Landscape Ecology **18:** 487–502, 2003.

Research article

Predicting the spatial distribution of ground flora on large domains using a hierarchical Bayesian model

Mevin B. Hooten[1,*], David R. Larsen[2] and Christopher K. Wikle[1]
*[1]Department of Statistics, University of Missouri, 222 Mathematical Sciences Building, Columbia, Missouri 65211, USA; [2]Department of Forestry, University of Missouri, Columbia, Missouri 65211, USA; *Author for correspondence (e-mail: hooten@stat.missouri.edu)*

Received 5 July 2002; accepted in revised form 28 March 2003

Key words: Bayesian statistics, Hierarchical Bayesian models, Landscape vegetation prediction, Spatial modeling, Missouri, USA, Ozark Highlands

Abstract

Accomodation of important sources of uncertainty in ecological models is essential to realistically predicting ecological processes. The purpose of this project is to develop a robust methodology for modeling natural processes on a landscape while accounting for the variability in a process by utilizing environmental and spatial random effects. A hierarchical Bayesian framework has allowed the simultaneous integration of these effects. This framework naturally assumes variables to be random and the posterior distribution of the model provides probabilistic information about the process. Two species in the genus *Desmodium* were used as examples to illustrate the utility of the model in Southeast Missouri, USA. In addition, two validation techniques were applied to evaluate the qualitative and quantitative characteristics of the predictions.

Introduction

The spatial pattern of ecological processes has historically been of interest to researchers (Pielou 1977; Turner 1989). Early spatial prediction efforts, in the spirit of landscape ecology, spawned from the desire to view ecological patterns on a spatial domain (Turner 1989; Borcard et al. 1992; Guisan et al. 1998). Such efforts were usually motivated by goals pertaining to understanding the biology of a species, species associations, species response to the environment, and the notion of preserving biodiversity (e.g., Gleason 1926; Erickson 1943; Whittaker 1956; Day and Monk 1974; Ernst 1978). Predicting spatially explicit ecological processes has evolved into a main theme within the discipline of landscape ecology (Forman and Godron 1986).

Several different methodologies for incorporating spatial structure and covariate information have recently been considered (e.g., Davis and Goetz 1990; Cressie 1993; Smith 1994; Cherrill et al. 1995; Franklin 1998; Zimmermann and Kienast 1999; Hoeting et al. 2000; Royle et al. 2001). These methods range from *ad hoc*, fixed-neighborhood, and decision based models to very complex statistical models. Still, statistically rigorous methods for combining species/environment relationships and explicit non-Gaussian spatial structure at large numbers of prediction locations are lacking. This can mainly be attributed to the computationally intensive nature of accounting for dependence structure on such spatial domains (Besag 1974; Ripley 1981; Cressie 1993). Computational potential is increasing at an exponential rate, thus providing abundant opportunity for innovation in conceptual and computational methods for dealing with high-dimensional problems. This paper presents an approach to landscape level process modeling on large domains while explicitly considering spatially correlated error.

Our goal is to combine information found within abiotic covariates and spatial dependence that may act as a surrogate for various biotic covariates, in order to provide spatial predictions for vegetation as well as the uncertainty in those predictions. This method

combines important features of two different approaches to ecological modeling, making it a very complete and robust alternative to conventional methods. Zimmermann and Kienast (1999) describe a method and justification for mapping individual species patterns using known environmental covariates, while Royle et al. (2001) describe methods useful in accounting for unknown ecological/biotic processes through spatial parameter estimation and modeling. Legendre (1993) emphasizes the importance of considering such effects and the biotic mechanisms they may be representing (i.e., seed dispersal, colonization, disturbance and competition). Although the aim of this project is not to infer which biotic mechanisms are imposing the spatial structure, the method we present accounts for the spatial random effects as well as species/environment relationships to model the distribution of a species on a landscape. Other similar methods have been proposed (e.g., Augustin et al. 1996; Hoeting et al. 2000; Lichstein et al. 2002), however the method presented in this paper maintains statistical rigor and provides a spatial representation of prediction error.

Constructing such a model within a statistically rigorous framework and utilizing distributional information with random parameters has made it possible to qualitatively evaluate predictions, assess model effectiveness, and provide an unparalleled amount of information at the pixel level. Knowledge of landscape level floristic information in addition to simulation-based data helps our method bridge the gap between the theoretical and applied realms of landscape modeling. We demonstrate this method using a dataset collected as part of a large-scale ecological project known as the Missouri Ozark Forest Ecosystem Project (MOFEP).

Study area

Southeastern Missouri, USA is home to a topographically complex section of the country known as the Ozark Highlands (Ozarks). Oklahoma and Arkansas share a portion of this ancient and environmentally heterogeneous area. More than 530 plant species have been documented in the Ozark Highlands (Grabner et al. 1997; Grabner 2000). Ecological and site defining characteristics supply variables that are correlated with plant diversity and individual species occurrence and abundance (Grabner et al. 1997; Grabner 2000). Once identified and quantified at a landscape level, these variables are useful for describing vegetation pattern.

A portion of the Current River Hills Subsection (an ecological subregion of the Ozark Highlands) is home to an extensive long-term ecological project known as the Missouri Ozark Forest Ecosystem Project (MOFEP). This project was designed to monitor and assess the short and long-term effects of common management practices on Ozark ecosystems. In this project, the Missouri Department of Conservation in conjunction with the University of Missouri and the USDA Forest Service collected data at 9 sites ranging in size from 265–530 ha (Figure 1). These sites were selected because they had minimal edge, were greater than 240 ha, and were largely free from anthropogenic manipulation for at least the past 40 years (Brookshire et al. 1997). The floristic data were collected at 10,368 specific locations throughout the 9 sites. This field dataset is one of the richest of its kind and provides a solid foundation from which to base landscape level predictions.

Ecological importance

The Ozarks provide a unique set of ecosystems of which many important ecological components are still not well understood. The triad of relationships between vegetation, edaphic characteristics, and nutrient cycling seem to be especially interesting to researchers.

Much of the Ozark Highlands consist of highly weathered, nutrient poor, ultisols and alfisols (Meinert et al. 1997). These edaphic factors provide an environment where leguminous nitrogen-fixing plants appear to succeed in dominating a majority of available understory growing space. Such relationships are partially documented but again not well understood (Grabner et al. 1997; Grabner 2000). This project focuses on modeling the distributions of two herbaceous plants in the genus *Desmodium* with hopes that in addition to providing a solid example for this new modeling technique, it may provide a better understanding of ecological processes in the Ozark Highlands.

This approach to landscape modeling may find application in several disciplines. Although the focus in this work is on developing a method to help us better understand realistic vegetation distributions on a landscape, other potential applications of this type of model may include but certainly are not limited to: (1) Spatio-temporal mapping of wildlife forage avail-

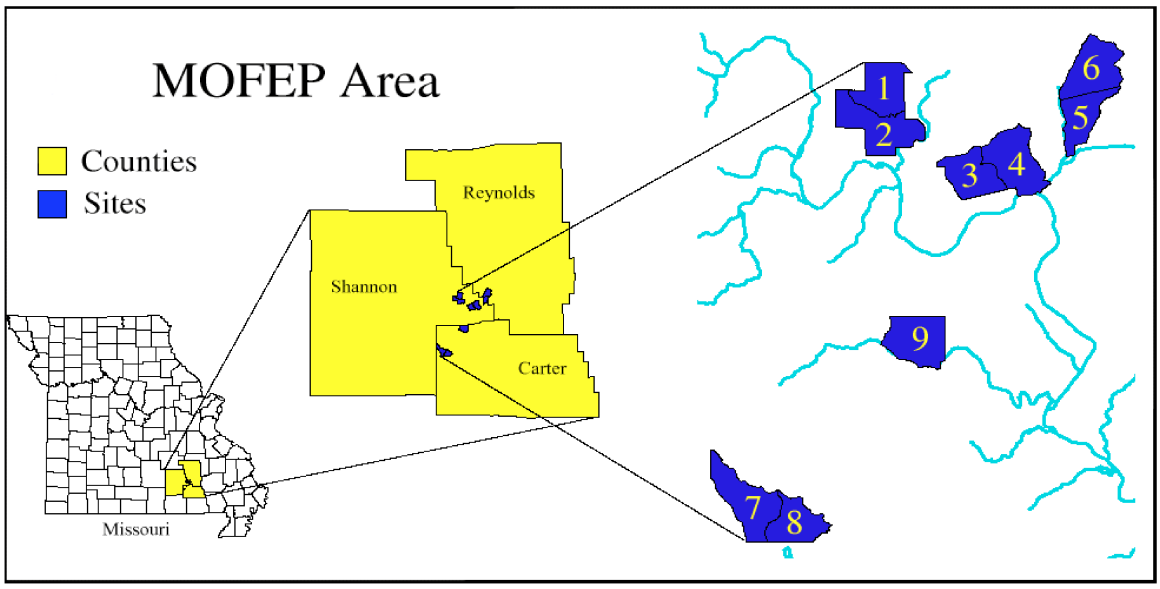


Figure 1. The 9 MOFEP sites are nested in the Ozark Highlands near the Current River.

ability, (2) Analysis of inter-species spatial interaction (competition), (3) Landscape level identification of areas susceptible to exotic invasion, (4) Analysis of spatial and temporal patterns of biodiversity. Scale is irrelevant to the methods presented here, thus the same technique could be implemented at the molecular and global levels.

