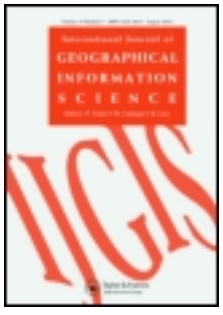

International Journal of Geographical Information Science

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tgis20

A multinomial logistic mixed model for the prediction of categorical spatial data

Guofeng Cao [a], Phaedon C. Kyriakidis [a] & Michael F. Goodchild [a]

[a] Department of Geography, University of California, Santa Barbara, CA, USA

Available online: 11 Nov 2011

To cite this article: Guofeng Cao, Phaedon C. Kyriakidis & Michael F. Goodchild (2011): A multinomial logistic mixed model for the prediction of categorical spatial data, International Journal of Geographical Information Science, DOI:10.1080/13658816.2011.600253

To link to this article: http://dx.doi.org/10.1080/13658816.2011.600253

International Journal of Geographical Information Science
i*First*, 2011, 1–16

Taylor & Francis
Taylor & Francis Group

A multinomial logistic mixed model for the prediction of categorical spatial data

Guofeng Cao*, Phaedon C. Kyriakidis and Michael F. Goodchild

Department of Geography, University of California, Santa Barbara, CA, USA

(*Received 26 April 2011; final version received 22 June 2011*)

In this article, the prediction problem of categorical spatial data, that is, the estimation of class occurrence probability for (target) locations with unknown class labels given observed class labels at sample (source) locations, is analyzed in the framework of generalized linear mixed models, where intermediate, latent (unobservable) spatially correlated Gaussian variables (random effects) are assumed for the observable non-Gaussian responses to account for spatial dependence information. Within such a framework, a spatial multinomial logistic mixed model is proposed specifically to model categorical spatial data. Analogous to the dual form of kriging family, the proposed model is represented as a multinomial logistic function of spatial covariances between target and source locations. The associated inference problems, such as estimation of parameters and choice of the spatial covariance function for latent variables, and the connection of the proposed model with other methods, such as the indicator variants of the kriging family (indicator kriging and indicator cokriging) and Bayesian maximum entropy, are discussed in detail. The advantages and properties of the proposed method are illustrated via synthetic and real case studies.

Keywords: categorical data; indicator kriging; GLMM; logistic regression; geostatistics

1. Introduction

Categorical spatial data, such as land use classes, vegetation species types, or categories of socioeconomic status, are all important information sources in spatial analysis, resource management, decision support systems, and planning in general. Such data types, which can be nominal, ordinal, or interval, exhibit spatial or spatiotemporal patterns with sharp boundaries and complex geometrical characteristics. This discrete (non-Gaussian) nature limits the applications of successful statistical methods that have been widely used for continuous variables including the celebrated kriging family of methods (Chilès and Delfiner 1999). A key task in statistical modeling of categorical spatial data is to estimate the joint probability mass function of a set of geo-referenced (spatially correlated) categorical variables, or, in a prediction/interpolation scenario, the posterior probability of class occurrence (class occurrence probability) at a target location (where the actual class is unknown) conditioned jointly to all known source data (class labels available at sample or source locations). In this article, we focus on the prediction/interpolation problem in spatially correlated categorical data and propose a spatial model within a framework of generalized

*Corresponding author. Email: cao@geog.ucsb.edu

ISSN 1365-8816 print/ISSN 1362-3087 online

http://dx.doi.org/10.1080/13658816.2011.600253
http://www.tandfonline.com

linear mixed models (GLMMs) to estimate the desired class occurrence probability for a target location.

In the past three decades, many approaches have been proposed from different perspectives to address such a problem. In traditional spatial analysis and geostatistics in particular, indicator kriging (IK) (Journel 1983) is the most frequently used method for estimating the posterior (conditional) probability of class occurrence at any target location. Same as other kriging variants, IK involves fitting a indicator variogram and then estimating the probability of occurrence of each class independently via kriging system. In spite of the easy implementation and wide applications, IK has been criticized for its well-known inherent problems: the probabilities of occurrence are not guaranteed to be between 0 and 1 and the sum of these probabilities may not be equal to 1. Most of the improvements over IK, indicator cokriging (ICK), and disjunctive kriging, for example, do not actually solve these inherent problems despite their high theoretical complexity and intimidating implementation difficulties (Chilès and Delfiner 1999; Emery 2006). A posterior correction method of the resulting conditional probabilities is often necessary through either a Gaussian transformation or a logistic regression model (Pardo-Igúzquiza and Dowd 2005).

From another perspective of statistics, a Bayesian maximum entropy (BME) approach has been proposed for categorical spatial data modeling (Christakos 1990; Bogaert 2002). Different from the kriging family of methods, BME is based on a joint multidimensional multinomial assumption in the desired categorical random field and estimates parameters (actually joint probability tables) by a nonsaturated log-linear model of main effects and interaction effects under certain marginal constraints (most often formulated in terms of bivariate joint probabilities). Thus, the BME approach automatically yields valid results without the inherent problems. But the heavy computation cost involved in the BME parameter estimation process limits the applications of this approach only to cases with a restricted number of categories and a small neighborhood. Most recently, a simplified variant of the BME approach (Allard *et al.* 2009) was proposed to address the heavy computational cost of categorical BME. Instead of modeling all the possible interactions (in terms of joint probabilities), this simplified approach only considers the pair-wise spatial interactions between the target location and its neighboring known source data (in a perspective of multi-point conditional probabilities). It has been shown that the simplified variant is a valid approximation of the BME solution (Allard *et al.* 2009). If the neighborhood of a target location is limited to only the four nearest neighbors along cardinal directions, this simplified BME will be equivalent to another recently proposed approach, the Markov chain random field (MCRF) (Li 2007), which models categorical spatial data with spatial transition probabilities (the so-called *transiogram*) and a single random Markov chain that could cover the whole categorical field. Both the simplified BME and MCRF work under a strict conditional independence assumption that may be inappropriate due to complex dependencies in a spatial context.

In addition to generative approaches (Bishop 2006) modeling two-point spatial interactions directly by spatial joint or transition probabilities, such as BME and MCRF, the use of spatially correlated latent variables to model geo-referenced non-Gaussian responses is another active research area. Most methods are developed within the convenient context of exponential family distributions and augment the observed data with latent variables (often assumed multivariate Gaussian) within the framework of GLMMs (Breslow and Clayton 1993). In such a spirit, Diggle *et al.* (1998) proposed GLMM-based methods for spatial count (with a log-linear link) and binary (with a logit link) data and coined the term model-based geostatistics. In such models, the posterior probability of introduced latent variables is not available in closed form owing to the non-Gaussian response variables and

the common approach to conduct inference on the latent variables is Markov chain Monte Carlo (MCMC) sampling, which has been criticized for convergence and computational time.

In this article, we follow the same paradigm of latent variables and propose a multinomial logistic mixed model specifically for spatially correlated categorical variables with multiple categorical outcomes. But instead of trying to sample the posterior probability of latent variables under the MCMC framework (Zhang 2002, Christensen 2004), or by using quasi-likelihood-based generalized estimating equations (Liang and Zeger 1986, Gotway and Stroup 1997), we directly approximate the posterior density (so-called *Laplace approximation*) based on a proposed approximation of the spatial covariance functions of the latent variables from the sample data. In this way, the sought-after class occurrence probability can be written as a function of covariance values between the target and source locations by applying the *Representer Theorem* (Kimeldorf and Wahba 1970) in the context of a reproducing kernel Hilbert space (RKHS). In addition, model parameters can be estimated through common iteratively gradient-based optimization methods, such as the Newton–Raphson method. The parameter estimation procedure of the proposed method, although consistent with the MCRF/BME assumption that the observed categorical spatial data are conditional independent given the target values, is not rigidly tied to this assumption as MCRF/BME are. In cases where this assumption is violated, the parameters of the proposed method will be adjusted accordingly to maximize the fit to the data and thus the strict assumption of conditional independence is relaxed. The implementation flowchart of the proposed spatial multinomial logistic mixed model is similar to that of traditional kriging, but compared with IK and ICK and other indicator variants of kriging family, the proposed model always generates consistent class occurrence probabilities for each target location.

