

A Non-stationary Spatial Generalized Linear Mixed Model Approach for Studying Plant Diversity

Anandamayee Majumdar,^{1,*} Corinna Gries² and Jason Walker³

¹ Department of Mathematics and Statistics, Arizona State University,
Tempe, AZ 85287, U.S.A.

² Global Institute of Sustainability, Arizona State University, Tempe, AZ
85287, U.S.A.

³ School of Life Sciences, Arizona State University, Tempe, AZ 85287, U.S.A

15 August, 2007

SUMMARY. We analyze the multivariate spatial distribution of plant species diversity, distributed across three ecologically distinct land uses—urban residential, urban non-residential, and desert. We model these data using a Spatial Generalized Linear Mixed Model (SGLMM). Here plant species counts are assumed to be correlated within and among the spatial locations. We implement this model across the Phoenix metropolis and surrounding desert. Using a Bayesian approach, we utilized the Langevin-Hastings Hybrid algorithm. Under a generalization of a spatial log Gaussian Cox model, the log-intensities of the species count processes follow Gaussian distributions. The purely spatial component corresponding to these log-intensities are jointly modeled using a cross-convolution approach, in order to depict a valid cross-

* *email:* ananda@math.asu.edu

correlation structure. We observe that this approach yields non-stationarity of the model ensuing from different land use types. We obtain predictions of various measures of plant diversity including plant richness and Shannon-Weiner diversity at observed locations. We also obtain a prediction framework for plant preferences in urban and desert plots.

KEY WORDS: cross convolution; cross covariance matrix; generalized linear mixed model; Langevin Hastings algorithm; log Gaussian Cox Model; Markov Chain Monte Carlo; Multivariate Spatial Model.

1. Introduction

In a fast developing landscape as in Phoenix, Arizona, urban ecologists aim to find associations of plant diversity with other factors like income per capita, land use, historic agricultural land use, and elevation (Hope, Gries, Zhu, Fagan, Redman, Grimm, Nelson, Martin and Kinzig, 2003). It is important to map various measures of plant diversity – in our case plant species richness, the Shannon-Weiner diversity metric and individual species intensities, and in order to understand the spatial pattern and the spatial variability of these diversity measures. As part of the Central Arizona-Phoenix Longterm Ecological Research (CAP LTER) project we conducted in the year 2005 a large scale, intensive survey of 204 plots. One objective was to compare plant diversity in the urban area with the surrounding open desert. Another aim was to investigate whether observations on socio-economic factors such as income per capita, historical agricultural use of land, land use at the present time and geophysical factors like elevation could be used for prediction of various plant diversity measures as well as make statistical inference about plant preferences at a specific desert/urban location.

The statistical objectives related to the project mentioned above are the following: (1) To establish a model for the multivariate plant species intensities. (2) To find diversity metrics – plant richness (total number of plant species in an observed frame), the Shannon-Weiner diversity metric, and plant preference (the combination of species likely to occur at a given location) – given land use, marker of historic agricultural land use, income per capita and elevation. (3) To predict plant intensity measures mentioned in (1) and diversity measures mentioned in (2) at newer points using the information obtained from the rest of the points. Whereas traditional geostatistical methods for prediction of spatial variables (Cressie, 1993) are based on an assumption of normality or normality under appropriate transformations, this is still inadequate for count data (Royle, Link and Sauer, 2002). Using a Bayesian framework, we analyze the spatial data using generalized linear mixed models for point processes as in Diggle, Tawn and Moyeed (1998) and Christensen and Waagepetersen (2002). As in Christensen and Waagepetersen (2002), we use an efficient Markov Chain Monte Carlo (MCMC) algorithm based on Langevin-Hastings updates.

The cross-convolution approach (Majumdar and Gelfand, 2007) is applied for modeling the covariance structure of log-intensities across various land use - desert, urban residential and urban non-residential for this multivariate count data. We prove that this approach can deliver non-stationarity of the multivariate spatial process of plant counts at the global level to distinguish between different land use, and preserve stationarity at the local level restricted to each land use. The cross-convolution approach is computationally less intensive because of its parsimony (Majumdar and Gelfand, 2007).

The plant richness intensity predictions and prediction intervals are obtained from quantiles of the predictive distribution. It turns out that low species intensities can be predicted fairly accurately. This is important for site-specific prediction on plant-species preference from an ecological perspective.

1.1 *Data Description*

The Phoenix metropolitan area of Central Arizona developed from a convergence of culture during the westward expansion of the Americas. During the latter half of the twentieth century, Phoenix has seen an exponential growth rate driven primarily by climate, cheap housing and water, the availability of jobs, and especially the commercial success of the air conditioner. Now harboring over 3.5 million people, Phoenix is the fifth largest and the second fastest growing metropolitan area in the nation.

As part of the Arizona-Phoenix Long Term Ecological Research project, ecological surveys were conducted at 204 sites covering 6400 km² encompassing the entire Phoenix metropolitan area. Permanent site locations were identified and field inventory was conducted between February and May, 2005. The sampling unit at each site was a 30m × 30m plot, in which all plant taxa were identified, collected, and archived at Arizona State University's herbarium. All perennials were counted within the plot in order to obtain density (perennials/900 m²). In the project described in Walker, Briggs, Dugan, Gries and Grimm (2006), one objective of the project was to investigate whether perennial occurrence could be predicted from observations of elevation, income per capita and knowledge of whether the land had been used for agriculture ever before. In our case, we used only those species that had non-zero counts in at least 10 or more spatial locations, and

those spatial locations for which complete covariate-information was available. Thus we finally had 58 species and 144 spatial locations. The upper panel of 1 displays the area under the survey with the black points indicating the points in our survey and the highlighted points (marked by “1”, “73”, and “120”) indicating the three points where prediction inference was carried out. The middle panel of Figure 1 displays the elevation surface of the entire area and the lower panel of Figure 1 displays the annual income per capita at these locations. The contours of this plot was created using the package ARC GIS..

Land use at each of the $n = 144$ sites was classified, and we use 3 main regional categories: desert ($n_1 = 71$), urban residential ($n_2 = 48$), and urban non-residential ($n_3 = 25$), $n = n_1 + n_2 + n_3$. Socio-economic variables such as income per capita and whether the land had been previously used for agriculture(1) or not(0) were obtained from the U.S. Census (2001) for the appropriate block group within which each survey point that was classified as residential was located and from historic landuse data.

[Figure 1 about here.]

2. Models

Generalized linear mixed models(GLMM) (Breslow and Clayton, 1993 ; Lee and Nelder, 1996) are extensions of Generalized Linear models(GLM)(McCullagh and Nelder, 1989). Generalized linear mixed models have been further generalized to the context of point processes to where the random effect is spatially correlated (Christensen, Møller and Waagepetersen, 2000). This allows for additional sources of variability due to unobservable, spatial random effects and specifically, has been applied quite widely in the context of a Bayesian

point-process models (for example, Christensen and Waagepetersen, 2002, Benes, Bodlak, Møller and Waagepetersen, 2002). Such random effects are sometimes modeled using a log Gaussian Cox process (Møller, Syversveen and Waagepetersen, 1998) for spatio-temporal count data (Brix and Møller, 2001) as well as spatio-temporal data where the spatial locations occur on a latticed grid and the time is sparsely distributed (Rasmussen, Møller, Aukema, Raffa and Zhu, 2006). We shall introduce a generalized spatial linear mixed model for count data, with each kind of land use giving rise to a different count-process, since plant-preferences in urban sites are different than that of desert sites. Also, within urban sites, it is quite likely that the plant diversity of the residential sites is different than that of non-residential sites. In general, we expect to see more variation in the urban residential sites compared to the other sites. The component count processes (related to the land use types) are assumed to be *spatially correlated* among themselves and are modeled using log Gaussian Cox processes, where the log-intensities corresponding to these processes have a joint Gaussian spatial distribution. Hence the cross-covariance matrix corresponding to this spatially non-homogeneous process (based on different kinds of land use) is modeled using a *non-stationary cross-covariance matrix* – in our case we use the model of *cross-convolution* (Majumdar and Gelfand, 2007) and show that while providing a valid covariance structure, it maintains stationarity at the local level – restricted on a specific land use, but ensures non-stationarity at the global level – to account for different land use types inherent to the data.