Material and methods

Field data

Before treatments were applied to MOFEP forest stands, a set of initial data were collected in order to assess pre-treatment conditions. The ground flora data gathered in 1995 were determined to be the most complete and representative, and therefore serve as our main set of field data. Originally intended for use with analysis of variance, these data are collected in a spatially non-random scheme and are made up of a total of 648 plots whereby each consists of four subplots that are in turn represented by four 1 m^2 quadrats (Figure 2). Vegetation in each quadrat is identified and quantified on a percent coverage basis (Grabner 1996).

Specifically for our modeling effort, we converted the dominance estimates to species presence/absence and then aggregated over the subplot in order to reduce unwanted change in spatial support related to a

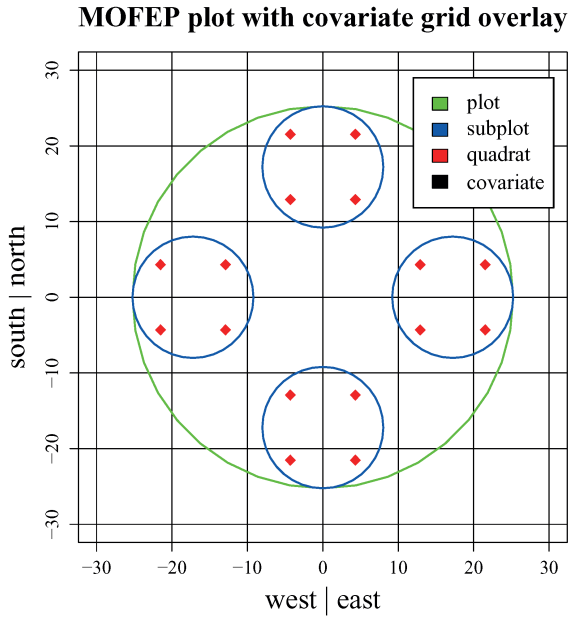


Figure 2. The conceptual MOFEP plot layout and covariate grid overlay (measurements in meters).

difference in measurement scales between the field and covariate data. The binary nature of the subplot aggregation is intuitive in that a species which is present in one of the four quadrats is registered as present for the subplot (hence coded as 1), while a species absent from all quadrats in a subplot is recorded as absent from the subplot (and hence coded as 0). The geographic location of the aggregate then

Table 1. All covariates with previous and current descriptions.

Covariate	Format	Type	Interval/Categories	Origin	Original Format
SWness[a]	grid	continuous	[−1–1]	DEM	grid
Rel. Elev.[b]	grid	continuous	[0–1]	DEM	grid
LTA[c]	grid	categorical	2	GIS layer	vector
VD[d]	grid	categorical	2	model	grid

[a]Similar to Beers aspect transform (Beers et al. 1966), (represents most "southwest" aspects as 1 and most "northeast" as −1)
[b]Relative Elevation (continuous measure of slope position, where 0 represents the bottom of the hill and 1 represents the top)
[c]Land Type Association (although this region of the Ozarks is made up of many LTA's, only two were deemed to be important to this Genus of plant)
[d]Variable depth soil (this is an ecological land type in the Ozarks and is speculated by researchers to be one of the most important variables influencing the growth of *Desmodium*)

becomes the center of the subplot. It is assumed that the four quadrats represent the grid cell in which the subplot center lies. The original dataset was consequently reduced to 2592 geographic locations with binary information for every species in the MOFEP region.

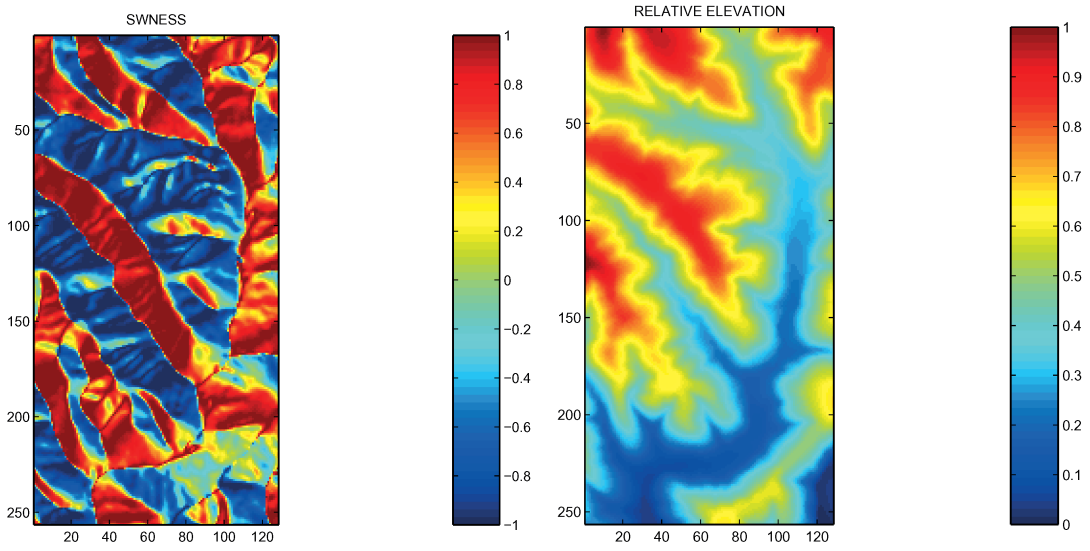
Covariate data

As discussed earlier, several environmental descriptors have varying degrees of influence on some plant species in the Ozark Highlands and when geographically quantified, are potentially useful as covariates in a predictive model of plant occurrence. The type and magnitude of influence is highly species dependent, therefore it was necessary to find diverse and representative covariates that may be related to the success of a given plant. Covariates were chosen based on availability, resolution, as well as previous published and unpublished analysis. The availability of certain potentially important covariates (e.g., micro-climate, specific soil type, animal influences, and forest canopy gaps) is limited due to the effort required in explicitly describing such information on a continuous landscape. Methods have been developed for simulating biophysical variables based on non-biotic environmental variables (i.e., Guisan and Zimmermann 2000), however the examples provided in this project focus on covariates at hand. Many potentially important covariates are available but do not share the same resolution as the proposed predictions. Such covariates would impose an unacceptable amount of error and were therefore omitted from this study. Table 1 displays the full list of covariates used in this project and their original formats as well as current descriptions.

Climatic covariate influences are partially absorbed by topographic variables upon which localized climatic features may be dependent in the Ozark Highlands. Although future models of this type could incorporate climate-related variables, inclusion in this project was not feasible owing to a lack of availability and resolution.

A Digital Elevation Model (DEM) is a very important component of this project because it provides several different types of information (e.g., slope, aspect, elevation, curvature) that are potentially important as covariates in a statistical model. The original DEM's used in this project were created and provided by Krystansky and Nigh (2000) at the Missouri Ecological Classification Project. All vector-based covariates were rastorized to a grid cell size of 10 m^2. The choice of this cell size was based on the following four reasons: (1) the bulk of grid-based covariates originated from a digital elevation model of the same resolution; (2) this area represents the finest resolution possible without being overwhelmed by measurement error; (3) the resolution is necessary to adequately describe Ozark landforms and subtle topography; and (4) it minimizes the difference in scales between field subplots and covariates.