The remainder of this article is organized as follows: In Section 2, the spatial multinomial logistic mixed model and the associated inference problems, such as parameter estimation and choice of the spatial covariance function of latent variables, are presented after an introduction of the working definitions and notations. Section 3 discusses the connections of the proposed method with other approaches, and case studies are provided in Section 4, followed by conclusions and discussion of future work in Section 5.

2. Model

2.1. Setting and notation

Consider a categorical random variable (RV) $C(\boldsymbol{x})$ ($\boldsymbol{x} \in R^d$) which can take one out of K mutually exclusive and collectively exhaustive class labels $c(\boldsymbol{x}) \in \{1, \ldots, K\}$, at any arbitrary location with coordinate vector $\boldsymbol{x}$. This RV $C(\boldsymbol{x})$ is assumed to follow a multinomial distribution, that is,

$$C(\boldsymbol{x}) \sim Mu(1, \boldsymbol{\pi}(\boldsymbol{x})) \tag{1}$$

where $Mu(\cdot, \cdot)$ indicates the multinomial distribution, $\boldsymbol{\pi}(\boldsymbol{x}) = [\pi_1(\boldsymbol{x}), \ldots, \pi_K(\boldsymbol{x})]^T$ is a vector of marginal probabilities of $\boldsymbol{x}$ for categories $\{1, \ldots, K\}$, respectively, and superscript T indicates transposition. It is obvious that $\sum_{k=1}^{K} \pi_k(\boldsymbol{x}) = 1$ holds. We have a set of observed such data at N different locations denoted as $\{c(\boldsymbol{x}_i); i = 1, \ldots, N\}$, which are assumed to be drawn i.i.d. from a fixed but unknown joint distribution $\mathcal{P}\{C(\boldsymbol{x}_1), \ldots, C(\boldsymbol{x}_N)\}$. We denote the coordinates of the observed locations and class labels by $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ and

$\boldsymbol{c} = \{c(\boldsymbol{x}_1), \ldots, c(\boldsymbol{x}_N)\}$, respectively, and together by $\mathcal{D} = \{\boldsymbol{x}, \boldsymbol{c}\}$ for notation simplicity. Given the observed data $\mathcal{D}$, as discussed in Section 1, we would like to predict the class occurrence probability for a given target location $\boldsymbol{x}^*$, that is, $P\{C(\boldsymbol{x}^*)|\boldsymbol{x}^*, \mathcal{D}\}$, while accounting for dependence information in a spatial context.

2.2. *Multinomial logistic mixed model for categorical spatial data*

The generalized linear model (GLM) (McCullagh and Nelder 1989) with a multinomial logit link function is a natural choice to model categorical responses if these class labels are assumed independent with each other, that is,

$$\log \frac{P\{C(\boldsymbol{x}_i) = k\}}{P\{C(\boldsymbol{x}_i) = k^*\}} = \beta_0^k \tag{2}$$

where β_0^k is a linear predictor and in this case it is only an intercept depending on the class label assigned to individual location $\boldsymbol{x}_i$ since we focus on spatial autocorrelation effects and no other explanatory variable is assumed to be available, and k^* is the base-line category which can be arbitrarily selected from $\{1, \ldots, K\}$. One way to extend the GLM to accommodate spatial dependence information is through GLMM (Breslow and Clayton 1993), which models such dependencies by introducing latent (unobservable) variables or random effects in the linear predictor. For the categorical data case in this article, we assume there are K intermediate, latent variables $u(\boldsymbol{x}, k), k = 1, \ldots, K$ for each location $\boldsymbol{x}$ (see Figure 1 for an illustration). By accounting for these latent variables, Equation (2) becomes

$$\log \frac{P\{C(\boldsymbol{x}_i) = k\}}{P\{C(\boldsymbol{x}_i) = k^*\}} = \beta_0^k + u(\boldsymbol{x}_i, k) \tag{3}$$

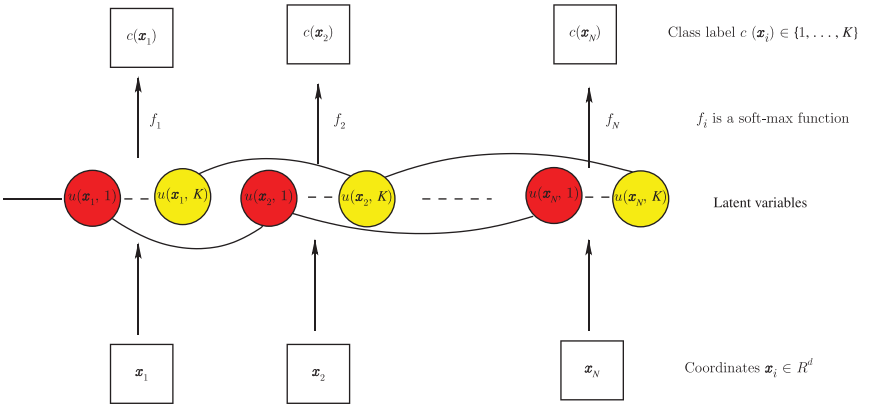

Figure 1. An illustration of a generalized linear mixed model (GLMM) for categorical spatial data. Circles represent the latent variables and square boxes represent the locations and associated observed class labels. Circles with same color indicate that these latent variables share the same mean and covariance function. The link between the circles represents interaction or dependency, and in this illustration we assume that latent variables for different categories are independent with each other, hence no links exist between red and yellow circles. Given these latent variables, the multinomial logistic (soft-max) function $f_i(\cdot)$ is employed to obtain the probability of class label $k \in \{1, \ldots, K\}$ for a location $\boldsymbol{x}_i$.

We furthermore assume that $\boldsymbol{u}(\boldsymbol{x}, k) = [u(\boldsymbol{x}_1, k), \ldots, u(\boldsymbol{x}_N, k)]^T$ is a Gaussian random field (GRF) specified by a mean function $\boldsymbol{\mu}_k(\boldsymbol{u}(\boldsymbol{x}); \boldsymbol{\theta})$ and a positive definite covariance function $\sigma_k(\boldsymbol{u}(\boldsymbol{x}_i), \boldsymbol{u}(\boldsymbol{x}_j); \boldsymbol{\theta})$, that is, $P(\boldsymbol{u}(\boldsymbol{x}, k)) = \mathbb{N}(\boldsymbol{\mu}_k(\boldsymbol{x}), \boldsymbol{\Sigma}_k; \boldsymbol{\theta})$, where $\boldsymbol{\Sigma}_k$ is the Gram matrix with elements $\boldsymbol{\Sigma}_{kij} = [\sigma_k(u(\boldsymbol{x}_i), u(\boldsymbol{x}_j); \boldsymbol{\theta})]$ and $\boldsymbol{\theta}$ is vector of parameters for the mean function and covariance function. Without losing generality and for notational convenience, we assume that a mean function $\boldsymbol{\mu} \equiv 0$ hereafter. Another assumption that is typically made is that $\boldsymbol{u}(\cdot, k)$ and $\boldsymbol{u}(\cdot, k')$ are uncorrelated if $k \neq k'$, which means that the K latent GRFs are independent with each other; under such an assumption, we have $\sigma(u(\boldsymbol{x}_i, k), u(\boldsymbol{x}_j, k'); \boldsymbol{\theta}) = 0$ for $k \neq k'$ and $\sigma(u(\boldsymbol{x}_i, k), u(\boldsymbol{x}_j, k); \boldsymbol{\theta}) = \sigma_k(u(\boldsymbol{x}_i), u(\boldsymbol{x}_j); \boldsymbol{\theta})$ otherwise, where σ_k is a covariance function specific to class k. In GRF settings, this can be further written as $\sigma_k(u(\boldsymbol{x}_i), u(\boldsymbol{x}_j); \boldsymbol{\theta}) = \sigma_k(\boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\theta})$. By accounting for the K latent GRFs, one can denote the covariance function between two data locations $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ as $\sigma(\boldsymbol{x}_i, \boldsymbol{x}_j)$ and the covariance matrix of $\boldsymbol{u}$ is $1/\lambda\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is the Gram matrix with $\boldsymbol{\Sigma}_{ij} = [\sigma(\boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\theta})]$ and λ is an adjustable regularization parameter.

