8), e02369. Copyright [Year] by the Ecological Society of America. Reprinted with permission.

$n = 300$, 1 km^2 sites was evenly split across the two dominant forest types, spruce-fir and lodgepole pine. Additional habitat covariates were collected at each site, and a description of these is included in Appendix B.

Passive infrared camera traps (Reconyx PC800, Holmen, Wisconsin, USA) were deployed near the center of each site. Cameras were approximately 0.5 m above the ground and pointed toward a lure tree 4–5 m away[46]. The lure tree was baited with minimal amounts of peanut butter and commercial rabbit lure. This setup was designed to maximize detections of both large and small-bodied mammals in the local community while minimizing attraction of individuals from outside the sampling region of the site. The sampling regions were likely not closed to immigration/emigration; thus, we interpret elevated detections at a site as more individuals using, as opposed to occupying, that site[47]. For additional details regarding the sampling design and study area see Ivan et al.[41].

Model Fitting

We fit the Royle-Nichols joint species model to the multispecies camera trap data binned into 20 two-day occasions because simulations showed this was the number of replications needed to identify a quadratic effect of occasion on individual detection probability. Not all cameras were operational for the entire 40 day sampling period, and thus the number of occasions varied from 7-20. We discarded four sites at which the camera was operational for less than one occasion. We also discarded another 12 sites that had been infested by bark beetles for more than 10 years. Ivan et al.[41] truncated the bark beetle infestation covariate

at 10 years because estimates of response curves beyond 10 years would be unreliable with so few sites. The final sample size was $n = 284$ sites.

For consistency, our joint occupancy model included the 13 species for which Ivan et al.[41] performed a single species analysis; several rare species were excluded from analysis due to insufficient detections. We note, however, that these rare species parameters may be identifiable in the joint model as has been the case in previous studies[2,48–52]. Ivan et al.[41] used a sequential procedure similar to that described in Lebreton et al.[53] to select the covariates in the occupancy and detection processes for each species. We adopted their detection model and used the same covariates but a different set of basis functions for YSO. Ivan et al.[41] treated YSO as a grouping variable and considered probability of use response curves that allowed for cubic associations and delayed responses to bark beetle infestation. Multiple response curves were model averaged to produce predictive YSO response curves for each species. We used orthogonal polynomial basis functions for the YSO variable in the species intensity models. The basis functions included a linear (YSO1) and quadratic (YSO2) effect. Appendix E provides a full description of the intensity and detection models.

We fit the model using a MCMC algorithm. To improve mixing and predictive ability, we regularized the coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ with slightly informative priors: $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, 10\boldsymbol{I})$ and $\boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$[54]. We specified a vague prior of $\boldsymbol{\Sigma}^{-1} \sim \text{Wishart}(15, (15\boldsymbol{I})^{-1})$ for the species variance-covariance matrix[55]. We used Gibbs sampling based on conjugate priors for parameters $\boldsymbol{\Sigma}_{spp}$ and $\boldsymbol{\beta}$ and Metropolis-Hastings updates for $\boldsymbol{N}$, $\boldsymbol{\lambda}$, and $\boldsymbol{\alpha}$. Derivations of the conjugate full-conditional distributions are provided in Appendix C with details about the Metropolis-Hastings updates. We tuned the Metropolis-Hastings updates so that acceptance rates varied between 20-40% for $\boldsymbol{\alpha}$, $\boldsymbol{N}$, and $\boldsymbol{\lambda}$. All continuous covariates were scaled to have mean 0 and variance 1. We ran the MCMC algorithm for $L = 20000$ iterations, and the first 5000 iterations were discarded as burn-in. We fit both the unrestricted and restricted models expressed in (9) and (12), respectively. The *Results* section presents inference for regression coefficients from the unrestricted joint species distribution model and contrasts it with that from Ivan et al.[41] single species distribution models to highlight inferential discrepancies caused by community confounding. Inference from the restricted joint species distribution model is omitted because it was similar in sign and significance to Ivan et al.[41] single species distribution models but differed in effect size because habitat covariates model intensity rather than occupancy.

Code Availability

All algorithms and code for fitting the unrestricted and restricted joint species distribution models are available in the Supplementary Information files. All MCMC algorithms and analyses were coded in R 4.0.3.

