geographical analysis

Geographical Analysis (2013) **45**, 419–440

Spatial Distribution of Trees and Landscapes of the Past: A Mixed Spatially Correlated Multinomial Logit Model Approach for the Analysis of the Public Land Survey Data

Eun-Hye Yoo[1], Bruce W. Hoagland[2], Guofeng Cao[3], Todd Fagin[2]

[1]Department of Geography, University at Buffalo, Buffalo, NY, USA, [2]Oklahoma Natural Heritage Inventory and Department of Geography and Environmental Sustainability, University of Oklahoma, Norman, OK, USA, [3]Department of Geography, University of Illinois at Urbana-Champaign, Urbana, IL, USA

Public Land Survey (PLS) data have been widely used in landscape studies of forest and woodlands in the pre- and early-European-settled Midwestern and Western United States. We aim to reconstruct presettlement forest vegetation at a finer spatial resolution than available from the PLS data using environmental covariates (slope, aspect, geology, and soil type) and the spatially correlated structure of witness tree data. To accommodate various data obtained from multiple sources while explicitly taking into account their spatial structures, we adopt a mixed spatially correlated multinomial logit model within the framework of a generalized linear mixed model. The application of the proposed model is illustrated using the three most abundant tree taxa from PLS data in the Arbuckle Mountains of south-central Oklahoma. To assess the influence of each source of information on the spatial prediction, we considered four variant multinomial/spatial models and evaluated their relative predictive power using a validation technique. The probabilistic information about the spatial distribution of tree species obtained from different models reveals the need to integrate information about witness tree data as well as environmental covariates, and the nature of tree species; that is, a tendency to cluster in space to share environmental conditions in the reconstruction of the presettlement forest vegetation surface.

Introduction

Public Land Survey (PLS) data from the General Land Office (GLO) have been widely used to study the composition and structure of forests and woodlands in the pre- and early-European-settled Midwestern and Western United States (Galatowitsch 1990; Schulte and Mladenoff 2001).

Correspondence: Eun-Hye Yoo, Department of Geography, University at Buffalo, Buffalo, NY 14261-0055
e-mail: eunhye@buffalo.edu

Submitted: March 12, 2012. Revised version accepted: December 4, 2012.

doi: 10.1111/gean.12018

PLS data consist of two components useful for landscape analysis: township plats, on which surveyors mapped general landscape characteristics, and surveyor notes, in which surveyors recorded information about so-called witness trees encountered along survey lines (White 1983; Whitney and DeCant 2001). Although the earliest research using PLS data investigated spatial changes in land cover types on the GLO plats (Fassett 1944; Curtis 1956), the witness tree records have been the primary focus of many subsequent ecological studies (Brown 1998; He et al. 2000; Manies and Mladenoff 2000; Cogbill, Burk, and Motzkin 2002; Wang 2007; Fagin and Hoagland 2011).

Although researchers have utilized PLS data for analysis of the spatial patterns of tree species, quantifying the areal extent of selected woody taxa from these data is problematic because of the coarse sampling structure of the data. Surveyors recorded witness trees at 0.8 km (0.5 mile) intervals, which correspond to the intersections of section lines and quarter-section points. In an effort to overcome these spatial limitations in PLS data, researchers have attempted to convert discrete PLS point data into continuous surfaces for more accurate landscape scale studies (Batek et al. 1999; He et al. 2000; Manies and Mladenoff 2000; Wang and Larsen 2006; Wang 2007; Fagin and Hoagland 2011).

In general, two broad approaches have been taken. The first involves analyzing the relationship between individual woody plant taxa and a set of environmental covariates using various environmental regression models (He et al. 2007; Franklin and Miller 2009; Fagin and Hoagland 2011). Implicit in this approach is the assumption that sets of relevant environmental covariates are available at the proper spatial and temporal resolution to analyze historical species–environment relationships adequately. In practice, such data may not be readily available, and those readily available may be omitted in the model fitting process. In addition, the relationship between environmental covariates and tree species occurrences often is assumed to be consistent over the entire study region, whereas its influences may be regionally or spatially heteroscedastic.

The second approach focuses solely on the spatial patterning of the witness tree data independent of any environmental covariates that may or may not influence the distributional patterns (Delcourt and Delcourt 1996; Manies and Mladenoff 2000; Wang and Larsen 2006). In this approach, the spatial autocorrelation of tree species occurrences often is the interest of the study, particularly with respect to the spatial scale at which an analysis is conducted. The discrepancy between the scale of the data (point support) and the scale of analysis (nonpoint support) has drawn the attention of researchers, as is evidenced by the development of spatial statistical models designed to resolve scale issues utilizing spatial autocorrelation in tree species occurrences (He et al. 2000; Manies and Mladenoff 2000; Friedman, Reich, and Frelich 2001; Wu 2004; Wang and Larsen 2006; Yoo and Trgovac 2011). However, spatial autocorrelation alone may not be sufficient to reconstruct fine-scale interaction between environmental conditions and tree species occurrences (Manies and Mladenoff 2000; Rathbun and Black 2006).

In this article, we propose a model that incorporates both environmental covariates and spatial structure (which may act as a surrogate for various biotic covariates in the spatial prediction model for vegetation) to reconstruct presettlement forest vegetation at a finer spatial resolution than available from the PLS data alone. Given that the genus and/or species name of a tree recorded at each point takes one of multiple categories, a multinomial generalized linear model (GLM) is a natural choice for analysis. The use of the multinomial model in the spatial distribution of tree species allows us to generalize existing models in which tree data are modeled individually, species by species, ignoring the presence or absence of other tree species. The proposed multinomial model simultaneously takes into account more than one tree species in a

consistent manner. A spatial dependence structure of tree taxa occurrences and the spatial structure of environmental covariates are explicitly incorporated within a framework of generalized linear mixed models (GLMMs) in the absence of independence.

In practice, various ways exist to accomplish the GLMM modeling (Bolker et al. 2009). For example, Diggle, Tawn, and Moyeed (1998) employ GLMMs for spatially dependent non-Gaussian variables observed in a continuous region and present a formal framework for parameter inference in a geostatistical setting using the Markov chain Monte Carlo (MCMC) technique. We follow the same paradigm of incorporating the latent spatial process through random effects, but we achieve computational simplicity by using a direct approximation approach (Cao, Kyriakidis, and Goodchild 2011). The covariance function of the latent spatial process is directly approximated by a mixture of the spatial covariance functions of observed data, and a posterior density function is represented via a multinomial logit linear combination of these approximated covariances. Furthermore, we extend the model of Cao, Kyriakidis, and Goodchild (2011) to include auxiliary variables to account for the influence of environmental covariates on the witness tree data.

We illustrate our approach using the three most abundant tree taxa in the 1870's PLS data from the Arbuckle Mountains of south-central Oklahoma: post oak (*Quercus stellata*), black oak (*Quercus velutina*), and elm (*Ulmus* spp.). We assess the effects of environmental covariates and spatially dependent structure on the witness tree model by comparing spatial predictions obtained from four models: (1) a multinomial logit model (nonspatial environmental regression, Model I); (2) an indicator kriging model (a traditional geostatistical model without environmental covariates, Model II); (3) a mixed spatially correlated multinomial logit model without environmental covariates (a GLMM without predictors, Model III); and (4) a mixed spatially correlated multinomial logit model with covariates (an environmental regression model with spatially correlated random effects, Model IV). The covariates used for the two environmental regression models include soils, geological substrates, aspect, and slope. The goal of model comparisons is twofold: to test the effect of spatial autocorrelation of the tree species occurrences by comparing Model I and the other three spatial models, and to assess the effects of environmental covariates on model performance by comparing purely spatial models (Models II and III) with environmental regression models (Models I and IV). Particularly, the comparison between Models III and IV allows us to measure a direct impact of environmental conditions on a spatial model.

For a model assessment, we use cross-validation, where a randomly selected subset (10%) of observed data is temporarily discarded and the class occurrences at the same locations are predicted using the remaining sample data. The cross-validation exercise is iteratively performed, and its result is summarized by the correct classification rate.

Methods

This section opens with a description of the study area, with a consideration of climate, soil, and potential natural vegetation. We then provide an overview of the components of our modeling approach—a mixed spatially correlated multinomial logit model—followed by a description of model fitting and prediction.

Study area and data set

The Arbuckle Mountains constitute a spatially heterogeneous region that increases in elevation from approximately 229 m to 411 m along an east-west axis (Dale 1956) (Fig. 1). The Arbuckle

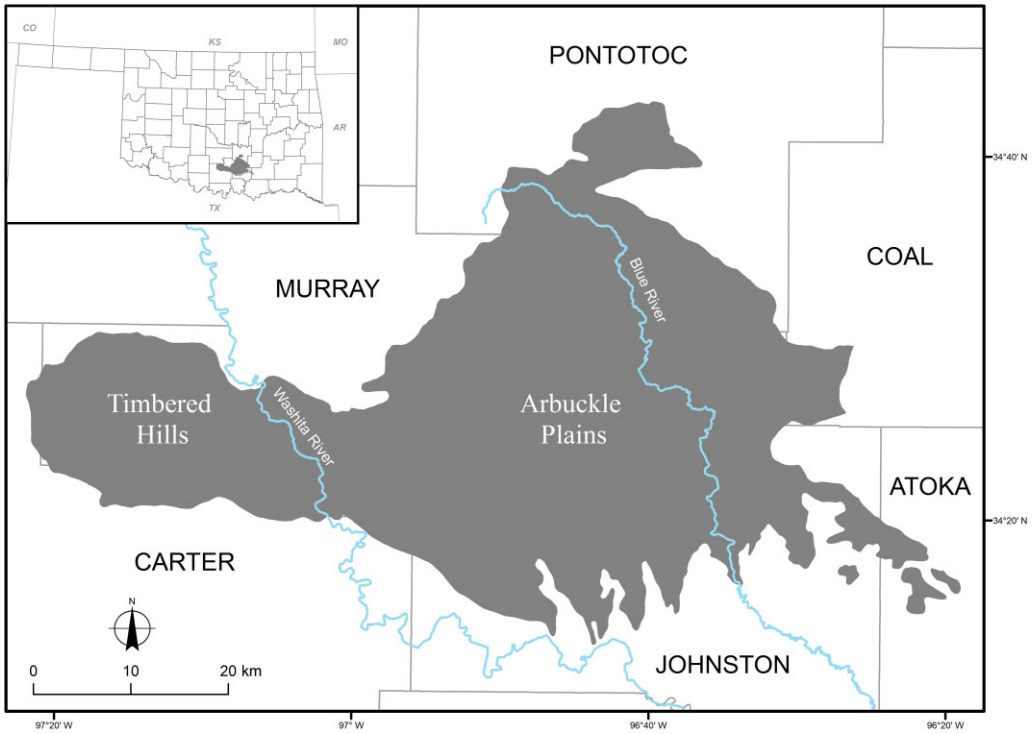

Figure 1. The study area of the Arbuckle Mountains encompasses the south-central part of the state of Oklahoma, United States.