2.1 Spatial Generalized Linear Mixed Models

Our model is a special case of a spatial GLMM, which is the framework proposed by Diggle et al. (1998) for modeling of non-Gaussian spatial data. For a spatial GLMM, the random variables $Y_{i_j}^{(k)}$ are mutually independent across $i_j = 1, \dots, n_j; j = 1, 2, 3$ given W . For each i_j , the distribution $[Y_{i_j}^{(k)}|W]$ has a density $f(\cdot; \lambda_{i_j}^{(k)})$ which depends on the conditional mean $\lambda_{i_j}^{(k)} = E[Y_{i_j}^{(k)}|W]$. Moreover, $\lambda_{i_j}^{(k)}$ is related to the linear predictor $w_j(s_{i_j}) + X_{i_j}(s_{i_j})^T \beta_j^{(k)}$ by a strictly increasing link function g so that

$$\lambda_{i_j}^{(k)} = g^{-1}(w_j(s_{i_j}) + X_{i_j}(s_{i_j})^T \beta_j^{(k)}) = g^{-1}(\mu_j^{(k)}(s_{i_j})) \quad (1)$$

The family of Spatial GLMMs are thus a flexible class of models for, e.g., spatially correlated count, binary and positive data.

The Gaussian field is assumed to be characterized by the covariance function $\{C_{jj'}, j, j' = 1, 2, 3\}$ given by

$$C_{jj'}(u) = E[w_j(s)w_{j'}(s')]$$

depends only on the distance $u = \|s - s'\|$ between locations $s, s' \in S$. The cross covariance function is modeled using a cross-convolution approach:

$$C_{jj'}(s) = C_j \star C_{j'}(s) = \int_{\mathbb{R}^2} C_j(s - v)C_{j'}(v)dv \quad (2)$$

where the above defines a valid positive semi-definite function(Theorem 1 below), as in traditional geostatistics Cressie (1993), and the $C_j(\cdot), j = 1, 2, 3$ are distinct and valid isotropic, stationary covariance functions defined on \mathbb{R}^2 ,

$$C_j(\|s\|) = \sigma_j \rho(\phi_j \|s\|), \quad s \in S. \quad (3)$$

where $\rho(\cdot)$ is a known correlation function and the parameter $(\sigma_1, \sigma_2, \sigma_3, \phi_1, \phi_2, \phi_3) \in [0, \infty]^6$ is unknown; here ϕ_j is a correlation scale parameter, and σ_j is the standard deviation corresponding to the j -th Gaussian process.

As in Diggle et al. (1998), we use a Bayesian approach to inference, where priors are imposed on the unknown regression parameters and the covariance parameters of the unknown random field.

2.2 Model for Plant Richness

Let $S \in \mathfrak{R}^2$ denote the field where the plants are observed. One could view the unobserved locations of the plant positions as a point process, and a key summary of the species occurrence is the species intensity $\lambda(\cdot)$. For species k and i_j -th location $s_{i_j} \in S$ $i_j = 1, \dots, n_j$ corresponding to the j -th land use type, and some small area δ around s_{i_j} , $\lambda_j^{(k)}(s_{i_j})\delta$ is the approximate expected number of plant species in the region. For each $s_{i_j} \in S$, we model $\lambda_j^{(k)}(s_{i_j})$ as

$$\lambda_j^{(k)}(s_{i_j}) = \exp(w_j(s_{i_j}) + X_{i_j}(s_{i_j})^T \beta_j^{(k)}) = \exp(\mu_j^{(k)}(s_{i_j})) \quad (4)$$

where $X_{i_j}(s_{i_j})$ is the covariate vector associated with s_{i_j} , $\beta_j^{(k)} \in \mathfrak{R}^p$ is a vector of regression parameters, and $w_j(s_{i_j})$ is a random effect that serves to model sources of variation not included in $X_{i_j}(s_{i_j})^T \beta_j^{(k)}$, pertaining to the j -th land use type. The random effects $W = \{w_j(s_{i_j}) : s_{i_j} \in S; i_j = 1, \dots, n_j; j = 1, 2, 3.\}$ are assumed to form a correlated zero-mean Gaussian field, so that it is possible to model correlation in the plant species intensities due to e.g., the reproduction mechanism of the plant species (in the desert area), or unobserved spatial covariates such as plant-preferences (in the urban area).

Let $Y_{i_j}^{(k)}$ denote the plant species count centered at location s_{i_j} . We model $Y_{i_j}^{(k)}$ conditionally on W , as independent Poisson distributed random variables with conditional mean $\lambda_{i_j}^{(k)} \approx A\lambda_j^{(k)}(s_{i_j})$. Here $A = 900 \text{ m}^2$ is the area of each individual plot where the survey was carried out. The resulting Poisson log-normal model is an example of a spatial GLMM with Poisson error distribution and a log link function(as discussed in Section 2.1).

The log-normal model can also be seen as an approximation to the distribution of counts under a log Gaussian Cox process (Møller et al., 1998 and Brix and Møller, 2001) with log-normal intensity surface $\{\lambda_j^{(k)}(s_{i_j})\}$ given in (4). Under a log Gaussian Cox process and given $\lambda_j^{(k)}(\cdot)$, the counts $Y_{i_j}^{(k)}$ are independent Poisson distributed with mean $\int_{C_A} \lambda_j^{(k)}(s-v)dv$, where C_A denotes a circle with center zero and area A .

2.3 Non-stationarity Using the Model of Cross-convolution

The following result is the underlying result we shall use to show the non-negative-definiteness and hence validity of the “cross-covariance” matrix defined in 2 and 3.

Theorem 1: Consider any finite set of points $\mathbf{s}_1, \dots, \mathbf{s}_n$ in R^d . Let \tilde{C} be an $nk \times nk$ matrix with $k \times k$ blocks $\tilde{C}_{ij} = C(s_i - s_j)$ where for the (l, k) -th term in the sub-,matrix \tilde{C}_{ij} is a cross-convolved function $C_l \star C_k(s_i - s_j)$ evaluated at $(s_i - s_j)$. Here C_l and C_k are isotropic functions on R^d . Then \tilde{C} is non-negative definite and hence a valid cross-covariance structure on s_1, \dots, s_n . (Majumdar and Gelfand, 2007).

Corollary 1: Let $P = \{\text{set of all possible subsets of rows of } \tilde{C}\}$ and let $P' = \{\text{set of all possible subsets of columns of } \tilde{C}\}$. Let $M_{\pi, \pi'}$ denote the matrix obtained from deleting the set $\pi \in P$ of rows and the set $\pi' \in P'$

of columns of the matrix \tilde{C} . It follows from Theorem 1 that M_{π} is non-negative definite for every π .

Corollary 2: The “cross-covariance” matrix defined in 2 and 3 has precisely the form of M_{π} for some π and is hence non-negative definite.

Now since for each land use type the spatial variance and spatial dependence are different, so the model we get is a non-stationary model at the global level, and stationary, when restricted to any land use, at the local level.