Results

Ivan et al.[41] fit single species distribution models to infer changes in mammalian use of stands impacted by the bark beetle epidemic. The impact of bark beetle damage was measured by years since initial infestation (YSO) and severity of outbreak quantified by mean overstory mortality (DeafConif). Moose, elk, and mule deer exhibited positive associations with bark beetle activity; red squirrels, golden-mantled ground squirrels, chipmunks, and coyotes exhibited negative associations; American

martens, black bears, snowshoe hares, and porcupines showed no associations; and red foxes and yellow-bellied marmots showed mixed associations. Inference from the joint model, Figures 2-4, was largely consistent with that of the single species models. All three ungulates species had a positive association with severity but the YSO effects Ivan et al.[41] noted for mule deer and moose were not observed (Figure 2). Both red and golden-mantled ground squirrels had negative associations with severity. Furthermore, use of subalpine stands decreased as a function YSO for red squirrels and coyotes. We did not, however, observe any significant associations between YSO and stand use for golden-mantled ground squirrels and chipmunks, contrary to the inference from the single species models. Consistent with inference from the single species models, red foxes had a negative association with severity but positive association with YSO1, whereas yellow-bellied marmots had a positive association with severity but negative association with YSO1.

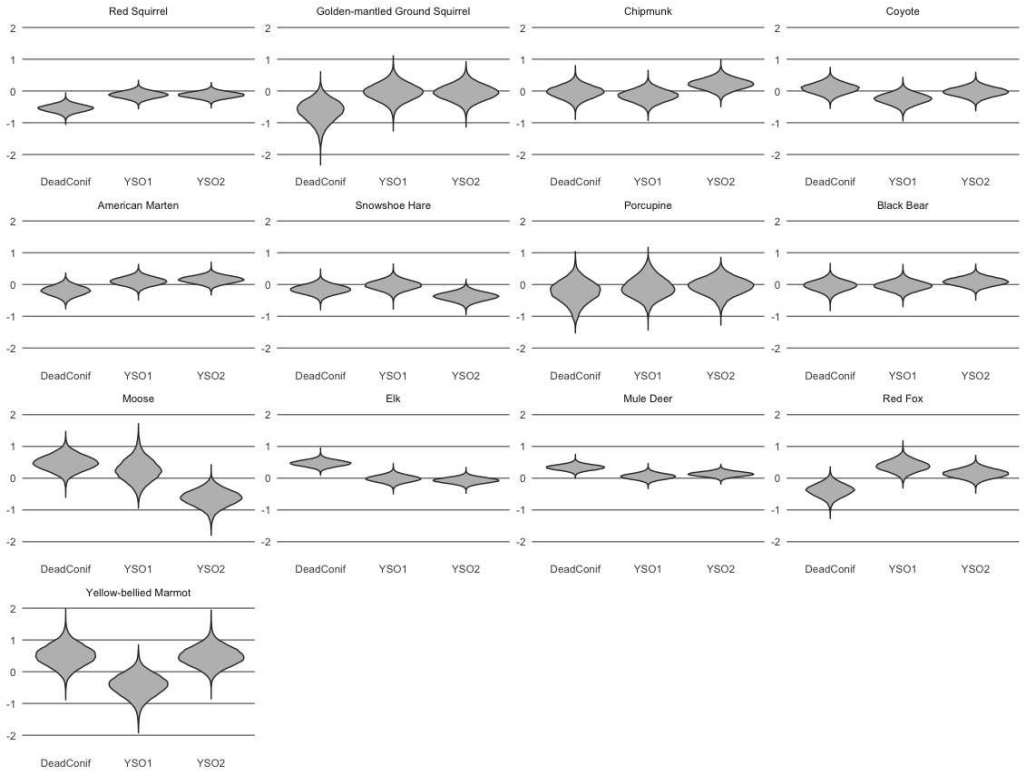

Figure 2. Violin plot for marginal posterior histograms of infestation regression parameters. Estimates are from the unrestricted joint species distribution model. DeadConif is the overstory mortality percentage, a proxy for severity of bark beetle infestation. YSO1 is the linear effect of the number of years since a site was infested with bark beetles. YSO2 is the quadratic effect.

The inferential discrepancies between the joint and single species models are indicative of community confounding. Covariates that were important in the single species model no longer indicate an effect when accounting for the community. For example, the posterior distribution of the effect pertaining to the indicator of whether the site was in a federally designated wilderness (WILD) overlapped zero in the joint species distribution model for Yellow-bellied Marmots (Figure 3). Ivan et al.[41] single species distribution model, however, identified the covariate as helpful quantity for predicting the distribution of the Yellow-bellied Marmot. The covariate could have been important in the single species model as a surrogate measure of other species that the American marten avoids or pursues. After we accounted for the interspecies dependence in the joint model (Figure 4), the WILD variable was no longer helpful as a predictor, although the contribution to that signal on species intensity was still accounted for.