Table 1 Description of Environmental Covariates

	Type	Interval/Categories	Source
Slope (X_1)	Continuous	[0, 58]	USGS
Aspect (X_2)	Categorical	9	USGS
Geology (X_3)	Categorical	7	USGS
Soil (X_4)	Categorical	5	NRSC STATSGO

Mountains were formed in conjunction with the Ouachita orogeny in the Pennsylvanian era. The region experienced considerable faulting and folding, which produced an anticline and exposed late Cambrian to middle Mississippian limestone sedimentary rocks (Fairchild, Hanson, and Davis 1990; Suneson 1997). Precambrian rhyolite and granite form extensive outcrops, though the largest of these are in the eastern portion of the study area (Ham 1969; Fairchild, Hanson, and Davis 1990).

As evidenced in earlier studies (Rathbun and Black 2006; He et al. 2007; Fagin and Hoagland 2011), environmental conditions to which tree species respond play an important role when reconstructing the spatial distribution of forest vegetation in a landscape. In our study, we selected four predictors that have varying degrees of influence on each tree species. Table 1 lists the covariates used in the current study and provides a brief description of each variable (see Fig. 2(b)–(e) for their spatial distributions). While each environmental covariate has a nontrivial

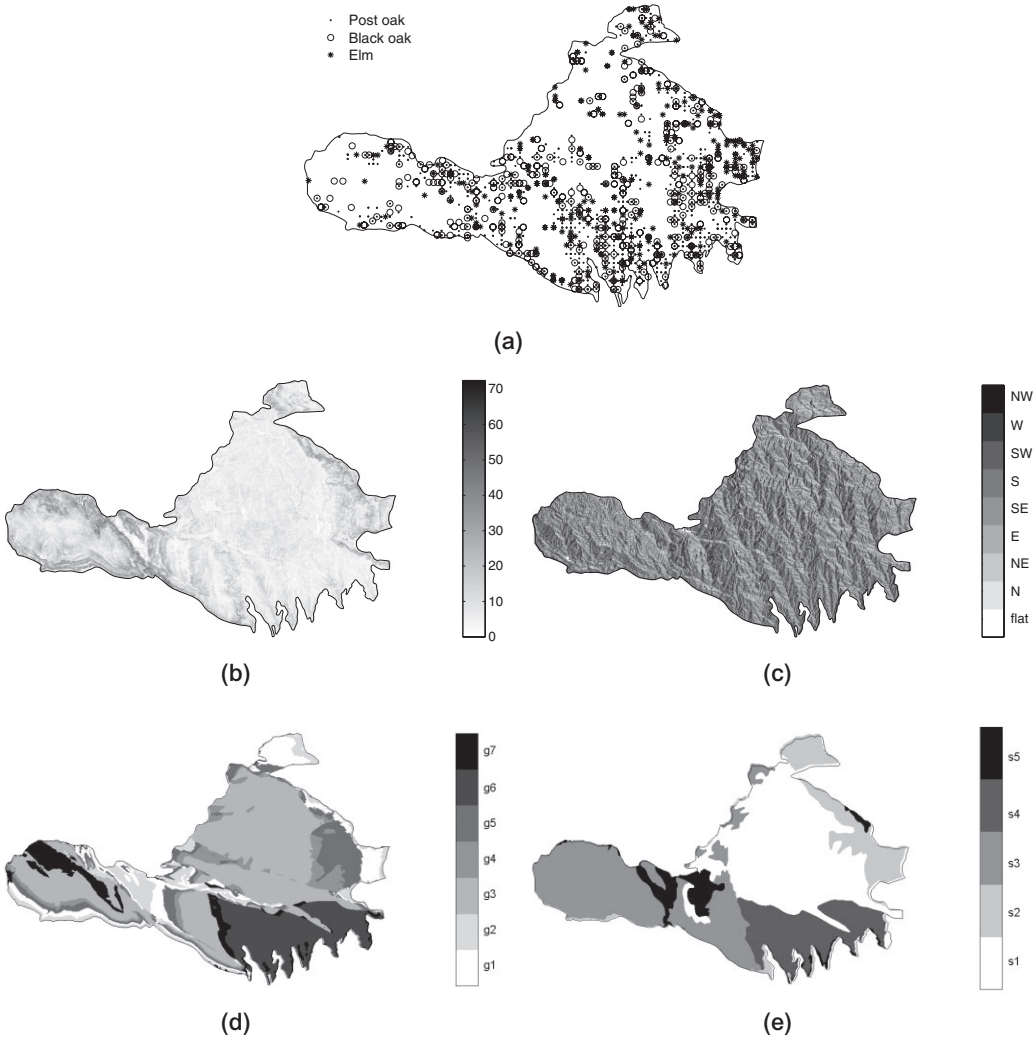

Figure 2. (a) Survey locations of the three most abundant tree species: post oak, black oak, and elm. (b) Slope at 30 × 30 m spatial resolution. (c) Aspect: N, NE, E, SE, S, SW, W, NW, flat. (d) Geology types: shale/limestone (g1), shale (g2), limestone (g3), Oil Creek sandstone (g4), Bromide sandstone (g5), granite (g6), other type (g7). (e) Soil types: Shidler-Scullin-rock outcrop-Lula-Claremore (s1), Shidler-rock outcrop (s2), Kiti-rock outcrop (s3), Chigley-Agan-rock outcrop (s4), Reinach–McLain–Dale (s5).

relation with the selected tree taxa, the set of four does not represent the complete range of abiotic factors important for tree distributions.

The predominant surface formations in Oklahoma are Permian and Pennsylvanian sandstones, making the Arbuckle Mountains a calcareous island in a sea of sandstone. The limestone and dolomite substrates, which constitute 69% of the surface rocks, produce clay soils that foster calciophilic species such as black prairie clover (*Dalea frutescens*), Ashe juniper (*Juniperus ashei*), seep muhly (*Muhlenbergia reverchonii*), and bastard oak (*Quercus sinuata* var. *breviloba*). Shallow soils characterize areas where granite and rhyolite are common (16.4%).

Although the soils are derived from either calcareous or igneous substrates, they share one feature: shallow depth. The most extensive soil type in the Arbuckle Mountains is the Shidler-Scullin-Lula-Claremore-rock outcrop complex, a silty clay loam that covers the greatest aerial extent (37.7%). Depth averages 20.5 cm, with slightly acidic to mildly alkaline pH. It occurs primarily on fractured limestone on 2%–8% slopes, and individual patches range from 16.2 to 121.4 ha (Bogard 1973; Burgess 1977; Watterson, Bogard, and Moebius 1984). The Kiti-rock outcrop complex is a clay loam soil that occupies 27.4% of the study area and consists of moderately alkaline loam of very shallow, well-drained soils. Patch size ranges from 40 to 242.8 ha on steep slopes (8%–20%) (Bogard 1973; Burgess 1977). The Chigley-Agan-rock outcrop complex consists of gravelly to sandy loam, 5%–12% slope over conglomerate. It constitutes 15.7% of soils in the Arbuckle Mountains in patches ranging from 2.0 to 20.2 ha. Soil pH ranges from medium acidic to neutral (Burgess 1977; Watterson, Bogard, and Moebius 1984). The remaining soil types occur in small patches scattered throughout the study area. The most extensive of these are bottomland soils derived from alluvium along large streams. The Reinach–McLain–Dale association constitutes the majority of these soils and occurs on the floodplains of the Washita and Blue rivers. These are sandy-loamy soils that are deep and nearly level, and that are neutral to slightly alkaline (Burgess 1977).

The slope and aspect data at each survey location were derived from a digital elevation model with a spatial resolution of 30 × 30 m. Although the elevation range of the Arbuckle Mountains is insufficient for true altitude effects, the spatial complexity of the locale results in the production of numerous microhabitats. These deep ravines and closed canyons foster many plant species that prefer mesic conditions. Likewise, the rugged topography creates natural firebreaks that allow fire-intolerant plant species to thrive.

A mixed spatially correlated multinomial logit model

Information extracted from the witness tree data includes point-to-plant distance (from the intersection of section lines or quarter-section points), bearing from the point to plant, diameter of the tree, and an identification of the tree by common name. In most cases, the common names can be cross-referenced with current taxonomic resources to determine the proper genus and/or species name. For modeling purposes, we use only the identification of the tree species and their locations from the GLO data. The species name assigned to each witness tree is represented by a categorical random variable (RV), $C(\mathbf{u})$, which may take one of several discrete values $k \in \{1, 2, 3, 4\}$ depending on the species name recorded at a survey site $\mathbf{u}$ in the study domain D. The three most abundant tree taxa—post oak, black oak, and elm—are coded as 1, 2, and 3, respectively, and other tree species are coded as 4.