The prediction domain chosen is nested within MOFEP sites one and two (Figure 4) and is represented by two land type associations, two parent material types, areas of variable depth soil, all aspects, variable relief and elevations (Figure 3). The prediction domain ($\approx$ 328 ha) consists of a [256 × 128] grid resulting in 32,768 prediction locations. Of the 216 subplots that fall within the prediction domain, *Desmodium glutinosum* is present on 67 and *Desmodium nudiflorum* is present on 159. All further covariate and prediction images will be shown on this domain. These gridded covariates (Figure 3) are ex-



(a) swness: -1 = NE, 1 = SW

(b) relative elevation: 0 = low, 1 = high

(c) lta: 0 = lta 2, 1 = lta 4

(d) variable depth: 1 = variable depth soil

Figure 3. Covariates used in model.

pected to account for a majority of the variation in the distributions of these two species.

Basic model

A hierarchical Bayesian framework was used to implement a generalized linear mixed model (hereafter GLMM). Many other vegetation prediction projects have used generalized linear models, especially with logistic link functions (e.g., Smith 1994; Franklin 1998; Guisan et al. 1998; Zimmermann and Kienast 1999). These probablistic models are convenient because of their intuitive nature and ease of implementation. While it is possible to introduce a spatial random effect in the GLM framework, thus obtaining the GLMM, problems with estimation and implementa-

tion arise when predicting on large domains (where correlation is described as a function of distance, not just neighborhood). A GLMM implemented through a Bayesian approach provides the ability to deal with uncertainty related to spatial correlation (Clayton 1997; Gilks et al. 1997; Royle et al. 2001).

Conceptually, hierarchical modeling is based on a simple result from probability. Consider three random variables X, Y, Z with the joint probability distribution $[X, Y, Z]$.* This distribution can always be factored into component distributions such as: $[X, Y, Z] = [Z \mid Y, X][Y|X][X]$. Thus, there is a hierarchical structure to the distributions. The idea is that the joint distribution $[X, Y, Z]$ may be very complicated and difficult to specify, but a series of conditional models are simpler to specify.

In Bayesian modeling, one is interested in updating a prior belief about a random variable given data. For example, assume we have observations Z of some process Y. We are interested in updating our prior distribution $[Y]$ with data to get the posterior distribution $[Y|Z]$. To do this, we need information about how the data depend on the process, $[Z|Y]$, which is often referred to as the "likelihood". Bayes' Theorem provides this updating:

$$[Y|Z] = \frac{[Z|Y][Y]}{[Z]} \tag{1}$$

Often, there are conditioning parameters, say θ, that describe the distributions. These can also be given prior distributions and one can utilize the hierarchical decomposition within the Bayesian framework. For example:

$$[Y, \theta|Z] \propto [Z|Y, \theta][Y|\theta][\theta]. \tag{2}$$

The challenging aspect of such models in complicated settings is the specification of the component models and the evaluation of the posterior. Very often the normalizing constant in (2) cannot be found analytically and Monte Carlo methods must be considered. Recent advances in Markov Chain Monte Carlo (MCMC) have dramatically increased the complexity of models that can be considered by this approach (Gilks et al. 1997).

* Note that throughout the paper we use the bracket notation "[]" to refer to a probability distribution. A conditional distribution, say A conditioned on B, is denoted as $[A|B]$.

Hierarchical spatial model

In our problem of predicting plant species presence or absence we can think of the data as representing some binary random variable. Specifically, $Y_i = 1$ if a species is present at spatial location s_i $(i = 1,\ldots,m)$ and $Y_i = 0$ if it is absent. If we let Y_i be a Bernoulli random variable then $\mathrm{E}(Y_i) = E(Y_i = 1) = p_i$. Our goal is to model this mean response in terms of some linear function of the n_x covariates X_{ij}, $(i = 1,\ldots,m)$,$(j = 1,\ldots,n_x)$, e.g., $p_i = f(\mathbf{x}_i'\beta)$, where $\mathbf{x}_i' = (x_{i1},\ldots,x_{in_x})'$ and β is the associated vector of parameters. Common examples of $f()$ with binary variables include the logistic function and the probit function. In general, these ideas fall under the category of generalized linear models (GLMs) in statistics (e.g., McCulloch 1994). In our case, we know that the chosen covariates can't possibly explain all of the variability in the mean response. Thus, we consider adding random effects, η_i, as surrogates for unobserved biological and environmental covariates. That is, $p_i = \mathrm{f}(\mathbf{x}_i'\beta + \eta_i$. With the random effect, this model is now known as a generalized linear mixed model (GLMM) (e.g., McCulloch 1994). One must then specify a distribution for the η_i. We expect these random effects to be correlated in space.

Traditional likelihood-based approaches for estimation of GLMMs with correlated random effects can be difficult to implement with complicated spatial effects and large prediction domains. Diggle et al. (1998) showed that one could easily account for such structures in a hierarchical Bayesian setting. However, their approach, like the traditional approach, cannot easily be implemented on very large prediction domains. Thus, we outline a hierarchical Bayesian probit model (i.e., $f()$ is the cumulative normal distribution function denoted by $\Phi()$) with spatial random effects that can be implemented on very high dimensional prediction domains, as encountered in landscape ecology.

Albert and Chib (1993) showed that the introduction of a latent (hidden) process to the probit formulation greatly facilitates Bayesian computation. We follow their approach, with two critical modifications: (1) Introduction of spatial random effects and, (2) the spectral parameterization of those effects. Our model is described below.

Let:

$$Y_i = \left\{ \begin{array}{l} 1 \text{ if } Z_i > 0, \text{species is present at spatial location } i \\ 0 \text{ if } Z_i \leqslant 0, \text{species is absent at spatial location } i, \end{array} \right\} \quad (3)$$

where Z_i is a latent spatial process with distribution $Z_i \sim N(\mathbf{x}_i'\beta + \eta_i, 1)$, and η_i is a spatially correlated Gaussian process with mean zero. Then, it can be shown that $p_i = E(Y_i) = \Phi\mathbf{x}_i'\beta + \eta_i)$. Although the data only apply to Z_i, $(i = 1,\ldots, m)$ we will ultimately be interested in predictions of p_j $(j = 1,\ldots,n)$ which may include many sites in addition to where we have data ($n \gg m$). Note that our covariates are available at all m locations as well. Thus, we consider the multivariate normal distribution $\mathbf{Z}|\beta, \eta \sim \mathrm{N}(\mathbf{X}\beta + \eta, \mathbf{I})$ where $\mathbf{Z}$ is $n \times 1$, β is $p \times 1$, $\mathbf{x}$ is $n \times p$, η is $n \times 1$, and $\mathbf{I}$ is an $n \times n$ identity matrix. In the hierarchical spirit we must also specify distributions for η and β. We select $\eta \sim N(0, \sigma_\eta^2 \mathrm{R}_\eta)$ and $\beta \sim N(\beta_o, \Sigma_\beta)$, where $\mathbf{R}_\eta$ is a spatial correlation matrix. In principle, an MCMC algorithm can be developed for this model analogous to Albert and Chib (1993). However, when n is of order 10^5 or more (as in our example), such algorithms are not feasible. Thus, we make one additional modification. We note that the spatial random effects can be rewritten $\eta = \Psi\alpha$, where Ψ is an $n \times n$ matrix of orthonormal basis functions (e.g., Fourier basis functions, wavelets, etc.) and, α are the associated spectral coefficients. Specifically, as in Wikle (2002) we let Ψ be Fourier basis functions so that $\Psi'\eta = \alpha$ is a discrete Fourier transform and $\eta = \Psi\alpha$ is the associated inverse transform. There are extremely efficient algorithms for computing these transforms, even in high dimensions (e.g., the Fast Fourier Transform). Thus, we write $\mathbf{Z}|\beta,\alpha \sim \mathrm{N}(\mathbf{x}\beta + \Psi\alpha, \mathbf{I})$ where $\alpha \sim N(\alpha_0, \sigma_\alpha^2 \mathrm{R}_\alpha)$. In addition to fast computation, this has the advantage that for stationary processes the Fourier spectral coefficients are approximately independent (e.g., Shumway and Stoffer 2000). Thus, σ_α is a diagonal matrix which further simplifies computation. Finally, we don't know *a priori* the spatial dependence in η (and thus α) or the variance component σ_η^2 (and thus σ_α^2), although we have some understanding of these features through preliminary data analysis. Consequently, we account for our understanding and our uncertainty about such understanding by specifying informed prior distributions on parameters. We allow $\sigma_\alpha^2 \sim IG(q_\alpha, r_\alpha)$ where $IG()$ refers to an inverse gamma distribution. Based on preliminary data analysis it was assumed that the spatial dependence for η can be modeled well by an isotropic exponential correlation function, $r_\eta(d) = exp(-\theta d)$, where d is the distance between two locations and θ is the spatial dependence parameter, $\theta > 0$. As discussed in Wikle (2002), if one knows $r_\eta(d)$ then one can specify $\mathbf{R}_\alpha(\theta)$ as well. Thus, we let $\theta \sim \mathrm{UNIF}(u_1, u_2)$ to finish out the model hierarchy, where UNIF(a,b) refers to a uniform distribution on the continuous closed interval between a and b. The hierarchical model is summarized as follows:

$$Y_i = \left\{ \begin{array}{l} 1 \text{ if } Z_i > 0, \text{species is present at spatial location } i \\ 0 \text{ if } Z_i \leqslant 0, \text{species is absent at spatial location } i \end{array} \right\} \quad (4)$$

$$Z|\beta, \alpha \sim N(X\beta + \Psi\alpha, I) \quad (5)$$

$$\beta|\beta_0, \textstyle\sum_\beta \sim N(\beta_0, \sum_\beta) \quad (6)$$

$$\alpha|\alpha_0, \sigma_\alpha^2, \theta \sim N(\alpha_0, \sigma_\alpha^2 R_\alpha(\theta)) \quad (7)$$

$$\sigma_\alpha^2|q_\alpha, r_\alpha \sim IG(q_\alpha, r_\alpha) \quad (8)$$

$$\theta|u_1, u_2 \sim UNIF(u_1, u_2) \quad (9)$$

Where, β_o, Σ_β, α_o, q_α, r_α, u_1, and u_2 are specified.

Coding the model

The process of iterative sampling from the posterior distribution is laborious, therefore computational efficiency is critical in implementing such a model. A programming language with spectral-transform capabilities and sparse matrix storage running on a dual-processor computer with large memory was used to perform most of the intensive operations. Efficient software, fast processors and disk space are not enough to implement the models proposed here. Therefore, the need for a mathematically efficient algorithm is the key for including a valid spatial com-

ponent using large data sets over extensive prediction domains ([256 × 128] or 32,768 prediction locations). Discussion of such algorithms can be found in Hooten (2001) and Wikle et al. (2001), Royle et al. (2001), Wikle (2002).

Validation methods

Validation is critical in a hierarchical modeling project because the complexity of a high dimensional hierarchical GLMM makes it more difficult to get statistical goodness-of-fit measures. Therefore, innovative methods must be devised to evaluate the effectiveness of a hierarchical model. Neter et al. (1996) mention that there are three conventional ways to validate a regression model: (1) collection of new data to check the model and its predictive ability; (2) comparison of results with theoretical expectations, earlier empirical results, and simulation results; (3) use of holdout sample to check the model and its predictive ability. Ideally, one would want to use a seperate but temporally similar data set for assessing the predictive power of the model at locations where data were not originally present. This would be especially helpful since this project is aimed at predicting continuously across a spatial domain. Independent data sets are available for selected MOFEP regions, however the prediction grid chosen for this project does not encompass such regions. It is not feasible to collect another dataset because the presence of vegetation has been affected by temporal variability and disturbance from management practices. Therefore, comparison of predictions with independent data is not an option for model validation.

The comparison of results with theoretical expectations offers promising insight into realistic natural processes. Simulation offers theoretical justification for implementing a spatial component within the model, however it is not a direct validation of the predictions.

The most promising method for model validation may be the third approach suggested by Neter et al. (1996). By withholding a portion of the original data, the results can be validated by the randomly withheld sample. In a dataset with only 216 observations, it is difficult to get a sufficiently large sample to be effective in assessing model accuracy. However, the dataset can be iteratively split into two sets, a model-building set and a validation or prediction set (i.e., cross-validation), for a series of model runs. Upon each iteration the validation set is compared with the predictions gained from modeling with the model-building set. The comparisons are reported in contingency tables and also the independence in predicted probabilities given the true data are tested for significance with a chi-squared test. Zar (1984) provides the following formulation of a χ^2 test statistic:

Table 2. D. glutinosum and D. nudiflorum (Threshold=0.5).

	Real	Predicted		
		P	A	total
D. glutinosum				
	P	18	15	33
	A	10	57	67
	total	28	72	100
D. nudiflorum				
	P	64	7	71
	A	16	13	13
	total	80	20	100

$$\chi^2 = \frac{n\left(\left|f_{1,1}f_{2,2} - f_{1,2}f_{2,1}\right| - \frac{n}{2}\right)^2}{(C_1)(C_2)(R_1)(R_2)} \qquad (10)$$

Where the values are from Table 2. Where "P" refers to presence and "A" to absence. The use of such a test assumes a null hypothesis that the predicted occurrence of a species is independent of the true occurrence of a species. The alternate hypothesis is that the predicted occurrence of a species is associated with the true occurrence of a species. Chi-squared tests are commonly used for assessing the accuracy of binary regression models. The chi-squared test also assumes independent observations which isn't the case here. However, it is assumed that since the spatial dependence is not overly large that the random selection process gives fairly independent samples. Nevertheless, the p-values in this setting are only considered guides and are not used for inferential decisions. It is important to note that such tests applied in similar situations are subject to arbitrary threshold values and although informative, may not always provide adequate validation for models of this type. Some studies have devised certain methods for calibrating such thresholds (e.g., Franklin 1998). This project utilized chi-squared tests in conjunction with boxplots for predicted probabilities versus real occurrence to suggest that a given threshold is reasonable.

The primary benefit of using a rigorous statistical

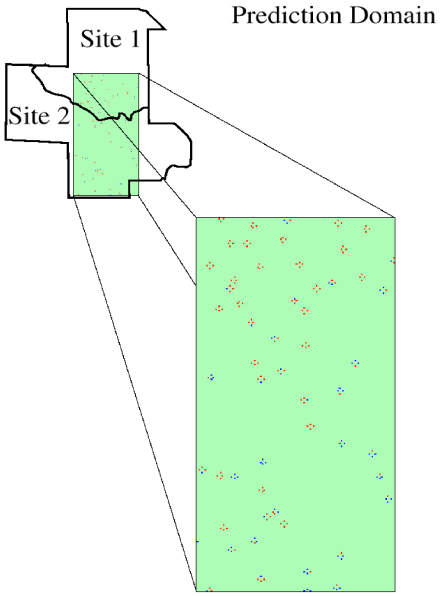


Figure 4. The prediction domain spanning MOFEP sites 1 and 2. Red and blue pixels represent the subplot locations.

approach to modeling is the ability to use a model-based validation. In this hierarchical framework there is a marginal posterior distribution for p_i corresponding to each element of the spatial prediction grid (Figure 5). Just as a grid cell posterior mean can be reported by averaging the posterior predictions for parameters of interest, a measure of error about the posterior mean (standard deviation or variance) can be reported as well. It follows that these measures can be presented in the form of an image and represent the spatial distribution of prediction error on the domain. Such images provide a reasonable approach with which to evaluate model accuracy and add to the overall validation of the model (Royle et al. 2001).

Results

Predictions

An area within sites 1 and 2 (Figure 1) consisting of nearly 330 hectares ($\approx$33,000 grid cells) was chosen to test the model. Of the 2592 MOFEP subplots, 216 fall within the aforementioned area (see Figure 4) and these are utilized in this form of the model. Approximately 32% of the data locations contained *Desmodium glutinosum* while 74% of the sampled locations contained *Desmodium nudiflorum*. The data locations only make up 0.66% of the prediction grid and are therefore very sparse within prediction domain.

By predicting at a discrete and contiguous number of locations (grid cells in this case) over a spatial domain, maps that display information about the posterior distribution can be created. Maps based on predicting the $E(Y_i)$ process, such as those shown in this section, can be viewed as probabilistic maps of species occurrence, such that map intensities can range from 0 to 1 (1 being 100% probability of species occurrence).