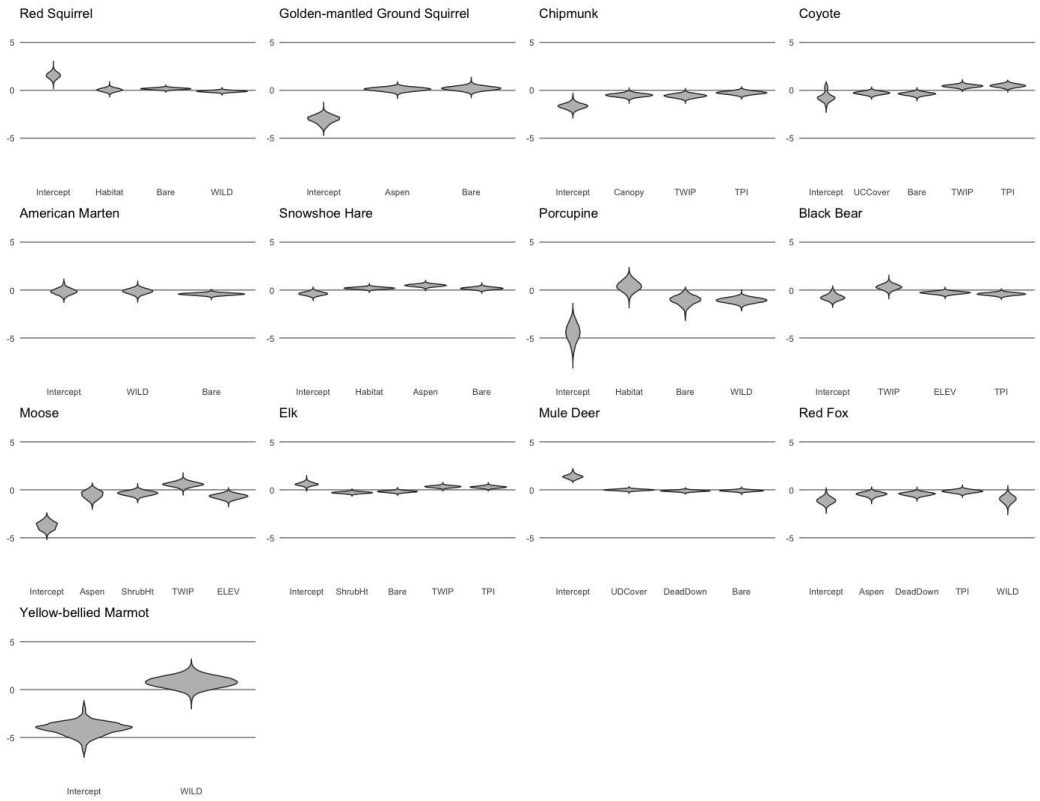

Figure 3. Violin plot for marginal posterior histograms of base habitat regression parameters. Estimates are from the unrestricted joint species distribution model. Apprendix B provides a description of habitat covariates.

The restricted regression approach we described in the *Community Confounding* section estimates the unconditional habitat effects, which would be similar to those obtained from a single species model. Irrespective of whether inference on conditional or unconditional effects is desired, it may be helpful to compute measures of community confounding to understand which fixed effects will vary between an unrestricted and restricted model. For example, severity (DeadConif) was the most confounded covariate for yellow-bellied marmots (Table 1). Inference on the effect of this covariate varied across the two approaches; the posterior distribution of the regression coefficient did not overlap zero in the unrestricted model 0.53 (0.03, 1.03) but did contain zero for the restricted model 0.33 (-0.14, 0.78) (90% credible intervals).

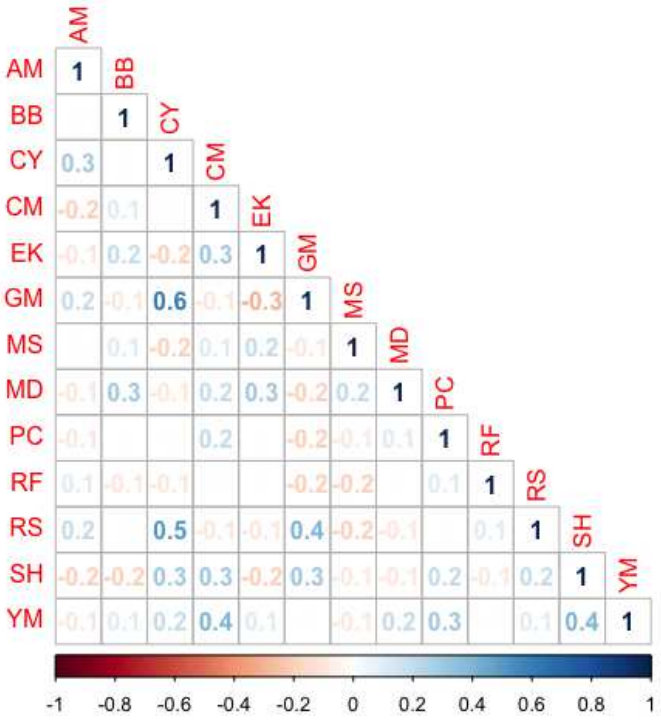

Figure 4. Posterior mean of species correlation matrix. Estimates are from the unrestricted joint species distribution model. AM = American Marten, BB = Black Bear, CY = Coyote, CM = Chipmunk spp., Ek = Elk, GM = Golden-mantled Ground Squirrel, MS = Moose, MD = Mule Deer, PC = Porcupine, RF = Red Fox, RS = Red Squirrel, SH = Snowshoe Hare, YM = Yellow-bellied Marmot.