A set of species labels of witness tree data obtained at $n = 2{,}561$ survey locations is viewed as a realization of the joint RVs $\{C(\mathbf{u}_i), i = 1, \ldots, n\}$. That is, $C(\mathbf{u}_i)$ becomes an indicator variable $\delta_k(\mathbf{u}_i)$ that takes the value 1 if the survey record indicates the kth tree species occurrence ($C(\mathbf{u}_i) = k$) at the ith survey site, and 0 otherwise. Let $\pi_k(\mathbf{u}_i)$ denote the probability, *Prob*, of the kth tree species occurrence at the ith survey site,

$$\pi_k(\mathbf{u}_i) = Prob\{C(\mathbf{u}_i) = k\}, \quad k = 1, \ldots, 4. \tag{1}$$

Assuming that the response categories are mutually exclusive and collectively exhaustive, the sum of the probabilities $\pi_k(\mathbf{u}_i) > 0$ for each category adds to 1; that is, $\sum_{k=1}^{4} \pi_k(\mathbf{u}_i) = 1$. The probability distribution of RV $C(\mathbf{u}_i)$ is given by the multinomial distribution with parameter m and

the vector of corresponding response probabilities $\boldsymbol{\pi}(\mathbf{u}_i) = [\pi_k(\mathbf{u}_i), k = 1, \ldots, 4]$; that is, $C(\mathbf{u}_i) \sim$ *Multinomial*$(m, \boldsymbol{\pi}(\mathbf{u}_i))$ with m independent observations $m = \sum_{k=1}^{4} m_k = 1$ where $m_k = 1$, $m_{k'} = 0$, $\forall k, k' = 1, \ldots, 4, k \neq k'$. The probability mass function of such a multinomial RV is given as follows (Agresti 2007):

$$Prob\{m_1, m_2, m_3, m_4\} = \frac{m}{m_1! \cdots m_4!} [\pi_1(\mathbf{u}_i)]^{m_1} \cdots [\pi_4(\mathbf{u}_i)]^{m_4}. \tag{2}$$

Perhaps the simplest approach to model such a tree species occurrence probability as given in equation (1) is to establish a linear relationship between non-Gaussian-distributed response variables and environmental covariates via a link function within a GLM framework. Logistic regression, in particular, is one of the best established frameworks to describe how tree species occurrence probabilities depend on a $(1 \times q)$ vector of environmental covariate values $\mathbf{x}_i = [x_l(\mathbf{u}_i), l = 1, \ldots, q]$ at a survey site $\mathbf{u}_i$ (Austin and Cunningham 1981; Margules, Nicholls, and Austin 1987; Franklin and Miller 2009). We assume that the log-odd $\eta_k(\mathbf{u}_i)$ of each tree species occurrence relative to a baseline or reference category (other type, $k = 4$) may be described by a linear model as follows:

$$\eta_k(\mathbf{u}_i) = \log\left(\frac{\pi_k(\mathbf{u}_i)}{\pi_4(\mathbf{u}_i)}\right) = \alpha_k + \mathbf{x}_i \boldsymbol{\beta}_k, \quad k = 1, 2, 3, \tag{3}$$

where the parameter α_k is an intercept, and $\boldsymbol{\beta}_k = [\beta_{kl}, l = 1, \ldots, q]^T$ is a $(q \times 1)$ vector of regression coefficients for the kth tree species associated with q covariates. This model is analogous to a logistic regression specification, except that the probability distribution of the response is multinomial instead of binomial, which yields more than one equation for each tree species. This multinomial logit model also may be written in terms of the probability of species occurrences instead of log-odds as follows:

$$\pi_k(\mathbf{u}_i) = \frac{\exp\{\eta_k(\mathbf{u}_i)\}}{1 + \sum_{k'=1}^{3} \exp\{\eta_{k'}(\mathbf{u}_i)\}}. \tag{4}$$

The multinomial logit model (Model I) for each tree species (post oak, black oak, or elm) in equation (3) is location independent, with its predicted probability at any location $\mathbf{u} \in D$ being determined by prevailing environmental conditions. The effect of the environmental conditions on the tree species of interest in the study area is modeled by parameters $\boldsymbol{\beta}_k$.

As noted in previous studies (Hooten, Larsen, and Wikle 2003; Rathbun and Black 2006; He et al. 2007; Wang 2007; Yoo and Trgovac 2011), the local abundance of a tree species is characterized by spatial clusters. Such spatially structured variations of a tree species distribution might result from omitted predictors. It also might be a natural reflection of spatial correlation or spatial random effects of ecological processes at a landscape level. One can accommodate such spatial random effects in a multinomial logit regression model by introducing latent variables. We can extend a GLM in equation (4) to a GLMM by including random effects, where the spatially correlated structure of tree species is taken into account by a linear predictor.

Let $S_k(\mathbf{u})$ denote a spatially correlated latent variable for the kth species with the collection of such latent variables $\{S_k(\mathbf{u}), \mathbf{u} \in D\}$ within a study domain D constituting a Gaussian random field (GRF), which is characterized by a mean function $\mu_k(\mathbf{u}) = E\{S_k(\mathbf{u})\}$ and a positive definite

covariance function $\sigma_{S_k}(\mathbf{u}, \mathbf{u}') = Cov\{S_k(\mathbf{u}), S_k(\mathbf{u}')\}$ (Chilès and Delfiner 1999). Under the stationarity assumption, the mean function of such a spatial Gaussian process is a constant $\mu_k(\mathbf{u}) = \mu_k$ for all $\mathbf{u} \in D$, and $\sigma_{S_k}(\mathbf{u}, \mathbf{u}') = \sigma_{S_k}(\mathbf{h})$ where $\mathbf{u}' = \mathbf{u} + \mathbf{h}$; that is, the covariance function depends only on a separation vector difference between any two locations $\mathbf{u}, \mathbf{u}' \in D$. Without loss of generality, we further assume a zero mean for $S_k(\mathbf{u})$, $\mu_k = 0$, for $k = 1 \ldots, 4$. In any such process, the spatial RVs at any collection of locations follow a multivariate Gaussian distribution with a mean vector $\mathbf{0}$ and a covariance matrix $\mathbf{\Sigma}_{S_k} = [\sigma_{S_k}(\mathbf{u}, \mathbf{u}'), \mathbf{u}, \mathbf{u}' \in D]$, where the covariance function $\sigma_{S_k}(\mathbf{u}, \mathbf{u}')$ can be modeled as a function of hyperparameters $\boldsymbol{\theta}_k = \{nugget, range, sill\}$ that completely characterize the spatial dependence in the underlying process; that is, $cov\{S_k(\mathbf{u}), S_k(\mathbf{u}')\} = \sigma_{S_k}(\mathbf{u}' - \mathbf{u}; \boldsymbol{\theta}_k)$ (Diggle, Tawn, and Moyeed 1998; Schabenberger and Gotway 2005).

Assuming that the underlying distribution of tree species occurrences in the study region depends on characteristics of the environmental conditions as well as the spatial dependence present in tree data, the following general model may be posited by combining the multinomial and the spatially correlated random effects:

$$\eta_k(\mathbf{u}_i) = \alpha_k + \mathbf{x}_i^T \boldsymbol{\beta}_k + S_k(\mathbf{u}_i), \quad k = 1, \ldots, 3, \tag{5}$$

where $\{S_k(\mathbf{u}), \mathbf{u} \in D\} \sim MVN(\mathbf{0}, \mathbf{\Sigma}_{S_k})$ denotes a zero mean stationary GRF, and the GRF being independent among the K tree species, that is, $Cov\{S_k(\mathbf{u}), S_{k'}(\mathbf{u})\} = 0$, if $k \neq k'$.

Model fitting and spatial prediction

Model fitting and statistical inference for the GLMMs in equation (5) are challenging, but not formidable, particularly when the latent (directly unobservable) random effects are spatially correlated. MCMC sampling-based methods are commonly resorted to in order to infer the distribution of the correlated latent variables and estimate the model parameters (Liang and Zeger 1986; Diggle, Tawn, and Moyeed 1998; Hooten, Larsen, and Wikle 2003; Wikle 2003; Christensen 2004; Schabenberger and Gotway 2005; He et al. 2007). However, such methods tend to have convergence issues and suffer from an extremely heavy computational burden. We adopt the approximation method following Cao, Kyriakidis, and Goodchild (2011) to incorporate the spatial effects in the witness tree modeling, while alleviating the previously mentioned issues, and further extend it to accommodate environmental covariates.

Unlike multinomial logit models, the mixed spatial multinomial logit models need to be evaluated per prediction location. At a prediction location $\mathbf{u}_0$, the conditional probability distribution of the kth tree species occurrence is

$$\pi_{k|\mathbf{c},\mathbf{S}}(\mathbf{u}_0) = Prob\{C(\mathbf{u}_0) = k | \mathbf{c}, \mathbf{S}\} = \int \underbrace{P\{C(\mathbf{u}_0) = k | \mathbf{c}, \mathbf{S}\}}_{\text{likelihood}} \underbrace{P\{\mathbf{c}|\mathbf{S}\}}_{\text{prior prob.}} d\mathbf{S}, \tag{6}$$

where $\mathbf{c} = [\delta_k(\mathbf{u}_i), i = 1, \ldots, n, k = 1, \ldots, 4]^T$ refers to the indicator-coded multinomial tree species data at n survey sites, and $\mathbf{S} = [S_k(\mathbf{u}_i), i = 1, \ldots, n, k = 1, \ldots, 4]^T$ denotes specific knowledge about the random field.

Typically, the conditional probability estimate $\hat{\pi}_{k|\mathbf{c},\mathbf{S}}(\mathbf{u}_0)$ at the prediction location $\mathbf{u}_0$ is obtained by computing the full joint probability distribution $Prob\{\cap_k C(\mathbf{u}_i) = k, i = 0, \ldots, n, k = 1, \ldots, 4\}$ across data locations and a prediction location, but we use its estimate informed from sample data. More specifically, we assume that the posterior probability distribution of the latent

variable $S_k(\mathbf{u})$ is a multivariate Gaussian RV, and observed data, given such multivariate Gaussian latent variables, are conditionally independent. Then, the posterior distribution of latent variables can be determined by the conditional posterior for fixed effects parameters α_k and $\boldsymbol{\beta}_k$ and the conditional posterior for the random effects $\boldsymbol{\theta}_k$. These unknown model parameters, $\{\alpha_k, \boldsymbol{\beta}_k, \boldsymbol{\theta}_k\}$ for $k = 1, \ldots, 4$, typically are obtained by integrating the likelihood function for a GLMM with respect to the latent random field $\mathbf{S}$, while the calculus-based integral in equation (6) is computationally intractable. To achieve this goal, we relied on a numerical method for approximation, such as a Laplacian approximation (Raudenbush, Yang, and Yosef 2000; Rue, Martino, and Chopin 2009).