The MCMC algorithm (Gibbs sampler) was run for 10,000 iterations after a burn-in period of 2000 iterations. Resulting parameter distributions were summarized in the form of histograms. Figure 7 shows the marginal posterior histograms for the parameters, β, in the model for both species. It can be inferred that the covariates are indeed important factors influencing the occurrence of these species because the distributions for β generally do not overlap zero. In the case of Land Type Association for *Desmodium glutinosum*, a slight overlap of zero is evident. This suggests that the Land Type Association covariate is perhaps less important than the other covariates when modeling this species.

Posterior prediction means, $\hat{p}_i$, were obtained from the Gibbs sampler for each unit in the prediction domain and plotted as an image (Figures 8 and 10). Realizations from the posterior distribution of p_i graphically portray the variablilty in distributions (Figures 9 and 11).

The simultaneous visual analysis of the prediction images and posterior predictions of the β parameters reveals important ecological differences in the response of the two species to certain covariates. Worthy of special attention are the opposite responses of the two species to the relative elevation and variable depth soil covariates.

Validation

Two primary methods were used for assessing the model accuracy: (1) testing for independence between predicted occurrence and real occurrence using cross-validation; (2) presenting maps of prediction error in the form of marginal posterior standard deviations for pixel means.

A threshold value was used to create binary values from the mean probability images so that when the probability is greater than the threshold the species is

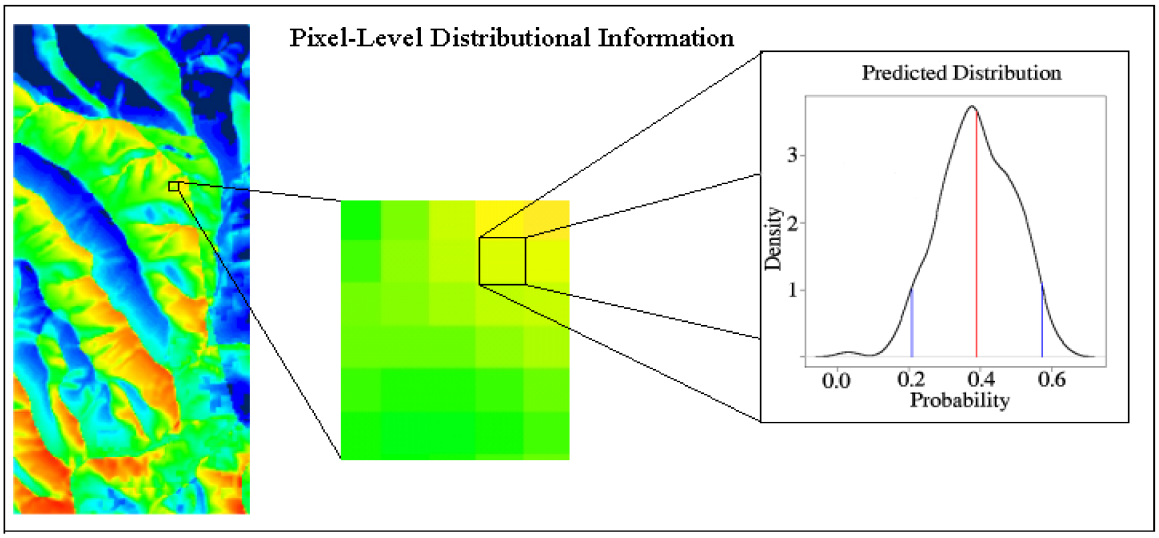
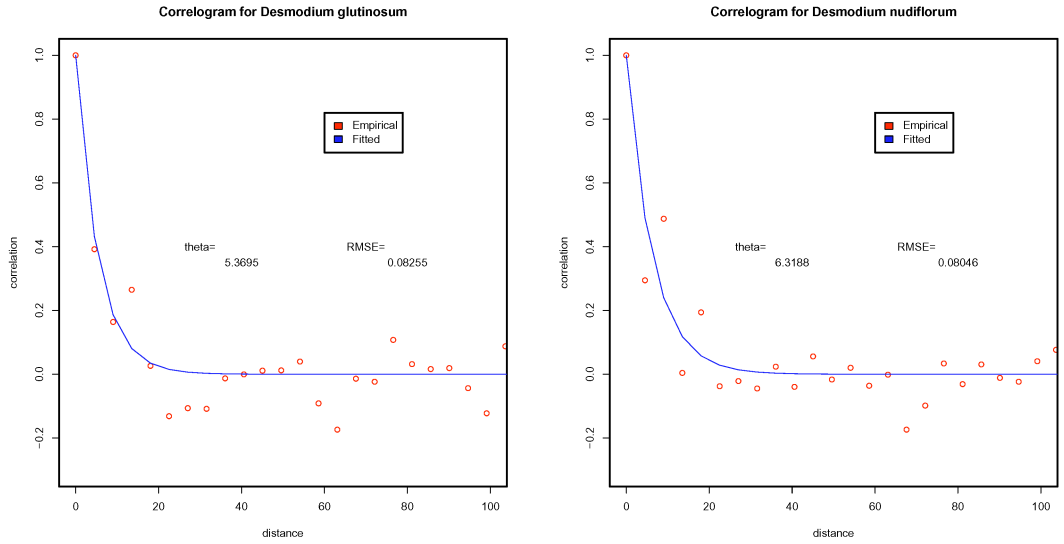


Figure 5. Graphical illustration of the distributional information available at the pixel-level provided by the posterior distribution.

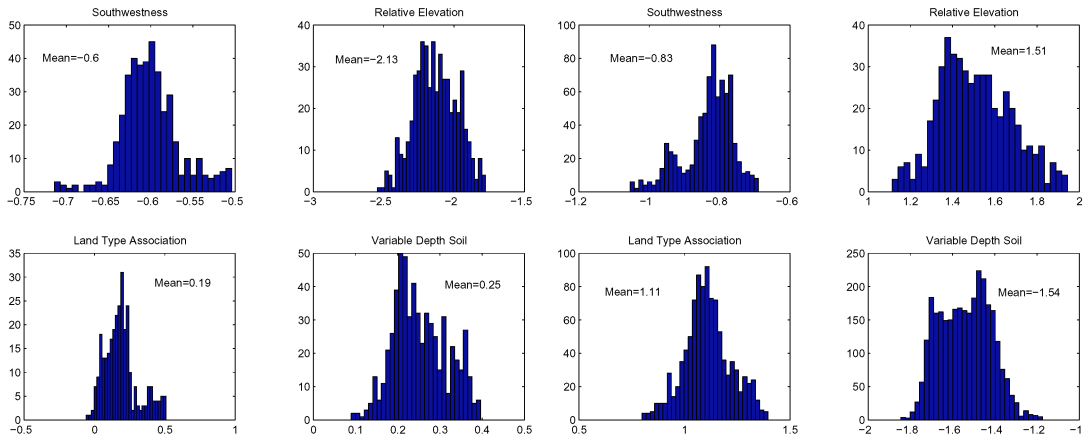


(a) (b)

Figure 6. Empirical correlograms for the two species and exponentially fitted lines with parameter theta and root mean squared error of fitted model.

thought to be present and when it is less than the threshold the species is thought to be absent. Conventionally, a probability threshold of 0.5 is used, however this may not always be the most reasonable threshold for these predictions because the data may not encompass the threshold. Determining if the model is distinguishing between those subplots where the species is present and those where it is absent is sufficient. Therefore chosing a threshold between the two distributions of predictions is the most reasonable approach. Boxplots showing the differentiation of predicted probabilities for *Desmodium glutinosum* and *Desmodium nudiflorum* are shown in Figure 12. Such boxplots suggest that a 50% threshold is a reasonable cutoff for the predicted probabilities in this case.

Approximately 50% of the original data were randomly withheld in an iterative fashion and the model output was aggregated in order to test for independence as discussed in section 2.7. The two-by-two



(a) *D. glutinosum* (b) *D. nudiflorum*

Figure 7. Histograms for β in the model.