Table 1. Posterior means of coefficients of determination for each species. Only maximum of posterior means shown for each species.

Species	Covariate	$\mathrm{E}\left(R^2|\boldsymbol{Y}\right)$
Black Bear	TWIP	0.62
American Marten	DeadConif	0.54
Coyote	DeadConif	0.53
Snowshoe Hare	DeadConif	0.53
Red Fox	DeadConif	0.52
Elk	DeadConif	0.51
Moose	DeadConif	0.51
Porcupine	DeadConif	0.51
Yellow-bellied Marmot	DeadConif	0.51
Golden-mantled Ground Squirrel	DeadConif	0.49
Red Squirrel	Bare	0.47
Chipmunk	DeadConif	0.46
Mule Deer	YSO1	0.44

Discussion

Species distributions are shaped by habitat as well as competition and mutualism within the community, and these effects are likely confounded. Because habitat and interspecies effects operate simultaneously and can be equally influential, restricted regression that gives priority to the habitat effects may not be appropriate always. Alternative approaches for adjusting the estimates of confounded effects have their own caveats. Consider a simpler case where the latent intensities, $\boldsymbol{\lambda}$, of the K species in our community were known. We could construct K regression models for predicting each species intensity as follows:

$$\boldsymbol{\lambda}_k = \boldsymbol{X}_k\boldsymbol{\beta}_k + \boldsymbol{\Lambda}_{-k}\boldsymbol{\eta}_k + \boldsymbol{\varepsilon}, \tag{15}$$

where $\boldsymbol{\Lambda}_{-k} = (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_{k-1}, \boldsymbol{\lambda}_{k+1}, \ldots, \boldsymbol{\lambda}_K)$ is a matrix of the $K-1$ other species intensities. If $\boldsymbol{X}_k$ and $\boldsymbol{\Lambda}_{-k}$ were highly collinear, principal component analysis (PCA) might be applied. PCA decomposes the variation explained by $\boldsymbol{X}_k$ and $\boldsymbol{\Lambda}_{-k}$ into $p = p_1 + p_2$ principal components, $\boldsymbol{\Gamma}_k = \left(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_p\right)$, where p_1 and p_2 are the number of columns of $\boldsymbol{X}_k$ and $\boldsymbol{\Lambda}_{-k}$ respectively. The regression model

$$\boldsymbol{\lambda}_k = \boldsymbol{W}_k\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \tag{16}$$

$$\boldsymbol{W}_k = (\boldsymbol{X}_k, \boldsymbol{\Lambda}_{-k})\,\boldsymbol{\Gamma}_k, \tag{17}$$

improves stability and can recover the posterior means and variances of $\boldsymbol{\beta}_k$ and $\boldsymbol{\eta}_k$ in (15). In practice, inference on $\boldsymbol{\beta}_k$ and $\boldsymbol{\eta}_k$ is often adjusted by truncating off the last $p-r$, for $r < p$, eigenvectors of $\boldsymbol{\Gamma}_k$ and employing the new design matrix

$$\boldsymbol{W}_k^\star = (\boldsymbol{X}_k, \boldsymbol{\Lambda}_{-k})\,\boldsymbol{\Gamma}_k^\star, \tag{18}$$

$$\boldsymbol{\Gamma}_k^\star = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_r)\,. \tag{19}$$

By retaining only the first r principal components, the smallest sources of variation are ignored in the estimation of $\boldsymbol{\beta}_k$ and $\boldsymbol{\eta}_k$. Jeffers[56] implemented this approach truncating off the last 7 of 13 principal components to adjust the estimates of regression coefficients relating various tree characteristics to maximum compressive strength. Other studies have selected a subset of principal components based on their strength of association with the response variable[57–60]. In some cases, the coefficient estimates from these reduced rank approaches appeared more tenable than those from the full rank specifications based on known physical relationships between the predictors and response. Thus, PCA regression can offer both computational and inferential improvements. PCA regression, however, does not explicitly address confounding, and interpretation of the adjusted regression coefficients is unclear.

