The Annals of Applied Statistics
2019, Vol. 13, No. 4, 2312–2340
https://doi.org/10.1214/19-AOAS1259

A LATENT DISCRETE MARKOV RANDOM FIELD APPROACH TO IDENTIFYING AND CLASSIFYING HISTORICAL FOREST COMMUNITIES BASED ON SPATIAL MULTIVARIATE TREE SPECIES COUNTS[1]

BY STEPHEN BERG, JUN ZHU, MURRAY K. CLAYTON, MONIKA E. SHEA AND DAVID J. MLADENOFF

University of Wisconsin-Madison

The Wisconsin Public Land Survey database describes historical forest composition at high spatial resolution and is of interest in ecological studies of forest composition in Wisconsin just prior to significant Euro-American settlement. For such studies it is useful to identify recurring subpopulations of tree species known as communities, but standard clustering approaches for subpopulation identification do not account for dependence between spatially nearby observations. Here, we develop and fit a latent discrete Markov random field model for the purpose of identifying and classifying historical forest communities based on spatially referenced multivariate tree species counts across Wisconsin. We show empirically for the actual dataset and through simulation that our latent Markov random field modeling approach improves prediction and parameter estimation performance. For model fitting we introduce a new stochastic approximation algorithm which enables computationally efficient estimation and classification of large amounts of spatial multivariate count data.

1. Introduction. In this paper we consider analyzing historical tree species composition data and mapping forest ecological communities of keen interest in a variety of ecological disciplines, including environmental history and landscape ecology. Sound modeling and analysis of historical vegetation using novel statistical methodology is useful for multiple purposes, including to aid ecological restoration efforts by providing reference landcover information at restoration sites and to assess landscape changes over time (Schulte, Mladenoff and Nordheim (2002), Shea, Schulte and Palik (2014)). If an area is known to have historically supported a particular vegetation profile, this could indicate that restoration to the historically supported vegetation type may be more ecologically appropriate (Egan (2005)).

The historical public land survey (PLS) contains informative data for studies of past forest composition. The PLS database for the state of Wisconsin is particularly noteworthy for both its spatial extent (approximately 150,000 km^2) and its

Received August 2018; revised March 2019.

[1] Supported by a USGS CESU project and an NSF MSB project under grant No. DEB1241868.

Key words and phrases. Markov random field, historical ecology, stochastic approximation, maximum likelihood, mixture models.

high resolution (survey points at roughly half mile intervals across the entire state). The survey was initially conducted to assess land values and facilitate the sale of land, but the collated and digitized PLS data currently provide the only precise, statewide record of the natural ecosystems that were present in Wisconsin just prior to major Euro-American settlement (Schulte and Mladenoff (2001)). The database is derived from surveyor notebooks from the original U.S. PLS, conducted across the United States from the late 1700's to the early 1900's. The Wisconsin portion of the survey was conducted from 1832 to 1866 (Liu et al. (2011)). Surveyors demarcated the land into square mile sections and placed a post as a survey marker at each section corner and at each half-mile point. At each survey point the protocol required that they record several environmental characteristics, including the species of two to four "witness" trees.

Here, we consider the resulting tree species composition data from the Wisconsin PLS and aggregate the observed tree species counts within an overlaid grid of cells for analysis. An illustration of this type of data is shown in Figure 1. We also consider the identification of community subpopulation structure in the PLS relating to recurring assemblages of tree species which are described in ecological literature as forest communities (Barnes et al. (2010)). Community subpopulations are a common feature of tree species composition data such as in the PLS database. We model forest community subpopulations via the classification of each grid cell with the forest community type most representative of that cell. Our modeling goal is twofold. On the one hand we would like to use tree species composition data to identify discrete assemblages of species corresponding to forest communities in the state of Wisconsin prior to the major environmental disturbances accompanying Euro-American settlement. On the other hand we would like to classify cells in the survey region with the forest community type to which they most likely belong.

To achieve our goal of accurately modeling and mapping forest community subpopulations in the PLS survey, we develop an approach wherein forest communities across the survey region are described by discrete, spatially correlated

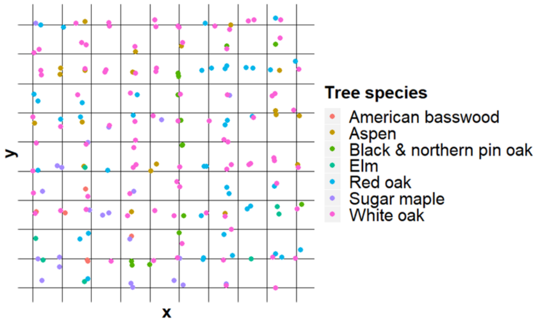


FIG. 1. *Data from a* 10 *by* 10 *km subregion of the Wisconsin Public Land Survey dataset. The overlaid grid cells are* 1 *km by* 1 *km. Tree species are recorded at multiple survey points within each grid cell.*

latent variables. The observed tree counts are described by community dependent multinomial distributions. Thus, the observed tree species compositions in the PLS dataset are assumed to result from a set of multiple underlying forest community types which occur in a spatially correlated fashion across the survey region. Our approach allows us to describe, indirectly but flexibly, spatial correlation between observations in nearby grid cells and also to address unobserved structure due to distinct forest community subpopulations.

Our analysis of tree species composition in the PLS dataset is unique among previous literature in that it explicitly accounts both for spatial correlation effects between nearby observations as well as for latent forest community structure. Tree species composition in the PLS was also studied in Paciorek et al. (2016), using, for example, a latent conditional autoregressive (CAR) model to account for spatial correlation with the goal of providing estimation of tree species composition in the PLS survey region. While the posterior predictions of forest composition in Paciorek et al. (2016) capture spatial covariance between the occurrence of related tree species, these predictions do not explicitly identify or map forest communities. A dissimilarity-based clustering approach as taken in Schulte, Mladenoff and Nordheim (2002) allows forest communities to be identified and mapped, but this approach does not explicitly model spatial correlation in the occurrence of forest community types across the study region.

In our work the tree species composition data come in the form of tree species count vectors, so that the data in each areal unit are multivariate, with counts of zero for absent tree species. As such, in contrast to most work in spatial clustering, our response variable is both multivariate and unordered. We do not constrain the forest community types to appear in spatially contiguous blocks. We also do not expect any ordinal relationship between the forest community types. In certain other common settings the term “spatial cluster” may refer to a spatially contiguous block of areal units where a response variable such as disease risk or rate is unusually high relative to other areal units. Frequently, the analysis goal in these settings is to identify “hotspots” of a disease and any associated risk factors (see, e.g., Gangnon and Clayton (2003), Lawson (2010), Waller (2009)). Constraints are also sometimes imposed so that each disease rate cluster only appears in a single contiguous block of areal units (see, e.g., Knorr-Held and Raßer (2004)).

We estimate the parameters of our model via maximum likelihood (ML), and we develop a new Markov chain Monte Carlo (MCMC) stochastic approximation (SA) method to do this. Our MCMC-SA method is related to but differs from the direct expectation-maximization (EM) approach (Dempster, Laird and Rubin (1977)). First, instead of performing the full M-step, we take a gradient step for the Markov random field parameters and a modified EM step for the other parameters in the model. Such EM algorithms with partial updates in the M-step are sometimes termed generalized EM algorithms (Dempster, Laird and Rubin (1977)). In general, performing the full M-step of the EM algorithm requires inverting between the mean parameterization and the natural parameterization of the complete data distribution (see, e.g., Fort and Moulines (2003)). In the Markov random field setting, performing this inversion is challenging, and it requires an MCMC sampling step nested within each EM algorithm iteration, as in Forbes and Fort (2007). We additionally apply regularization penalties to ensure that maxima of our objective function do not occur at the boundary of the parameter space (see, e.g., Chen (2017), Hong et al. (2017), Städler, Bühlmann and van de Geer (2010)). Finally, in our latent Markov random field model, the spatial dependence between the latent forest community types makes computing the log-likelihood challenging, and we use a path integration approach to accurately compute log likelihoods on holdout data (see, e.g., Section 6.2 in Neal (1993), or Gelman and Meng (1998)).

While MCMC methods may in general be slow, our MCMC-SA method is feasible even for relatively big data like the PLS dataset due to a computationally efficient implementation of the sampling. Additionally, in the case study of the PLS dataset, we achieve significant improvement in prediction performance using our method relative to an alternative approach that does not account for spatial dependence. A simulation study further shows that our MCMC-SA method can recover the true parameters under the correct model specification and outperform some competing methodology. Though our application in this paper focuses on identification and classification of forest communities across space, our methodology can be readily modified for use in other ecological community identifications or other settings such as medical image segmentation of tissue types.

The remainder of the paper is organized as follows. In Section 2 we propose a multinomial model with a latent discrete Markov random field for the PLS dataset. In Section 3 we develop a maximum likelihood approach to estimate the model parameters and propose a stochastic approximation procedure to compute these estimates. In Section 4 we apply our model and estimation method to analyze and interpret the PLS dataset. In Section 5 results are presented from a simulation study, followed by conclusions and a discussion in Section 6. We also provide a supplemental article containing additional technical details (Berg et al. (2019)).

2. Model.

Our observed data consist of the counts of each tree species within an overlaid grid of cells. We assume that each cell has a latent forest community type with an associated multinomial probability distribution governing the species composition for each type of forest community. We also assume conditional independence between observed trees given the latent forest community types which in turn are assumed to follow a Markov random field. Thus, our model is a mixture of multinomial distributions, where the types are spatially correlated.

2.1. *Latent model.*

The spatial grid of cells are assumed to be labeled with one of K possible types, in our case K different forest communities. Corresponding to each grid cell is a spatial neighborhood of adjacent grid cells. We view our approach as agnostic regarding the underlying origin of the spatial dependence in the dataset. For example, an influential environmental covariate may occur in discrete patches across a map, causing certain forest community types to appear or disappear in these areas. Additionally, local within- and between-community interactions may cause spatial patterning on the observed grid. The approach here attempts to mimic and to account for the observed spatial correlation structure rather than to exactly replicate the true data generating process.

For notation we refer to random variables with capitals and realizations in lowercase. When referring to a probability density for a discrete random vector Z depending on a parameter vector θ, we use the shorthand $p(z|\theta)$ for $p(Z = z|\theta)$. For a vector z, we use z_i to denote the ith entry of z. For a matrix $\mathbf{A}$ we use the notation $\mathbf{A}_j$ to denote the jth column of $\mathbf{A}$, and $\mathbf{A}_{ij}$ to denote the element in the ith row and the jth column of $\mathbf{A}$. We denote the set of spatial neighbors of a cell i by the set $N(i) = \{i' : i' \text{ is a neighbor of cell } i\}$, where the neighbors are defined so that $i \notin N(i)$. We use the notation $i' \sim i$ to indicate that $i' \in N(i)$. Additionally, the neighborhoods are assumed to be symmetric, so that if cell i is a neighbor of cell i', then cell i' is a neighbor of cell i.