3. Posterior Simulation and Prediction

In this section, we discuss Bayesian inference and prediction for spatial GLMM's. We focus on the Poisson log-normal model, but the techniques are readily applicable to other spatial GLMM's. For the $n = 144$ observations in the plant richness data set, we let $s_1, s_2, \dots, s_n \in S$ be the centers of the observation frames. We further let $W_j = (w_j(s_{i_1}), \dots, w_j(s_{i_{n_j}}))$ be the purely spatial gaussian vector corresponding to the j -th land use type and denote $W = (W_1, W_2, W_3)$ the $n \times 1$ unobserved purely spatial gaussian vector corresponding to the observed n locations. Denote the $n \times 1$ observed realization of $Y = (Y_1(s_1), \dots, Y_1(s_{n_1}), Y_2(s_{n_1+1}), \dots, Y_2(s_{n_1+n_2}), Y_3(s_{n_1+n_2+1}), \dots, Y_3(s_n))^T$ by $y = (y_1, y_2, \dots, y_{n_1}, y_{n_1+1}, \dots, y_{n_1+n_2}, y_{n_1+n_2+1}, \dots, y_n)^T$. For locations s_{n+1}, \dots, s_{n+q} , $q \geq 0$ of interest for prediction, we let $W^* = (w_{j_{s_{n+1}}}(s_{n+1}), \dots, w_{j_{s_{n+q}}}(s_{n+q}))^T$ where $j_{s_{n+l}}$ is the land use type corresponding to the location s_{n+l} , $l = 1, \dots, q$. Now W^* and Y are conditionally independent given $(W, \beta, \phi_1, \phi_2, \phi_3, \sigma_1^2, \sigma_2^2, \sigma_3^2)$. Our Bayesian inference is thus separated into, first, the posterior simulation of $(W, \beta, \phi_1, \phi_2, \phi_3, \sigma_1^2, \sigma_2^2, \sigma_3^2)$ given $Y = y$, and second, the prediction of Y .

3.1 Posterior Simulation for Spatial GLMMs

Briefly, a Metropolis-Hastings iteratively generates an ergodic Markov chain that yields the posterior sample. In each step, a proposal is generated for an update of the current state of the chain. The update is then accepted or rejected according to a certain acceptance probability. The most commonly used algorithm is Gaussian random walk Metropolis, where the proposal distribution is Normal with mean equal to the present state. For details on Markov Chain Monte Carlo, we refer to Roberts and Rosenthal (1998).

Christensen and Waagepetersen (2002) use a so-called fixed scan hybrid algorithm, where transformed random effects corresponding to the W are updated simultaneously using *truncated Langevin-Hastings updates*. The truncated Langevin-Hastings update utilizes the gradient of the log posterior and can, as exemplified in Christensen and Waagepetersen (2002), lead to much reduced Monte Carlo errors compared with the standard alternative of a random walk Metropolis update. Furthermore, it is easy to implement the Langevin-Hastings update for any spatial GLMM with a canonical link function. We use a similar method in our analysis for updating each of the spatial components in the model.

3.1.1 Reparameterization Let $\tilde{C} = C(\phi_1, \phi_2, \phi_3, \sigma_1^2, \sigma_2^2, \sigma_3^2)$ denote the covariance matrix of W , and let $\tilde{C}^{\frac{1}{2}}$ be the square root so that $\tilde{C} = \tilde{C}^{\frac{1}{2}}(\tilde{C}^{\frac{1}{2}})^T$, where the square root, e.g., may be obtained by Cholesky factorization. We can then assume that $W = \tilde{C}^{\frac{1}{2}}\Gamma$, where Γ follows an n -dimensional standard multivariate Normal distribution. Posterior simulations of W can be obtained by transforming MCMC samples of the conditional distribution of Γ given

$Y = y$. It is demonstrated in Møller et al. (1998) that this reparameterization of the random effect can lead to better mixing properties of the MCMC algorithm. In Section 4, β is *a priori* Normal with mean μ_b and covariance matrix Σ_b , so that a reparameterization $\beta = \mu_b + \Sigma_b^{\frac{1}{2}}\epsilon$, $\epsilon \sim N_{3 \times p \times K}(0, I)$, can be applied to β also.

3.1.2 The hybrid Metropolis algorithm In our MCMC algorithm, $\gamma, \epsilon, \log(\phi_j), \log(\sigma_j^2), j = 1, 2, 3$ are updated in turn, in each scan using either truncated Langevin-Hastings or random walk Metropolis updates as discussed below.

As in Section 4, suppose the priors for $\phi_j, \sigma_j^2, \beta$ are independent with prior densities $\pi_{\phi_j}, \pi_{\sigma_j^2}, \pi_{\beta}$, respectively, where π_{β} is the Normal density described in section 3.1.1. The log posterior density of $(\Gamma, \epsilon, \beta, \{\sigma_j^2, \phi_j, j = 1, 2, 3\})$ given $Y = y$ is

$$\begin{aligned} f(\gamma, \epsilon, \beta, \{\sigma_j^2, \phi_j, j = 1, 2, 3\}) &= \text{const}(y) + \sum_{j=1}^3 (\log(\pi_{\phi_j}) + \log(\pi_{\sigma_j^2})) + \frac{1}{2} \|\epsilon\|^2 \\ &\quad + \frac{1}{2} \|\gamma\|^2 + \sum_{k=1}^K \sum_{j=1}^3 \sum_{i_j=0}^{n_j} y_{i_j}^{(k)} \log(\lambda_{i_j}^{(k)}) - \lambda_{i_j}^{(k)} \end{aligned}$$

with $\lambda_{i_j}^{(k)} = \exp(w_j(s_{i_j}) + X_{i_j}(s_{i_j})\beta_j^{(k)})$ where $(w_1(s_1), \dots, w_1(s_{n_1}), \dots, w_3(s_{n_1+n_2+1}), \dots, w_3(s_n)) = W = \tilde{C}^{\frac{1}{2}}\gamma$ and $\beta = \mu_b + \Sigma_b^{\frac{1}{2}}\epsilon$. Let $\mu_{i_j}^{(k)} = \log(\lambda_{i_j}^{(k)}) = w_j(s_{i_j}) + X_{i_j}(s_{i_j})\beta_j^{(k)}$. The truncated Langevin-Hastings proposal distribution for γ is $N(\gamma + (d/2)\nabla(\gamma)^{trunc}, dI)$, where dI is the user-specified proposal variance,

$$\nabla(\gamma)^{trunc} = -\gamma + \tilde{C}^{T\frac{1}{2}}\{y_{i_j}^{(k)} - \lambda_{i_j}^{(k)} \wedge H\}_{i_j,k}$$

is obtained by truncating the gradient (with respect to γ) of the log-target density, and $0 < H < \infty$ is a truncation constant. A similar simple form is available for any GLMM with g equal to the canonical link function (see McCullagh and Nelder (1989)). A theoretical study on MCMC algorithms with truncated Langevin-Hastings updates for conditional simulations in GLMMs for fixed model parameters is given in Christensen, Møller and Waagepetersen (2001). For ϵ , we similarly use a truncated Langevin-Hastings update, where the truncated gradient for ϵ takes the form $-\epsilon + \Sigma_b^{1/2} X^T \{y_{i_j}^{(k)} - \lambda_{i_j}^{(k)} \wedge H\}$. The derivative of $\log f(\gamma, \epsilon, \beta, \{\sigma_j^2, \phi_j, j = 1, 2, 3\})$ with respect to ϕ_j is complicated, since it involved differentiating the matrix $\tilde{C}^{\frac{1}{2}}$. In addition, Langevin-Hastings updates are most useful for updating high-dimensional quantities Roberts and Rosenthal (1998). Therefore, for the one-dimensional parameters $\log(\phi_j)$ and $\log(\sigma_j^2)$, we use standard Normal random-walk Metropolis updates. An R implementation of the MCMC algorithm described above can be downloaded from <http://spatio-temporal.asu.edu/PlantDiversity-Rcodes.html>.