As mentioned above, the key task in this article is to estimate the class occurrence probability function for a target location given the source data, that is, $P\{C(\boldsymbol{x}^*)|\boldsymbol{x}^*, \mathcal{D}\}$. By introducing latent variables and applying Bayes' rule, the predictive function can be given by

$$P\{C(\boldsymbol{x}^*)|\boldsymbol{x}^*, \mathcal{D}\} = \int P\{C(\boldsymbol{x}^*)|\boldsymbol{x}^*, \boldsymbol{u})P(\boldsymbol{u}|\mathcal{D}\}d\boldsymbol{u} \tag{4}$$

where $P\{\boldsymbol{u}|\mathcal{D}\}$ is the posterior probability of the latent variable $\boldsymbol{u}$ and can be written as

$$P\{\boldsymbol{u}|\mathcal{D}\} \propto P\{\boldsymbol{c}|\boldsymbol{u}\}P\{\boldsymbol{u}|\boldsymbol{x}\} = P\{\boldsymbol{c}|\boldsymbol{u}\} \exp\{-\frac{\lambda}{2}\boldsymbol{u}^T\Sigma^{-1}\boldsymbol{u}\} \tag{5}$$

The exponential component of the right-hand side of Equation (5) comes from the assumption that $P\{\boldsymbol{u}|\boldsymbol{x}\}$ is multivariate Gaussian. Based on a conditional independence assumption of $\boldsymbol{c}$ given latent variables $\boldsymbol{u}$ (as Figure 1 illustrates), the posterior distribution of $\boldsymbol{u}$ can be further written as

$$P\{\boldsymbol{u}|\mathcal{D}\} \propto \prod_{i=1}^{N} p\{c(\boldsymbol{x}_i)|\boldsymbol{u}(\boldsymbol{x}_i)\} \exp\{-\frac{\lambda}{2}\boldsymbol{u}^T\Sigma^{-1}\boldsymbol{u}\} \tag{6}$$

The links between an observable category $c(\boldsymbol{x}_i)$ and the latent variables $\boldsymbol{u}(\boldsymbol{x}_i)$ have been described in Equation (3). Based on this link function, one can easily obtain an explicit function of class probability:

$$P\{c(\boldsymbol{x}_i) = k|\boldsymbol{u}(\boldsymbol{x}_i)\} = \frac{\exp\{\beta_0^k + u(\boldsymbol{x}_i, k)\}}{\sum_{k'=1}^{K} \exp\{\beta_0^{k'} + u(\boldsymbol{x}_i, k')\}} \tag{7}$$

Given this multinomial logistic function (soft-max function) of Equation (7), one can obtain the distribution of the latent variables of Equation (5) and eventually the desirable predictive probability function for a target location $\boldsymbol{x}^*$ in Equation (4). Unfortunately, in most cases, the integral over the latent distribution $P(\boldsymbol{u}|\mathcal{D}\}$ (expectation of $\boldsymbol{u}$) in the predictive probability function of Equation (4) is computationally intractable since there are $N \times K$ latent variables $u(\boldsymbol{x}_i)$ that need to be integrated out. A common approximation is

to replace the integral by the value of the integrand at the mode of the posterior distribution where Equation (5) is maximal, the so-called *Laplace approximation* (Williams and Barber 2002), that is, maximum a posteriori (MAP) estimation of $\boldsymbol{u}$. Thus, Equation (4) is approximated by

$$P\{C(\boldsymbol{x}^*)|\boldsymbol{x}^*, \mathcal{D}\} \approx P\{C(\boldsymbol{x}^*)|\boldsymbol{x}^*, \boldsymbol{u}_{\text{MAP}}, \mathcal{D}\} \text{ where } \boldsymbol{u}_{\text{MAP}} = \underset{\boldsymbol{u}}{\text{argmax}}\, P\{\boldsymbol{u}|\mathcal{D}\} \tag{8}$$

Since the multinomial logistic function of Equation (7) is non-Gaussian, the posterior distribution over the latent values in Equation (5) and the predictive distribution of Equation (8) cannot be written in analytical form. To find $\boldsymbol{u}_{\text{MAP}}$, one can take the logarithm of the posterior density of Equation (5), as

$$\log P\{\boldsymbol{u}|\mathcal{D}\} = \sum_{i=1}^{N} p\{c(\boldsymbol{x}_i)|u(\boldsymbol{x}_i)\} - \frac{\lambda}{2}\boldsymbol{u}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{u} + \rho \tag{9}$$

where ρ is a constant to account for the normalized information and does not influence the search of $\boldsymbol{u}$ maximizing Equation (9) and is therefore dropped for notational simplicity. If the parameter vector $\boldsymbol{\theta}$ for the covariance function is assumed to be known, the negative of Equation (9) fulfills the conditions of the Representer Theorem in Scholkopf and Smola (2001), a generalization of what was originally proposed in Kimeldorf and Wahba (1970).

Theorem 1 (Representer Theorem) *Let $\mathcal{H}$ be a RKHS with a kernel δ: $\mathcal{X} \times \mathcal{X} \to \mathcal{R}$. For any function G: $\mathcal{R}^n \to \mathcal{R}$, and any nondecreasing function Ω: $\mathcal{R} \to \mathcal{R}$, if the optimization problem can be well defined as*

$$\boldsymbol{J}^* = \min_{f\in\mathcal{H}} \boldsymbol{J}(f) := \min_{f\in\mathcal{H}} \{\Omega||f||^2_{\mathcal{H}} + G(f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n))\}$$

then there are $\alpha_1, \ldots, \alpha_n \in \mathcal{R}$, such that $f(\cdot) = \sum_{i=1}^{n} \alpha_i \delta(\boldsymbol{x}_i, \cdot)$ achieves $\boldsymbol{J}(f) = \boldsymbol{J}^*$.