Let n denote the total number of grid cells, and $z \in \Omega = \{1, \dots, K\}^n$ denote a vector of n (unobserved) forest community types. The random vector of forest community types Z is assumed to follow a Potts-type model with a vector of parameters $\eta \in \mathbb{R}^K$:

$$p(z|\eta) = \exp\left\{\sum_{i=1}^{n}\sum_{k=1}^{K-1} \eta_k I(z_i = k) + \eta_K \sum_{i=1}^{n} \sum_{\substack{i' \in N(i) \\ i' > i}} I(z_i = z_{i'}) - \xi(\eta)\right\}, \tag{1}$$

where z_i refers to the forest community type for cell i, and

$$\xi(\eta) = \sum_{z' \in \Omega} \exp\left\{\sum_{i=1}^{n}\sum_{k=1}^{K-1} \eta_k I(z'_i = k) + \eta_K \sum_{i=1}^{n} \sum_{\substack{i' \in N(i) \\ i' > i}} I(z'_i = z'_{i'})\right\}$$

is a normalizing constant ensuring that $p(z|\eta)$ is a probability density (Wu (1982)). In (1), for $k < K$, the parameter η_k controls the probability of the kth type relative to the baseline type K. The spatial correlation parameter η_K controls the strength of interactions between the types and, when $\eta_K = 0$, the types are spatially independent across grid cells.

We define a length K vector $T(z)$ of sufficient statistics with the kth entry

$$T(z)_k = \begin{cases} \displaystyle\sum_{i=1}^{n} I(z_i = k); & (k < K), \\ \displaystyle\sum_{i=1}^{n} \sum_{\substack{i' \in N(i) \\ i' > i}} I(z_i = z'_i); & (k = K). \end{cases} \tag{2}$$

This allows us to rewrite the model (1) more succinctly as

$$p(z|\eta) = \exp\{\eta^T T(z) - \xi(\eta)\}$$

which belongs to the exponential family with the natural parameter vector η (Shao (2003)).

For boundary conditions in lattice data models, there are several approaches to specifying the neighborhood of the cells on the boundary of the lattice. We use the so-called “free” boundary conditions, where boundary cells simply have fewer neighbors than internal cells (see, e.g., Comets and Gidas (1992)). Other approaches attempt to ensure that each cell has the same number (usually 4, for the square lattice) of neighbors. For example, in “toroidal” boundary conditions cells on one side or corner of the lattice are connected to cells on the opposing side or corner of the lattice.

2.2. *Data model.* Given the forest community types, we specify our model for the conditional distribution of the observed tree species counts. For notation we let the integer $M > 0$ denote the number of tree species in the dataset. For the PLS case study the $M = 33$ most common species are used. We denote by $\mathbf{Y}_i$ the length M vector of tree counts in cell i and use $\mathbf{Y} \in \mathbb{Z}^{M\times n}$ to denote the matrix of count vectors for the entire dataset. Thus, $\mathbf{Y}_{mi}$ is the count of trees of species m in cell i. We let q_i denote the total number of trees observed within the ith cell. That is, $q_i = \sum_{m=1}^{M} \mathbf{Y}_{mi}$.

We assume that each of the K forest community types is associated with a distinct multinomial distribution over the M tree species. Conditional on the latent type $Z_i = k$, the tree species of individual trees within a grid cell are assumed to be independent multinomials with sample size 1, so that the count vector $\mathbf{Y}_i$ follows a multinomial distribution with sample size q_i and species probability parameters depending on the kth forest community type. Additionally, the species of individual trees are assumed to be independent across grid cells and, thus, the counts $\mathbf{Y}_i$ are also independent across grid cells, both conditional on the latent forest community types. However, when the spatial correlation parameter $\eta_K \neq 0$, the latent types are spatially correlated which induces spatial correlation among the tree counts $\mathbf{Y}_i$.

We parameterize the species distribution for each forest community type k using a species probability vector $\boldsymbol{\mu}_k \in \mathscr{M}$, where $\mathscr{M}$ refers to the (open) probability simplex defined by $\mathscr{M} = \{\mu = [\mu_1, \dots, \mu_M]^T : \sum_m \mu_m = 1; \mu_m > 0, \forall m\}$. We also define the species probability matrix $\boldsymbol{\mu}$ with column vectors $\boldsymbol{\mu}_k$ by $\boldsymbol{\mu} = [\boldsymbol{\mu}_1 \quad \boldsymbol{\mu}_2 \quad \dots \quad \boldsymbol{\mu}_K] \in \mathscr{M}^K$. The $\boldsymbol{\mu}_{mk}$ element of the $\boldsymbol{\mu}$ matrix is equal to the probability that a tree in a grid cell is species m, given that the forest community type of that grid cell is k.

By the conditional independence of tree species between and within grid cells, the conditional density of the observed tree counts given the forest community types Z and the species probability matrix $\boldsymbol{\mu}$ is

$$p(\mathbf{y}|z, \boldsymbol{\mu}) = \prod_{i=1}^{n} p(\mathbf{y}_i|z_i, \boldsymbol{\mu}) = \prod_{i=1}^{n} C_i \prod_{m=1}^{M} \boldsymbol{\mu}_{m,z_i}^{\mathbf{y}_{mi}}, \tag{3}$$

where $\boldsymbol{\mu}_{m,z_i}$ is the mth entry of column z_i of $\boldsymbol{\mu}$, and the factor $C_i = (\prod_{m=1}^{M} \mathbf{y}_{mi}!)^{-1} q_i!$ counts the number of possible ways of assigning species to each tree in the ith grid cell with the species counts $\mathbf{y}_i$.

In summary, our full data generating mechanism comprises two steps:

1. Draw the forest community types Z according to the density in (1).
2. Conditioning on the forest community types $Z = z$ from step 1, draw the tree species counts $\mathbf{Y}$ according to the density in (3).

Define $R(\mathbf{y}, z) \in \mathbb{R}^{M \times K}$ to be a matrix of statistics with the (m,k)th element $R(\mathbf{y}, z)_{mk} = \sum_{i=1}^{n} \mathbf{y}_{mi} I(z_i = k)$ summarizing the total number of species m trees in the grid cells that belong to the kth type of forest community. Then, the complete data density for $(\mathbf{Y}, Z)$ is

$$\begin{aligned} & p(\mathbf{y}, z|\eta, \boldsymbol{\mu}) \\ & \quad = p(z|\eta) p(\mathbf{y}|z, \boldsymbol{\mu}) \\ & \quad = \exp\left\{\eta^T T(z) - \xi(\eta) + \sum_{m=1}^{M} \sum_{k=1}^{K} \log(\boldsymbol{\mu}_{mk}) R_{mk}(\mathbf{y}, z) + \sum_{i=1}^{n} \log(C_i)\right\}. \end{aligned} \tag{4}$$

It is sometimes convenient to write the parameter vector η and the parameter matrix $\boldsymbol{\mu}$ using a single vector parameter θ. Conversely, we may also need to obtain η and $\boldsymbol{\mu}$ from the corresponding vector θ. Thus, we define a vectorization operator $\text{vec}(A) : \mathbb{R}^{M \times K} \to \mathbb{R}^{MK}$, $\boldsymbol{\mu} \to [\boldsymbol{\mu}_1^T, \ldots, \boldsymbol{\mu}_K^T]^T$ for viewing the parameter matrix $\boldsymbol{\mu}$ as a vector. Then, we define $\theta \in \mathbb{R}^{K+MK}$ by $\theta = [\eta^T, \text{vec}(\boldsymbol{\mu})^T]^T$. We use Θ to denote the parameter space for θ, and we use $\eta(\theta) \in \mathbb{R}^K$ and $\boldsymbol{\mu}(\theta) \in \mathbb{R}^{M \times K}$ to denote the η and $\boldsymbol{\mu}$ associated with θ. When it is clear, we simply write η or $\boldsymbol{\mu}$ rather than $\eta(\theta)$ or $\boldsymbol{\mu}(\theta)$.

3. Method.

3.1. *Maximum regularized likelihood estimation.* Here, we estimate the parameter θ via maximum likelihood. For the model described in (1) and (3), the observed data log likelihood when $\mathbf{Y} = \mathbf{y}$ is

$$\ell(\theta) = \log\left\{\sum_{z \in \Omega} p(\mathbf{y}, z|\theta)\right\}. \tag{5}$$

The consistency of the maximum likelihood estimate, $\hat{\theta} = \text{argmax}_\theta \ell(\theta)$, is shown in an increasing domain asymptotics setting, under identifiability assumptions on θ and when the estimation is constrained to a compact parameter space (Comets and Gidas (1992)).

The observed data log-likelihood (5) may exhibit unwanted behavior, such as maxima on the boundary of the parameter space, which is common in latent variable models. Such behavior can occur even in simple settings, such as a mixture of normal densities with component-specific variances, where it is possible to achieve an arbitrarily high likelihood by setting the mean of one of the components to a single data point and by sending the variance of that component toward 0 (see, e.g., Chen (2017), Section 3.2). In our work there is apparent convergence of entries of the tree species probability matrix, $\boldsymbol{\mu}_{mk}$, to 0, which seems to occur mostly for the rarer tree species, while there is no observed convergence of components of the parameters associated with the forest community types, η, to the boundary of $\mathbb{R}^K$.

To guarantee the convergence of our estimation procedures to points within the parameter space Θ, we impose weakly informative prior penalties on the observed data log likelihood (5) (see, e.g., Chen (2017), Hong et al. (2017), Städler, Bühlmann and van de Geer (2010)). In particular Kushner and Yin (1997) added "soft penalties" to ensure that the objective function is well behaved and the iterates from a stochastic approximation procedure remain bounded, which we use here to optimize a regularized log-likelihood function,

$$\ell_{\text{pen}}(\theta) = \ell(\theta) + \rho_1(\eta) + \rho_2(\boldsymbol{\mu}), \tag{6}$$

where $\rho_1(\eta)$ and $\rho_2(\boldsymbol{\mu})$ are penalty functions.

For each component of η, we apply a Logistic$(0, \sigma)$ prior density with $\sigma > 0$. That is, $\rho_1(\eta) = \sum_{k=1}^{K} \log f_\sigma(\eta_k)$, where for $k = 1, \dots, K$,

$$f_\sigma(\eta_k) = \sigma^{-1}\left[\exp\{\eta_k/(2\sigma)\} + \exp\{-\eta_k/(2\sigma)\}\right]^{-2}. \tag{7}$$

In the PLS case study and in the simulation studies, we use $\sigma = 1$.

For each column of $\boldsymbol{\mu}$, we put a Dirichlet$(\alpha 1_M)$ prior with $\alpha > 1$, where 1_M is a vector of M 1's, so that

$$\rho_2(\boldsymbol{\mu}) = (\alpha - 1) \sum_{k=1}^{K} \sum_{m=1}^{M} \log(\boldsymbol{\mu}_{mk}). \tag{8}$$

For an integer $\alpha > 1$, $\rho_2(\boldsymbol{\mu})$ can be viewed as adding to the dataset some pseudodata corresponding to $\alpha - 1$ grid cells for each forest community type in which one tree from each of the M species is observed. We use $\alpha = 2$ as the regularization parameter.