The class-occurrence probability at a prediction location $\mathbf{u}_0$ is approximated by replacing the integral in equation (6) by the maximum a posteriori (MAP) estimate $\mathbf{s}_k^* = \arg\max_{\mathbf{s}} P\{\mathbf{S}|\mathbf{c}\}$, which takes the form of a weighted linear combination of covariance values between any observation point and the prediction location; that is, $S_k^*(\mathbf{u}_0) = \sum_{i=1}^{n} \omega_k(\mathbf{u}_i)\sigma_{S_k}(\mathbf{u}_i, \mathbf{u}_0; \boldsymbol{\theta}_k)$ (Cao, Kyriakidis, and Goodchild 2011). The weight $\omega_k(\mathbf{u}_i)$ for the kth category at the ith datum location $\mathbf{u}_i$ carries the information about the presence/absence of the kth tree species at all n data locations while accounting for redundancy. Therefore, the predicted kth tree species occurrence probability at a location $\mathbf{u}_0$ is rewritten using the MAP estimate of $S_k^*(\mathbf{u}_0)$ as

$$\hat{\pi}_k(\mathbf{u}_0) = \frac{\exp\{\hat{\eta}_k(\mathbf{u}_0)\}}{1 + \sum_{k'=1}^{3} \exp\{\hat{\eta}_{k'}(\mathbf{u}_0)\}} = \frac{\exp\left\{\hat{\alpha}_k + \mathbf{x}_0^T \hat{\boldsymbol{\beta}}_k + S_k^*(\mathbf{u}_0)\right\}}{1 + \sum_{k'=1}^{3} \exp\left\{\hat{\alpha}_{k'} + \mathbf{x}_0^T \hat{\boldsymbol{\beta}}_{k'} + S_{k'}^*(\mathbf{u}_0)\right\}}, \quad k = 1, \ldots, 3, \quad (7)$$

where $\mathbf{x}_0 = [x_l(\mathbf{u}_0),\ l = 1, \ldots, q]$ denotes the vector of environmental covariate values at the prediction location $\mathbf{u}_0$, and $\hat{\alpha}_k$ and $\hat{\boldsymbol{\beta}}_k$ denote the maximum likelihood estimates of regression model coefficients.

In the absence of environmental covariates (Model III), the covariance matrix for the zero mean stationary latent spatial field $\{S_k(\mathbf{u}_i),\ i = 1, \ldots, n\}$ for each type of tree species ($k = 1, \ldots, 4$) could be obtained using a mixture approach (Cao, Kyriakidis, and Goodchild 2011). That is, under the assumption of a common covariance function $\sigma_S(\mathbf{u}_i, \mathbf{u}_j; \boldsymbol{\theta})$ shared by all tree species, we can estimate the log-odds $\eta_k^*(\mathbf{u}_0)$ for the kth tree species as a weighted linear combination of the common covariance values at n survey sites. In the mixture approach, the common covariance values are estimated as a linear combination of estimated indicator covariograms $\hat{\sigma}_{S_k}(\mathbf{u}_i, \mathbf{u}_j; \boldsymbol{\theta}_k)$ for each species weighted by its global proportion $\tau_{k'}$ for each tree taxa as

$$\eta_k^*(\mathbf{u}_0) = \sum_{i=1}^{n} \omega_k(\mathbf{u}_i)\hat{\sigma}_S(\mathbf{u}_0, \mathbf{u}_i; \boldsymbol{\theta}) = \sum_{i=1}^{n} \omega_k(\mathbf{u}_i) \sum_{k'=1}^{4} \tau_{k'} \hat{\sigma}_{S_{k'}}(\mathbf{u}_0, \mathbf{u}_i; \boldsymbol{\theta}_{k'}), \quad (8)$$

where the sum of the global proportions equals 1; that is, $\sum_{k'=1}^{4} \tau_{k'} = 1$.

However, to take into account the spatial effects of selected covariates in the proposed model (Model IV), we further extend the preceding mixture approach in equation (8) to include environmental covariates as an external drift. Thus, the log-odds estimate $\hat{\eta}_k(\mathbf{u}_0)$ for Model IV is

$$\hat{\eta}_k(\mathbf{u}_0) = \hat{\alpha}_k + \sum_{l=1}^{q} \hat{\beta}_{l,k} x_l(\mathbf{u}_0) + \sum_{k'=1}^{4} \sum_{i=1}^{n} \omega_{i,k',k} \hat{\sigma}_{S_{k'}}(\mathbf{u}_0, \mathbf{u}_i; \theta_{k'}), \quad (9)$$

where $\hat{\omega}_{i,k',k}$ denotes a weight for the ith datum with respect to the k'th species applied to the latter's covariance function, denoted as $\hat{\sigma}_{S_{k'}}(\mathbf{u}_0, \mathbf{u}_i; \boldsymbol{\theta}_{k'})$, specified by the parameters $\boldsymbol{\theta}_{k'}$,

$k' = 1, \ldots, 4$. For each $\hat{\sigma}_{S_{k'}}(\mathbf{u}_0, \mathbf{u}_i; \boldsymbol{\theta}_{k'})$, we follow the covariogram fitting procedure that is commonly used in geostatistics, which first computes the empirical covariances based on observed data and then fits the desirable covariance function parameters $\boldsymbol{\theta}_{k'}$ through weighted least squares methods. With the covariogram parameter estimates $\hat{\boldsymbol{\theta}}_{k'}$, one can readily obtain the vectors of fixed effects parameters α_k, $\boldsymbol{\beta}_k$ and the vector of weights for covariograms $\boldsymbol{\omega}_k = [\omega_{i,k',k}, i = 1, \ldots, n, k, k' = 1, \ldots, 4]$ by maximizing the likelihood or minimizing a loss function. Consider an $(n \times 1)$ vector of indicators $\mathbf{j}(\mathbf{u}_i) = [j_1(\mathbf{u}_i), \ldots, j_4(\mathbf{u}_i)]^T$ that represents which class $\mathbf{u}_i$ belongs to, where $j_k(\mathbf{u}_i) = 1$ if $c(\mathbf{u}_i) = k$, and zero otherwise. The loss function $\ell(\boldsymbol{\Theta})$ for unknown parameters $\boldsymbol{\Theta} = [\boldsymbol{\Theta}_k, k = 1, \ldots, 4]$ for each species can be written as

$$\ell(\boldsymbol{\Theta}) = -\sum_{i=1}^{n}\left[\mathbf{j}(\mathbf{u}_i)^T(\Sigma(\mathbf{u}_i, \cdot)\boldsymbol{\Theta})^T - \log\sum_{k'=1}^{4}\exp\{\Sigma(\mathbf{u}_i, \cdot)\boldsymbol{\Theta}_{k'}\}\right], \tag{10}$$

where $\boldsymbol{\Theta}_k = \left[\boldsymbol{\omega}_k^T\ \alpha_k\ \boldsymbol{\beta}_k^T\right]^T$ denotes the $(n + q + 1) \times 1$ vector of unknown parameters for the kth tree species, and $\boldsymbol{\Sigma}(\mathbf{u}_i, \cdot)$ is the ith row of the covariance matrix $\boldsymbol{\Sigma}$, which consists of $(n \times n)$ data-to-data covariance matrix $\boldsymbol{\Sigma}_S = \left[\sigma_{S_k}(\mathbf{u}_i, \mathbf{u}_j), i, j = 1, \ldots, n, k = 1, \ldots, 4\right]$ and the $(n \times (q + 1))$ design matrix $\mathbf{F} = [x_l(\mathbf{u}_i), i = 1, \ldots, n, l = 0, \ldots, q]$ as $\boldsymbol{\Sigma} = [\boldsymbol{\Sigma}_S\ \mathbf{F}]$.

Because of the large number of unknown parameters, minimization of the logistic loss function given by equation (10) may suffer from overfitting problems. As an alternative, we adopt the so-called grouped least absolute shrinkage and selection operator (LASSO) regularization (Yuan and Lin 2006), where the unknown parameters $\boldsymbol{\Theta}$ are grouped according to the covariates, and regularization parameters are assigned to each group. With this regularization, the loss function in equation (10) can be written as a combination of the L_1 norm,[1] which produces a sparse solution by penalizing the regression coefficients to zero, and the L_2 norm,[2] which yields soft penalization on the coefficients. The minimization of grouped LASSO takes the form of a constrained convex optimization problem, which we solved using a recently proposed efficient algorithm, namely the *projected Quasi-Newton method*, to obtain optimal parameter estimates (Schmidt et al. 2009).

Results and discussion

This section consists of a data and model output summary, followed by results for spatial predictions and model validation. It includes a detailed discussion of model performance and comparison of statistical techniques. The ecological response of the three species to abiotic variables is interpreted as well.

Data summary

We used a total of 2,561 witness tree data points obtained from the 1870's survey. For modeling the distributions of trees, we focused on the three most abundant taxa—post oak (48.0%), black oak (20.2%), and elm (12.8%)—and recategorized the rest of the species as other type (19%).

As shown in Fig. 2(a), post oak is the most abundant tree species, with a strong concentration in the southern portion of the study area, while black oak is more evenly distributed, with a few clusters in the central and east regions of the study area. In contrast to the two oaks, the elm's spatial distribution is concentrated in the northeast side of the study area. The spatial distributions of all three species show the presence of spatial clusters but with different intensities.

Environmental covariates used in the current study consist of slope, aspect (N, NE, E, SE, S, SW, W, NW, flat), geology surface material (shale/limestone, shale, limestone, Oil Creek

Table 2 Prediction Models

		Environmental excluded	Covariates included
Spatial structure	Excluded	—	I
	Included	II and III	IV

sandstone, Bromide sandstone, granite, other type), and soil type (Shidler-Scullin-rock outcrop-Lula-Claremore, Shidler-rock outcrop, Kiti-rock outcrop, Chigley-Agan-rock outcrop, Reinach–McLain–Dale). Categorical covariates, such as aspect, geology, and soil type, are transformed into a set of binary 0–1 indicator variables.