Table 3. D. glutinosum

		Predicted		
Threshold = 0.5		P	A	total
	P	18	15	33
Real	A	10	57	67
	total	28	72	100

Table 4. D. nudiflorum

		Predicted		
Threshold =0.5		P	A	total
	P	64	7	71
Real	A	16	13	29
	total	80	20	100

contingency tables for both species are shown below. The chi-squared test for *Desmodium glutinosum* yielded a χ^2 value of 15.30 and a p-value of < 0.0001 (calculated using Table 3). A χ^2 value of 13.63 and a p-value of 0.0002 was found for *Desmodium nudiflorum* (calculated using Table 4).

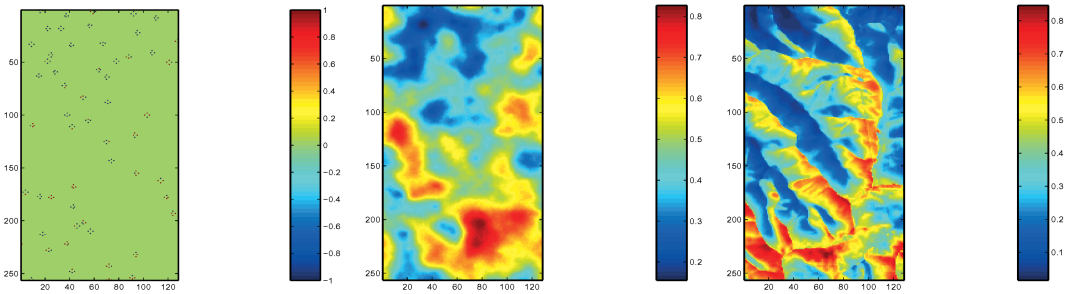
The second part of the validation is based on maps of prediction error presented in Figure 13. These maps, created using the marginal posterior standard deviations for pixel means, offer a rigorous spatially based measure of model accuracy. The results of such mapping efforts are similar for both species, and show that the prediction error is greatest at locations farthest from the data, as expected.

Discussion

Results of the modeling were informative. A pattern of spatial arrangement in the data was evident and expected. As mentioned, this is likely due to the effect of influential spatial processes in environmental covariates. However if none of these covariates are explicitly known on the domain, a pure spatial prediction could be gained from the geographic location and value of the data alone.

Patterns in the predicted images for the posterior mean of the spatial process somewhat resembled that of the images for covariate-based predicted means. This similarity occurs because the spatial process is trying to absorb any spatial correlation in the unknown covariates. Differences between the two predictions illustrate the need for spatially explicit covariates in the model.

The modeled residual spatial random field (η) given in Figures 9 and 11 can be thought of as the "missing covariate(s)" and will ultimately improve the accuracy of the model. Images given in the same figures illustrate that the η-process differs by species as suggested in the exploratory analysis (Figure 6). From an analytical standpoint, it is clear that the process is most evident in neighborhoods about the data. From an ecological standpoint, these neighborhoods likely have an implicit biological meaning. The evaluation and interpretation of a given species η-process could be extensive and is therefore beyond the scope of this project. It is recognized however, that the re-

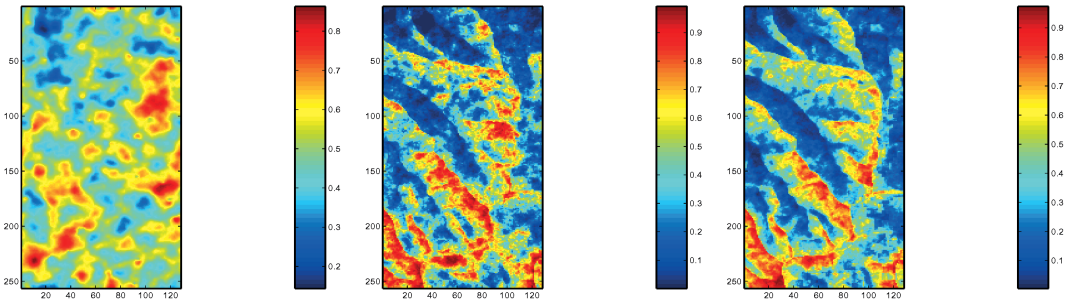


(a) Data location (1 = present, -1 = absent)

(b) Predicted posterior mean with no covariates

(c) Predicted posterior mean with covariates

Figure 8. Data locations and posterior mean with and without covariates for *Desmodium glutinosum*.

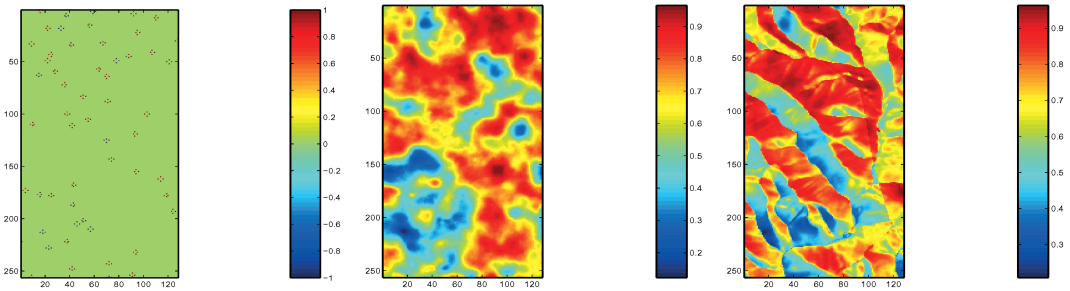


(a) Residual spatial process mean

(b) Posterior realization 1

(c) Posterior realization 2

Figure 9. Residual spatial process and two realizations for *Desmodium glutinosum*.

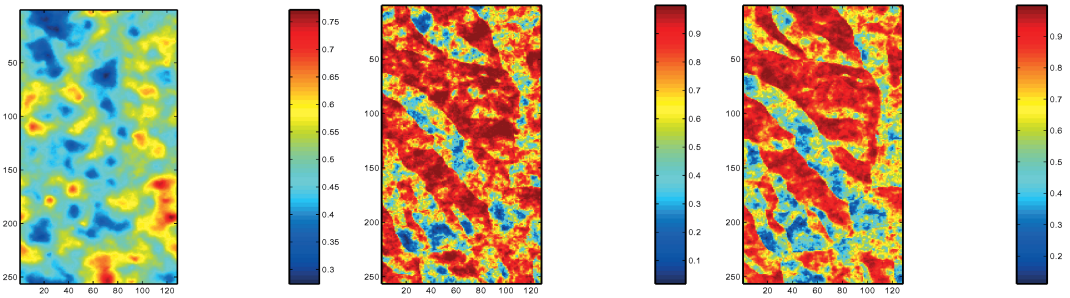


(a) Data location (1 = present, -1 = absent)

(b) Predicted posterior mean with no covariates

(c) Predicted posterior mean with covariates

Figure 10. Data locations and posterior mean with and without covariates for *Desmodium nudiflorum*.