3.2. *Modified EM algorithm.* To optimize the penalized likelihood in (6), we derive a modified EM algorithm. The computations required by both the η and $\boldsymbol{\mu}$ updates involve expectations over all possible K^n configurations of the forest community types. When the spatial correlation parameter $\eta_K \neq 0$, we approximate the exact updates by a stochastic procedure, which we describe in Section 3.3. When the spatial correlation parameter $\eta_K = 0$, it is possible to compute the expectations exactly, and we derive the EM updates for the spatially independent model in Section B.1 of Supplement A (Berg et al. (2019)).

Since the forest community types are unobserved, the problem of estimating the parameters $\theta = [\eta^T, \text{vec}(\boldsymbol{\mu})^T]^T$ falls naturally into the missing data framework and the expectation-maximization (EM) algorithm is a possible solution

(Dempster, Laird and Rubin (1977)). In each iteration of our modified EM algorithm, a surrogate function is constructed in the E-step and is based on the current parameter value θ^{cur}:

$$\begin{aligned} Q(\theta|\theta^{\mathrm{cur}}) &= \rho_1(\eta) + \rho_2(\boldsymbol{\mu}) + \sum_{z\in\Omega} p(z|\mathbf{y},\theta^{\mathrm{cur}})\log\{p(z,\mathbf{y}|\theta)\} \\ &= \Big[\rho_1(\eta) + \sum_{z\in\Omega} p(z|\mathbf{y},\theta^{\mathrm{cur}})\log\{p(z|\eta)\}\Big] \\ &\quad + \Big[\rho_2(\boldsymbol{\mu}) + \sum_{z\in\Omega} p(z|\mathbf{y},\theta^{\mathrm{cur}})\log\{p(\mathbf{y}|z,\boldsymbol{\mu})\}\Big] \\ &\equiv Q_1(\eta|\theta^{\mathrm{cur}}) + Q_2(\boldsymbol{\mu}|\theta^{\mathrm{cur}}). \end{aligned} \tag{9}$$

In the M-step the parameter value θ^{new} for the next iteration is obtained by maximizing the Q-function over θ. This process is repeated iteratively by setting $\theta^{\mathrm{cur}} = \theta^{\mathrm{new}}$ and then maximizing the new Q-function again. Under suitable conditions any limit point of such an EM algorithm is guaranteed to be a stationary point of the log likelihood (Wu (1983)).

The surrogate Q-function (9) takes an average over the complete data log likelihood $\log\{p(z,\mathbf{y}|\theta)\}$ with respect to the conditional distribution $p(z|\mathbf{y},\theta^{\mathrm{cur}})$ of the types, given the observed data $\mathbf{y}$ and evaluated at θ^{cur}, whereas the penalty functions $\rho_1(\eta)$ and $\rho_2(\boldsymbol{\mu})$ remain unchanged. The Q-function (9) can also be shown to minorize the regularized log likelihood (6), in the sense that

$$\ell_{\mathrm{pen}}(\theta) - \ell_{\mathrm{pen}}(\theta^{\mathrm{cur}}) > Q(\theta|\theta^{\mathrm{cur}}) - Q(\theta^{\mathrm{cur}}|\theta^{\mathrm{cur}}).$$

Thus, increasing the value of the Q-function guarantees an even greater increase in the value of the regularized log likelihood (6). The implementation detail for maximizing the Q-function is given as follows:

Update η: First, we deal with the maximization of the Q_1-function in (9),

$$Q_1(\eta|\theta^{\mathrm{cur}}) = \sum_{z\in\Omega} p(z|\mathbf{y},\theta^{\mathrm{cur}})\log\{p(z|\eta)\} + \rho_1(\eta). \tag{10}$$

Since $p(z|\eta)$ is in the exponential family with sufficient statistic $T(z)$, we have $\partial\log\{p(z|\eta)\}/\partial\eta = T(z) - E\{T(z')|\eta\}$ (Shao (2003)) and

$$\begin{aligned} &\frac{\partial Q_1(\eta|\theta^{\mathrm{cur}})}{\partial\eta}\bigg|_{\eta=\eta^{\mathrm{cur}}} \\ &\quad = \frac{\partial\rho_1(\eta)}{\partial\eta}\bigg|_{\eta=\eta^{\mathrm{cur}}} + \sum_{z'\in\Omega} p(z'|\mathbf{y},\theta^{\mathrm{cur}})[T(z') - E\{T(z)|\eta^{\mathrm{cur}}\}] \\ &\quad = \frac{\partial\rho_1(\eta)}{\partial\eta}\bigg|_{\eta=\eta^{\mathrm{cur}}} + E\{T(z)|\mathbf{y},\theta^{\mathrm{cur}}\} - E\{T(z)|\eta^{\mathrm{cur}}\}. \end{aligned} \tag{11}$$

Thus, the gradient of $Q_1(\eta|\theta^{\text{cur}})$ has a convenient representation in terms of the conditional and marginal distributions at $\theta = \theta^{\text{cur}}$. Furthermore, it can be shown that $\frac{\partial \ell_{\text{pen}}(\theta)}{\partial \eta}|_{\theta=\theta^{\text{cur}}} = \frac{\partial Q_1(\eta|\theta^{\text{cur}})}{\partial \eta}|_{\eta=\eta^{\text{cur}}}$.

Finding the η, which maximizes $Q_1(\eta|\theta^{\text{cur}})$ in the M-step, would require inverting, at every iteration, between the exponential family natural parameter, η, and the exponential family mean parameter, $E\{T(z)|\eta\}$. For Markov random field models, this inversion would require a sequence of MCMC draws and is a challenging computational problem (Forbes and Fort (2007)). Thus, we elect to instead use a gradient ascent update for the η component of θ:

$$\eta^{\text{new}} = \eta^{\text{cur}} + c^{-1} \frac{\partial Q_1(\eta|\theta^{\text{cur}})}{\partial \eta} = \eta^{\text{cur}} + c^{-1} \frac{\partial \ell_{\text{pen}}(\theta)}{\partial \eta}, \tag{12}$$

where c is a fixed constant stepsize chosen to ensure reasonable convergence behavior.

Update $\boldsymbol{\mu}$*:* In contrast to η, the update for $\boldsymbol{\mu}$ has a convenient representation in terms of the conditional distribution $p(z|\mathbf{y}, \theta^{\text{cur}})$, because by (3), we have

$$\begin{aligned} Q_2(\boldsymbol{\mu}|\theta^{\text{cur}}) &= \rho_2(\boldsymbol{\mu}) + \sum_{z \in \Omega} p(z|\mathbf{y}, \theta^{\text{cur}}) \log\{p(\mathbf{y}|z, \boldsymbol{\mu})\} \\ &= \rho_2(\boldsymbol{\mu}) + \sum_{z \in \Omega} p(z|\mathbf{y}, \theta^{\text{cur}}) \sum_{i=1}^{n} \sum_{k=1}^{K} \{\log(\boldsymbol{\mu}_k)^T \mathbf{y}_i\} I(z_i = k) \\ &\quad + \sum_{i=1}^{n} \log(C_i) \qquad (13) \\ &= \rho_2(\boldsymbol{\mu}) + \sum_{i=1}^{n} \sum_{k=1}^{K} \{\log(\boldsymbol{\mu}_k)^T \mathbf{y}_i\} P(z_i = k|\mathbf{y}, \theta^{\text{cur}}) + \sum_{i=1}^{n} \log(C_i) \\ &= \sum_{k=1}^{K} Q_2^k(\boldsymbol{\mu}_k) + \sum_{i=1}^{n} \log(C_i), \end{aligned}$$

where $\sum_{i=1}^n \log(C_i)$ does not depend on $\boldsymbol{\mu}$ and

$$Q_2^k(\boldsymbol{\mu}_k) = \sum_{m=1}^{M} (\alpha - 1) \log(\boldsymbol{\mu}_{mk}) + \sum_{i=1}^{n} \sum_{m=1}^{M} P(z_i = k|\mathbf{y}, \theta^{\text{cur}}) \mathbf{y}_{mi} \log(\boldsymbol{\mu}_{mk}).$$

It is shown in Section B.2 of Supplementary Material (Berg et al. (2019)) that the maximizer $\boldsymbol{\mu}^{\text{new}}$ of (13) has entries

$$\boldsymbol{\mu}_{mk}^{\text{new}} = \{\alpha - 1 + N_{mk}\} / \{M(\alpha - 1) + N_k\}, \tag{14}$$

where $N_{mk} = \sum_{i=1}^n P(z_i = k|\mathbf{y}, \theta^{\text{cur}}) \mathbf{y}_{mi}$ and $N_k = \sum_{m=1}^M N_{mk}$.

In a standard EM update for $\boldsymbol{\mu}$, we have $\boldsymbol{\mu}_k^{\text{new}} = \boldsymbol{\mu}_k^{\text{cur}} + (\boldsymbol{\mu}_k^{\text{new}} - \boldsymbol{\mu}_k^{\text{cur}})$. Here, it is more convenient to use an altered version, because $p(z|\mathbf{y}, \theta^{\text{cur}})$ is known only up to a normalizing constant, and the $\boldsymbol{\mu}$ update must be approximated by MCMC. The quantities related to $p(z|\mathbf{y}, \theta^{\text{cur}})$ appear in both the numerator and denominator of (14) and, thus, it is challenging to estimate the EM update for $\boldsymbol{\mu}$ in an unbiased fashion based only on a single draw from $p(z|\mathbf{y}, \theta^{\text{cur}})$. Thus, we propose a "short-step" for updates:

$$\tilde{\boldsymbol{\mu}}_k^{\text{new}} = \boldsymbol{\mu}_k^{\text{cur}} + \gamma_k(\boldsymbol{\mu}_k^{\text{new}} - \boldsymbol{\mu}_k^{\text{cur}}), \tag{15}$$

where

$$\gamma_k = \{M(\alpha - 1) + N_k\} \Big/ \left\{M(\alpha - 1) + \sum_{i=1}^{n} q_i\right\}. \tag{16}$$

Since $N_k < \sum_{i=1}^n q_i$ for all $\theta \in \Theta$, we have $\gamma_k < 1$ for all θ^{cur}. On the other hand $\gamma_k \geq \{M(\alpha - 1)\}/\{M(\alpha - 1) + \sum_{i=1}^n q_i\} > 0$. Thus, the $\tilde{\boldsymbol{\mu}}^{\text{new}}$ update results from taking a shortened EM step starting from $\boldsymbol{\mu}^{\text{cur}}$. For the product $\gamma_k \boldsymbol{\mu}_k^{\text{new}}$ in (15), the numerator of γ_k cancels with the denominator of $\boldsymbol{\mu}_k^{\text{new}}$ in (14). The denominator of γ_k depends on the number of tree species M, the regularization parameter α and the number of trees in the dataset $\sum_{i=1}^n q_i$. Thus, $\gamma_k \boldsymbol{\mu}_k^{\text{new}}$ depends on $p(z|\mathbf{y}, \theta^{\text{cur}})$ only through the numerator of the $\boldsymbol{\mu}_k^{\text{new}}$ update in (14) which can be estimated based on a single draw from $p(z|\mathbf{y}, \theta)$ (see Section 3.3).