By the application of Theorem 1, the maximizer $\boldsymbol{u}_{\text{MAP}}$ of Equation (9) is guaranteed to be of the form:

$$u(\boldsymbol{x}_i, k)_{\text{MAP}} = \sum_{j=1}^{N} \beta_j^k \sigma(\boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\theta}) \tag{10}$$

To ease the discussion, for each location $\boldsymbol{x}_i$, we introduce an indicator vector $\boldsymbol{j}(\boldsymbol{x}_i) = [j_1(\boldsymbol{x}_i), \ldots, j_K(\boldsymbol{x}_i)]^T$ to represent which class $\boldsymbol{x}_i$ belongs to, where $j_k(\boldsymbol{x}_n) = 1$ if $c(\boldsymbol{x}_n) = k$, 0 if not. By applying the representation of u in Equation (10), one can rewrite Equation (9) as a function of β, denoted as $\mathcal{L}(\beta)$:

$$\mathcal{L}(\beta) = \sum_{i=1}^{N} \{\boldsymbol{j}(\boldsymbol{x}_i)^T (\Sigma(\boldsymbol{x}_i, \cdot)\boldsymbol{\beta})^T - \log \sum_{k'=1}^{K} \exp\{\boldsymbol{\Sigma}(\boldsymbol{x}_i, \cdot)(\boldsymbol{\beta}^{k'})\}\} - \frac{\lambda}{2}\sum_{k=1}^{K} (\boldsymbol{\beta}^k)^T \boldsymbol{\Sigma}\boldsymbol{\beta}^k \tag{11}$$

where $\boldsymbol{\beta} = [\boldsymbol{\beta}^1, \ldots, \boldsymbol{\beta}^K]$ and each of $\boldsymbol{\beta}^k$ is a $N \times 1$ vector of weights for the observed data for class k and $\boldsymbol{\Sigma}(\boldsymbol{x}_i, \cdot)$ indicates the i-th row of the covariance matrix $\boldsymbol{\Sigma}$.

If the covariance matrix $\boldsymbol{\Sigma}$ is assumed to be known, one can find the optimal $\boldsymbol{\beta}$ by solving the system of equations $(\partial\mathcal{L}(\boldsymbol{\beta}))/(\partial\boldsymbol{\beta}) = 0$. Since $\mathcal{L}(\boldsymbol{\beta})$ is a nonlinear function of $\boldsymbol{\beta}$,

the Fisher scoring or Newton–Raphson iterative algorithm, available in common statistics software, can be applied to maximize Equation (11) with respect to $\boldsymbol{\beta}$.

From a perspective of regularization theory, $\mathcal{L}(\boldsymbol{\beta})$ can be viewed as a penalized likelihood, which is often used in regression to address *overfitting* when the dimension of the explanatory variables is very high. The first component on the right-hand side term of Equation (11) is the multinomial logistic likelihood and $\boldsymbol{\Sigma}$ in the second component can be viewed as the regularization matrix to penalize large values of $\boldsymbol{\beta}$ (Zhu and Hastie 2005). If $\lambda = 0$ in Equation (11), this is equivalent to assuming that each categorical datum is independent with each other (no spatial dependencies). If $\lambda = \infty$ on the other hand, the influence of the observed data will diminish to zero and the $\mathcal{L}(\boldsymbol{\beta})$ will be fully controlled by prior (latent GRFs).

Once the optimal $\boldsymbol{\beta}$ is found, the predictive probability for a target location $\boldsymbol{x}^*$ can be obtained as

$$\hat{P}\{C(\boldsymbol{x}^*) = k|\mathcal{D}\} = \frac{\exp\{\beta_0^k + \sum_{i=1}^{N} \beta_i^k \sigma(\boldsymbol{x}^*, \boldsymbol{x}_i; \boldsymbol{\theta})\}}{\sum_{k'=1}^{K} \{\exp\{\beta_0^{k'} + \sum_{i=1}^{N} \beta_i^{k'} \sigma(\boldsymbol{x}^*, \boldsymbol{x}_i; \boldsymbol{\theta})\}\}} \tag{12}$$

2.3. *Covariance functions for latent variables*

In the discussion above, spatial dependence in categorical spatial data is modeled through the covariance function, or kernel function in general, of the latent GRFs. Since the introduced latent variables cannot be observed directly, one usually resorts to the MCMC method to infer the distribution of latent variables (Christensen 2004). Similar problems appear in the truncated (pluri)gaussian simulation method (Armstrong *et al.* 2003), where the occurrence probability for a certain class is obtained through a integral of latent GRFs based on the associated class proportion and threshold values. To determine the covariance function of each latent GRF, iterative methods including sequential simulation (Dowd *et al.* 2003) and Gibbs sampler (Emery 2007) were proposed. From another perspective, Diggle *et al.* (1998) analytically approximated the latent covariance function by a linear function of an observed covariance. In this article, an *ad hoc* approximation of the latent covariance function is proposed, and together with the adjustable regularization parameter λ, the covariance matrix of the latent GRFs is specified.

In the geostatistics literature, the covariance function $\sigma(\boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\theta})$ is usually written as $\sigma(\boldsymbol{x}_i - \boldsymbol{x}_j; \boldsymbol{\theta}) = \upsilon \times \sigma(\boldsymbol{h}/a)$ (the so-called *covariogram*) under a stationary random field assumption, where $\sigma(\cdot) : \mathcal{R} \to \mathcal{R}$ is a monotonously decreasing function with $\sigma(0) = 1$ and $\lim_{h\to\infty} \sigma(h) = 0$, $\boldsymbol{h}$ is a vector in R^d, υ is the variance or scale parameter, and a is the so-called *range* to represent the influence of this covariance function and here $\boldsymbol{\theta} = \{\upsilon, a\}$. An arbitrary function of $\boldsymbol{x}_i - \boldsymbol{x}_j$ is not, in general, a valid covariogram and several commonly used eligible covariograms have been studied extensively (Chilès and Delfiner 1999). Such covariograms, together with kriging variants, are the fundamental tools to explore and model spatial dependencies or spatial similarities. For univariate Gaussian cases, the elementary properties of covariograms, such as positive definiteness and symmetry, as well as their connection with the implied spatial patterns have been well studied (Chilès and Delfiner 1999, Stein 1999, Lantuejoul 2002). But for categorical and multivariate cases as in this article, the above construct is usually problematic. It is possible to define a GRF with multiple auto- and cross-covariograms (Goulard and Voltz 1992,

Goovaerts 1997), but it is still not clear in general how the covariogram should be defined. So the covariogram is usually built independently for each class. In this article, we adopt a *mixture of covariograms*; that is, a linear combination of indicator covariograms for each class weighted by their respective class proportions:

$$\sigma(\boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \sigma_k^{\text{obs}}(\boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\theta}_k) \tag{13}$$

where π_k is the class proportion of class k and $\sum_{k=1}^{K} \pi_k = 1$, and $\sigma_k^{\text{obs}}(\cdot, \cdot)$ is an indicator covariance function for observations with class label k, specified with parameters $\boldsymbol{\theta}_k$. The positive definiteness of each $\sigma_k^{\text{obs}}(\boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\theta}^k)$ guarantees that their linear combination $\sigma(\boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\theta})$ is also a positive definite function. For each $\sigma_k^{\text{obs}}(\cdot, \cdot)$, we follow the covariogram fitting procedure that is commonly used in geostatistics, which computes the empirical indicator covariances first based on observed data and fits the desirable covariance function through least squares methods.

2.4. Summary

In this section, the proposed model for categorical spatial data is developed within the GLMM framework. Given an observation neighborhood centered on a target location with unknown class label, the prediction of class occurrence probability for that location is represented as a multinomial logistic function of covariances between target and source locations. Parameter estimation and an *ad hoc* specification of eligible spatial covariance functions are discussed. The flowchart of the proposed method for computing the predictive class occurrence probability function is similar with the kriging family of methods, which includes computation of empirical covariances, covariance function fitting, and estimation of weights. Connections between the proposed method and the above-mentioned methods including simplified BME and IK/ICK will be discussed in Section 3.