3.2 Bayesian Prediction

Here we consider prediction of functionals of $(\lambda_{n+1}^{(1)}(x_{n+1}), \dots, \lambda_{j_{n+1}}^{(k)}(x_{n+1}), \dots, \lambda_{j_{n+q}}^{(1)}(x_{n+q}), \dots, \lambda_{j_{n+q}}^{(K)}(x_{n+q}))^T$, where $\lambda_i^{(k)}, i = n+1, \dots, n+q, k = 1, \dots, K$ are given by (4). The predictive distribution of $\lambda^* = (\lambda_{n+1}^{(1)}(x_{n+1}), \dots, \lambda_{j_{n+1}}^{(k)}(x_{n+1}), \dots, \lambda_{j_{n+q}}^{(1)}(x_{n+q}), \dots, \lambda_{j_{n+q}}^{(K)}(x_{n+q}))^T$ is given by

$$[\lambda^*|y] = E \{[g^{-1}(W^* + X^*\beta)]|y\}, \quad (5)$$

where $X^* = (x_{n+1}, \dots, x_{n+q})^T$, $W^* = (w_{n+1}, \dots, w_{n+q})^T$, and g^{-1} is applied coordinatewise in (5). The observed data are conditionally independent of W^* given $(W, \beta, \{\phi_j, \sigma_j^2; j = 1, 2, 3\})$. An algorithm to obtain a sample $(\lambda^{*(m)})^l$ from the predictive distribution of $[\lambda^*|y]$ is obtained follows:

1. Sample $(W^l, \beta^l, \{\phi_j^l, \sigma_j^{2l}; j = 1, 2, 3\})$ from the posterior distribution $[W, \beta, \{\phi_j, \sigma_j^2; j = 1, 2, 3\}|y]$ using the MCMC algorithm in Section 3.

2. For each $l = 1, \dots, L$, and for the m -th species, $m = 1, \dots, M$ $\lambda^{*(m)l} = g^{-1}(W_m^{*l} + X^*\beta^l)$, where W_m^{*l} are sampled independently from the multivariate Normal distribution $[W_m^*|W^l, \beta^l, \{\phi_j^l, \sigma_j^{2l}; j = 1, 2, 3\}]$.

Consider for the moment just one location, say, x_{n+1} . Common practice in traditional spatial MCMC methods is to use $E[\lambda_{j_{n+1}}^{(k)}(x_{n+1})|y]$ as a predictor for $\lambda_{j_{n+1}}^{(k)}(x_{n+1})$ and $E[\lambda_{j_{n+1}}^{(k)}(x_{n+1})|y] \pm 1.96(\text{var}[\lambda_{j_{n+1}}^{(k)}(x_{n+1})|y])^{1/2}$ as an approximate 95% prediction interval. However, when $g(\lambda) = \log(\lambda)$ and a vague prior is used for the variance parameters σ_j^2 , then the mean and variance of $[\lambda_{j_{n+1}}^{(k)}(x_{n+1})|W]$ may become infinite (De Oliveira, Kedem and Short, 1997), whereby also $E[\lambda_{j_{n+1}}^{(k)}(x_{n+1})|y]$ becomes infinite. Further, when the predictive distribution is skewed, the use of $E[\lambda_{j_{n+1}}^{(k)}(x_{n+1})|y] \pm 1.96(\text{var}[\lambda_{j_{n+1}}^{(k)}(x_{n+1})|y])^{1/2}$ as a prediction interval is not appropriate. Here we use the median $\text{med}[\lambda_{j_{n+1}}^{(k)}(x_{n+1})|y]$ as predictor for $\lambda_{j_{n+1}}^{(k)}(x_{n+1})$ and $[\zeta_{2.5\%}(x_{n+1}), \zeta_{97.5\%}(x_{n+1})]$ as the prediction interval, where $\zeta_p(x_{n+1})$ is the p quantile of $[\lambda_{j_{n+1}}^{(k)}(x_{n+1})|y]$. Also, prediction of the observations $Y^* = (Y_{j_{n+1}}^{(k)}(x_{n+1}), \dots, Y_{j_{n+q}}^{(k)}(x_{n+q}))^T$ may be of interest (See Section 4.5). The predictive distribution for $Y_{j_{n+1}}^{(k)}(x_{n+1})$ is given by the probabilities

$$\begin{aligned} p_{j_{n+1}}^{(k)}(z|y) &= P(Y_{j_{n+1}}^{(k)}(x_{n+1}) = z|y) \\ &= E(E[f(z; \lambda_{j_{n+1}}^{(k)}(x_{n+1})|W, \beta, \{\phi_j, \sigma_j^2; j = 1, 2, 3\})|y]) \end{aligned}$$

where $f(\cdot; \lambda)$ is the Poisson probability mass function parameterized by the mean λ . As an approximate $1 - p$ predictive set for $Y_{j_{n+1}}^{(k)}(x_{n+1})$, we use $\{z : p/2 \leq \sum_{l=0}^{z-1} p_{j_{n+1}}^{(k)}(l|y) + p_{j_{n+1}}^{(k)}(z|y)/2 \leq 1 - p/2\}$, $0 < p < 1$.

4. The plant diversity data application

For the perennial species data, we use the elevation(in meters), indicator function (0-1) measuring whether or not the land was ever used for agriculture , annual income per capita as the explanatory variables. All these variables are standardized by first subtracting the mean and then dividing by the standard deviation to reduce collinearity among the covariates. This standardizing also improves the mixing of the MCMC algorithm. Thereby 3 explanatory variables $x_1(s)$, $x_2(s)$ and $x_3(s)$ are obtained for each $s \in S$. We further include an intercept parameter $\beta_{j_0}^{(k)}$ corresponding to the j -th land use and k -th species. $\beta_j^{(k)} = (\beta_{j_0}^{(k)}, \dots, \beta_{j_3}^{(k)})^T$ is the regression coefficient vector for a spatial point corresponding to the j -th land use and k -th species so that $x(s) = (1, x_1(s), x_2(s), x_3(s))^T$.

The perennial species data is modeled using Poisson log-Normal model specified in Section 2.1, with the exponential correlation function $\rho(u) = \exp(-u)$ as “building blocks” of the non-stationary matrix \tilde{C} mentioned in 3. We do not believe that much information concerning the type of correlation function is available from our data set, as very close observations do not occur, and the Gaussian field is not observed. For the sake of parsimony, we therefore do not include uncertainty concerning the choice of correlation model in our inference.

4.1 Priors for the plants diversity data

We use a Normal prior i.e., a $N(\mu_b, \sigma_b^2 I)$ prior for the $3 \times p \times K$ vector β of regressors. We anticipate the values of the posterior distribution to be in a certain range of the empirical estimate μ_b of β and hence by picking a large $\sigma_b^2 = 10$, we ensure that this prior does not overtly influence the mixing of the MCMC. The μ_b is also used as starting value for the β value in the MCMC. For the σ_j^2 , we use Inverse Gamma priors, which is commonly used for variance parameters. We employ the Inverse Gamma(α_σ, c_σ) distribution, with shape and scale parameters $\alpha_\sigma = 2$, $c_\sigma = 1$ respectively, so that the prior mean is set to 1 and the variance is set to infinity. Likewise, we employ the Inverse Gamma(α_τ, c_τ) distribution for the error variances τ_j^2 , with shape and scale parameters $\alpha_\tau = 2$, $c_\tau = 1$ respectively, so that the prior mean is set to 1 and the variance is set to infinity. For the correlation parameters ϕ_j , we use the conventional Gamma priors with shape and scale parameters α_ϕ, c_ϕ respectively, where $\alpha_\phi = 1$, $c_\phi = 1$. This choice sets the prior mean and variance both to be 1, and hence matches with the range of our anticipated value of ϕ_j , since we know that the range of the data and the ϕ_j parameter are related in an approximate relationship $\phi_j \approx 4.5/range$ (Majumdar and Gelfand, 2007).

4.2 Model Choice

Model choice is investigated with respect to the specification of $\mu_{i_j}^{(k)}$ in (2). All the species attributes, i.e. $\beta_j^{(k)}$ were significant and so the regression coefficients were retained in all the models.