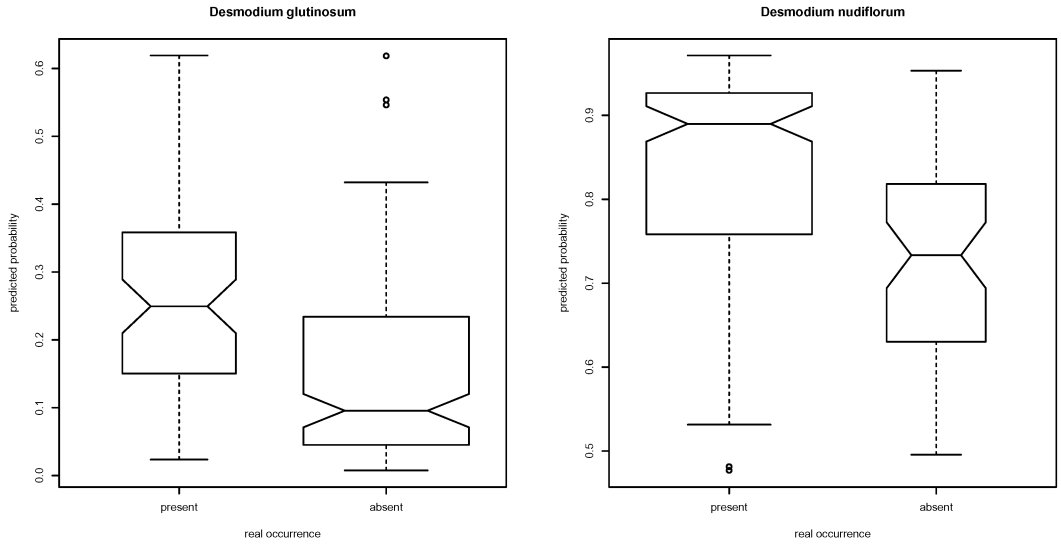


(a) Residual spatial process mean

(b) Posterior realization 1

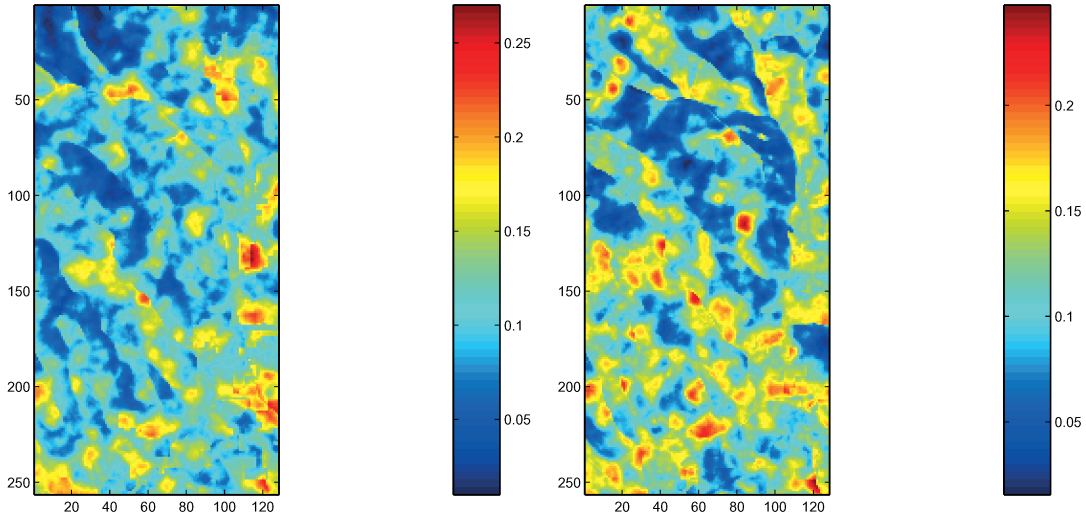
(c) Posterior realization 2

Figure 11. Residual spatial process and two realizations for *Desmodium nudiflorum*.



(a) *Desmodium glutinosum* (b) *Desmodium nudiflorum*

Figure 12. Predicted probabilities vs. real occurrence. Widths are relative to the amount of data while notches represent a 95% confidence interval for the median.



(a) (b)

Figure 13. Standard deviation map for the prediction mean of *D. glutinosum* and *Desmodium nudiflorum*.

sidual process is probably a synthesis of complex interactions, both biological and environmental.

The predicted mean images for each plant graphically depict relationships between the vegetation and environment (Figures 8 and 10). Mathematically, the grid cell intensities of such maps are based on rigorous relationships and represent the best possible predictions available. Ecologically, the predictions are based on a combination of environmental factors and surrogates for biological factors, making them a very complete and robust reflection of the species response on a continuous spatial domain. Theoretically, this response might be interpreted as the spatial prediction of vegetation occurence or perhaps suitable habitat. It is suggested that the interpretation be flexible so it may individually suit a specific project.

Predicted maps of the mean posterior probabilities take the form of the covariates. Therefore the predicted images (represented by a grid) could be utilized in a GIS for an endless number of applications. Additionally, output from the proposed model could be used as covariates in another model (such as a model predicting wildlife habitat, fire susceptibility, forage, …). Importantly, measures of uncertainty related to these covariates are then available (such information has traditionally been unavailable).

The distributional nature of the posterior predictions is difficult to portray with a map of prediction means because the mean response of a species at a given time and location is represented by a single value. However, if a realization is taken from each pixel distribution on a contiguous domain, the result is one of infinitely many maps of possible real species response. Such a map, as presented in Figures 9 and 11 can be thought of as a snapshot of a stochastic process. These maps have much to offer as foundations for theoretical projects investigating patch dynamics, realistic vegetation patterning, and environmental simulation.

The purpose of this project is to provide methods that are robust enough to be applied to a number of different applications. Of course a specific scenario was chosen to test and present the proposed methods and as such the validation of this model applied to the specific scenario has the purpose of illustrating certain advantages of the technique.

The chi-squared tests provided a way to evaluate the ability of the model to distinguish between those areas where a species should be present or absent. The tabular results of the tests showed that the specified threshold of 0.5 is a reasonable classification criteria for these data. However, as mentioned earlier, statistics that are dependent on such analyst-based classifications are subjective. However, they still provide a way to test the differentiating power of the model. Chi-squared tests used here prove that the model is distinguishing between presence and absence for both species. Resulting p-values allow the rejection of a null hypothesis that predictions and actual data are independent.

It was mentioned earlier that one advantage of using a statistical model over a more *ad hoc* approach, is the ability to provide a model-based estimate of prediction error. In a spatial setting this information is especially valuable for it can provide an error estimate at each prediction location. In the form of a map, this information can be compared to the location of field data, covariate effects, residual spatial effects, and ultimately the predictions. The practice of reporting error for spatially explicit data is becoming more popular (e.g. Justice and Running 1998) and it is convenient that the approach presented here automatically provides a way to incorporate such information.

It is common for maps of marginal posterior standard deviations to illustrate that the model is good at predicting near the locations of the actual data, and gains some error as it is applied further from the data (Royle et al. 2001). The maps presented in Figure 13 illustrate this expected patterning of prediction error. Overall, the standard deviations of the posterior mean predictions are quite small and suggest that the model is doing a good job predicting occurrence for both species (especially near the locations of the data). Such maps have the potential to be quite useful in other models where it is appropriate to specify uncertainty related to prior information. Ultimately this chain of utilizing the output from one model to formulate priors or data in new models could prove to be useful in modeling more complex systems.

Conclusions

The purpose of this project was to develop and test a robust methodology for realistically accounting for uncertainties in modeling natural processes at a landscape level. Specifically, this problem was considered from a hierarchical Bayesian perspective. This approach has allowed the incorporation of a critical spatial component into a generalized linear mixed model. Similar modeling methodologies have been proposed, however many are currently focused on small spatial domains because of mathematical and computational limitations. The formulations presented here are aimed at predicting natural processes on larger domains.

The methodology presented here was tested using part of a landscape level dataset collected in the Missouri Ozarks as part of the Missouri Ozark Forest Ecosystem Project. Relationships between vegetation occurrence and known environmental features were investigated in an exploratory analysis. Residual spatial random effects in the data were observed and extensively analyzed via several methods. Simulations were conducted to insure that the proposed methods of informing a spatial prior were indeed reasonable.

An approach for constructing generalized linear models was modified to include a spatial parameter. The coding of this model was optimized using a spectral transformation that invariably simplified the formulation of the marginal posterior distributions. The integrations required to analytically find the necessary joint distributions were intractible, therefore the model was implemented using an iterative process known as Markov Chain Monte Carlo that allowed samples to be drawn from the posterior distribution through a Gibbs sampler.

Parameter distributions, posterior mean prediction images, maps of the residual spatial process, and realizations were provided in order to show the range of information available through this approach. Different methods for validating such models were presented and applied to the modeling results for *Desmodium glutinosum* and *Desmodium nudiflorum*. Both validation methods suggested that this model is performing quite well overall and is especially accurate at predicting near the original data.

As mentioned earlier, probabilistic results of the proposed model using the example dataset could be viewed as the spatial prediction of a species on a continuous domain at the time of data collection or as the realized niche of the given species. It was not the intent of the project to determine which description is most appropriate but to provide the most robust and informative predictions about the ecological process.

Finally, these methods provide a partial answer to a problem alluded to by Clark et al. (2001), where they state that the failure to appropriately account for the various sources of ecological uncertainty will result in misleading inference and even though recent techniques in atmospheric science are available, ecology seems to be a late adopter. The intent of this paper was not to make inference about the biology of a certain species, however the two specific examples discussed are meant to show how the methodology could be applied in an ecological setting. The next step might be for ecologists to apply this or a combination of this and other methods to a specific area of interest or concern in order to make management decisions or scientific observation.