3. Connections with other methods

3.1. Indicator (co)kriging

Similar to other kriging variants, IK is a linear predictor based on indicator data. It has many variants depending on the assumptions on the local mean value or prior probabilities of class occurrence. In this discussion, we focus on simple IK, in which the local mean value is assumed to be known. Following Goovaerts (1997), the dual form of simple IK can be written as

$$[\hat{P}\{C(\boldsymbol{x}^*) = k|\mathcal{D}\}]_{\text{IK}} = \pi_k + \sum_{j=1}^{N} w_{j,k}^{IK} \sigma_k(\boldsymbol{x}_j - \boldsymbol{x}^*; \boldsymbol{\theta}) \tag{14}$$

where each $w_{j,k}^{\text{IK}}$ is the dual IK weight pertaining to the auto-covariance function $\sigma_k(\boldsymbol{x}_i - \boldsymbol{x}^*; \boldsymbol{\theta}^k)$, which can be found by solving the dual simple IK system (Goovaerts 1997). ICK extents IK by taking the interclass dependencies into account via cross-covariance functions:

$$[\hat{P}\{C(\boldsymbol{x}^*) = k|\mathcal{D}\}]_{\text{ICK}} = \pi_k + \sum_{j=1}^{N}\sum_{k'=1}^{K} w_{j,kk'}^{\text{ICK}}\sigma_{kk'}(\boldsymbol{x}_i - \boldsymbol{x}^*) \tag{15}$$

where each $w_{j,kk'}^{\text{ICK}}$ is the dual ICK weight pertaining to the indicator covariance $\sigma_{kk'}(\boldsymbol{x}_i - \boldsymbol{x}^*)$, an auto-covariance if $k = k'$ and a cross-covariance otherwise. Both methods have been criticized over the years. A major drawback of IK and ICK is that they tend to generate inconsistent results, which do not necessarily meet the basic probability constraints. The proposed method can be taken as a 'generalized linear model' of IK/ICK, since if the $u(\boldsymbol{x}_i, k)$ in the logit form of Equation (3) is replaced with its representer form of Equation (10), one can write

$$\log\frac{P\{C(\boldsymbol{x}^*) = k)\}}{P\{C(\boldsymbol{x}^*) = k_0)\}} = \beta_0^k + \sum_{j=1}^{N}\beta_j^k\sigma(\boldsymbol{x}^*, \boldsymbol{x}_j) \tag{16}$$

Through a multinomial logit link function between the observed class labels and the latent variables, the proposed method yields consistent class occurrence probabilities.

3.2. Bayesian maximum entropy

BME, which provides another perspective for statistical modeling of spatial data (Christakos 1990), has been extended to model categorical spatial data through directly modeling joint multivariate probabilities of class occurrence (Bogaert 2002). BME actually decomposes the joint probability into a log-linear model of main effects (a univariate probability) and interaction effects (multivariate probability, but most often only bivariate) directly and estimates these effects (multiple probability tables) by maximizing entropy under marginal constraints (bivariate marginals in observed data will not be changed). For a dataset with N locations and K classes, there are $N^2 \times K^2$ parameters to be estimated. The computational cost of this method restricts its application to small datasets with few categories. A simplified variant of BME was recently proposed (Allard *et al.* 2009) to address this heavy computational burden. Instead of working on joint probabilities, this simplified method models the conditional probability through a combination of interaction effects (bivariate distributions) between the target location and its neighbors.

$$[\hat{P}\{C(\boldsymbol{x}^*) = k|\mathcal{D}\}]_{\text{BME}} = \frac{\pi_k \prod_{i\in\mathcal{N}(\boldsymbol{x}^*)} P\{c(\boldsymbol{x}_i)|C(\boldsymbol{x}^*) = k\}}{\sum_{k'=1}^{K=K} \pi_{k'} \prod_{i\in\mathcal{N}(\boldsymbol{x}^*)} P\{c(\boldsymbol{x}_i)|C(\boldsymbol{x}^*) = k'\}} \tag{17}$$

where $\mathcal{N}(\boldsymbol{x}^*)$ is a neighborhood of the target location $\boldsymbol{x}^*$. If class k_0 is selected as a base-line category and Equation (17) can be rewritten as

$$[\hat{P}\{C(\boldsymbol{x}^*) = k|\mathcal{D}\}]_{\text{BME}} = \frac{\exp\left\{\log\frac{\pi_k}{\pi_{k_0}} + \sum_{i=1}^{N}\log\frac{P\{c(\boldsymbol{x}_i)|C(\boldsymbol{x}^*)=k\}}{P\{c(\boldsymbol{x}_i)|C(\boldsymbol{x}^*)=k_0\}}\right\}}{\sum_{k'=1}^{K}\exp\left\{\log\frac{\pi_{k'}}{\pi_{k_0}} + \sum_{i=1}^{N}\log\frac{P\{c(\boldsymbol{x}_i)|C(\boldsymbol{x}^*)=k'\}}{P\{c(\boldsymbol{x}_i)|C(\boldsymbol{x}^*)=k_0\}}\right\}} \tag{18}$$

If we let $\mathcal{N}(\boldsymbol{x}^*) = \{1, \ldots, N\}$, $\beta_0^{k'} = \log \frac{\pi_{k'}}{\pi_{k_0}}$ and $\beta_i^{k'} \sigma(\boldsymbol{x}^*, \boldsymbol{x}_i; \boldsymbol{\theta}) = \frac{P\{c(\boldsymbol{x}_i)|C(\boldsymbol{x}^*)=k'\}}{P\{c(\boldsymbol{x}_i)|C(\boldsymbol{x}^*)=k_0\}}$, Equations (12) and (18) are equivalent. The simplified BME estimates the bivariate probability or transition probability as a function of distance first and then predicts the class occurrence probability under the strict conditional independence assumption. This assumption may be inadequate for spatial data due to the existence of complex dependencies in a spatial context. The fitting procedure of the proposed model of Equation (11) is consistent with the conditional independence assumption that the observed data are independent with each other given a class label $c(\boldsymbol{x}_i^*)$ is assigned to a target location. But the proposed model is not rigidly tied to this strict assumption since the parameters β can be adjusted accordingly to maximize the fit to the observed data in cases whereby this assumption does not hold. Cao *et al.* (in press) is another effort to relax this strict independence assumption by applying the *Tau model* (Journel 2002; Krishnan 2008), which is a general probabilistic paradigm to combine diverse sources information while accounting for information redundancies, in a spatial context. For binary variables, one can easily use the same substitution trick as above to find the relationship between the fusion weights τ_i in the Tau model (Cao *et al.* in press) and β in this proposed model.

4. Case study

In this section, the performance of the proposed method in reproducing spatial patterns is investigated and compared to the most recent simplified BME approach based on a synthetic and a real-world case studies. For an arbitrary location, both of these two methods are implemented for computing class occurrence probabilities based on observed data and the class label with maximum predicted probability is assigned to this location (MAP). Following Cao *et al.* (in press), a set of auto- and cross-transiogram functions are used for quantifying spatial patterns implied in categorical spatial data. Thus, in addition to correct classification rates, a common criterion for evaluating a prediction algorithm, the performance of the two methods is evaluated by how well they can reproduce the transiograms in the reference data.