For model selection we adopt the computationally convenient Deviance Information Criterion(DIC) (Speigelhalter, Best, Carlin and Van Der Linde,

2002). This criterion is sensitive to choice of parameterization (Section 8, Spiegelhalter et al. (2002)). We considered two parameterizations. One treats $\theta = \{\lambda_{i_j}^{(k)}\}$ as the *natural set* of parameters, and the other treats the $\theta = \{\mu_{i_j}^{(k)}\}$ as the set of parameters. Other choices are possible, including treating $\{\beta_j^{(k)}\}$ as the set of parameters. These different procedures give different answers for p_D , the effective degrees of freedom, but, at least, for the two choices we tried, the DIC's were similar and the ordering of the models the same (smaller is better).

Table 1 provides a summary of the model comparison using DIC. Comparison of the full model (1) with the model (2) (in which the regression coefficients for non residential and residential urban points are the same for each species), model (3) (in which the regression coefficients for urban, non residential and residential points are the same for each species) and model (4) (where spatial variance, error variance and decay parameters are the same for all land-use – thus giving rise to a Stationary setting) shows that under each criterion, the full model is the best. We confine ourselves to analyses under this model for the remainder of the paper.

4.3 Computation

For the posterior simulations, we use 100,000 scans of the hybrid algorithm, where $\tilde{C}^{\frac{1}{2}}$ is computed using Cholesky factorization, the truncation constant in the gradients of γ and ϵ is $H = 50$, and we obtain a Monte-Carlo sample of size 10,000 by subsampling every 10-th scan.

Theoretical results in Roberts, Gelman and Gilks (1997) and Roberts and Rosenthal (1998) suggest that one should tune the proposal variances to obtain acceptance rates around 0.23 for random-walk updates and 0.57 for

Langevin-Hastings updates. The overall acceptance rates for the updates of γ , ϵ and σ_1^2 , σ_2^2 , σ_3^2 , τ_1^2 , τ_2^2 , τ_3^2 , ϕ_1 , ϕ_2 , ϕ_3 were 0.70, 0.67, 0.34, 0.34, 0.34, 0.40, 0.39, 0.39, 0.43, 0.42, 0.42, respectively.

4.4 *Posterior distribution*

Posterior 95% credibility intervals for the $\{\phi_j, \sigma_j^2, \tau_j^2, j = 1, 2, 3\}$ are in Table 2. The regression coefficient parameters β_1 , β_2 , β_3 corresponding to standardized values of $\log(\text{elevation})$, $\log(\text{income})$, and an indicator function showing whether a land was ever used in agriculture respectively for the 38 perennial species used in this study, are shown in Figure 2. This figure has three rows for each of three land use-types and three columns, each for the three covariates. Each panel in this figure displays the regression coefficient from the first through the 38th species. We note that the regression coefficients have very small posterior variance (thereby yielding almost a dot-like representation with respect to the scale of the Figure) and are somewhat varied over the 38 species. The regression coefficient for elevation varies over the 38 species most of all for the nonresidential points (Panel (c)). The regression coefficient for the indicator function (whether a land was ever used for agriculture) also varies over the 38 species most of all for the nonresidential points (Panel (c)).

[Figure 2 about here.]

4.5 *Prediction of Species Intensity and Species Preference*

We are interested in mapping the intensity $\lambda_{i_j}^{(k)}$ of two perennial species, *Larrea tridentata* (one of the most common sonoran desert species) and *Penisetum setaceum* (one of the few introduced urban species that also thrives in the desert). The log of the median and the standard deviation spatial

surfaces are given in the 1st and the 2nd columns of Figure 3 respectively, for *Larrea tridentata* in the upper panel and for *Pennisetum setaceum* in the lower panel. Note that lighter means lower value. In Figure 3, the highest intensity of *Larrea tridentata* is seen to be in the desert area, as expected. Also, *Larrea tridentata* is clearly a more abundant species in all land use when compared to *Pennisetum setaceum*, which occurs rarely (indicated by very low log intensities). The prediction errors $Y_{i_j} - med[\lambda_{i_j}^{(k)}]$ versus $\log(med[\lambda_{i_j}^{(k)}])$ of *Larrea tridentata* are displayed in the upper panel and 3rd column of Figure 3 and for *Pennisetum setaceum* in the lower panel and 3rd column of Figure 3. We observe that the prediction uncertainty is large whenever $med[\lambda_{i_j}^{(k)}]$ is large – as expected for Poisson likelihood.

[Figure 3 about here.]

For testing model prediction, we choose three spatial locations from each of the three land use types. The observed values and the 95% prediction intervals are presented for these locations, s_{j_i} , $i = 1, \dots, 3$ – one from the desert (spatial location 1), one from urban residential (spatial location 73) and another from a non-residential urban location (spatial location 120). The observed count is denoted with a circle. The predicted interval is denoted with a line that joins the upper and lower bound of the interval. For the desert location in the upper panel of Figure 4, the first two species clearly highly dominate in count. Although the actual values are smaller than the lower bounds of the predicted intervals, the prediction intervals indicate a high value of those two species. Since we are more interested in predicting the non-zero counts in terms of species preference, these two prediction intervals

point towards the two species that highly dominate over others at this location. Note that all the rest of the species have very short prediction intervals and most of the observed counts are covered by these intervals (most of the counts being equal to zero). From the same point of view, we fail to see such precision of inference for the residential point (Figure 4, middle panel) and the non-residential point (Figure 4, lower panel), since many of the prediction intervals are wide, even when the actual count is actually zero (though the observed count is covered by most of these prediction intervals as well). The loss of precision follows from the larger variation and hence less “predictability” in the urban area. However, many of these observed zero counts are predicted using prediction intervals restricted to zero only. In terms of precision, we see that prediction at the non-residential point yields more precision than the residential point. Hence these prediction intervals rule out quite a few species from the most likely set of species that could dominate at a given point in the range of our study-area. We thus note that these prediction intervals are more “conservative” for inference about species-preference when the non-zero counts are of interest.

[Figure 4 about here.]

4.6 *Inference on Species Diversity and Species Richness*

There is a considerable literature on diversity measures. See the summary discussion in Kempton (2002). For illustration, we work with the Shannon-Weiner form of index which in our case takes the form

$$\exp \left\{ - \sum_{k=1}^K \left(\frac{\lambda_j(s_{i_j})^{(k)}}{\sum_k \lambda_j(s_{i_j})^{(k)}} \log \frac{\lambda_j(s_{i_j})^{(k)}}{\sum_k \lambda_j(s_{i_j})^{(k)}} \right) \right\} \quad (6)$$

where K is the number of species. Note that (6) is maximized at $\frac{\lambda_j(s_{i_j})^{(k)}}{\sum_k \lambda_j(s_{i_j})^{(k)}} = \frac{1}{K}$ and equals K in this case. It is minimized if $\frac{\lambda_j(s_{i_j})^{(k)}}{\sum_k \lambda_j(s_{i_j})^{(k)}} \rightarrow 1$ for some k and tends to 1 in this case. Hence (6) is scaled to the number of species.

The interpretation which is attached to (6) is that it will be large if many species are equally likely to co-occur in location s_{i_j} . In our case the Shannon-Weiner diversity metric is given in the posterior median and posterior standard deviation surfaces (Figure 5, 1st row, 1st and 2nd columns, respectively). We note that higher diversity is noticed in the urban area.

[Figure 5 about here.]

Another important aspect of inference on species diversity is species richness. The observed species richness in location s_{i_j} is $\sum_{k=1}^K \mathbf{1}(Y_{i_j}^{(k)} > 0)$. Again, this is purely a descriptive summary. Regression models have been used to explain these observed richness values using environmental features and enable interpolation to unobserved sites. See Guisan and Zimmerman (2000) in this regard. Under our model, the analogue at location s_{i_j} is the posterior distribution of $E\left(\sum_{k=1}^K \mathbf{1}(Y_{i_j}^{(k)} > 0) | y\right) = \sum_{k=1}^K P(Y_{i_j}^{(k)} > 0 | y) = \sum_{k=1}^K E(1 - e^{-\lambda_{i_j}^{(k)}} | y) = K - \sum_{k=1}^K E(e^{-\lambda_{i_j}^{(k)}} | y)$.