The set $\mathscr{M}$ is convex, and from (14), $\boldsymbol{\mu}^{\text{new}} \in \mathscr{M}^K$. From convexity, when $\boldsymbol{\mu}^{\text{new}} \in \mathscr{M}^K$ and $\boldsymbol{\mu}^{\text{cur}} \in \mathscr{M}^K$, (15) implies $\tilde{\boldsymbol{\mu}}^{\text{new}} \in \mathscr{M}^K$ as well. Additionally, the update in (15) preserves the ascent property of the EM algorithm. By concavity of the log function, Q_2^k is concave, so that $Q_2^k(\tilde{\boldsymbol{\mu}}_k^{\text{new}}) \geq \gamma_k Q_2^k(\boldsymbol{\mu}_k^{\text{new}}) + (1 - \gamma_k) Q_2^k(\boldsymbol{\mu}_k^{\text{cur}}) \geq Q_2^k(\boldsymbol{\mu}_k^{\text{cur}})$. When $\boldsymbol{\mu}_k^{\text{new}} \neq \boldsymbol{\mu}_k^{\text{cur}}$, the inequalities are strict. Since $Q_1(\eta|\theta^{\text{cur}}) + Q_2(\boldsymbol{\mu}|\theta^{\text{cur}})$ minorizes $\ell_{\text{pen}}(\theta)$, any increase in the value of Q_2 implies an increase in the value of $\ell_{\text{pen}}(\theta)$.

3.3. *Stochastic approximation.*

To update θ, we devise a stochastic approximation procedure $\theta^{\text{new}} = \theta^{\text{cur}} + g(\theta^{\text{cur}})$, where

$$g(\theta^{\text{cur}}) = g\left(\begin{bmatrix} \eta^{\text{cur}} \\ \operatorname{vec}(\boldsymbol{\mu}^{\text{cur}}) \end{bmatrix}\right) = \begin{bmatrix} c^{-1} \dfrac{\partial Q_1(\eta|\theta^{\text{cur}})}{\partial \eta}\Big|_{\eta = \eta^{\text{cur}}} \\ \operatorname{vec}(\tilde{\boldsymbol{\mu}}^{\text{new}}) - \operatorname{vec}(\boldsymbol{\mu}^{\text{cur}}) \end{bmatrix} \tag{17}$$

and $g(\theta)$ is to be estimated based on MCMC samples. Stochastic approximation approaches are useful when the function $g(\cdot)$ is difficult or impossible to evaluate, but $g(\cdot)$ can be approximated by an estimate $G(\theta, \mathbf{z})$, where $\mathbf{z}$ is a random variable drawn from a distribution π_θ, such that π_θ and $G(\cdot, \cdot)$ satisfy, for each θ,

$$\int G(\theta, \mathbf{z}) \pi_\theta(d\mathbf{z}) = g(\theta) \tag{18}$$

(see, e.g., Benveniste, Métivier and Priouret (1990), Kushner and Yin (1997), Robbins and Monro (1951)).

The update in (17) is a combination of a gradient ascent update for η and a short-step update for $\boldsymbol{\mu}$, from which an iterate sequence may be constructed in the following way. Starting from an initial parameter $\theta^{(0)}$ and initial $\mathbf{z}^{(0)}$, we obtain draws $\mathbf{z}^{(t+1)}$ from $\pi_{\theta^{(t)}}(\cdot)$ and set $\theta^{(t+1)} = \theta^{(t)} + \epsilon^{(t+1)} G(\theta^{(t)}, \mathbf{z}^{(t+1)})$. The sequence of stepsizes $\{\epsilon^{(t)}\}$ is deterministic and generally satisfies conditions such as $\epsilon^{(t)} \downarrow 0$, $\sum_{t=1}^{\infty} (\epsilon^{(t)})^2 < \infty$ and $\sum_{t=1}^{\infty} \epsilon^{(t)} = \infty$ (see, e.g., Benveniste, Métivier and Priouret (1990), Kushner and Yin (1997)). Here, we use $\epsilon^{(t)} = t^{-1}$.

From (11) the gradient of $Q_1(\eta|\theta^{\text{cur}})$, and hence the gradient of the observed log likelihood, with respect to the η parameter can be computed based on the difference of two expectations of $T(z)$. The first expectation is taken with respect to the conditional distribution $p(z|\mathbf{y}, \theta^{\text{cur}})$, and the second expectation is taken with respect to the marginal distribution $p(z|\eta^{\text{cur}})$, while the gradient of the logistic prior $\rho_1(\eta)$ can be computed analytically. The $\boldsymbol{\mu}$ update $\tilde{\boldsymbol{\mu}}^{\text{new}} - \boldsymbol{\mu}^{\text{cur}}$ in (15) can be computed by taking the expectation of the function

$$\begin{aligned} &H_\alpha(\boldsymbol{\mu}^{\text{cur}}, z)_{mk} \\ &\quad = \frac{(\alpha - 1)(1 - M\boldsymbol{\mu}_{mk}^{\text{cur}}) + \sum_{i=1}^{n} I\{z_i = k\}(\mathbf{y}_{mi} - q_i \boldsymbol{\mu}_{mk}^{\text{cur}})}{M(\alpha - 1) + \sum_{i=1}^{n} q_i}, \end{aligned} \tag{19}$$

with respect to $p(z|\mathbf{y}, \theta)$. Thus, the update $g(\theta^{\text{cur}})$ in (17) can be written as an integration with respect to the density

$$\pi_{\theta^{\text{cur}}}(\mathbf{z}) = p(\mathbf{z}_1|\mathbf{y}, \theta^{\text{cur}}) p(\mathbf{z}_2|\eta^{\text{cur}}), \tag{20}$$

where $\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2)$ denotes an ordered pair of configurations $\mathbf{z}_1, \mathbf{z}_2 \in \Omega$. From (20) $\mathbf{z}_1$ and $\mathbf{z}_2$ are drawn independently under $\pi_{\theta^{\text{cur}}}(\cdot)$. The probability density $\pi_{\theta^{\text{cur}}}(\cdot)$ takes as its argument an element $\mathbf{z}$ of the sample space $\Omega^2 = \Omega \times \Omega$.

Since integrals with respect to $\pi_\theta(\cdot)$ require intractable sums over all possible type configurations $\mathbf{z}_1$ and $\mathbf{z}_2$, we estimate $g(\theta)$ based on approximate draws from $\pi_\theta(\cdot)$. While it is difficult to sample directly from $p(\mathbf{z}_1|\mathbf{y}, \theta)$ and $p(\mathbf{z}_2|\eta)$ due to the spatial correlation between the types z_i, it is possible to use Markov chain transition kernels (specifically, Gibbs sampling kernels) to approximate draws from these distributions (Geman and Geman (1984)). Due to the conditional independence between grid cells in the conditional distribution $p(\mathbf{y}|z, \boldsymbol{\mu})$, both $p(z|\mathbf{y}, \theta)$ and $p(z|\eta)$ are Markov random field densities with simple conditional distributions at each cell given the rest of the cells. Thus, it is possible to construct Gibbs sampling transition kernels $P_{1,\theta}(\cdot, \cdot) : \Omega \times \Omega \to [0, 1]$ and $P_{2,\theta}(\cdot, \cdot) : \Omega \times \Omega \to [0, 1]$, so that the stationary distributions of $P_{1,\theta}(\cdot, \cdot)$ and $P_{2,\theta}(\cdot, \cdot)$ are $p(z|\mathbf{y}, \theta)$ and $p(z|\eta)$, respectively. In order to approximately sample from π_θ, we run a Markov chain using the transition kernel $P_\theta(\mathbf{z}, \mathbf{z}') : \Omega^2 \times \Omega^2 \to [0, 1]$ defined by $P_\theta(\mathbf{z}, \mathbf{z}') = P_{1,\theta}(\mathbf{z}_1, \mathbf{z}_1') P_{2,\theta}(\mathbf{z}_2, \mathbf{z}_2')$. Detailed formulas for the Gibbs samplers are given in Section A.1 of the Supplementary Material (Berg et al. (2019)) (see also, e.g., Gaetan and Guyon (2010)). From the definition of $P_\theta(\cdot, \cdot)$, we see that the

Algorithm 1: Stochastic modified EM

```
Initialize parameter $\theta_0 \in \Theta$, configuration $\mathbf{z}_0 \in \Omega^2$, number of iterations $T$
for $t = 1$ to $T$ do
    Draw $\mathbf{z}_t \in \Omega^2$ according to $P_{\theta_{t-1}}(\mathbf{z}_{t-1}, \cdot)$
    $\epsilon_t = t^{-1}$
    $\theta_t = \theta_{t-1} + \epsilon_t G(\theta_{t-1}, \mathbf{z}_t)$
Return $\theta_T$
```

updates to the Markov chain for the conditional distribution $p(z|\mathbf{y}, \theta)$ are independent from the updates to the Markov chain for the marginal distribution $p(z|\eta)$.

In our stochastic modified EM procedure we choose a stepsize c and define the function $G(\cdot, \cdot) : \Theta \times \Omega^2 \to \Theta$ by

$$G(\theta, \mathbf{z}) = G\left(\begin{bmatrix} \eta \\ \operatorname{vec}(\boldsymbol{\mu}) \end{bmatrix}, \mathbf{z}\right) = \begin{bmatrix} c\left\{\dfrac{\partial \rho_1(\eta)}{\partial \eta} + T(\mathbf{z}_1) - T(\mathbf{z}_2)\right\} \\ H_\alpha(\boldsymbol{\mu}, \mathbf{z}_1) \end{bmatrix}. \tag{21}$$

Then, we find parameters $\hat{\theta}$ maximizing the penalized likelihood $\ell_{\text{pen}}(\theta)$ via the procedure given in Algorithm 1.

Implementation details, including a discussion of the choice of the stepsize c, are given in Section A.2 of the Supplementary Material (Berg et al. (2019)).

4. Case study: Historical forest communities based on public land survey data. The Wisconsin PLS dataset is a historical survey of trees, conducted primarily from 1832 to 1866 (Schulte and Mladenoff (2001)). The dataset has been commonly used in ecological studies of forest composition prior to and concurrent with Euro-American settlement. As described in the Introduction, surveyors from the PLS walked along a 1 mile by 1 mile grid-like pattern across the state and recorded the species of 2–4 representative trees at survey points every half-mile (Figure 1). The dataset is large, both in terms of the number of trees observed (328,499), distributed roughly uniformly across the state, as well as in terms of the spatial extent (145,000 square kilometers). Additionally, the tree species count data at each grid cell are highly multivariate, and for small enough grid cells most tree species counts are 0, since only 2–4 trees were observed at each survey point and the survey points are at least half a mile away from each other.

For data analysis a spatial grid of cells is first overlaid on the survey region (the state of Wisconsin). For each grid cell i, a count vector $\mathbf{y}_i$, of length $M = 33$ species, is constructed from the trees observed at survey points within that cell. Grid cells are not required to contain any trees. For our spatially correlated model the forest community type probability at any cell takes into account tree information from nearby adjacent and nonadjacent grid cells containing trees. We compare three grid resolutions—4 km by 4 km, 2 km by 2 km and 1 km by 1 km grid cells, resulting in 9469 cells, 37,134 cells and 146,851 cells, respectively. For each grid resolution each cell is assumed to have a single forest community type. We use a first-order spatial neighborhood structure with up to four nearest neighbors. That is, two points with integer lattice coordinates (i, j) and (i', j') are neighbors when $|i - i'| + |j - j'| = 1$. Next, we fit the spatially correlated multinomial mixture models via the stochastic modified EM procedure in Algorithm 1. For comparison we fit spatially independent multinomial mixture models via the standard EM algorithm, a derivation of which is given in Section B.1 of the Supplementary Material (Berg et al. (2019)).