Statistical models and model estimation

Reconstructed spatial distributions of presettlement vegetation abundance depend on multiple factors, including the spatial structure of witness tree data and environmental conditions. While the proposed mixed spatial multinomial logit model can accommodate various data simultaneously, other model specifications may not. The four statistical models used to describe presettlement vegetation reconstruction can be categorized into an aspatial environmental regression model in equation (3) and spatially explicit models, with and without environmental covariates (see Table 2).

A multinomial logit regression model (Model I) with more than two outcomes is a generalization of the standard logistic regression model for two outcomes. That is, the counts in different types of tree species have a multinomial distribution, which specifies the probability for each possible way a witness tree species observed at a survey location $\mathbf{u}_i (i = 1, \ldots, n)$ can fall in one of the four categories ($k = 1, \ldots, 4$). Alternatively, someone may collapse all levels of outcomes to binary per tree species so that the presence/absence of each tree species is independently modeled. This view aligns with a traditional geostatistical model (Model II), the following indicator kriging specification with the use of spatial structure information (Chilès and Delfiner 1999):

$$\hat{\pi}_k(\mathbf{u}_0) = \sum_{i=1}^{n} \lambda_i(\mathbf{u}_0)\left[\delta_k(\mathbf{u}_i) - \hat{\tau}_k\right] + \hat{\tau}_k, \quad k = 1, \ldots, 4, \tag{11}$$

where $\delta_k(\mathbf{u}_i)$ denotes the kth tree species being observed at the ith survey location $\mathbf{u}_i$, and $\hat{\tau}_k$ denotes the estimated global proportion of the kth tree taxa, which is assumed to be constant across the study region. The indicator kriging weights for the prediction location $\mathbf{u}_0$ are denoted by $\lambda_i(\mathbf{u}_0)$.

Mixed spatial effects models (Models III and IV) are linked to Model I to accommodate multicategorical outcomes of a response variable but are also related to Model II explicitly to take into account spatial structure of tree data. The major difference between mixed spatially correlated models and indicator kriging lies in the representation of witness tree data. Model II is a species-specific spatial prediction model whereas Model III can accommodate more than one species at a time. That is, the probability of post oak occurrence at a prediction location may be linked to black oak or elm occurrence probabilities in Model III, but not in Model II. However, this does not mean that Model III can accommodate species co-occurrences. The unit of analysis

Table 3 Estimated Parameters (Standard Errors) of the Multinomial Logit Models

		Post oak	Black oak	Elm
Intercept		–0.20 (0.531)	**–1.72**[†] (0.000)	**–2.81** (0.000)
Slope (X_1)		–0.00 (0.947)	–0.00 (0.928)	–0.02*[‡] (0.061)
Aspect (X_2)	N	–1.14 (0.112)	–25.99 (1.000)	0.74 (0.357)
	NE	0.27 (0.300)	0.04 (0.907)	0.83* (0.051)
	E	0.35 (0.164)	0.03 (0.917)	**1.22** (0.002)
	SE	0.41* (0.086)	0.43 (0.104)	**1.16** (0.002)
	S	0.27 (0.242)	0.12 (0.650)	**1.16** (0.002)
	SW	0.41* (0.078)	–0.03 (0.911)	**1.34** (0.000)
	W	0.11 (0.618)	0.19 (0.456)	**1.26** (0.001)
	NW	0.09 (0.694)	0.07 (0.784)	**1.31** (0.001)
Geology (X_3)	g1	–0.15 (0.535)	**0.93** (0.005)	0.59 (0.135)
	g2	**0.84** (0.001)	**1.41** (0.000)	0.94 (0.022)
	g3	0.09 (0.733)	**1.19** (0.000)	**1.22** (0.002)
	g4	0.33 (0.234)	**1.02** (0.005)	0.60 (0.163)
	g5	**0.56** (0.029)	**1.21** (0.000)	**1.03** (0.011)
	g6	0.49 (0.150)	**0.92** (0.035)	0.50 (0.313)
Soil (X_4)	s1	0.43* (0.058)	**0.77** (0.004)	0.62* (0.053)
	s2	0.45* (0.056)	0.50* (0.075)	**1.33** (0.000)
	s3	0.00 (0.991)	0.31 (0.217)	0.06 (0.836)
	s4	**1.81** (0.000)	**1.65** (0.000)	**1.64** (0.001)
Deviance		6050		

[†]Bold type indicates the significance level <0.05.
[‡]Asterisk indicates that the significance level is 0.10.

for all four prediction models is point support because each datum is associated with a unique survey location. No two tree species coexist at a single location with such a framework. Scaling up the unit of analysis to nonpoint support would allow us to apply a multivariate logistic regression model to model species co-occurrences (Ovaskainen, Hottola, and Siitonen 2010).

The environmental regression model (Model I) reveals some interesting associations between tree distributions and the covariates, particularly those for the geology and soil composition of the study area. Table 3 summarizes the multinomial logit regression results. Neither slope nor aspect has a high correlation with two of the three most abundant tree taxa, with elm being the exception, which might be due to a relatively flat landscape for the study area. In contrast, both geology and soil composition have a strong correlation with tree occurrences. In particular, black oak has a statistically significant relationship across all categories of geological type. These results reveal that Bromide sandstone (g5) is a statistically significant predictor across all three tree taxa, with varying degrees of magnitude. The major soil type in the southeast part of the study region, Chigley-Agan-rock outcrop (s4), shows a positive relationship with all three tree taxa, where its correlation with post oak is stronger than with the others. Both Shidler-Scullin-rock outcrop-Lula-Claremore (s1) and Shidler-rock outcrop (s2) have a relatively strong correlation with all three taxa. We considered only the main effects of environmental variables because the examination of residuals shows the presence of heteroscedasticity when interaction between elevation and geology or soil type is included. While there is no single

Table 4 Sensitivity of the Multinomial Logit Model Fit to Covariate Inclusion

Covariates*	Deviance statistics
X_1, X_2, X_3, X_4	6,050
X_2, X_3, X_4	6,055
X_1, X_3, X_4	6,097
X_1, X_2, X_4	6,118
X_1, X_2, X_3	6,120
X_1, X_2	6,333
X_1, X_3	6,167
X_1, X_4	6,166
X_2, X_3	6,130
X_3, X_4	6,102
X_1	6,382
X_2	6,359
X_3	6,177
X_4	6,173

*See Table 1 for the definition of covariates.

goodness/lack-of-fit test for a multinomial logit regression, deviance statistics often are used for a global test of model fit (Goeman and Cessie 2006; Agresti 2007). The deviance statistic for the full model (with four environmental covariates) is 6,050, which indicates that the model is adequate. The sensitivity of a multinomial logit model to covariates is assessed by comparing deviance statistics as shown in Table 4. The deviance statistic increases as a subset of covariates is used, and the statistic is high when either soil or geology is dropped. Despite the statistically significant association between elm distribution and aspect shown in Table 3, the aspect has a minimum effect on the global model fit. We suspect that this result might be due to the uneven distribution of species; that is, elm takes only 12.8%. In summary, Model I explains some but not substantial variation of tree species occurrences whereas the spatial autocorrelations of deviance residuals shown in Fig. 3(a) indicate that substantial spatial random effects are present in the residuals.

The three spatial models accommodate the spatially correlated structure in the PLS data through a covariance matrix of a latent spatial process. Indicator covariograms estimated for individual tree species in Fig. 3(b) show that the spatial distributions of all three tree taxa have a semivariogram range between 0.7 and 2.5 km. This range most likely is influenced by the 0.8 km spacing of the sample points because local-scale variability (less than 0.8 km) of the tree species distributions is not measured using witness tree data alone. The indicator covariograms also reveal differences in the microscale variability of each tree species' distribution, quantified by the semivariogram nugget effects. Post oak has the highest variability (0.22), followed by black oak (0.16) and elm (0.11), which reflects the class proportion in the data: a higher class proportion results in higher spatial variability at a microscale.

Spatial predictions and model validation

Presettlement patterns of tree species abundance are retrieved from the four witness tree prediction models over a grid cell of 30 × 30 m. Fig. 4 presents the predicted species occurrence

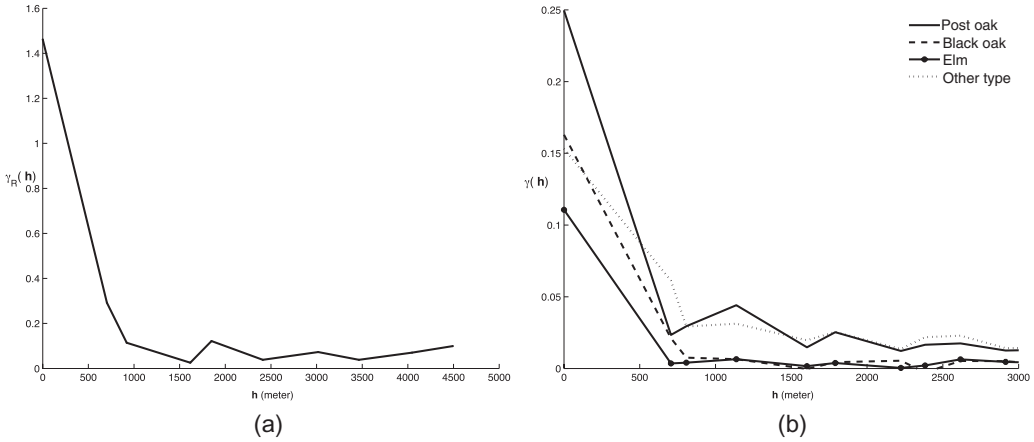

Figure 3. (a) Empirical covariograms of the multinomial regression residuals (deviance residuals). (b) Empirical indicator covariograms of the three most abundant tree taxa and other type.

probabilities for post oak, black oak, and elm (in each column) from the four prediction models (in each row). Using witness tree data alone, prediction at such a fine resolution could result in the loss of prediction accuracy due to the sparse sampling scheme of the PLS data. Because tree data in Oklahoma were collected at 0.8 km (quarter-mile) intervals, typical in the PLS system (He et al. 2000), only a fraction of the point observations (approximately 0.07% of point pairs) are less than 0.4 km apart. When sufficient data are unavailable, a reliable variogram is hard to establish, and consequently, the quality of spatial predictions at small scales is deteriorated. However, given that all four environmental covariates considered are available at such a fine resolution (30×30 m), we represent the landscape heterogeneity of the reconstructed presettlement of historic vegetation at a finer resolution than available with the PLS data.