In the implementation of the proposed method, the empirical covariograms are computed first and the 'true' covariograms are fitted by a nonlinear fitting procedure. The overall covariogram is obtained by the mixture of these auto-covariograms (Equation (13)) and λ is obtained through cross-validation by maximizing the correct estimation rate. For the simplified BME approach (Equation (17)), due to lack of the theoretical models of transiograms as covariograms in kriging, the Nadaraya–Watson kernel smoothing regression method (Nadaraya 1964) is used to interpolate the empirical transiograms. In other words, for $p_{k|k'}(h)$, the transiogram from class k' to class k, we have empirical transiogram values $p_1, \ldots, p_L$ for lag distance $h_1, \ldots, h_L$, respectively. Then $p_{k|k'}(h^*)$, the transiogram from class k' to class k for an arbitrary lag distance h^*, can be given by

$$p_{k|k'}(h^*) = \frac{\sum_{i=1}^{L} \kappa(h_i - h^*) p_i}{\sum_{i=1}^{L} \kappa(h_i - h^*)} \tag{19}$$

where $\kappa(\cdot)$ is a kernel function with a bandwidth r. A Gaussian kernel function $\kappa(\Delta h_i) = \exp\{-(\Delta h_i / r)^2\}$ and $\Delta h_i = ||h_i - h^*||$ is used in this experiment. Higher values of r lead to smoother results. Similar with λ, r can be obtained by cross-validation. Since the empirical

transiogram values are usually obtained by scanning all observed data, thus given a distance h, these empirical transiograms should meet the basic constraints of transiogram values, that is, $\sum_{k=1}^{K} p_{k|k'}(h) = 1$. Based on such eligible input, it can be easily shown that output of Equation (19) meets probability constraints naturally. In summary, the flowchart for both methods can be summarized as follows:

- Scanning the sample data for empirical transiograms (for BME) and indicator covariograms (for the proposed method).
- For the simplified BME, the best neighborhood ($\mathcal{N}$ in Equation (18)) and the bandwidth of the kernel function (r in Equation (19)) are obtained by cross-validation. For the proposed method, Equation (13) is used for the mixture of covariagrams; λ is obtained through cross-validation. Based on these, the parameters of the proposed method β are estimated through iterative methods.
- Given a target location x^*, the class occurrence probability $P(C(\boldsymbol{x}^*)|\mathcal{D})$ is computed through these two methods respectively (Equation (17) for the simplified BME and Equation (12) for the proposed method).
- The class label with the maximum probability is assigned to the target location x^*.

A Matlab implementation of the described procedures for both these two methods is available in http://www.geog.ucsb.edu/ cao/research.html

4.1. Synthetic case study

In this synthetic case study, the reference map is a realization of truncated multivariate Gaussian simulation with three categories and 64×64 dimension (Figure 2a). A dataset was created by randomly sampling 576 locations from this image (Figure 2b) and class proportions for each class are [0.3, 0.45, 0.25]. The most striking spatial pattern in the reference map is that patch size is rather small and class 1 tends to be embedded in class 2, class 2 tends to be embedded in class 3, and adjacencies between class 1 and class 3 are rare. Based on the sample class labels (Figure 2b), both the proposed method and the simplified BME are implemented to obtain the class occurrence probabilities for each target location, and the respective results are given in Figure 2c and d. For the simplified BME approach, a neighborhood of 20 nearest points is chosen and $r = 0.1$. For the proposed method, a Gaussian covariance function is used to model empirical covariogram values and $\lambda = 0.011$. Correct classification rate is the most often used criterion to compare the performance of prediction algorithms. Based on the sample dataset, the proposed method correctly classifies 61% of the target locations while the simplified BME only has 51% correct classification rate. From the two prediction maps (Figure 2c and d), we can see that the simplified BME approach tends to generate larger patches than the reference map, which is partially due to the strict conditional independence assumption (Cao *et al*. in press), while the proposed method tends to generate patches of similar size with those in the reference map and clear boundaries between classes. This is confirmed in the resulting probability map for class 1 (Figure 3) in which the predicted probability values of the proposed method (Figure 3b) are more distinguishable than those of the simplified BME approaches (Figure 3a). Looking at the corresponding transiograms (Figure 4), we can see that the proposed method generates closer transiograms (green solid lines) to the reference transiograms (red solid lines) than those of the simplified BME approach (blue solid lines), which indicates that the proposed method reproduces better the reference adjacencies between different classes.

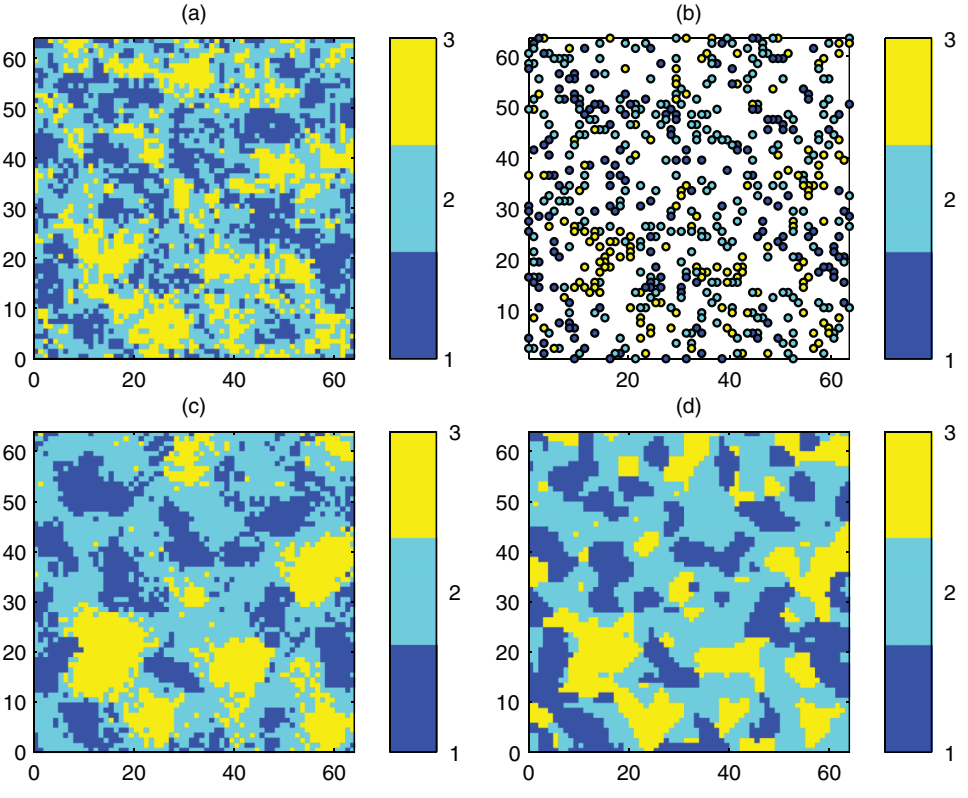

Figure 2. (a) Reference map with dimension 64 × 64 and three categories; (b) a point dataset with 576 points sampled from (a); (c) prediction map of the simplified BME approach; (d) prediction map of the proposed method.

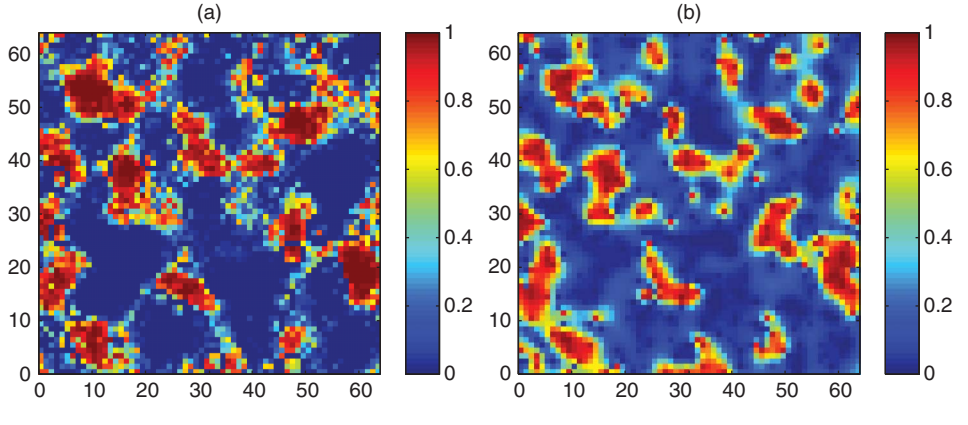

Figure 3. (a) Predictive probability for class 1 obtained from the simplified Bayesian maximum entropy (BME) approach; (b) predictive probability for class 1 obtained from the proposed approach