Using the posterior mean across spatial location we can create a posterior potential richness surface by plotting $K - \sum_{k=1}^K E(e^{-\lambda_{i_j}^{(k)}} | y)$ versus s_{i_j} and the posterior median richness and the posterior standard deviation of richness are given in the 2nd row of Figure 5, in the 1st and 2nd columns, respectively. Under our modeling, species richness can only be inferred within the domain of the study and is only relative to the set of species which have been modeled.

ACKNOWLEDGEMENTS

The authors would like to thank Steven S. Carroll for sampling design; L. Dugan, D. Gonzales, E. Holmes, Q. Stewart, J. Walker, M. Feldner, S. Whitcomb, and X. Zhuo for field and lab assistance; Salt River Project for the donation of helicopter time; Cities of Phoenix, Scottsdale and Tempe, Maricopa County Parks, Tonto National Forest, Arizona State Lands Department, Sky Harbor Airport and all the private property owners involved for giving permission to access their land. This work was funded by National Science Foundation Grants # DEB-9714833 and DEB-0423704 (the Central Arizona-Phoenix Long-Term Ecological Research Project) and the NSF Bio-complexity in the Environment Program (EAR-0322065) and the NSF grant #SES-0604373 supported by the Methodology, Measurement, and Statistics Program, the Statistics and Probability Program, and a consortium of federal statistical agencies.

RÉSUMÉ

REFERENCES

- Benes, V., Bodlak, K., Møller, J. and Waagepetersen, R. P. (2002). Bayesian analysis of log Gaussian Cox process models for disease mapping. Technical Report R-02-2001, Department of Mathematical Sciences, Aalborg University.
- Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9–25.

- Brix, A. and Møller, J. (2001). Space-time multi type log Gaussian Cox processes with a view to modelling weed data. *Scandinavian Journal of Statistics* **28**, 471 – 488.
- Christensen, O. F., Møller, J. and Waagepetersen, R. (2000). Analysis of spatial data using generalized linear mixed models Langevin-type Markov Chain Monte Carlo. Technical Report R-00-2009, Department of Mathematical Science, Aalborg University.
- Christensen, O. F., Møller, J. and Waagepetersen, R. (2001). Geometric ergodicity of Metropolis Hastings algorithms for conditional simulation in generalised linear mixed models. *Methodology and Computing in Applied Probability* **3**, 309 – 327.
- Christensen, O. F. and Waagepetersen, R. (2002). Bayesian prediction of spatial count data using generalized linear mixed models. *Biometrics* **58**, 280 – 286.
- Cressie, N. A. C. (1993). *Statistics for Spatial Data, revised edition*. Wiley, New York.
- De Oliveira, V., Kedem, B. and Short, D. A. (1997). Bayesian prediction of transformed Gaussian random fields. *Journal of the American Statistical Association* **92**, 1422 – 1433.
- Diggle, P., Tawn, J. A. and Moyeed, R. A. (1998). Model-based geostatistics (with discussion). *Applied Statistics* **47**, 299 – 350.
- Guisan, A. and Zimmerman, N. E. (2000). Predictive habitat distribution models in Ecology. *Ecological Modeling* **135**, 147 – 186.
- Hope, D., Gries, C., Zhu, W., Fagan, W. F., Redman, C. L., Grimm, N. B., Nelson, A. L., Martin, C. and Kinzig, A. (2003). Socioeconomics drive

- urban plant diversity. *Proceedings of the National Academy of Sciences* **100**, 8788 – 8792.
- Kempton, R. A. (2002). Species diversity. In El-shaarwari, A. H. and Piegorsch, W. A., editors, *Encyclopaedia of Environmetrics*, volume 4, pages 2086 – 2092.
- Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models. *J.R. Statist.Soc. B* **58**, 619 – 678.
- Majumdar, A. and Gelfand, A. E. (2007). Spatial modeling for environmental data using convolved covariance functions. *Mathematical Geology* **39**. to appear.
- McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models, 2nd edition*. Chapman and Hall, London.
- Møller, J., Syversveen, A. R. and Waagepetersen, R. (1998). Log Gaussian Cox processes. *Scandinavian Journal of Statistics* **25**, 451 – 482.
- Rasmussen, J. G., Møller, J., Aukema, B. H., Raffa, K. F. and Zhu, J. (2006). Bayesian inference for point processes observed at sparsely distributed times. Technical Report R-2006-24, Department of Mathematical Sciences, Aalborg University.
- Roberts, G. O., Gelman, A. and Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Annals of Applied Probability* **7**, 110 – 120.
- Roberts, G. O. and Rosenthal, J. (1998). Optimal scaling of discrete approximations to Langevin diffusions. *J. R. Statist. Soc. B* **60**, 255 – 268.
- Royle, J. A., Link, W. A. and Sauer, J. (2002). Statistical mapping of count survey data. In Scott, J., Heglund, P., Morrison, M., Haufler,

J., Raphael, M., Wall, W. and Samson, F., editors, *Predicting species occurrences: issues of accuracy and scale*, pages 625–628, Washington DC. Island Press.

Speigelhalter, D. J., Best, N. G., Carlin, B. P. and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *J. R. Statist. Soc. B* **64**, 1 – 34.

Walker, J. S., Briggs, J. M., Dugan, L., Gries, C. and Grimm, N. B. (2006). Is there an urban ecological succession?(submitted for publication).

APPENDIX

Proof of Corollary 2

To see this, observe that with the notations of Theorem 1, $k = p$, the number of land use categories, and the “cross-covariance” matrix defined in 2 and 3 is the matrix $M_{\pi} \pi$ where π correspond to those set of rows/columns corresponding to all possible $C_l \star C_l(s_i - s_j)$ where $\{s_i, s_j\}$ is not contained in land use type l .

[Table 1 about here.]

[Table 2 about here.]

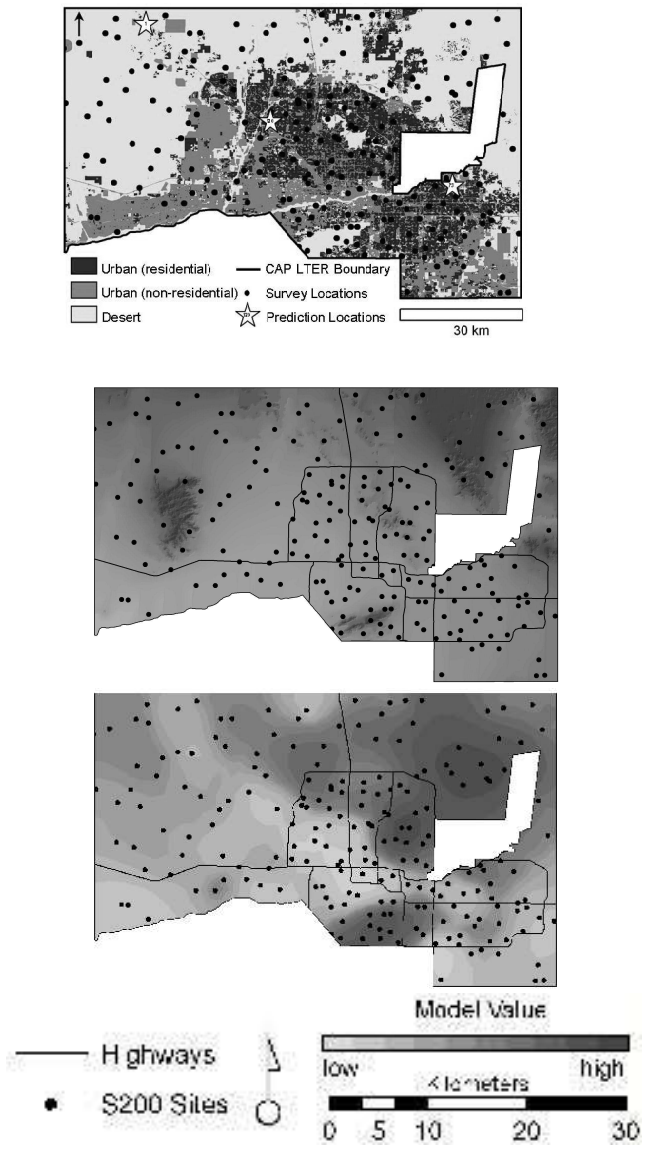


Figure 1. Map of landuse and spatial locations (upper panel), log(elevation) measured in meters(mid panel) and log(income per capita) measured in dollars(lower panel)