Fig. 4(a), (d), (g), and (j) presents the four reconstructed species occurrence probabilities for post oak. The Model I prediction of post oak occurrence probabilities in Fig. 4(a) shows the effects of geology and soil conditions, while the Model II predictions in Fig. 4(d) clearly reveal the spatial configuration of the PLS data (the presence of post oak denoted by dark shade, and absence denoted by white shade). All three spatial model predictions show a mix of dark and white shades with varying sizes and different magnitudes: Model II yields higher post oak occurrence probabilities than the other three models across the study area, with minimum spatial continuity (the pattern changes rapidly), whereas Model III and IV predictions tend to smooth the high probabilities by accommodating other information, such as the presence of the other two taxa and the local environmental conditions to which the prediction location belongs. Spatial effects models (Models II and III) rely only on the witness tree data and their spatial structure, but apparently their alternative representation (data coding) leads to different prediction maps. Furthermore, the predictions obtained with Model II reveal some limitations of the traditional indicator approach; that is, predicted probabilities are not necessarily within [0, 1]. The circle and cross symbols in Fig. 4(d), (e), and (f) for post oak, black oak, and elm indicate the locations whose predicted probabilities are above 1 and below 0, respectively. In addition, no guarantee exists that the sum of predicted probabilities at each prediction location is equal to 1. Model IV predictions in Fig. 4(j) indicate that the southern portion of the study area is the post oak concentration region. Spatial concentration of high occurrence probabilities in this region also

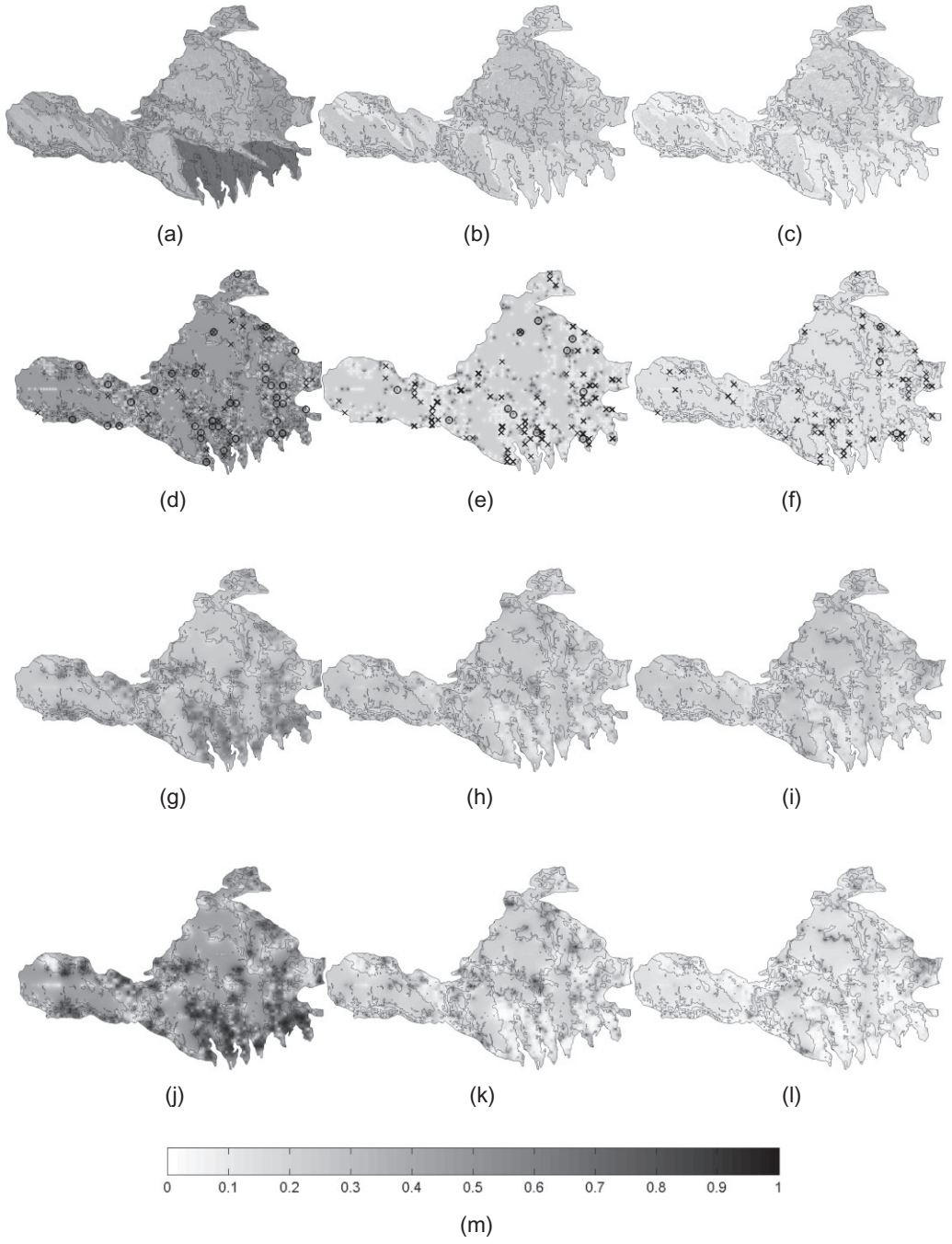

Figure 4. Predicted species occurrence probabilities: each column per tree species and each row per model. Predicted probabilities of post oak presented in the first column are obtained from (a) Model I, (d) Model II, (g) Model III, and (j) Model IV, respectively. The second (b, e, h, k) and third (c, f, i, l) columns tabulate predicted probability maps of black oak and elm, respectively. (m) The legend shared by the 12 maps. The symbols circle (o) and cross (x) in (b), (e), and (h) denote Model II-predicted probabilities greater than 1 and negative, which is meaningless, respectively.

gradually shifts to nearby areas. The favorable environmental conditions for post oak, such as shale and Bromide sandstone, in addition to the PLS survey data might create such large spatial clusters, but the kernel approach used in the mixed spatially correlated multinomial model results in a smoothly varying probability surface.

Fig. 4(b), (e), (h), and (k) presents the black oak occurrence probability predictions. All four model predictions for black oak are lower than the predicted probability of post oak occurrence, except for small clusters of high probability zones in the central and northern portions of the study area (see Fig. 4(e) and (k)). The surface formation favorable to post oak, such as shale/limestone and Bromide sandstone, does not have the same effects on black oak. In contrast, the soil type Shidler-Scullin-rock outcrop-Lula-Claremore shows a strong correlation with black oak, as shown in Fig. 4(b). The indicator kriging model (Model II) prediction in Fig. 4(e) shows high probabilities of black oak occurrences at small scales with negative estimates of some probabilities (denoted by cross symbols). These negative probabilities are due to the PLS tree data; that is, the presence or absence of black oak across a number of small patches. Both Models III and IV smooth such abrupt changes, where Model IV in Fig. 4(k) enhances the higher and lower probability zones informed by environmental conditions.

Fig. 4(c), (f), (i), and (l) portrays the elm occurrence probabilities obtained from the four prediction models. All four models reconstruct its relatively high concentration in the northeast part of the study area, which is dominated by the Shidler-rock outcrop soil type. In general, the occurrence probabilities for elm are low, which reflects its small proportion of PLS data (12%). Similar to predictions for the two oaks, Model II produces unacceptable probabilities; that is, out of the range [0, 1]. The Model II predictions of elm occurrence probability in Fig. 4(c) are not as smooth as those of the two oaks as shown in Fig. 4(d) and (e). This difference is due to the spatial structure model used for elm, such as a shorter semivariogram range than those of the two oaks (see Fig. 3(b)).

In summary, all four models reconstruct the dominant distribution of post oak in the south, small clusters of black oak in the north, and a small patch of elm concentrated in the northeast corner of the study area. The difference between the indicator kriging model (Model II) and the mixed spatially correlated models (Models III and IV) lies in the degree of smoothness of high and low probabilities: Model II tends to depict the spatial patterns based on the witness tree data specific to a single species whereas mixed multinomial models tend to integrate various sources of information and yield a smoother pattern. All spatial models allow one to identify areas with higher prediction uncertainties (0.3 and 0.4), which are due to their distant location from survey points and less significant relationships with environmental covariates.

The predicted class-occurrence probabilities for the three tree taxa are further processed to identify a class with the maximum posterior probability (to recover the tree species that is likely to have been present at the prediction location). Fig. 5 depicts the species prediction maps obtained from the four models, which are overlaid with the forest-woodland boundary delineated from the 1870's PLS plats by Fagin (2009). When the three dominant tree taxa occurrence probabilities are small, a pixel is classified as other type. Model I associates other type with Shidler-Scullin-rock outcrop-Lula-Claremore while the other spatial models predict other-type species in areas where observed data are scarce.