4.1. *Choice of K and model validation.* We use a cross-validation procedure to determine the number of forest community types to use and to assess the quality of the spatially correlated mixture models relative to the spatially independent mixture models. In particular we generate a testing dataset by randomly selecting 20 percent of the trees from the full set of surveyed trees. The remaining 80 percent of the trees are placed in a training dataset. We then create training and testing datasets $\mathbf{y}_{\text{train}}$ and $\mathbf{y}_{\text{test}}$ for each grid resolution (1 km, 2 km and 4 km) from these training and testing trees. We also consider five total numbers of forest community types $K = 8, 12, 16, 20$, or 24. For each combination of grid resolution and number of forest community types, we fit each of the models on the training dataset $\mathbf{y}_{\text{train}}$, starting from three random initial parameter values to mitigate the multimodality of the likelihood.

We examine two log-likelihood based measures of prediction performance, using the same training and testing datasets across models fit for different grid resolutions and numbers of forest community types to ensure the likelihoods are comparable among different models. We first compute, for each of the fitted models at each grid resolution, a holdout log likelihood

$$\ell_{\text{holdout}}(\hat{\theta}) = \log\left\{\sum_{z\in\Omega} p(\mathbf{y}_{\text{test}}|z, \hat{\theta})p(z|\hat{\theta})\right\}. \tag{22}$$

We focus our model assessment on holdout log likelihoods rather than on the errors of estimated coefficients, because the true data-generating parameters are unknown for the real data. Next, we compute a predictive log likelihood

$$\ell_{\text{pred}}(\hat{\theta}) = \log\{p(\mathbf{y}_{\text{test}}|\mathbf{y}_{\text{train}}, \hat{\theta})\} = \log\left\{\sum_{z\in\Omega} p(\mathbf{y}_{\text{test}}|z, \hat{\theta})p(z|\mathbf{y}_{\text{train}}, \hat{\theta})\right\}. \tag{23}$$

In contrast to the holdout log likelihood $\ell_{\text{holdout}}(\hat{\theta})$ that is marginal on the testing dataset, the predictive log likelihood $\ell_{\text{pred}}(\hat{\theta})$ measures the quality of predictions of the testing dataset, conditional on the training dataset. Since our maps of the study area are ultimately based on the conditional distribution $p(z|\mathbf{y}, \hat{\theta})$, the predictive log likelihood is a relevant performance metric. To ensure that these likelihoods are comparable across different grid resolutions, we drop the grid-resolution dependent factors C_i in (3). The log likelihoods without the constants C_i are equal to the log likelihoods of the individual trees, before being aggregated into counts. Unlike the tree species counts $\mathbf{y}_i$ for each cell, which vary by the grid resolution, the log likelihood of the collection of individual trees has the same interpretation across grid resolutions.

For the spatially correlated models the holdout log likelihood in (22) is difficult to compute, and we use path integration, also known as thermodynamic integration or path sampling (Gelman and Meng (1998), Neal (1993)). We describe the path integration procedure in Section A.3 of the Supplementary Material (Berg et al. (2019)). The predictive log likelihood (23) is also difficult to compute for the spatially correlated models. By the fact that

$$\begin{aligned}\log\{p(\mathbf{y}_{\text{test}}|\mathbf{y}_{\text{train}},\hat{\theta})\} &= \log\{p(\mathbf{y}_{\text{train}},\mathbf{y}_{\text{test}}|\hat{\theta})\} - \log\{p(\mathbf{y}_{\text{train}}|\hat{\theta})\}\\ &= \log\{p(\mathbf{y}|\hat{\theta})\} - \log\{p(\mathbf{y}_{\text{train}}|\hat{\theta})\},\end{aligned} \tag{24}$$

we write $\ell_{\text{pred}}(\hat{\theta})$ as the difference between the two marginal likelihoods in (24) and use path integration to compute these two marginal log likelihoods separately.

Table 1 displays the holdout and predictive log likelihoods obtained from the spatial and independent models for the different grid resolutions and numbers of forest community types. Intuitively, we expect it to be easier to predict the held out trees after having seen spatially nearby training trees. A comparison of the

TABLE 1

Values of holdout log likelihood (ℓ_{holdout}) *and predictive log likelihood* (ℓ_{pred}) *for the Wisconsin Public Land Survey case study for either spatially independent models or the spatially correlated models*, *different numbers of forest community types* (K), *and the grid resolution* (1 *km*, 2 *km or* 4 *km*), *averaged over three runs from random initial starting parameters*, *and normalized by the number of trees in the testing dataset*

		$\ell_{\text{holdout}}(\hat{\theta})$			$\ell_{\text{pred}}(\hat{\theta})$		
Model	K	1 km	2 km	4 km	1 km	2 km	4 km
Independent	8	−2.77	−2.6	−2.37	−2.11	−2.15	−2.18
	12	−2.76	−2.58	−2.35	−2.04	−2.09	−2.13
	16	−2.76	−2.57	−2.33	−2	−2.06	−2.1
	20	−2.75	−2.57	−2.32	−1.98	−2.03	−2.08
	24	−2.75	−2.57	−2.32	−1.96	−2.02	−2.07
Spatial	8	−2.23	−2.21	−2.22	−2.03	−2.13	−2.19
	12	−2.18	−2.16	−2.2	−1.96	−2.08	−2.17
	16	−2.15	−2.15	−2.2	−1.91	−2.05	−2.15
	20	−2.15	−2.15	−2.18	−1.9	−2.03	−2.14
	24	−2.15	−2.15	−2.18	−1.88	−2.04	−2.14

marginal and conditional log likelihoods in Table 1 bears this out; the predictive log likelihoods $\ell_{\text{pred}}(\hat{\theta})$ are always larger than the holdout log likelihoods $\ell_{\text{holdout}}(\hat{\theta})$.

At all grid resolutions and numbers of forest community types, the spatial model performs better based on holdout log likelihood than the corresponding spatially independent model. Additionally, the highest (best) spatially independent holdout log likelihood is lower than the holdout log likelihood from even the worst spatially correlated model. For the spatially independent models the holdout log likelihoods for models with fixed numbers of forest community types decrease as the grid resolution becomes finer, while the holdout log likelihoods for the spatial models with fixed number of forest community types are more similar across the grid resolutions.

In contrast to the holdout log likelihoods, the predictive log likelihoods for both the spatially correlated and independent models improve as the grid resolution becomes finer. Additionally, the predictive log likelihoods increase monotonically at each grid resolution as more forest community types are added to the model. The largest (best) predictive log likelihood is obtained for a 1 km spatial model with 24 forest community types. The spatially independent models sometimes achieve higher predictive log likelihoods at the 2 km and 4 km grid resolutions, but the best predictive log likelihoods out of all the models are attained by spatial models at the 1 km resolution.

Finally, model fits from different initializations on the PLS dataset, where the true data generating mechanism is unknown, were qualitatively similar with some variability in the fitted forest communities. For a fixed number of forest community types, the correlation parameter estimates are typically similar across the grid resolutions. For example, for the 16-community models, the smallest spatial correlation parameter estimates are 1.615, 1.610 and 1.549, whereas the largest are 1.631, 1.619 and 1.596, for the grid resolutions 1 km, 2 km and 4 km, respectively.

4.2. *Ecological interpretation.* After model fitting, the forest community classifications at each grid cell are determined from sitewise maximum a posteriori (MAP) estimates using Gibbs sampling. Maps of these classifications are shown in Figures 2–3, which indicate that the spatially correlated models tend to produce more spatially smooth classification maps than the spatially independent models, particularly for the smaller grid resolutions, as is expected. A key to the tree species abbreviations in these figures is given in Table 2.

The predictive log likelihoods for both the spatially correlated and independent models improve as the grid resolution decreases from 4 km to 1 km (Table 1), which suggests that the tree data are more likely to come from the same forest community type within smaller grid cells, and that the larger grid cells aggregate trees from multiple forest community types. This pattern is consistent with ecological observations of forest patch size in the region (Mladenoff et al. (1993)).

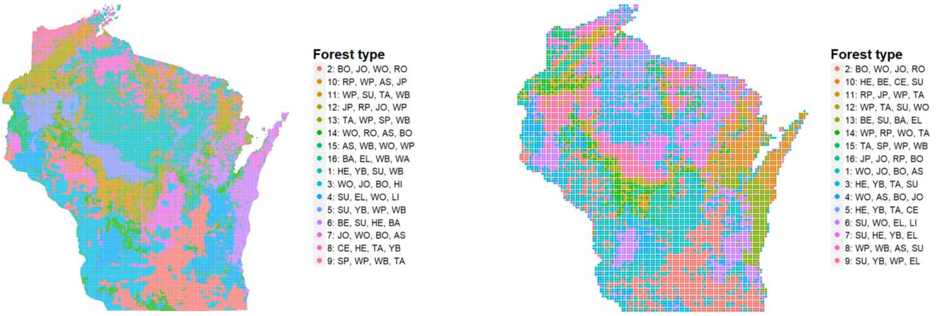
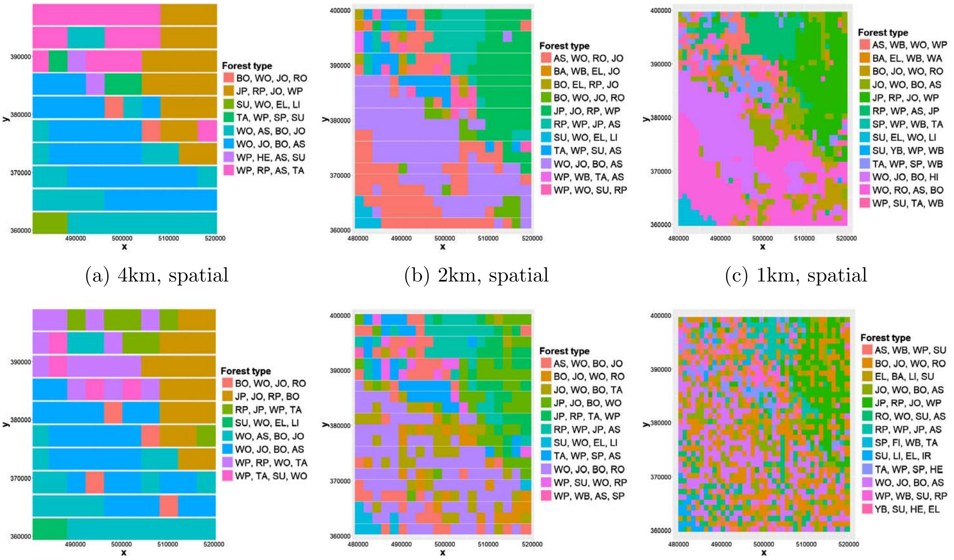


FIG. 2. *Forest community classifications for the public land survey case study from the* 16-*community spatially correlated* (*left*) *and spatially independent* (*right*) *models with the highest holdout log likelihoods which occurred at the* 1 *km and* 4 *km grid resolutions*, *respectively. A key to the tree species abbreviations is given in Table* 2.