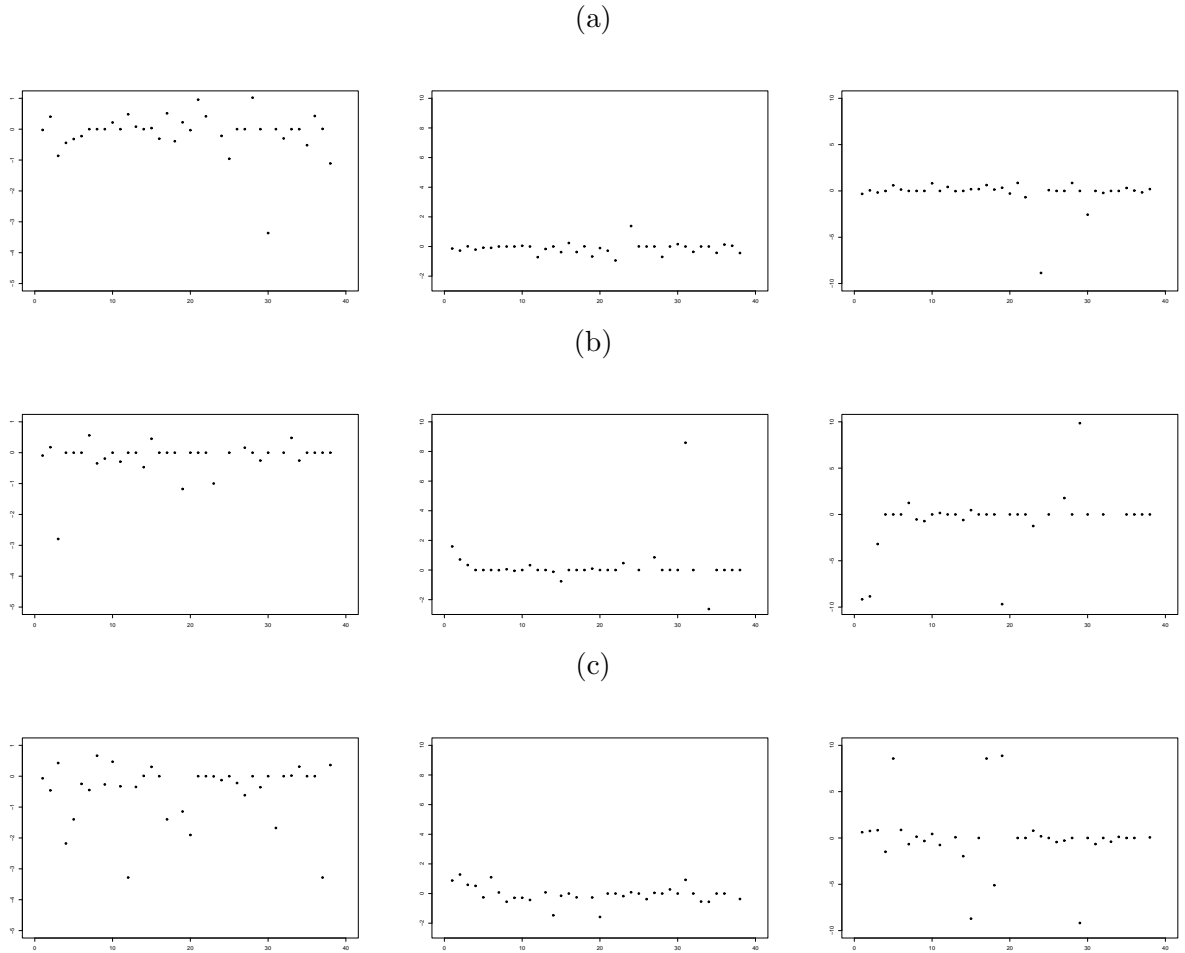


Figure 2. Posterior 95% credibility intervals for regression coefficients $\beta_{elevation}$ (left), β_{income} (middle) and $\beta_{ever\ in\ ag.}$ (right) for (a) desert (b) residential (c) nonresidential land use across 38 plant species.

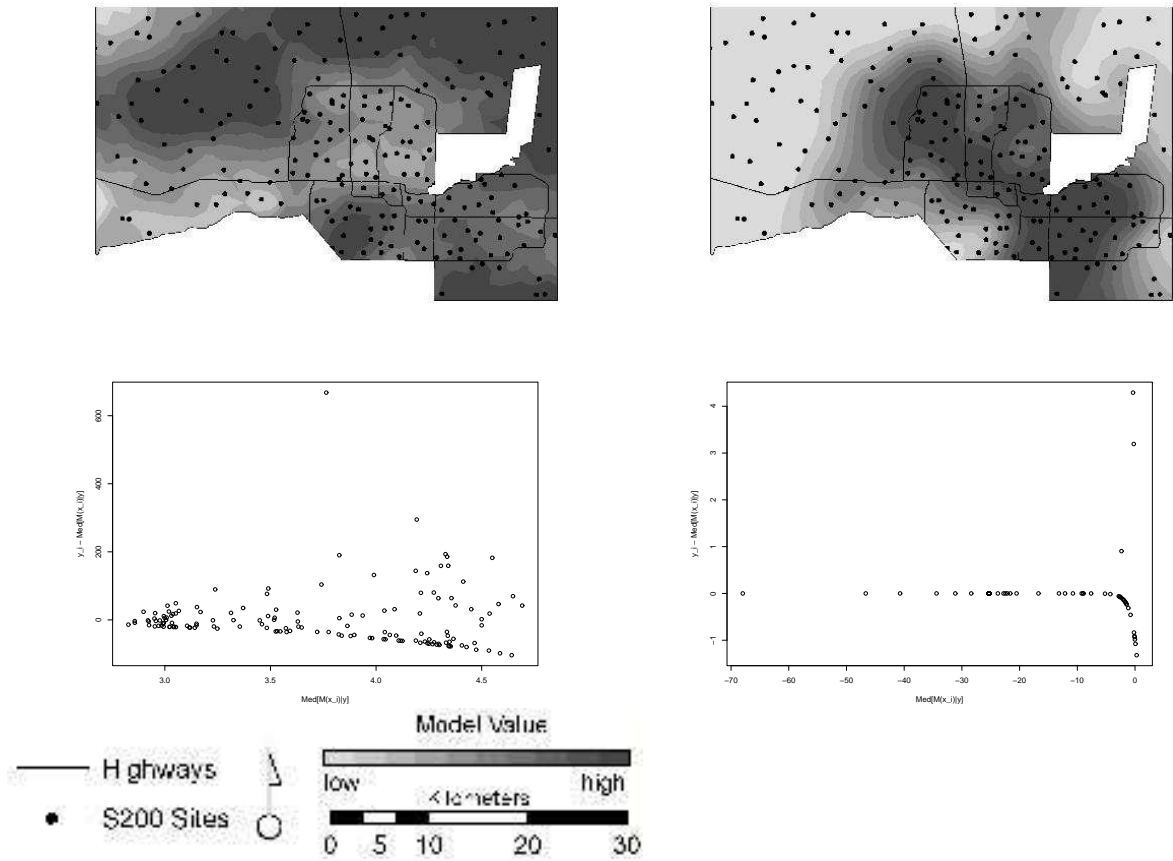
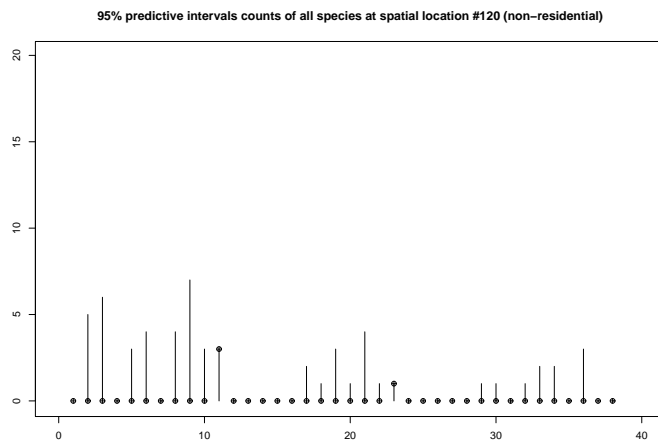
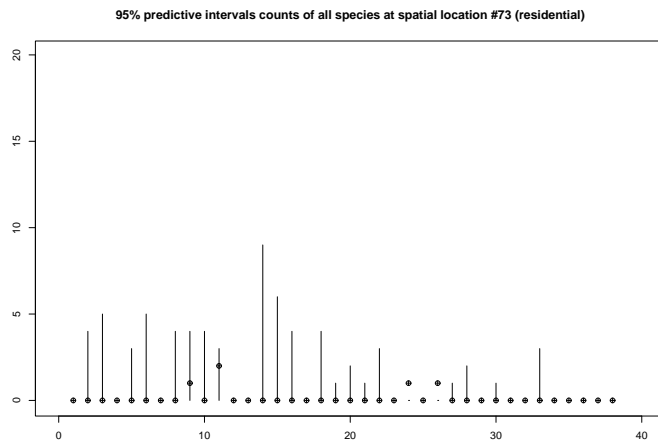
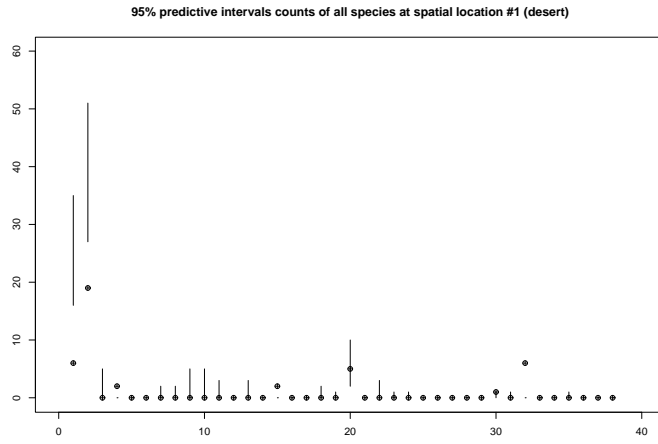