Another dubious approach for adjusting parameter estimates in a confounded model is model averaging[61,62]. Consider an ensemble of S models with posteriors $[\boldsymbol{\theta}|\boldsymbol{y},\mathscr{M}_1],\ldots,[\boldsymbol{\theta}|\boldsymbol{y},\mathscr{M}_S]$ and a particular parameter θ_l present in all S models. The marginal posterior distribution of θ_l across all models is given by

$$[\theta_l|\boldsymbol{y}] = \sum_{s=1}^{S} [\theta_l|\boldsymbol{y},\mathscr{M}_s] P(\mathscr{M}_s|\boldsymbol{y}), \tag{20}$$

where $P(\mathscr{M}_s|\boldsymbol{y})$ is the posterior model probability of model s. One pitfall of model averaging is that the contribution of the posterior distribution of θ_l in model $\mathscr{M}_s$ is weighted by the posterior model probability $P(\mathscr{M}_s|\boldsymbol{y})$, which is not necessarily indicative of the parameter's importance in the model[63]. Furthermore, the interpretation of the parameter θ_l often depends on which of the other $\boldsymbol{\theta}_{-l}$ parameters are included in the model[64]. Some models in the ensemble may include parameters confounded with θ_l while others do not. Averaging θ_l across the ensemble makes no explicit adjustments for confounding and complicates inference[63–65]. Restricted regression is appealing in that, unlike PCA regression and model averaging, confounding is addressed by attributing all contested sources of variation to the fixed effects.

Recently, concerns regarding the coverage properties of the fixed effects estimator under restricted regression have been expressed[31,66]. For example, Zimmerman and Ver Hoef[66] showed that applying any restricted regression method to a SGLMM leads to frequentest coverage of the fixed effects that is lower than the corresponding non-spatial model. Similarly, Khan and Calder[31] found that when fitting a restricted version of the SGLMM with an intrinsic conditional autoregressive prior, credible intervals of the fixed effects from the restricted model were generally nested inside those yielded by the non-spatial model. Given these results, both Zimmerman and Ver Hoef[66] and Khan and Calder[31] recommended reverting to inference from the non-spatial model, rather than that of the restricted SGLMM, when inference from the unrestricted SGLMM appears untenable. This practice may not be appealing for multispecies models in which removing the species random effects would prevent inference on interspecies dependence. Because the random effect $\boldsymbol{\eta}$ is rarely of interest in spatial applications, there has been little investigation on the inferential impacts of restricted regression on $\boldsymbol{\eta}$. Such investigation, however, may be helpful in determining the appropriateness of restricted regression for multispecies models.

In summary, we specified a joint species distribution model that accounts for interspecies dependence at the intensity level. We examined how inference from the joint model differed from that of the single species models in Ivan et al.[41]. Confounding in

multispecies models is unique from spatial and time series applications in that the random effect is almost always of inferential interest, and hence, adjustments to the regression coefficients, $\boldsymbol{\beta}$, and random effects, $\boldsymbol{\eta}$, should both be considered. Using unconditional effects may not be appropriate in all settings, but alternative methods for adjusting parameter estimates, like PCA regression and model averaging, are also not universally applicable.

There may not exist a general remedial method for handling confounding in multispecies models, but confounding measures should be investigated because they can provide new insights into ecological systems. For example, we discovered that the severity of bark beetle outbreak was heavily confounded with the species random effects (Table 1). This suggests that effects of bark beetle outbreaks are complex with components potentially related to both the mammalian community and habitat. Consequently, restoration of the subalpine forest may not coincide with recovery of the mammalian community. Changes in mammalian use of subalpine stands impacted by bark beetles may be temporary reactions spurred by changes to the intensities of competitors, predators, and mutualists. Thus, management practices aimed at restoring previous habitat conditions might not immediately trigger recovery of species negatively impacted by the outbreak. Understanding the implications of any particular management strategy rely on our ability to disentangle habitat and community effects.

Experimental methods and modeling techniques for alleviating confounding have been proposed in ecology. Hefley et al.[67] showed that replicate populations can help disentangle confounded fixed and random effects. In the context of multispecies models, replication involves analyzing several communities simultaneously, which is often infeasible. Hefley et al.[67] also recommended explicit population models rather than phenomenological regression-based models for analysis of temporally confounded count data. Similarly, Fieberg et al.[65] advocated for mechanistic models guided by causal diagrams for analyzing temporally confounded animal movement data. An avenue of future research for multispecies modeling is to compare inference from phenomenological regression-based models, such as the one proposed here, with that of models that explicitly include ecological mechanisms such as competitive exclusion, mutualism, and predation. Because community and temporal confounding have the same mathematical framework, mechanistic models are a promising solution for confounded multispecies data.

References

1. Altwegg, R. & Nichols, J. D. Occupancy models for citizen-science data. *Methods Ecol. Evol.* **10**, 8–21 (2019).

2. Hui, F., Warton, D., Foster, S. & Dunstan, P. To mix or not to mix: Comparing the predictive performance of mixture models vs. separate species distribution models. *Ecology* **94**, 1913–9, DOI: 10.1890/12-1322.1 (2013).