We focus our ecological interpretation on the spatially correlated model with 1 km grid resolution and 16 forest community types which has the highest holdout log likelihood out of the 1 km models. This model also has the highest holdout log likelihood for the 16 forest community models across grid resolutions. Table 3 summarizes the forest communities for this model and indicates that species associations within these 16 forest communities are consistent with ecological ex-



(a) 4km, spatial (b) 2km, spatial (c) 1km, spatial

(d) 4km, independent (e) 2km, independent (f) 1km, independent

FIG. 3. *A comparison of* 16-*community spatial and independent models over a* 40 *km by* 40 *km subsection of the Wisconsin survey region for grid resolutions* 4 *km*, 2 *km and* 1 *km*. *A key to the tree species abbreviations is given in Table* 2.

TABLE 2
The tree species abbreviations, names, and counts for the Wisconsin Public Land Survey case study

Abbreviation	Name	Count	Abbreviation	Name	Count
AL	Alder	100	LI	American basswood	7520
AS	Aspen	12,029	RM	Red maple	1475
BA	Black ash	5957	RO	Red oak	5228
BE	American beech	7586	RP	Red pine	9925
BO	Bur oak	34,065	SO	Swamp white oak	207
BU	Butternut	534	SP	Spruce	6048
BW	Black walnut	113	SU	Sugar maple	32,718
CE	White cedar	8297	TA	Tamarack	19,741
CH	Black cherry	454	WA	White ash	2119
CO	Eastern cottonwood	122	WB	Paper birch	11,770
EL	Elm	11,090	WI	Willow	346
FI	Balsam fir	4441	WM	Silver maple	550
HE	Eastern hemlock	26,369	WO	White oak	33,170
HI	Shagbark hickory	1198	WP	Eastern white pine	21,717
IR	Ironwood	4076	YB	Yellow birch	22,008
JO	Black & northern pin oak	26,058	ZZ	No trees	464
JP	Jack pine	11,004			

TABLE 3
Summaries of the 16 *estimated forest community types for the Wisconsin Public Land Survey case study under the model selected based on cross validation, including the counts of grid cells which are classified as each forest community type, the top four tree species in each forest community and the corresponding four largest estimated species probabilities. A key to the tree species abbreviations is given in Table* 2

Forest type	Count	Top Species	Species Probabilities
1	23,174	HE, YB, SU, WB	0.352, 0.223, 0.154, 0.048
2	19,373	BO, JO, WO, RO	0.72, 0.137, 0.117, 0.007
3	16,121	WO, JO, BO, HI	0.478, 0.267, 0.209, 0.013
4	12,380	SU, EL, WO, LI	0.282, 0.151, 0.131, 0.121
5	10,746	SU, YB, WP, WB	0.342, 0.201, 0.089, 0.071
6	8427	BE, SU, HE, BA	0.39, 0.138, 0.096, 0.065
7	7370	JO, WO, BO, AS	0.643, 0.177, 0.091, 0.031
8	7219	CE, HE, TA, YB	0.283, 0.175, 0.133, 0.094
9	7075	SP, WP, WB, TA	0.179, 0.167, 0.161, 0.146
10	6013	RP, WP, AS, JP	0.489, 0.19, 0.077, 0.061
11	5720	WP, SU, TA, WB	0.66, 0.055, 0.046, 0.042
12	5506	TA, WP, SP, WB	0.821, 0.036, 0.026, 0.015
13	5420	JP, RP, JO, WP	0.753, 0.087, 0.052, 0.024
14	4865	WO, RO, AS, BO	0.484, 0.221, 0.069, 0.067
15	4256	AS, WB, WO, WP	0.62, 0.063, 0.045, 0.043
16	3186	BA, EL, WB, WA	0.236, 0.161, 0.073, 0.064

pectation for the survey region (Curtis (1959)). Similarly, the maps of the most likely forest community for each grid cell generally meet expectations (Curtis (1959), Finley (1976)). Although one of the models at the 2 km grid resolution has a higher holdout log likelihood than the model we discuss here; the difference in holdout likelihoods was small (−2.143 for the 2 km model vs. −2.144 for the 1 km model), while the improvement in predictive log likelihood from 2 km to 1 km is more substantial (−2.04 for the 2 km model, vs. −1.91 for the 1 km model).

Among the oak communities, bur oak (BO) is the highest probability species in the community (forest community type 2 in Table 3) that is most likely to occur in the region that was historically oak savanna, mainly in topographically gentle sites (Curtis (1959)). While all oak species in the survey region are fire adapted, bur oak is the most fire tolerant (Peterson and Reich (2001)). Its dominance in flatter areas could be due to increased frequencies of prairie fires passing through (Shea, Schulte and Palik (2014), Stambaugh and Guyette (2008)). A more mixed oak community (forest community type 3), dominated by white oak (WO) with a high probability of black/jack oak (JO) and bur oak, was most likely to occur in a more topographically diverse, historically savanna region; the topography likely allowed for more diverse fire patterns and species assemblages (Shea, Schulte and Palik (2014)). The community dominated by black/jack oak (forest community type 7) had highest probability in regions with dry soils; of the oak species in Wisconsin, black and jack oak are the most drought tolerant, so their dominance on these sites is ecologically sensible (Curtis (1959), Shea, Schulte and Palik (2014)). While the other oak species likely were restricted to sunny savannas, the white oak-red oak (RO) community (forest community type 14) may have existed as a closed canopy community in southern Wisconsin. White oak and red oak are the more shade tolerant oaks (Curtis (1959)).

The three pine species in Wisconsin occur in several communities; three of which are each dominated by the three species. The separation of the three species is expected, because while they are all associated with drier site conditions (Curtis (1959)), they are each differently adapted to drought and fire and, especially for jack pine (JP) and red pine (RP), often form monospecific stands depending on fire frequency (Burns and Honkala (1990), Radeloff et al. (1999)). White pine (WP) has greater than 0.1 probability in the red pine dominated community (forest community type 10) as well as in a community (forest community type 9) with similar probabilities of spruce (SP), paper birch (WB) and tamarack (TA). Compared to the other pine species, white pine grows on a range of sites including those with richer soil and has intermediate shade tolerance which allows it occur on a variety of sites and even intergrade with northern mesic forest community types (Burns and Honkala (1990), Curtis (1959), Fahey, Lorimer and Mladenoff (2012)). Given the widespread nature of white pine, it is not surprising that it has high probability of occurring in more than one community, including forest community type 9 which has species combinations that are possible on sites with recent disturbance or sites that are refuges from fire (Fahey, Lorimer and Mladenoff (2012)).

In northern Wisconsin, mesic forest occurs on sites with rich and moist, but well drained, soils and is mainly dominated by eastern hemlock (HE), sugar maple (SU), yellow birch (YB) and American beech (BE) (Curtis (1959)). The cluster results separate this forest type into four communities, and probabilities of each forest community type seem to vary geographically, depending on the range boundaries of several species (Curtis (1959), Davis, Schwartz and Woods (1991)). Beech dominates one community; sugar maple and hemlock are other high-probability species in the community (forest community type 6) which is most likely to occur east of beech's range boundary in eastern Wisconsin.

In northern Wisconsin, forest community type 8 is the most likely community, where hemlock has the highest probability along with white cedar, yellow birch and sugar maple. White cedar (CE) is most abundant in far northern Wisconsin (Curtis (1959)). South of that a different community is more likely to occur (forest community type 1) with high probability of hemlock, yellow birch and sugar maple. West and south of the range of hemlock, forest community type 5 is most likely to occur; in that community hemlock is absent, and sugar maple and yellow birch dominate.

The remaining communities also align with expected forest types. In southern Wisconsin community type 4 is southern mesic forest which is most likely in known closed forest areas as expected (Curtis, 1959, Mladenoff et al.). Community type 16 is wet-mesic forest in both north and south (Curtis (1959)). Forest community type 13 is a tamarack wetland, and forest community type 15 is northern dry/dry-mesic sites that are recently disturbed and dominated by aspen (AS) (Curtis (1959)).

4.3. *Model diagnostic and implementation validation.* In addition to the log likelihoods, we consider an absolute deviation measure of discrepancy between the observed and predicted proportions of tree species. To compute this measure of discrepancy, we overlay a grid of n 20 km by 20 km cells on the state of Wisconsin and compute the discrepancy

$$D = (Mn)^{-1} \sum_{i=1}^{n} \sum_{m=1}^{M} |\bar{p}_{mi} - \hat{p}_{mi}|,$$

where i indexes the 20 km by 20 km grid cells, $\bar{p}_{mi} = \mathbf{y}_{mi,\text{test}}/q_i^{\text{test}}$ denotes the empirical proportion of testing species m trees in the ith grid cell and $\hat{p}_{mi}$ denotes the corresponding predicted proportion under a given model. The discrepancy D measures the average difference between the observed and predicted proportions

TABLE 4

Values of ℓ_1 discrepancy (D) on the testing dataset on a 20 km by 20 km grid for the spatially independent and dependent models with different numbers of forest community types (K) and grid resolutions (1 km, 2 km, 4 km) in the Wisconsin Public Land Survey case study

Model	K	1 km	2 km	4 km
Independent	8	0.0153	0.0136	0.013
	12	0.014	0.0118	0.012
	16	0.0132	0.0111	0.0111
	20	0.013	0.0107	0.0108
	24	0.0129	0.0104	0.0103
Spatial	8	0.0125	0.0132	0.014
	12	0.011	0.0115	0.0131
	16	0.00994	0.011	0.0127
	20	0.00957	0.0106	0.0123
	24	0.00933	0.0106	0.0123

in the 20-km by 20-km grid cells. For the mixture models the predicted species proportions for the ith grid cell $\hat{p}_{mi}$ are

$$\hat{p}_{mi} = \sum_{z \in \Omega} p(z|\mathbf{y}, \hat{\theta}) \sum_{k=1}^{K} I(z_i = k)\hat{\boldsymbol{\mu}}_{mk}.$$

We compute these predicted species probabilities analytically for the spatially independent models but via MCMC for the spatially correlated models.

From Table 4, the overall differences between the predicted and observed species proportions are small, indicating good fit between the observed and predicted species proportions. The best performing models with respect to the measure D achieve an average absolute deviation of about 0.01 between the observed and predicted proportion for each of the 33 species. The deviations for the spatially correlated models decrease as the grid resolution becomes finer, in contrast to the deviations for the spatially independent models which increase as the grid resolution becomes finer. The overall pattern for the absolute deviations, as the number of categories and the grid resolutions change, is similar to the pattern for the predictive log likelihoods in Table 1.