Figure 3. From top to bottom, posterior median $\text{med}[\lambda_{ij}^{(k)}]$ and prediction errors $(Y_{ij} - \text{med}[\lambda_{ij}^{(k)}])$ versus $\log(\text{med}[\lambda_{ij}^{(k)}])$ for *Larrea tridentata* (left panel) and *Pennisetum setaceum* (right panel)



29
Figure 4. 95% posterior predictive intervals for the counts of each of 38 species in 3 hold-out locations: (a) Desert (b) Residential (c) Nonresidential

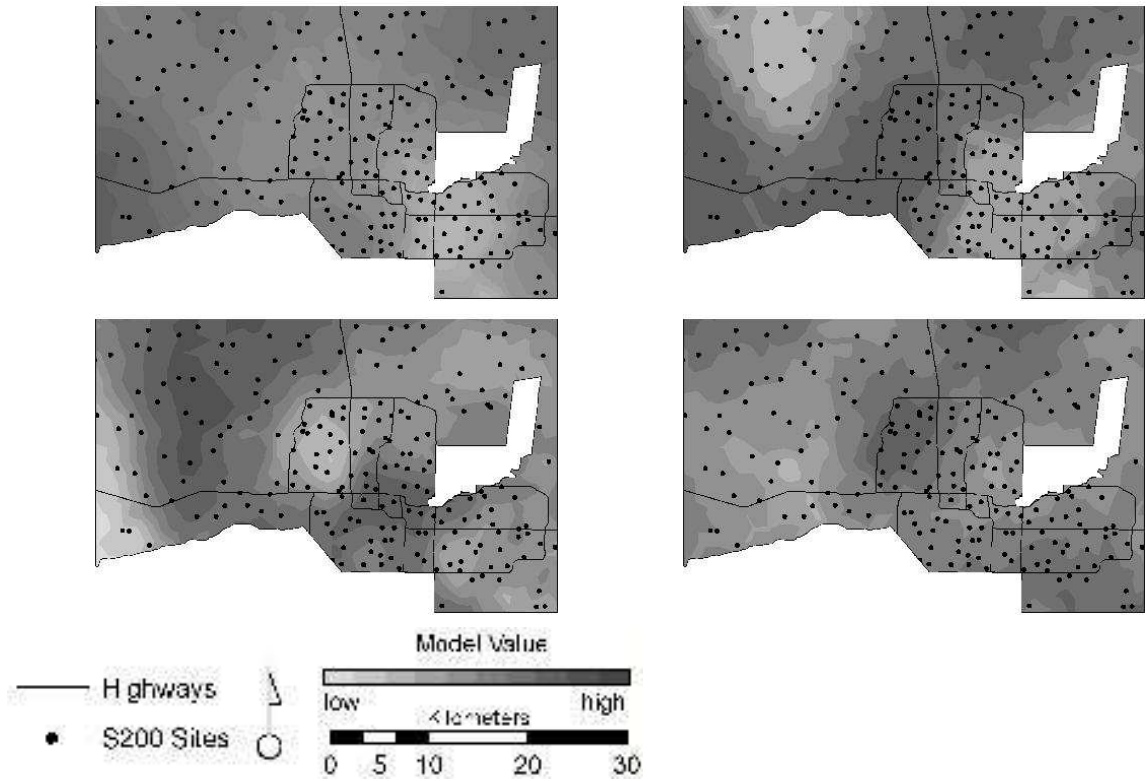


Figure 5. Posterior predicted median of Shannon Weiner Diversity metric(SWDM)(top left); posterior predicted standard deviation of SWDM(top right); posterior predicted median of species richness(top left); Posterior predicted standard deviation of species richness(top right);

Table 1
Model Comparison with DIC

models	# independent parameters	$\theta = \{\lambda_j^{(k)}\}$		$\theta = \{\log(\lambda_j^{(k)})\}$	
		p_D	DIC	p_D	DIC
(1) Full	465	-63591.4	-125559.8	-638748.7	-700717.0
(2) $\beta_{residential}^{(k)}$ = $\beta_{nonresidential}^{(k)}$ $k = 1, \dots, K$	313	0.0053	-61590.16	0.0057	-61590.16
(3) $\beta_{desert}^{(k)}$ = $\beta_{residential}^{(k)}$ = $\beta_{nonresidential}^{(k)}$ $k = 1, \dots, K$	161	0.0252	-46159.14	0.0370	-46159.12
(4) Stationary	459	-63512.8	-125460.2	-638727.5	-700648.3

Table 2

Posterior median and 95% credibility intervals for variance and spatial correlation.

Parameters	Quantiles		
	2.5%	50%	97.5%
Decay parameter			
ϕ_{desert}	0.026	0.635	3.897
$\phi_{residential}$	0.022	0.688	3.722
$\phi_{nonresidential}$	0.028	0.724	3.798
Spatial variance			
σ_{desert}^2	0.173	0.590	4.099
$\sigma_{residential}^2$	0.180	0.614	4.145
$\sigma_{nonresidential}^2$	0.178	0.595	4.562
Error variance			
τ_{desert}^2	0.178	0.567	3.907
$\tau_{residential}^2$	0.191	0.585	3.043
$\tau_{nonresidential}^2$	0.189	0.579	3.518