3. Warton, D. *et al.* So many variables: Joint modeling in community ecology. *Trends Ecol. & Evol.* **30**, 766–779, DOI: 10.1016/j.tree.2015.09.007 (2015).

4. Tobler, M. W. *et al.* Joint species distribution models with species correlations and imperfect detection. *Ecology* **100**, e02754, DOI: https://doi.org/10.1002/ecy.2754 (2019). https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1002/ecy.2754.

5. Estevo, C. A., Nagy-Reis, M. B. & Nichols, J. D. When habitat matters: Habitat preferences can modulate co-occurrence patterns of similar sympatric species. *PloS one* **12**, e0179489 (2017).

6. Steen, D. A. *et al.* Snake co-occurrence patterns are best explained by habitat and hypothesized effects of interspecific interactions. *J. Animal Ecol.* **83**, 286–295 (2014).

7. Wisz, M. S. *et al.* The role of biotic interactions in shaping distributions and realised assemblages of species: Implications for species distribution modelling. *Biol. Rev.* **88**, 15–30 (2013).

8. Royle, J. & Nichols, J. Estimating abundance from repeated presence-absence data or point counts. *Ecology* **84**, 777–790 (2003).

9. MacKenzie, D. I. *et al.* Estimating site occupancy rates when detection probabilities are less than one. *Ecology* **83**, 2248–2255 (2002).

10. Orrock, J. L., Pagels, J. F., McShea, W. J. & Harper, E. K. Predicting presence and abundance of a small mammal species: The effect of scale and resolution. *Ecol. Appl.* **10**, 1356–1366 (2000).

11. Cingolani, A. M., Cabido, M., Gurvich, D. E., Renison, D. & Díaz, S. Filtering processes in the assembly of plant communities: Are species presence and abundance driven by the same traits? *J. Veg. Sci.* **18**, 911–920 (2007).

12. Dibner, R. R., Doak, D. F. & Murphy, M. Discrepancies in occupancy and abundance approaches to identifying and protecting habitat for an at-risk species. *Ecol. Evol.* **7**, 5692–5702, DOI: https://doi.org/10.1002/ece3.3131 (2017). https://onlinelibrary.wiley.com/doi/pdf/10.1002/ece3.3131.

13. Bascompte, J. Mutualistic networks. *Front. Ecol. Environ.* **7**, 429–436 (2009).

14. Van Dam, N. How plants cope with biotic interactions. *Plant Biol.* **11**, 1–5 (2009).

15. Tobler, M. W., Zúñiga Hartley, A., Carrillo-Percastegui, S. E. & Powell, G. V. N. Spatiotemporal hierarchical modelling of species richness and occupancy using camera trap data. *J. Appl. Ecol.* **52**, 413–421, DOI: 10.1111/1365-2664.12399 (2015). https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/1365-2664.12399.

16. Broms, K. M., Hooten, M. B. & Fitzpatrick, R. M. Model selection and assessment for multi-species occupancy models. *Ecology* **97**, 1759–1770, DOI: 10.1890/15-1471.1 (2016). https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1890/15-1471.1.

17. Rota, C. T. *et al.* A multispecies occupancy model for two or more interacting species. *Methods Ecol. Evol.* **7**, 1164–1173, DOI: 10.1111/2041-210X.12587 (2016). https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.12587.

18. Maphisa, D. H., Smit-Robinson, H. & Altwegg, R. Dynamic multi-species occupancy models reveal individualistic habitat preferences in a high-altitude grassland bird community. *PeerJ* **7**, e6276 (2019).

19. Clark, A. E. & Altwegg, R. Efficient Bayesian analysis of occupancy models with logit link functions. *Ecol. Evol.* **9**, 756–768 (2019).

20. Broms, K. M., Johnson, D. S., Altwegg, R. & Conquest, L. L. Spatial occupancy models applied to atlas data show Southern Ground Hornbills strongly depend on protected areas. *Ecol. Appl.* **24**, 363–374 (2014).

21. Johnson, D. S., Conn, P. B., Hooten, M. B., Ray, J. C. & Pond, B. A. Spatial occupancy models for large data sets. *Ecology* **94**, 801–808 (2013).

22. Reich, B. J., Hodges, J. S. & Zadnik, V. Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics* **62**, 1197–1206 (2006).

23. Hodges, J. S. & Reich, B. J. Adding spatially-correlated errors can mess up the fixed effect you love. *The Am. Stat.* **64**, 325–334, DOI: 10.1198/tast.2010.10052 (2010). https://doi.org/10.1198/tast.2010.10052.