4.2. Real-world case study

To further investigate the performance of the proposed method in real cases, the lithology types in the well-known Jura dataset (Goovaerts 1997) are used, in which 5 rock types

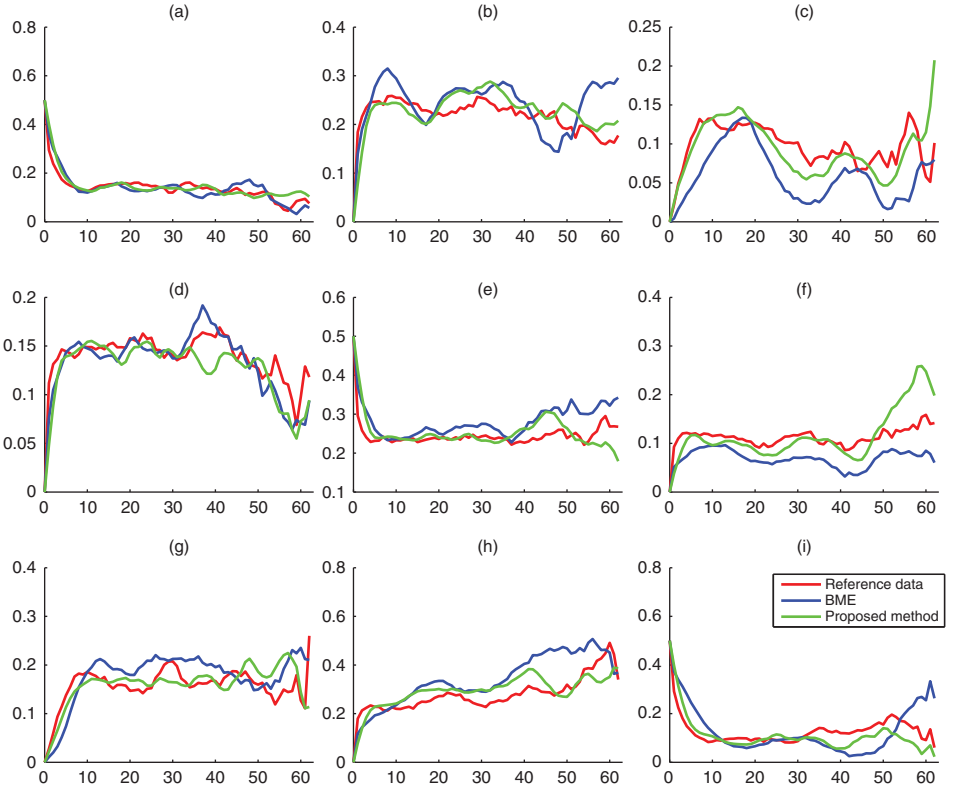

Figure 4. Comparison of the reproduced transiograms of the simplified approach (blue solid lines) and the proposed method (green solid lines) and the reference transiograms (red solid lines). (a) Class 1–1; (b) Class 1–2; (c) Class 1–3; (d) Class 2–1; (e) Class 2–2; (f) Class 2–3; (g) Class 3–1; (h) Class 3–2; (i) Class 3–3. BME, Bayesian maximum entropy.

are sampled at 359 locations and a subset of 259 samples is selected as training dataset (Figure 5a and b). Following Bel *et al*. (2009), category four (Portlandian) is recategorized as five (Quaternary) since there are only three sample points with this category. The class proportions of the these four categories are $[0.2046, 0.3282, 0.2432, 0.2239]$, respectively. In the implementation of the BME approach, omnidirectional transiograms, a Gaussian kernel with bandwidth $r = 0.05$ (Equation (19)) and a neighborhood with five nearest points are chosen. For the proposed method, a Gaussian model is used to fit the empirical covariogram values and $\lambda = 0.01$. From the prediction maps of these two methods (Figure 5(c) and (d)), we can see that the proposed method tends to generate clearer interclass boundaries (Figure 5d) than the BME approach (see the isolated nodes in Figure 5c). The BME approach tends to neglect small-scale features (Figure 5c) that the proposed method can successfully preserve. (Figure 5d). The correct classification rate of the BME approach is 86.63% (311 out of 359) and that of the proposed method is 90.53% (325 out of 359).

5. Conclusions and future work

In this article, a prediction/interpolation method for spatially distributed categorical variables with multiple outcomes, namely a spatial multinomial logistic mixed model, is

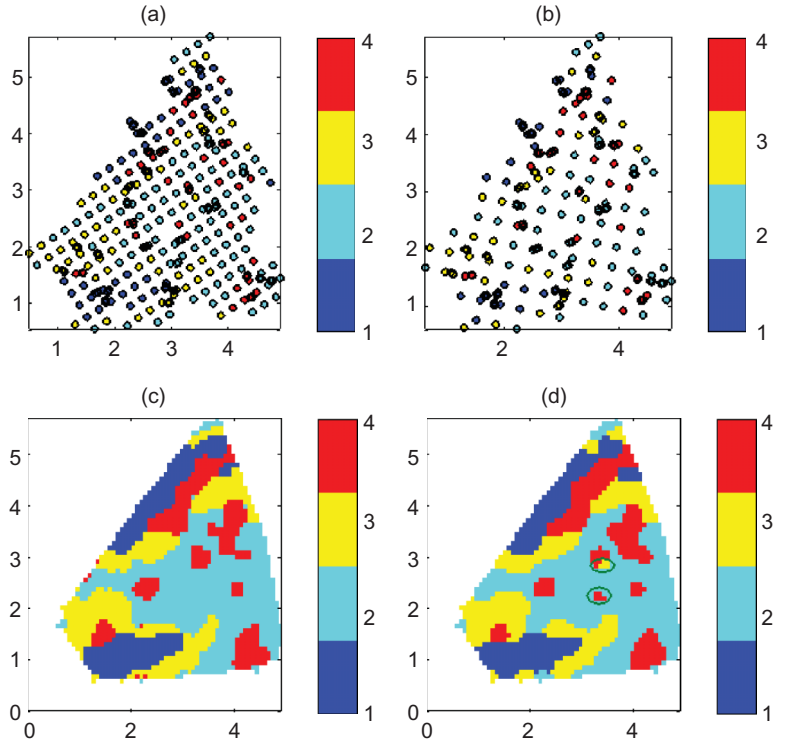

Figure 5. Jura lithology dataset and prediction maps: (a) lithology types at 395 sample locations; (b) prediction set; (c) prediction map of simplified BME; (d) prediction map of the proposed method. Please note the small yellow and red patches (visualized in the green ellipses) neglected in the results of simplified BME (c) are successfully recovered in the results of the proposed method (d).

proposed within the framework of GLMM, in which spatially correlated latent variables (multivariate Gaussian distributed) are assumed to account for spatial dependencies in categorical responses. Instead of using the most commonly used MCMC sampling to infer the assumed latent variables, an approach with convergence issues and heavy computational burden, we proposed an *ad hoc* method to approximate the analytically intractable posterior probability of the latent variables. The sought-after class occurrence probability function for a target location is written as a multinomial logistic linear combination of covariance values between the target and source data locations, which can be analogous to the dual form of kriging methods. The associated inference problems, such as parameter estimation and the specification of the covariance functions, are discussed in detail. The connections of the proposed method with IK/ICK and a most recent simplified version of categorical BME are highlighted. Synthetic and real-world case studies demonstrate the advantages of the proposed method over existing methods, such as simplified BME, a method closely related to the spatial Markov chain model. A Matlab implementation of the simplified BME, IK, and the proposed method, as well as the associated functions, such as the empirical transiograms/covariograms computation, covariogram fitting, and kernel transiograms interpolation, has been also developed.