Acknowledgements

We thank the following parties: NASA and the University of Montana for funding through the EOS Training Center Project. UMC Quantitative Silviculture Laboratory for computational machinery and software. Missouri Department of Conservation for unparalleled involvement in a Missouri landscape level project. Jennifer Grabner and Tim Nigh for providing field data and many helpful comments and suggestions. John Krystansky and B.J. Gorlinski for landscape grid and GIS data as well as many helpful comments and suggestions. Wikle's research was supported by a grant from the US Environmental Protection Agency's Science to Achieve Results (STAR) program, Assistance Agreement No. R827257-01-0.

References

Albert J. and Chib S. 1993. Bayesian analysis of binary and polychotomous response data. Journal of the American Statistical Association 88: 669–679.

Augustin N., Mugglestone M. and Buckland S. 1996. An autologistic model for the spatial distribution of wildlife. Journal of Applied Ecology 33: 339–347.

Beers T., Dress P. and Wensel L. 1966. Aspect transformation in site productivity research. Journal of Forestry 64: 691–692.

Besag J. 1974. Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistics Society 36: 192–236.

Borcard D., Legendre P. and Drapeau P. 1992. Partialling out the spatial component of ecological variation. Ecology 73: 1045–1055.

Brookshire B., Jensen R. and Dey D. 1997. The Missouri Ozark Forest Ecosystem Project: past, present, and future. In: Brookshire B. and Shifley S. (eds), Proceedings of the Missouri Ozark Forest Ecosystem Project symposium: an experimental approach to landscape research, number GTR NC-193. US Department of Agriculture, Forest Service, North Central Forest Experiment Station, St. Paul, Minnesota, USA, pp. 1–25.

Cherrill A., McClean C., Watson P., Tucker K., Rushton S. and Sanderson R. 1995. Predicting the distributions of plant species at the regional scale: A hierarchical matrix model. Landscape Ecology 10: 197–207.

Clark J., Carpenter S., Barber M., Collins S., Dobson A., Foley J. et al. 2001. Ecological Forecasts: An emerging imperative. Science 293: 657–660.

Clayton D. 1997. Markov Chain Monte Carlo in Practice chapter 16, Generalized linear mixed models. Chapman & Hall, New York, New York, USA, pp. 276–301.

Cressie N. 1993. Statistics for Spatial Data: Revised Edition. John Wiley and Sons, New York, New York, USA.

Davis F. and Goetz S. 1990. Modeling vegetation pattern using digital terrain data. Ecology 4: 69–80.

Day F. and Monk C. 1974. Vegetation pattern on a southern appalachian watershed. Ecology 55: 1064–1074.

Diggle P., Tawn J. and Moyeed R. 1998. Model-based geostatistics (with discussion). Applied Statistics 47: 299–350.

Erickson R. 1943. Population size and geographical distribution of *Clematis fremontii* var. *riehlii*. Annals of the Missouri Botanical Garden 30: 63–68.

Ernst W. 1978. Discrepancy between ecological and physiological optima of plant species: a re-interpretation. Oecologica Plantarum 13: 175–188.

Forman R. and Godron M. 1986. Landscape Ecology. John Wiley and Sons, New York, New York, USA.

Franklin J. 1998. Predicting the distribution of shrub species in southern California from climate and terrain-derived variables. Journal of Vegetation Science 19: 733–748.

Gilks W., Richardson S. and Spiegelhalter D. (eds) 1997. Markov Chain Monte Carlo in Practice. Chapman & Hall, New York, New York, USA.

Gleason H. 1926. The individualistic concept of plant association. Bulletin of the Torrey Botanical Club 53: 7–26.

Grabner J. 1996. MOFEP botany: pre-treatment sampling and data management protocol. MDC unpublished report.

Grabner J. 2000. Ground layer vegetation in the Missouri Ozark Forest Ecosystem Project: Pre-treatment species composition, richness, and diversity. In: Brookshire B. and Shifley S. (eds), Missouri Ozark Forest Ecosystem Project: Site history, Soils, Landforms, Woody and Herbaceous Vegetation, Down Wood, and Inventory Methods for the Landscape Experiment, number GTR NC-208. US Department of Agriculture, Forest Service, North Central Forest Experiment Station, St. Paul, Minnesota, USA, pp. 107–131.

Grabner J., Larsen D. and Kabrick J. 1997. An analysis of MOFEP ground flora: pre-treatment conditions. In: Brookshire B. and Shifley S. (eds), Proceedings of the Missouri Ozark Forest Ecosystem Project symposium: an experimental approach to landscape research, number GTR NC-193. US Department of Agriculture, Forest Service, North Central Forest Experiment Station, St. Paul, Minnesota, USA, pp. 169–197.

Guisan A., Theurillat J. and Kienast F. 1998. Predicting the potential distribution of plant species of plant species in an alpine environment. Journal of Vegetation Science 9: 65–74.

Guisan A. and Zimmermann N. 2000. Predictive habitat distribution models in ecology. Ecological Modelling 135: 147–186.

Hoeting J., Leecaster M. and Bowden D. 2000. An improved model for spatially correlated binary responses. Journal of Agricultural, Biological, and Environmental Statistics 5: 102–114.

Hooten M. 2001. Modeling and mapping the distribution of legumes in the Missouri Ozarks. Master's Thesis, University of Missouri, Columbia, Missouri, USA.

Justice C. and Running S. 1998. The Moderate Resolution Imaging Spectroradiometer (MODIS): land remote sensing for global change research. IEEE Transactions on Geoscience and Remote Sensing 96.

Krystansky J. and Nigh T. 2000. Missouri Ecological Classification Project, ELT Model.

Legendre P. 1993. Spatial autocorrelation: Trouble or new paradigm? Ecology 74: 1659–1673.

Lichstein J., Simons T., Shriner S. and Franzreb K. 2002. Spatial autocorrelation and autoregressive models in ecology. Ecological Monographs 72: 445–463.

McCulloch C. 1994. Maximum likelihood variance components estimation for binary data. Journal of the American Statistical Association 89: 330–335.

Meinert D., Nigh T. and Kabrick J. 1997. Landforms, geology, and soils of the MOFEP study area. In: Brookshire B. and Shifely S. (eds), Proceedings of the Missouri Ozark Forest Ecosystem Project symposium: an experimental approach to landscape research, volume GTR. US Department of Agriculture, Forest Service, North Central Forest Experiment Station, St. Paul, Minnesota, USA, pp. 169–197.

Neter J., Kutner M., Nachtsheim C. and Wasserman W. 1996. Applied Linear Statistical Models. 4th edn. WCB/McGraw-Hill, Boston, Massachusetts, USA.

Pielou E. 1977. Mathematical Ecology. John Wiley and Sons, New York, New York, USA.

Ripley B. 1981. Spatial Statistics. John Wiley and Sons Inc, New York, New York, USA.

Royle J., Link W. and Sauer J. 2001. *Predicting Species Occurences: Issues of Scale and Accuracy*, chapter Statistical mapping of count survey data. Island Press, Covello, California, USA.

Shumway R. and Stoffer D. 2000. Time series analysis and its applications. Spring-Verlag, New York, New York, USA.

Smith P. 1994. Autocorrelation in logistic regression modeling of species' distributions. Global Ecology and Biogeography Letters 4: 47–61.

Turner M. 1989. Landscape Ecology: The effect of pattern on process. Annual Review of Ecology and Systematics 20: 171–197.

Whittaker R. 1956. Vegetation of the Great Smoky Mountains. Ecological Monographs 26: 1–80.

Wikle C. 2002. *Spatial Cluster Modeling*, chapter Spatial Modelling of Count Data: A Case Study in Modelling Breeding Bird Survey Data on Large Spatial Domains. Chapman and Hall/CRC, pp. 199–209.

Wikle C., Milliff R., Nychka D. and Berliner L. 2001. Spatio-temporal hierarchical Bayesian modeling: Tropical ocean, surface winds. Journal of the American Statistical Association 96: 382–397.

Zar J. 1984. Biostatistical Analysis. 2nd edn. Prentice-Hall, Englewood Cliffs, New Jersey, USA.

Zimmermann N. and Kienast F. 1999. Predictive mapping of alpine grasslands in Switzerland: species versus community approach. Journal of Vegetation Science 10: 469–482.