24. Hughes, J. & Haran, M. Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *J. Royal Stat. Soc. Ser. B (Statistical Methodol.* **75**, 139–159, DOI: https://doi.org/10.1111/j.1467-9868.2012.01041.x (2013).

25. Hanks, E. M., Schliep, E. M., Hooten, M. B. & Hoeting, J. A. Restricted spatial regression in practice: geostatistical models, confounding, and robustness under model misspecification. *Environmetrics* **26**, 243–254, DOI: 10.1002/env.2331 (2015). https://onlinelibrary.wiley.com/doi/pdf/10.1002/env.2331.

26. Hefley, T. J., Hooten, M. B., Hanks, E. M., Russell, R. E. & Walsh, D. P. The bayesian group lasso for confounded spatial data. *J. Agric. Biol. Environ. Stat.* **22**, 42–59 (2017).

27. Bradley, J. R., Holan, S. H., Wikle, C. K. *et al.* Multivariate spatio-temporal models for high-dimensional areal data with application to longitudinal employer-household dynamics. *Annals Appl. Stat.* **9**, 1761–1791 (2015).

28. Murakami, D. & Griffith, D. A. Random effects specifications in eigenvector spatial filtering: A simulation study. *J. Geogr. Syst.* **17**, 311–331 (2015).

29. Thaden, H. & Kneib, T. Structural equation models for dealing with spatial confounding. *The Am. Stat.* **72**, 239–252 (2018).

30. Prates, M. O., Assunção, R. M., Rodrigues, E. C. *et al.* Alleviating spatial confounding for areal data problems by displacing the geographical centroids. *Bayesian Analysis* **14**, 623–647 (2019).

31. Khan, K. & Calder, C. A. Restricted spatial regression methods: Implications for inference. *J. Am. Stat. Assoc.* 1–13 (2020).

32. Paciorek, C. The importance of scale for spatial-confounding bias and precision of spatial regression estimators. *Stat. Sci. A review journal Inst. Math. Stat.* **25**, 107–125, DOI: 10.1214/10-STS326 (2010).

33. Dominici, F., McDermott, A. & Hastie, T. J. Improved semiparametric time series models of air pollution and mortality. *J. Am. Stat. Assoc.* **99**, 938–948 (2004).

34. Houseman, E. A., Coull, B. A. & Shine, J. P. A nonstationary negative binomial time series with time-dependent covariates: Enterococcus counts in Boston Harbor. *J. Am. Stat. Assoc.* **101**, 1365–1376 (2006).

35. Corbeil, R. R. & Searle, S. R. Restricted maximum likelihood (reml) estimation of variance components in the mixed model. *Technometrics* **18**, 31–38 (1976).

36. Hoeting, J. A., Leecaster, M. & Bowden, D. An improved model for spatially correlated binary responses. *J. Agric. Biol. Environ. Stat.* **5**, 102–114 (2000).

37. Tyre, A. J. *et al.* Improving precision and reducing bias in biological surveys: Estimating false-negative error rates. *Ecol. Appl.* **13**, 1790–1801, DOI: https://doi.org/10.1890/02-5078 (2003). https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1890/02-5078.

38. Clark, J. S. *et al.* High-dimensional coexistence based on individual variation: a synthesis of evidence. *Ecol. Monogr.* **80**, 569–608 (2010).

39. Scheffe, H. *The Analysis of Variance*, vol. 72 (John Wiley & Sons, 1959).

40. Hodges, J. S. & Clayton, M. K. Random effects old and new. *Stat. Sci.* (2011).

41. Ivan, J., Seglund, A., Truex, R. & Newkirk, E. Mammalian responses to changed forest conditions resulting from bark beetle outbreaks in the southern Rocky Mountains. *Ecosphere* **9**, DOI: 10.1002/ecs2.2369 (2018).

42. Chan, J. C.-C. & Jeliazkov, I. MCMC estimation of restricted covariance matrices. *J. Comput. Graph. Stat.* **18**, 457–480 (2009).

43. Rue, H. & Held, L. *Gaussian Markov Random Fields: Theory and Applications* (CRC press, 2005).

44. Hanson, T. E., Branscum, A. J., Johnson, W. O. *et al.* Informative *g*-priors for logistic regression. *Bayesian Analysis* **9**, 597–612 (2014).

45. Baddeley, A. *et al.* Spatial logistic regression and change-of-support in poisson point processes. *Electron. J. Stat.* **4**, 1151–1201, DOI: 10.1214/10-EJS581 (2010).

46. Blecha, K. A. Risk-reward tradeoffs in the foraging strategy of cougar (puma concolor): prey distribution, anthropogenic development, and patch selection. *Thesis. Colo. State Univ. Fort Collins, Color. USA* (2015).