The uncertainty associated with each species occurrence probability increases as the distance between prediction points and the data of species incidence increases. Consequently, the class with the maximum posterior probability becomes none of the three species of interest. The environmental regression models (Models I and IV) tend to overestimate post oak and to

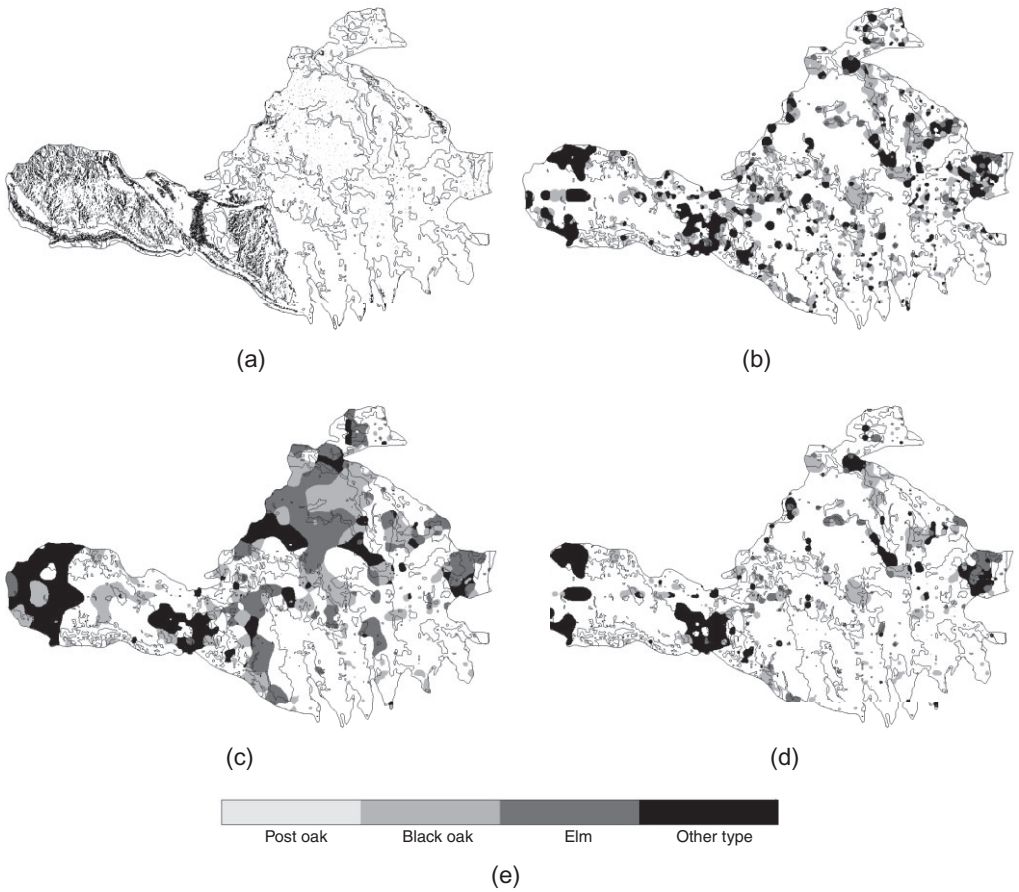

Figure 5. Predicted species occurrences: (a) multinomial logit model (Model I); (b) simple indicator kriging model (Model II); (c) mixed spatially correlated multinomial logit model with no covariates (Model III); (d) mixed spatially correlated multinomial logit model with covariates (Model IV).

underestimate the specie's diversity. The proportion of post oak occurrences obtained with Models I and IV are 90% and 81%, respectively, while those for the spatial effect models are 72% (Model II) and 55% (Model III). Assuming that the survey data reflect the true distribution of trees in the 1870s and that the data-driven global proportion of post oak is 48%, Model III outperforms the other specifications.

We further assessed the predictive power of the four models using cross-validation. We segregated the data into validation and training sets. Validation data consist of a subset (10%) of observed witness tree data, which are withheld in the model fitting process and later used to validate model outcomes. In other words, only the training data are used for model fitting. The size of the validation data set may be insufficient for an effective accuracy assessment of the models, and sectional bias might be involved in the data-splitting process. Therefore, we repeated the validation process iteratively (100 times) so that new sets of training and validation data were randomly selected with each iteration, and model accuracy was calculated based on new model fits and new validation data. At each iteration, the predicted probabilities are computed at newly

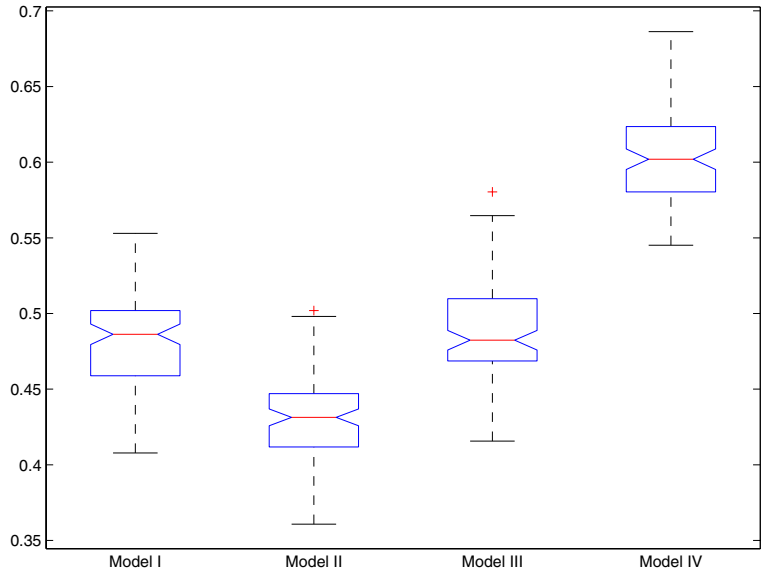

Figure 6. The correct classification ratio for the four prediction models. The line at the center of the boxes denotes the median correct classification ratio, and the edges of each box denote 75% and 25% quantiles. The extreme outliers are denoted by a plus (+) sign.

selected validation data locations (n = 255) as if the tree species at those locations are unknown using the remaining sample data (n = 2,292) with four prediction models. Finally, the class with the maximum posterior probability at the 255 locations is compared with the observed tree species. Per iteration, a correct classification ratio is calculated as the proportion of the total number of validation points with a correct classification. The result is summarized with boxplots (see Fig. 6). The mixed spatially correlated multinomial model with covariates (Model IV) is superior to the other three models. The other two multinomial models (Models I and III) have better prediction accuracy than the traditional indicator model (Model II). The average accuracy of Model IV is 60%, with a variability of 55%–70%, whereas the average prediction accuracy for the other models is below 50%. This variability of accuracy implies that the model prediction accuracy depends on the quality of observed data and the spatial configuration of prediction points. The spatial clustering of training data and the location of validation data play an important role in prediction accuracy. This result implies that the indicator coding of multicategorical data may cause the loss of prediction accuracy, and the synthesis of spatial dependence of witness tree data combined with environmental covariates yields the best prediction accuracy. Our results also indicate that the environmental regression model (Model I) performs as well as or better than the spatial effects models. However, caution should be exercised in generalizing our findings because the prediction and model fit of the environmental regression model depends on the influence of the selected environmental covariates on the witness tree data and the degree of spatial autocorrelation of the PLS data. Finally, the number of classes considered also is an important factor to consider in model comparisons.

Conclusion

In attempting to reconstruct presettlement forest vegetation using PLS witness tree data, we propose a model that incorporates both environmental covariates and the spatial dependence of

witness tree data. The proposed mixed spatially correlated multinomial logit model allows researchers to investigate the relationship between the location-specific relative abundance of witness tree species and the environmental conditions to which witness tree species respond within the context of spatial effects. One approach to incorporate such spatial effects, typically modeled by a variance–covariance model with complex structural dependence, in GLMMs is by introducing latent spatial process. Various techniques are available that allow us to incorporate such complex spatial dependence information within the paradigm of latent spatial variables, but we followed the work of Cao, Kyriakidis, and Goodchild (2011) due to its flexibility in terms of accommodating covariates and its efficiency in terms of incorporating spatial effects in these covariates.

The quantitative and qualitative evaluation of the proposed approach was conducted on the basis of a correct classification ratio and the reproduction of species diversity in the predicted class occurrence. To measure effectively the influence of spatial random effects and fixed effects (environmental covariates) on model performance, we designed a four-model comparison study. The four witness tree prediction models include (1) an aspatial environmental regression model (Model I); (2) a traditional indicator kriging model (Model II); (3) a mixed spatially correlated effects model with no predictors (Model III); and (4) a mixed spatially correlated effects model with environmental covariates (Model IV). Model predictions consist of two steps: first, species occurrence probability is predicted for a location from each of the four models; second, the set of species occurrence probabilities at each prediction location is further processed to identify the species with the maximum species occurrence probability.

In the cross-validation study, we compare the class of the observed tree species with the maximum species occurrence probability prediction at validation data locations, and determine the ratio of the total number of correct classifications. To avoid the selection bias of validation/training data, this evaluation was iteratively performed using a randomly selected training/validation data set. The results show that Model IV is superior to the other three models in terms of the correct classification ratio, and that Models I and III rank second. This result comes as no surprise in that the effects of the environmental covariates on the spatial distribution of the selected tree species are nontrivial, and substantial spatial effects are present in the deviance residuals.

From the species occurrence map for the study area obtained with the four prediction models, we assess model performance via the reproduction of species diversity. To assess species diversity, we compare the global proportion of each species. The result indicates that Model III performs the best, while the other three models tend to overestimate the most abundant tree species, post oak. Model III tends to smooth the species occurrence probability where observed data are unavailable whereas Model IV enhances the smooth probabilities with the support of environmental covariates. Both the strong association between post oak occurrences, geology, and soil information, and the large widely spread post oak data contribute to the overestimation by Model IV for post oak occurrence, which leads to underestimation of the other minor species.

In summary, the landscape structure of the study area determines the relative contributions of each factor—environmental conditions and the spatial autocorrelation present in tree occurrences—on the predictive power of a proposed model. In particular, the performance of the multivariate environmental regression models (Models I and IV) heavily depends on the quality of selected covariates. How strong or weak the relationship is between multicategorical witness tree data and the selected covariates determines the accuracy of model predictions. Meanwhile, a strong spatial autocorrelation in witness tree data alone may produce as accurate class

predictions as other models, as shown by both spatial effects models (Models II and III). Differences between a traditional indicator kriging model and a mixed spatially correlated multinomial model are discussed from both conceptual and practical perspectives. Indicator kriging is the simplest form of a mixed multinomial regression model where the response outcome has only one category. We also demonstrate the limitations of taking an indicator kriging approach where each species is taken separately in a case study. The predicted probabilities of species occurrences are not necessarily bounded between 0 and 1, nor is the sum of total predicted class-occurrence probabilities per species equal to 1. Furthermore, a loss of efficiency occurs in collapsing multiple categories in the sense that larger standard errors result, which is more severe, when observations are unevenly spread across the categories (Agresti 2007). The contribution of incorporating environmental conditions to model PLS witness tree species distributions is clearly demonstrated by comparing the predictive power of Model IV to that of Model III. Given that the influence of each source of information on the predictive power of a tree model is highly regional and landscape dependent, a flexible modeling approach should be employed in which various sources of information obtained from multiple sources are simultaneously integrated while their redundancy and correlation are taken into account.