Although the conditional independence assumption is implied in the proposed method, it is not rigidly tied to this assumption as the BME/MCRF approach is, and thus this

strict assumption is actually relaxed. When compared to generative approaches, such as BME/MCRF, which directly model pair-wise interactions through transition probabilities or bivariate joint probabilities, the proposed method can be seen as a discriminative model which assigns different weights to each source location; the more influence a sample location has over the target location, the larger weight this sample location will eventually have. This paradigm is more flexible than generative approaches when one moves beyond the spatial autocorrelation effects to account for indirect data from multiple sources. For example, suppose for each sample location $\boldsymbol{x}_i$, in addition to the observed categories $c(\boldsymbol{x}_i)$, there is another value $z(\boldsymbol{x}_i)$ of a continuous variable available, which is closely related to the categorical variable being modeled. If one wants to account for the information of $z(\boldsymbol{x}_i)$ via generative methods, the conditional probability of $z(\boldsymbol{x}_i)$ given $c(\boldsymbol{x}_i)$ needs to be modeled, and this is usually a tedious task. In the proposed method, however, the predictive class occurrence probability function in Equation (12) can be rewritten as

$$\hat{P}\{C(\boldsymbol{x}^*) = k|\mathcal{D}\} = \frac{\exp\{\beta_0^k + \sum_{i=1}^{N} \alpha_i^k z(\boldsymbol{x}_i) + \sum_{i=1}^{N} \beta_i^k \sigma(\boldsymbol{x}^*, \boldsymbol{x}_i; \boldsymbol{\theta})\}}{\sum_{k'=1}^{K} \{\exp\{\beta_0^{k'} + \sum_{i=1}^{N} \alpha_i^{k'} z(\boldsymbol{x}_i) + \sum_{i=1}^{N} \beta_i^{k'} \sigma(\boldsymbol{x}^*, \boldsymbol{x}_i; \boldsymbol{\theta})\}\}} \tag{20}$$

where α_i^k is the weight for $z(\boldsymbol{x}_i)$ pertaining to class k. The exact same estimation procedure can be applied to obtain the model parameters including α_i^k.

Similar to the dual form of kriging, a global neighborhood is used in the proposed method to capture all the necessary spatial pattern information. But as the data volume increases, the estimation of model parameters will become very computationally demanding. In this case, a partition strategy (Auñón and Gómez-Hernández 2000), which was originally proposed to address the same issue in dual kriging, could be adopted in the method proposed in this article.

Acknowledgment

We gratefully acknowledge the funding provided by the National Geospatial-Intelligence Agency (NGA) to support this research.

References

Allard, D., D'Or, D., and Froidevaux, R., 2009. Estimating and simulating spatial categorical data using an efficient maximum entropy approach. Avignon, France: Unit´e Biostatistique et Processus Spatiaux Institut National de la Recherche Agronomique, Technical Report No. 37.

Armstrong, M., *et al*., 2003. *Plurigaussian simulations in geosciences*. Berlin: Springer.

Auñón, J. and Gómez-Hernández, J., 2000. Dual kriging with local neighborhoods: application to the representation of surfaces. *Mathematical Geology*, 32 (1), 69–85.

Bel, L., *et al*., 2009. CART algorithm for spatial data: application to environmental and ecological data. *Computational Statistics & Data Analysis*, 53 (8), 3082–3093.

Bishop, C., 2006. *Pattern recognition and machine learning*. Vol. 4. New York: Springer.

Bogaert, P., 2002. Spatial prediction of categorical variables: the Bayesian maximum entropy approach. *Stochastic Environmental Research and Risk Assessment*, 16 (6), 425–448.

Breslow, N. and Clayton, D., 1993. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88 (421), 9–25.

Cao, G., Kyriakidis, P.C., and Goodchild, M.F., in press. Combining spatial transition probabilities for stochastic simulation of categorical fields. *International Journal of Geographical Information Science*.

Chilès, J. and Delfiner, P., 1999. *Geostatistics: modeling spatial uncertainty*. New York: Wiley Interscience.
Christakos, G., 1990. A Bayesian/maximum-entropy view to the spatial estimation problem. *Mathematical Geology*, 22 (7), 763–777.
Christensen, O., 2004. Monte Carlo maximum likelihood in model-based geostatistics. *Journal of Computational and Graphical Statistics*, 13 (3), 702–718.
Diggle, P., Tawn, J., and Moyeed, R., 1998. Model-based geostatistics. *Applied Statistics*, 47 (3), 299–350.
Dowd, P.A., Pardo-Iguzquiza, E., and Xu, E.C., 2003. Plurigau: a computer program for simulating spatial facies using the truncated plurigaussian method. *Computers & Geosciences*, 29, 123–141.
Emery, X., 2006. A disjunctive kriging program for assessing point-support conditional distributions. *Computers & Geosciences*, 32 (7), 965–983.
Emery, X., 2007. Simulation of geological domains using the plurigaussian model: new developments and computer programs. *Computers & Geosciences*, 33 (9), 1189–1201.
Goovaerts, P., 1997. *Geostatistics for natural resources evaluation*. New York: Oxford University Press.
Gotway, C. and Stroup, W., 1997. A generalized linear model approach to spatial data analysis and prediction. *Journal of Agricultural, Biological, and Environmental Statistics*, 2 (2), 157–178.
Goulard, M. and Voltz, M., 1992. Linear coregionalization model: tools for estimation and choice of cross-variogram matrix. *Mathematical Geology*, 24 (3), 269–286.
Journel, A., 2002. Combining knowledge from diverse sources: an alternative to traditional data independence hypotheses. *Mathematical Geology*, 34 (5), 573–596.
Journel, A.G., 1983. Non-parametric estimation of spatial distributions. *Mathematical Geology*, 15 (3), 445–468.
Kimeldorf, G. and Wahba, G., 1970. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41 (2), 495–502.
Krishnan, S., 2008. The tau model for data redundancy and information combination in Earth sciences: theory and application. *Mathematical Geosciences*, 40 (6), 705–727.
Lantuejoul, C., 2002. *Geostatistical simulation: models and algorithms*. Berlin: Springer Verlag.
Li, W., 2007. Markov chain random fields for estimation of categorical variables. *Mathematical Geology*, 39 (3), 321–335.
Liang, K. and Zeger, S., 1986. Longitudinal data analysis using generalized linear models. *Biometrika*, 73 (1), 13.
McCullagh, P. and Nelder, J., 1989. *Generalized linear models*. 2nd ed. Boca Raton, FL: Chapman & Hall, CRC.
Nadaraya, E., 1964. On estimating regression. *Teoriya Veroyatnostei i ee Primeneniya*, 9 (1), 157–159.
Pardo-Igúzquiza, E. and Dowd, P., 2005. Multiple indicator cokriging with application to optimal sampling for environmental monitoring. *Computers & Geosciences*, 31 (1), 1–13.
Scholkopf, B. and Smola, A.J., 2001. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. Cambridge, MA: MIT Press.
Stein, M., 1999. *Interpolation of spatial data: some theory for kriging*. Berlin: Springer Verlag.
Williams, C. and Barber, D., 2002. Bayesian classification with Gaussian processes, Pattern Analysis and Machine Intelligence. *IEEE Transactions*, 20 (12), 1342–1351.
Zhang, H., 2002. On estimation and prediction for spatial generalized linear mixed models. *Biometrics*, 58 (1), 129–136.
Zhu, J. and Hastie, T., 2005. Kernel logistic regression and the import vector machine. *Journal of Computational and Graphical Statistics*, 14 (1), 185–205.