47. MacKenzie, D. I. *et al. Occupancy Estimation and Modeling: Inferring Patterns and Dynamics of Species Occurrence* (Academic Press, 2006).

48. Guisan, A., Weiss, S. & Weiss, A. GLM versus CCA spatial modeling of plant species distribution. *Plant Ecol.* **143**, 107–122, DOI: 10.1023/A:1009841519580 (1999).

49. Ovaskainen, O. & Soininen, J. Making more out of sparse data: hierarchical modeling of species communities. *Ecology* **92**, 289–295 (2011).

50. Madon, B., Warton, D. I. & Araújo, M. B. Community-level vs species-specific approaches to model selection. *Ecography* **36**, 1291–1298, DOI: 10.1111/j.1600-0587.2013.00127.x (2013).

51. Ovaskainen, O., Abrego, N., Halme, P. & Dunson, D. Using latent variable models to identify large networks of species-to-species associations at different spatial scales. *Methods Ecol. Evol.* **7**, 549–555, DOI: 10.1111/2041-210X.12501 (2015).

52. Tikhonov, G., Abrego, N., Dunson, D. & Ovaskainen, O. Using joint species distribution models for evaluating how species-to-species associations depend on the environmental context. *Methods Ecol. Evol.* **8**, 443–452, DOI: 10.1111/2041-210X.12723 (2017).

53. Lebreton, J.-D., Burnham, K. P., Clobert, J. & Anderson, D. R. Modeling survival and testing biological hypotheses using marked animals: A unified approach with case studies. *Ecol. Monogr.* **62**, 67–118, DOI: 10.2307/2937171 (1992).

54. Hooten, M. B. & Hobbs, N. T. A guide to Bayesian model selection for ecologists. *Ecol. Monogr.* **85**, 3–28, DOI: https://doi.org/10.1890/14-0661.1 (2015). https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1890/14-0661.1.

55. Chung, Y., Gelman, A., Rabe-Hesketh, S., Liu, J. & Dorie, V. Weakly informative prior for point estimation of covariance matrices in hierarchical models. *J. Educ. Behav. Stat.* **40**, 136–157, DOI: 10.3102/1076998615570945 (2015). https://doi.org/10.3102/1076998615570945.

56. Jeffers, J. N. Two case studies in the application of principal component analysis. *J. Royal Stat. Soc. Ser. C (Applied Stat.* **16**, 225–236 (1967).

57. Hill, C. R., Fomby, T. B. & Johnson, S. R. Component selection norms for principal components regression. *Commun. Stat. Methods* **6**, 309–334 (1977).

58. Kung, E. C. & Sharif, T. A. Regression forecasting of the onset of the indian summer monsoon with antecedent upper air conditions. *J. Appl. Meteorol. Climatol.* **19**, 370–380 (1980).

59. Smith, G. & Campbell, F. A critique of some ridge regression methods. *J. Am. Stat. Assoc.* **75**, 74–81 (1980).

60. Jolliffe, I. T. A note on the use of principal components in regression. *J. Royal Stat. Soc. Ser. C (Applied Stat.* **31**, 300–303 (1982).

61. Roberts, H. V. Probabilistic prediction. *J. Am. Stat. Assoc.* **60**, 50–62, DOI: 10.1080/01621459.1965.10480774 (1965).

62. Leamer, E. E. *Specification Searches: Ad Hoc Inference with Nonexperimental Data*, vol. 53 (Wiley New York, 1978).

63. Cade, B. S. Model averaging and muddled multimodel inferences. *Ecology* **96**, 2370–2382 (2015).

64. Banner, K. M. & Higgs, M. D. Considerations for assessing model averaging of regression coefficients. *Ecol. Appl.* **27**, 78–93 (2017).

65. Fieberg, J., Ditmer, M. & Freckleton, R. Understanding the causes and consequences of animal movement: A cautionary note on fitting and interpreting regression models with time-dependent covariates. *Methods Ecol. Evol.* **3**, DOI: 10.1111/j.2041-210X.2012.00239.x (2012).

66. Zimmerman, D. L. & Ver Hoef, J. M. On deconfounding spatial confounding in linear models. *The Am. Stat.* 1–9 (2021).

67. Hefley, T. J., Hooten, M. B., Drake, J. M., Russell, R. E. & Walsh, D. P. When can the cause of a population decline be determined? *Ecol. Lett.* **19**, 1353–1362 (2016).

Acknowledgements

This research was funded by Colorado Parks and Wildlife.

Appendix

All appendices referenced in the manuscript are in the Supplementary Information files.

Author Contributions

J.V. and M.H. wrote the manuscript and designed the model. J.I. acquired the data. All authors reviewed the manuscript.

Data Availability

The data are available in the Supplementary Information files.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Appendix.pdf