For the spatial models, we also investigate an intuitive approximation of $p(\mathbf{y}_{\text{test}}|\mathbf{y}_{\text{train}}, \hat{\theta})$ which allows us to validate our path integration implementation. Under the assumption

$$p(z|\mathbf{y}_{\text{train}}, \hat{\theta}) \approx \prod_{i=1}^{n} p(z_i|\mathbf{y}_{\text{train}}, \hat{\theta}),$$

TABLE 5
Predictive log likelihood values on the testing dataset for the spatially correlated model, computed using path integration and the approximate method of (25), *for the Wisconsin Public Land Survey case study with different numbers of forest community types* (K) *and grid resolutions* (1 *km*, 2 *km*, 4 *km*)

Method	K	1 km	2 km	4 km
Path integral	8	−2.03	−2.13	−2.19
	12	−1.96	−2.08	−2.17
	16	−1.91	−2.05	−2.15
	20	−1.9	−2.03	−2.14
	24	−1.88	−2.04	−2.14
Approximate	8	−2.03	−2.14	−2.19
	12	−1.96	−2.08	−2.17
	16	−1.91	−2.05	−2.15
	20	−1.9	−2.03	−2.14
	24	−1.88	−2.04	−2.14

we have

$$\begin{aligned}\log\{p(\mathbf{y}_{\text{test}}|\mathbf{y}_{\text{train}})\} &= \log\left\{\sum_{z\in\Omega} p(\mathbf{y}_{\text{test}}|z,\hat{\theta})p(z|\mathbf{y}_{\text{train}},\hat{\theta})\right\} \\ &\approx \sum_{i=1}^{n}\log\left\{\sum_{k=1}^{K} p\{\mathbf{y}_{\text{test}}|z_i,\hat{\theta}\}p(z_i=k|\mathbf{y}_{\text{train}},\hat{\theta})\right\}.\end{aligned} \tag{25}$$

Using this approximation combined with MCMC draws from $p(z|\mathbf{y}_{\text{train}},\hat{\theta})$ to obtain empirical estimates of $p(z_i = k|\mathbf{y}_{\text{train}},\hat{\theta})$, we compute an approximation of the true predictive log likelihood $\ell_{\text{pred}}(\hat{\theta})$, denoted as $\ell_{\text{pred}}^{\text{approx}}(\hat{\theta})$. Table 5 suggests that the results from this approximate procedure agree very well with the results obtained via path integration in spite of the mostly different implementation details, providing evidence for the correctness of our path integral implementation.

5. Simulation study. We conduct a simulation study to evaluate the methodology applied to the PLS case study in Sections 2–4. We consider $g \times g$ grids of cells, where the grid size is $g = 50, 100, 200$ or 400 corresponding to $n = 2500$, 10,000, 40,000 or 160,000 grid cells, respectively. We also consider the effect of observing larger and smaller numbers of trees within each cell by conducting simulations at $q = 3$ or *six* trees observed per cell. For each combination of grid size (g) and number of trees per cell (q), 100 simulations are performed. There are $K = 8$ true forest community types, with associated probabilities given in the $\boldsymbol{\mu}$

matrix below where the $K = 8$ columns of $\boldsymbol{\mu}$ each sum to 1.

$$\boldsymbol{\mu} = \begin{array}{c} \begin{array}{cccccccc} \text{Class 1} & \text{Class 2} & \text{Class 3} & \text{Class 4} & \text{Class 5} & \text{Class 6} & \text{Class 7} & \text{Class 8} \end{array} \\ \begin{bmatrix} 0.186 & 0.126 & 0.049 & 0.264 & 0.036 & 0.212 & 0.031 & 0.403 \\ 0.228 & 0.177 & 0.086 & 0.139 & 0.465 & 0.016 & 0.015 & 0.057 \\ 0.016 & 0.015 & 0.016 & 0.026 & 0.064 & 0.022 & 0.016 & 0.054 \\ 0.089 & 0.016 & 0.035 & 0.299 & 0.022 & 0.041 & 0.235 & 0.021 \\ 0.026 & 0.018 & 0.015 & 0.024 & 0.134 & 0.016 & 0.220 & 0.015 \\ 0.019 & 0.092 & 0.103 & 0.016 & 0.016 & 0.065 & 0.045 & 0.027 \\ 0.044 & 0.015 & 0.015 & 0.016 & 0.016 & 0.016 & 0.016 & 0.125 \\ 0.015 & 0.195 & 0.016 & 0.016 & 0.111 & 0.021 & 0.019 & 0.062 \\ 0.028 & 0.133 & 0.199 & 0.059 & 0.040 & 0.109 & 0.049 & 0.018 \\ 0.036 & 0.015 & 0.025 & 0.015 & 0.016 & 0.360 & 0.025 & 0.021 \\ 0.017 & 0.017 & 0.016 & 0.046 & 0.015 & 0.057 & 0.026 & 0.021 \\ 0.039 & 0.016 & 0.017 & 0.030 & 0.015 & 0.017 & 0.016 & 0.015 \\ 0.223 & 0.094 & 0.015 & 0.015 & 0.016 & 0.016 & 0.016 & 0.016 \\ 0.016 & 0.016 & 0.374 & 0.016 & 0.018 & 0.016 & 0.016 & 0.027 \\ 0.017 & 0.054 & 0.017 & 0.018 & 0.016 & 0.019 & 0.257 & 0.117 \end{bmatrix} \end{array}$$

The simulated vectors of forest community types z have density

$$p(z|\eta) = \exp\{\eta^T T(z) - \xi(\eta)\},$$

where $\eta = [-0.060, -0.055, -0.039, -0.037, -0.024, -0.057, -0.004, 1.2]^T$ and $T(z)$ is defined as in (2). That is, the spatial correlation parameter $\eta_K = 1.2$. Given the forest community types $Z = z$, the tree count vectors $\mathbf{Y}_i$ are independent multinomials with sample sizes $q = 3$ or $q = 6$ trees at each grid cell. Since the regularized likelihood is invariant to permutations of the mixture categories, we use the permutation of categories that minimizes the mean squared errors (MSE),

$$\text{MSE} = \sum_{k=1}^{8} \sum_{m=1}^{15} (\widehat{\boldsymbol{\mu}}_{mk} - \boldsymbol{\mu}_{mk})^2/(MK),$$

for each simulation when assessing the estimation error. The MSE for the $\boldsymbol{\mu}$ matrix are reported in Table 6 for the stochastic modified EM algorithm described in Algorithm 1 (“Modified EM”), in comparison to the spatially independent model fit via the EM algorithm (“Independent EM”), the ordinary stochastic gradient as described in Younes (1989) (“Ordinary SG”) and a version of stochastic gradient with differently scaled stepsizes for the η and $\boldsymbol{\mu}$ parameter (“Rescaled SG”). Implementation details for ordinary and rescaled stochastic gradient are given in Section B.3 of the Supplementary Material (Berg et al. (2019)).

Table 6 suggests that the modified EM algorithm performs best at every setting, followed by independent EM. When only $q = 3$ trees are included in each cell, the MSEs for the $\boldsymbol{\mu}$ parameter from the spatially independent model are over

TABLE 6

Simulation mean squared error (MSE) for the species probability parameter matrix $\boldsymbol{\mu}$ using different algorithms for $q = 3, 6$ simulated trees per grid cell and for different numbers of grid cells n

Method	Trees per cell	$n = 50^2$	$n = 100^2$	$n = 200^2$	$n = 400^2$
Modified EM	$q = 3$	2e−04	4e−05	1e−05	2e−06
	$q = 6$	5e−05	1e−05	3e−06	7e−07
Independent EM	$q = 3$	4e−04	8e−05	2e−05	6e−06
	$q = 6$	2e−04	4e−05	2e−05	8e−07
Rescaled SG	$q = 3$	7e−04	5e−04	4e−04	5e−04
	$q = 6$	4e−04	4e−04	4e−04	3e−04
Ordinary SG	$q = 3$	0.003	0.003	0.003	0.003
	$q = 6$	0.002	0.002	0.002	0.002

twice that of the spatially correlated. When $q = 6$ trees are included at each cell, the performance of the spatially correlated and independent models are more similar, although the spatially correlated model still always performs better than the spatially independent model. For both models the MSE at each grid size is, as expected, lower when $q = 6$ trees are included than when $q = 3$ trees are included. For both the spatially correlated and independent models, the parameter estimates $\hat{\boldsymbol{\mu}}$ appear to be converging to the truth at about the rate of $\sqrt{n}$. The convergence occurs in spite of the fact that the likelihood is multimodal, while the fitting algorithms were randomly initialized. This suggests that the estimation procedure is robust to the choice of initialization. Interestingly, the rescaled stochastic gradient performs better than ordinary stochastic gradient but still performs worse than the independent EM algorithm.

The spatially independent model does not include the spatial correlation parameter η_K, so that Table 7 compares the bias, variance and mean squared errors for the spatial correlation parameter η_K only for the stochastic modified EM, ordinary stochastic gradient and rescaled stochastic gradient algorithms. Again, the stochastic modified EM algorithm performs better than either rescaled stochastic gradient or ordinary stochastic gradient. This difference is particularly pronounced for the larger grid sizes. For $g = 400$ and $q = 6$, the MSE for the stochastic modified EM algorithm is approximately 100 times smaller than the MSE for the rescaled stochastic gradient algorithm. As can be seen from Table 7, the component of MSE due to bias for the stochastic modified EM algorithm is very small relative to the component of MSE due to variance. Additionally, the MSE decreases monotonically as the grid size increases, as well as when more trees are observed within each cell. This suggests that our algorithm accurately recovers the spatial correlation parameter in the Potts distribution.

Since the spatially independent model does not include the spatial correlation parameter η_K, the estimates for η_k when $k < K$ (Table 8) for the independence

TABLE 7

Simulation bias, variance, and mean squared error (MSE) for the spatial correlation parameter η_K using different algorithms for $q = 3$, six simulated trees per grid cell and for different numbers of grid cells n

Method	Error	$n = 50^2$ $q = 3$	$n = 50^2$ $q = 6$	$n = 100^2$ $q = 3$	$n = 100^2$ $q = 6$	$n = 200^2$ $q = 3$	$n = 200^2$ $q = 6$	$n = 400^2$ $q = 3$	$n = 400^2$ $q = 6$
Modified EM	Bias	−0.002	−0.009	−0.002	−0.002	−0.001	−1e−05	−0.002	−6e−04
	Variance	9e−04	6e−04	2e−04	1e−04	4e−05	3e−05	8e−06	7e−06
	MSE	9e−04	7e−04	2e−04	1e−04	4e−05	3e−05	1e−05	7e−06
Rescaled SG	Bias	−0.03	−0.02	−0.02	−0.02	−0.02	−0.02	−0.02	−0.02
	Variance	0.002	0.001	5e−04	4e−04	2e−04	3e−04	2e−04	3e−04
	MSE	0.003	0.002	0.001	8e−04	7e−04	6e−04	6e−04	6e−04
Ordinary SG	Bias	−0.2	−0.2	−0.2	−0.2	−0.2	−0.2	−0.2	−0.2
	Variance	0.005	0.006	0.005	0.006	0.005	0.007	0.005	0.005
	MSE	0.04	0.03	0.04	0.03	0.05	0.03	0.05	0.03

TABLE 8

Simulation mean squared error (MSE) for the η_k parameters when $k < K$ using different algorithms for $q = 3, 6$ simulated trees per grid cell and for different numbers of grid cells n

Method	Trees per cell	$n = 50^2$	$n = 100^2$	$n = 200^2$	$n = 400^2$
Modified EM	$q = 3$	0.0031	0.00082	0.00017	5.9e−05
	$q = 6$	0.0031	0.00053	0.00011	2.6e−05
Independent EM	$q = 3$	0.24	0.066	0.041	0.03
	$q = 6$	0.082	0.041	0.027	0.026
Rescaled SG	$q = 3$	0.0099	0.0083	0.0089	0.0085
	$q = 6$	0.044	0.043	0.041	0.035
Ordinary SG	$q = 3$	0.62	0.66	0.69	0.61
	$q = 6$	1	1	1.2	1.1

model are expected to be biased relative to the true data generating η_k parameters, so that comparisons between the correlated and uncorrelated model estimates are less meaningful for these parameters. The stochastic modified EM algorithm performs best out of all the methods for every combination of grid size and number of trees per grid cell.