Future work will include (1) incorporating environmental covariates as a stochastic process to account for error-corrupted predictors; (2) evaluating model performance under the different scenarios using simulated data; (3) imposing additional constraints on the mixed spatially correlated model (e.g., coherence constraint to reproduce the global proportion); (4) taking into account spatial scale disparity, how to resolve the scale differences between tree data points, environmental covariates, and target prediction resolution; and (5) investigating species co-occurrence over nonpoint support.

Notes

1 Given an n-dimensional vector $\mathbf{x} = [x_1, \ldots, x_n]^T$, the L_p norm of $\boldsymbol{x}$ is defined as $|\mathbf{x}|_p = \left(\sum_i^n |x_i|^P\right)^{\frac{1}{P}}$. L_1 and L_2 norms are particularly defined by setting $P = 1$ and $P = 2$, respectively (Meyer 2000).

2 The L_2 norm, also called the Euclidean norm, relates to the Euclidean distance from the origin to the point **x**.

References

Agresti, A. (2007). *An Introduction to Categorical Data Analysis*. Hoboken, NJ: Wiley.

Austin, M., and R. Cunningham (1981). "Observational Analysis of Environmental Gradients." *Proceedings of the Ecological Society of Australia* 11, 109–19.

Batek, M., A. Rebertus, W. Schroeder, T. Haithcoat, E. Compas, and R. Guyette (1999). "Reconstruction of Early Nineteenth-Century Vegetation and Fire Regimes in the Missouri Ozarks." *Journal of Biogeography* 26(2), 397–412.

Bogard, G. (1973). *Soil Survey of Pontotoc County, Oklahoma*. Washington, DC: U.S. Department of Agriculture, Soil Conservation Service.

Bolker, B., M. Brooks, C. Clark, S. Geange, J. R. Poulsen, M. H. H. Stevens, and J. White (2009). "Generalized Linear Mixed Models: A Practical Guide for Ecology and Evolution." *Trends in Ecology and Evolution* 24, 127–35.

Brown, D. (1998). "Mapping Historical Forest Types in Baraga County, Michigan, USA as Fuzzy Sets." *Plant Ecology* 134(1), 97–111.

Burgess, D. (1977). *Soil Survey of Johnston County, Oklahoma*. Washington, DC: U.S. Department of Agriculture, Soil Conservation Service.

Cao, G., P. Kyriakidis, and M. Goodchild (2011). "A Multinomial Logistic Mixed Model for Prediction of Categorical Spatial Data." *International Journal of Geographical Information Science* 25(12), 2071–86.

Chilès, J. P., and P. Delfiner (1999). *Geostatistics: Modeling Spatial Uncertainty*. New York: Wiley.

Christensen, O. (2004). "Monte Carlo Maximum Likelihood in Model-Based Geostatistics." *Journal of Computational and Graphical Statistics* 13(3), 702–18.

Cogbill, C., J. Burk, and G. Motzkin (2002). "The Forests of Presettlement New England, USA: Spatial and Compositional Patterns Based on Town Proprietor Surveys." *Journal of Biogeography* 29(10-11), 1279–304.

Curtis, J. (1956). "The Modification of Mid-Latitude Grasslands and Forests by Man." In *Man's Role in Changing the Face of the Earth*, 721–36, edited by W. L. Thomas. Chicago, IL: University of Chicago Press.

Dale, E. (1956). "A Preliminary Survey of the Flora of the Arbuckle Mountains, Oklahoma." *Texas Journal of Science* 8(1), 41–73.

Delcourt, H., and P. Delcourt (1996). "Presettlement Landscape Heterogeneity: Evaluating Grain of Resolution Using General Land Office Survey Data." *Landscape Ecology* 11(6), 363–81.

Diggle, P., J. Tawn, and R. Moyeed (1998). "Model-Based Geostatistics." *Applied Statistics* 47(3), 299–350.

Fagin, T. (2009). "In Search of the Forest Primeval: Data-driven Approaches to Mapping Historic Vegetation." PhD dissertation, Department of Geography, University of Oklahoma, Norman.

Fagin, T., and B. Hoagland (2011). "Patterns from the Past: Modeling Public Land Survey Witness Tree Distributions with Weights-of-Evidence." *Plant Ecology* 212(2), 207–17.

Fairchild, R., R. Hanson, and R. Davis (1990). "Hydrology of the Arbuckle Mountains Area, South-Central Oklahoma." *Oklahoma Geological Survey Circular 91*. Norman: University of Oklahoma.

Fassett, N. (1944). "Vegetation of the Brule Basin, Past and Present." *Transactions of the Wisconsin Academy of Sciences, Arts, and Letters* 36, 33–56.

Franklin, J., and J. Miller (2009). *Mapping Species Distributions: Spatial Inference and Prediction*. New York: Cambridge University Press.

Friedman, S., P. Reich, and L. Frelich (2001). "Multiple Scale Composition and Spatial Distribution Patterns of the North-Eastern Minnesota Presettlement Forest." *Journal of Ecology* 89(4), 538–54.

Galatowitsch, S. (1990). "Using the Original Land Survey Notes to Reconstruct Presettlement Landscapes in the American West." *Great Basin Naturalist* 50, 181–91.

Goeman, J. J., and S. Cessie (2006). "A Goodness-of-Fit Test for Multinomial Logistic Regression." *Biometrics* 62(4), 980–5.

Ham, W. E. (1969). "Regional Geology of the Arbuckle Mountains, Oklahoma." *Oklahoma Geological Survey Guidebook 17*. Norman: University of Oklahoma.

He, H., D. Dey, X. Fan, M. Hooten, J. Kabrick, C. Wikle, and Z. Fan (2007). "Mapping Pre-European Settlement Vegetation at Fine Resolutions Using a Hierarchical Bayesian Model and GIS." *Plant Ecology* 11(6), 85–94.

He, H., D. Mladenoff, T. Sickley, and G. Guntenspergen (2000). "GIS Interpolations of Witness Tree Records (1839–1866) for Northern Wisconsin at Multiple Scales." *Journal of Biogeography* 27, 1031–42.

Hooten, M., D. Larsen, and C. Wikle (2003). "Predicting the Spatial Distribution of Ground Flora on Large Domains Using a Hierarchical Bayesian Model." *Landscape Ecology* 18, 487–502.

Liang, K.-Y., and S. Zeger (1986). "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika* 73(1), 13–22.

Manies, K., and D. Mladenoff (2000). "Testing Methods to Produce Landscape-Scale Presettlement Vegetation Maps from the U.S. Public Land Survey Records." *Landscape Ecology* 15(8), 741–54.

Margules, C., A. Nicholls, and M. Austin (1987). "Diversity of Eucalyptus Diversity of Eucalyptus Species Predicted by a Multivariable Environmental Gradient." *Oecologia* 71, 229–32.

Meyer, C. D. (2000). *Matrix Analysis and Applied Linear Algebra*. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM).

Ovaskainen, O., J. Hottola, and J. Siitonen (2010). "Modeling Species Co-Occurrence by Multivariate Logistic Regression Generates New Hypotheses on Fungal Interactions." *Ecology* 91(9), 2514–21.

Rathbun, S., and B. Black (2006). "Modeling and Spatial Prediction of Pre-Settlement Patterns of Forest Distribution Using Witness Tree Data." *Environmental and Ecological Statistics* 13(4), 427–48.

Raudenbush, S. W., M.-L. Yang, and M. Yosef (2000). "Maximum Likelihood for Generalized Linear Models with Nested Random Effects via High-Order, Multivariate Laplace Approximation." *Journal of Computational and Graphical Statistics* 9(1), 141–57.

Rue, H., S. Martino, and N. Chopin (2009). "Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71(2), 319–92.

Schabenberger, O., and G. Gotway (2005). *Statistical Methods for Spatial Data Analysis*. Boca Raton, FL: CRC Press.

Schmidt, M., E. Berg, M. Friedlander, and K. Murphy (2009). *Optimizing Costly Functions with Simple Constraints: A Limited-Memory Projected Quasi-Newton Algorithm*. Proceedings of the 12th International Conference on Artificial Intelligence and Statistics, 456–63, edited by D. Dyk and M. Welling. Clearwater Beach, FL.

Schulte, L., and D. Mladenoff (2001). "The Original US Public Land Survey Records: Their Use and Limitations in Reconstructing Presettlement Vegetation." *Journal of Forestry* 99(10), 5–10.

Suneson, N. (1997). *The Geology of the Eastern Arbuckle Mountains in Pontotoc and Johnston Counties, Oklahoma: An Introduction and Field-Trip Guide*. Open-file report (Oklahoma Geological Survey) 4–97. Norman: Oklahoma Geological Survey.

Wang, Y.-C. (2007). "Spatial Patterns and Vegetation-Site Relationships of the Presettlement Forests in Western New York, USA." *Journal of Biogeography* 34, 500–13.

Wang, Y.-C., and C. Larsen (2006). "Do Coarse Resolution U.S. Presettlement Land Survey Records Adequately Represent the Spatial Pattern of Individual Tree Species?" *Landscape Ecology* 21(7), 1003–17.

Watterson, A., V. Bogard, and G. Moebius (1984). *Soil Survey of Murray County, Oklahoma*. Washington, DC: U.S. Department of Agriculture.

White, C. (1983). *A History of the Rectangular Survey System*. Washington, DC: U.S. Department of the Interior, Bureau of Land Management.

Whitney, G., and J. DeCant (2001). "Government Land Office Surveys and Other Early Land Surveys." In *Historical Ecology Handbook*, 147–76, edited by D. Egan and E. Howell. Washington, DC: Island Press.

Wikle, C. (2003). "Hierarchical Bayesian Models for Predicting the Spread of Ecological Processes." *Ecology* 84(6), 1382–94.

Wu, J. (2004). "Effects of Changing Scale on Landscape Pattern Analysis: Scaling Relations." *Landscape Ecology* 19(2), 125–38.

Yoo, E.-H., and A. Trgovac (2011). "Scale Effects in Uncertainty Modeling of Presettlement Vegetation Distribution." *International Journal of Geographical Information Science* 25(3), 405–21.

Yuan, M., and Y. Lin (2006). "Model Selection and Estimation in Regression with Grouped Variables." *Journal of the Royal Statistical Society: Series B* 68, 49–67.