Finally, in our simulation study, the minimum number of trees in a dataset is 7500, while in the PLS case study, over 300,000 trees were observed. Thus, the "prior sample sizes" of trees from each forest community type are much smaller than the observed sample size, and we do not expect the prior penalties to substantially bias the estimation procedure. The simulation study results bear this out. Additionally, the η parameters are estimated in simulation with very little bias due to the regularization.

6. Conclusions and discussion.

In this work we have modeled forest communities on a landscape via a latent Markov random field model. The spatially correlated model outperformed the spatially independent model for parameter estimation in a simulation study and for prediction on the historical Wisconsin PLS dataset. The fitted models were sensible relative to prior ecological literature, and we provided ecological interpretation of the fitted models on the PLS dataset. We also proposed a stochastic approximation procedure for jointly estimating the forest community species compositions and the spatial correlation strength in our latent Markov random field model.

In Forbes et al. (2013), the spatial correlation structure includes additional parameters to allow the interaction strength to depend on the forest types. We achieved adequate results with a single spatial correlation parameter and leave the investigation of more sophisticated spatial correlation structures to future work. It would also be interesting to relate the forest community classifications to environmental covariates across the PLS survey area. Furthermore, in a spatial-temporal setting with forest successional dynamics, for example, the layout of the grid cells could include neighbors in space and time, and one might include additional correlation parameters to account for possible temporal dependence, as well as spatiotemporal interactions that differ by forest communities. Finally, our computational method can be used for conditional distributions different from the multinomial distribution. While we provide a computationally feasible method in this work, parameter estimation for noisily observed Markov random fields is still computationally challenging. We leave these for future research as well.

Acknowledgments. We thank the Editor, an anonymous Associate Editor and an anonymous referee for their insightful and constructive comments that have helped improve the manuscript.

SUPPLEMENTARY MATERIAL

Supplement to "A latent discrete Markov random field approach to identifying and classifying historical forest communities based on spatial multivariate tree species counts" (DOI: 10.1214/19-AOAS1259SUPP; .pdf). Contains additional description of computational methods.

REFERENCES

BARNES, B., ZAK, D., DENTON, S. and SPURR, S. (2010). *Forest Ecology*, 4th ed. Wiley, New York.

BENVENISTE, A., PRIOURET, P. and MÉTIVIER, M. (1990). *Adaptive Algorithms and Stochastic Approximations*. *Applications of Mathematics* (*New York*) **22**. Springer, Berlin. MR1082341

BERG, S., ZHU, J., CLAYTON, M. K, SHEA, M. E and MLADENOFF, D. J (2019). Supplement to "A latent discrete Markov random field approach to identifying and classifying historical forest communities based on spatial multivariate tree species counts." DOI:10.1214/19-AOAS1259SUPP.

BURNS, R. M. and HONKALA, B. H. (1990). *Silvics of North America* **2**. U.S. Department of Agriculture, Washington, DC.

CHEN, J. (2017). Consistency of the MLE under mixture models. *Statist*. *Sci*. **32** 47–63. MR3634306

COMETS, F. and GIDAS, B. (1992). Parameter estimation for Gibbs distributions from partially observed data. *Ann*. *Appl*. *Probab*. **2** 142–170. MR1143397

CURTIS, J. T. (1959). *The Vegetation of Wisconsin*: *An Ordination of Wisconsin Plant Communities*. Univ. Wisconsin Press, Madison, WI.

DAVIS, M. B., SCHWARTZ, M. W. and WOODS, K. (1991). Detecting a species limit from pollen in sediments. *J*. *Biogeogr*. **18** 653–668.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J*. *Roy*. *Statist*. *Soc*. *Ser*. *B* **39** 1–38. MR0501537

EGAN, D. (2005). *The Historical Ecology Handbook*: *A Restorationist's Guide to Reference Ecosystems*. Island Press.

FAHEY, R. T., LORIMER, C. G. and MLADENOFF, D. J. (2012). Habitat heterogeneity and life-history traits influence presettlement distributions of early-successional tree species in a late-successional, hemlock-hardwood landscape. *Landsc*. *Ecol*. **27** 999–1013.

FINLEY, R. W. (1976). The original vegetation cover of Wisconsin. *Wisconsin Department of Natural Resources*.

FORBES, F. and FORT, G. (2007). Combining Monte Carlo and mean-field-like methods for inference in hidden Markov random fields. *IEEE Trans. Image Process.* **16** 824–837. MR2460196

FORBES, F., CHARRAS-GARRIDO, M., AZIZI, L., DOYLE, S. and ABRIAL, D. (2013). Spatial risk mapping for rare disease with hidden Markov fields and variational EM. *Ann. Appl. Stat.* **7** 1192–1216. MR3113506

FORT, G. and MOULINES, E. (2003). Convergence of the Monte Carlo expectation maximization for curved exponential families. *Ann. Statist.* **31** 1220–1259. MR2001649

GAETAN, C. and GUYON, X. (2010). *Spatial Statistics and Modeling*. *Springer Series in Statistics*. Springer, New York. MR2569034

GANGNON, R. E. and CLAYTON, M. K. (2003). A hierarchical model for spatially clustered disease rates. *Stat. Med.* **22** 3213–3228.

GELMAN, A. and MENG, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statist. Sci.* **13** 163–185. MR1647507

GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI-6** 721–741.

HONG, C., NING, Y., WANG, S., WU, H., CARROLL, R. J. and CHEN, Y. (2017). PLEMT: A novel pseudolikelihood-based EM test for homogeneity in generalized exponential tilt mixture models. *J. Amer. Statist. Assoc.* **112** 1393–1404. MR3750863

KNORR-HELD, L. and RASSER, G. (2004). Bayesian detection of clusters and discontinuities in disease maps. *Biometrics* **56** 13–21.

KUSHNER, H. J. and YIN, G. G. (1997). *Stochastic Approximation Algorithms and Applications*. *Applications of Mathematics* (*New York*) **35**. Springer, New York. MR1453116

LAWSON, A. B. (2010). Hotspot detection and clustering: Ways and means. *Environ. Ecol. Stat.* **17** 231–245. MR2725781

LIU, F., MLADENOFF, D. J., KEULER, N. S. and MOORE, L. S. (2011). Broadscale variability in tree data of the historical Public Land Survey and its consequences for ecological studies. *Ecol. Monogr.* **81** 259–275.

MLADENOFF, D. J., SICKLEY, T. A., SCHULTE, L. A., RHEMTULLA, J. M. and BOLLIGER, J. (2009). Wisconsin's Land Cover in the 1800s. *Wisconsin Department of Natural Resources*.

MLADENOFF, D. J., WHITE, M. A., PASTOR, J. and CROW, T. R. (1993). Comparing spatial pattern in unaltered old-growth and disturbed forest landscapes. *Ecol. Appl.* **3** 294–306.

NEAL, R. M. (1993). Probabilistic Inference Using Markov Chain Monte Carlo Methods. Technical Report.

PACIOREK, C. J., GORING, S. J., THURMAN, A. L., COGBILL, C. V., WILLIAMS, J. W., MLADENOFF, D. J., PETERS, J. A., ZHU, J. and MCLACHLAN, J. S. (2016). Statistically-estimated tree composition for the northeastern United States at Euro-American settlement. *PLoS ONE* **11** 1–20.

PETERSON, D. W. and REICH, P. B. (2001). Prescribed fire in oak savanna: Fire frequency effects on stand structure and dynamics. *Ecol. Appl.* **11** 914–927.

RADELOFF, V. C., MLADENOFF, D. J., HE, H. S. and BOYCE, M. S. (1999). Forest landscape change in the northwestern Wisconsin Pine Barrens from pre-European settlement to the present. *Can. J. For. Res.* **29** 1649–1659.

ROBBINS, H. and MONRO, S. (1951). A stochastic approximation method. *Ann. Math. Stat.* **22** 400–407. MR0042668

SCHULTE, L. and MLADENOFF, D. (2001). The original US public land survey records: Their use and limitations in reconstructing pre-European settlement vegetation. *J. For.* **99** 5–10.

SCHULTE, L. A., MLADENOFF, D. J. and NORDHEIM, E. V. (2002). Quantitative classification of a historic northern Wisconsin (USA) landscape: Mapping forests at regional scales. *Can. J. For. Res.* **32** 1616–1638.

SHAO, J. (2003). *Mathematical Statistics*, 2nd ed. *Springer Texts in Statistics*. Springer, New York. MR2002723

SHEA, M. E., SCHULTE, L. A. and PALIK, B. J. (2014). Reconstructing vegetation past: Pre-Euro-American vegetation for the midwest Driftless area, USA. *Ecol. Restor.* **32** 417–433.

STÄDLER, N., BÜHLMANN, P. and VAN DE GEER, S. (2010). ℓ_1-penalization for mixture regression models. *TEST* **19** 209–256. MR2677722

STAMBAUGH, M. C. and GUYETTE, R. P. (2008). Predicting spatio-temporal variability in fire return intervals using a topographic roughness index. *For. Ecol. Manag.* **254** 463–473.

WALLER, L. A. (2009). Detection of clustering in spatial data. In *The SAGE Handbook of Spatial Analysis* (J. Fagerberg, D. C. Mowery and R. R. Nelson, eds.) 299–321 10. Sage, London.

WU, F. Y. (1982). The Potts model. *Rev. Modern Phys.* **54** 235–268. MR0641370

WU, C.-F. J. (1983). On the convergence properties of the EM algorithm. *Ann. Statist.* **11** 95–103. MR0684867

YOUNES, L. (1989). Parametric inference for imperfectly observed Gibbsian fields. *Probab. Theory Related Fields* **82** 625–645. MR1002904

S. BERG
J. ZHU
M. K. CLAYTON
DEPARTMENT OF STATISTICS
UNIVERSITY OF WISCONSIN-MADISON
MADISON, WISCONSIN 53706
USA
E-MAIL: saberg2@wisc.edu
jzhu@stat.wisc.edu
mkclayto@wisc.edu

M. E. SHEA
D. J. MLADENOFF
DEPARTMENT OF FORESTRY
AND WILDLIFE ECOLOGY
UNIVERSITY OF WISCONSIN-MADISON
MADISON, WISCONSIN 53706
USA
E-MAIL: mshea3@wisc.edu
djmladen@wisc.edu