Stoch Environ Res Risk Assess (2014) 28:1785–1799
DOI 10.1007/s00477-013-0842-7

ORIGINAL PAPER

A statistical framework of data fusion for spatial prediction of categorical variables

Guofeng Cao · Eun-hye Yoo · Shaowen Wang

Published online: 1 January 2014

Abstract With rapid advances of geospatial technologies, the amount of spatial data has been increasing exponentially over the past few decades. Usually collected by diverse source providers, the available spatial data tend to be fragmented by a large variety of data heterogeneities, which highlights the need of sound methods capable of efficiently fusing the diverse and incompatible spatial information. Within the context of spatial prediction of categorical variables, this paper describes a statistical framework for integrating and drawing inferences from a collection of spatially correlated variables while accounting for data heterogeneities and complex spatial dependencies. In this framework, we discuss the spatial prediction of categorical variables in the paradigm of latent random fields, and represent each spatial variable via spatial covariance functions, which define two-point similarities or dependencies of spatially correlated variables. The representation of spatial covariance functions derived from different spatial variables is independent of heterogeneous characteristics and can be combined in a straightforward fashion. Therefore it provides a unified and flexible representation of heterogeneous spatial variables in spatial analysis while accounting for complex spatial dependencies. We show that in the spatial prediction of categorical variables, the sought-after class occurrence probability at a target location can be formulated as a multinomial logistic function of spatial covariances of spatial variables between the target and sampled locations. Group least absolute shrinkage and selection operator is adopted for parameter estimation, which prevents the model from over-fitting, and simultaneously selects an optimal subset of important information (variables). Synthetic and real case studies are provided to illustrate the introduced concepts, and showcase the advantages of the proposed statistical framework.

Keywords Categorical data · Data fusion · Kernel methods · Geostatistics · LASSO

1 Introduction

With the continuing advancement of spatial data acquisition and dissemination technology, a large amount of spatial data from diverse sources often are available for many geographical or environmental research problems. In the mapping of tree species distribution, for example, measurements of environmental conditions, such as elevation, temperature, soil nutrients and moisture, are often available in addition to the witness tree data. These diverse environmental conditions are known to influence the tree species occurrences, and spatial distribution of each of them provides a partial yet insightful view to the distribution of tree species. It would be ideal to fuse these diverse partial information efficiently to achieve a

G. Cao (✉)
Department of Geosciences, Texas Tech University, Lubbock, TX, USA
e-mail: guofeng.cao@ttu.edu

E. Yoo
Department of Geography, State University of New York at Buffalo, Buffalo, NY, USA
e-mail: eunhye@buffalo.edu

S. Wang
Department of Geography and Geographic Information Science, University of Illinois at Urbana-Champaign, Champaign, IL, USA
e-mail: shaowen@illinois.edu

comprehensive view. These spatial information, however, often demonstrates incompatible heterogeneities with each other in terms of nature (continuous or categorical), intrinsic quality (soft or hard data), spatial scales, and sample locations. Together with complex spatial dependence and inter-dependence structures among spatial variables, these incompatibilities or heterogeneities render fusing these diverse sources of spatial information a rather challenging problem.

The principle of data fusion is generic and has been widely used in many disciplines; it is basically to integrate multiple sources of information at best in order to achieve a better inference over what each single source could provide. Despite the simplicity in statements, the precise objectives of data fusion and ways to achieve them are diverse in different fields, and usually tied with specific applications (Bogaert and Fasbender 2007). In the context of spatial prediction of categorical variables, we describe a statistical and computational framework for efficiently fusing multiple sources of spatially distributed data, while explicitly accounting for the (inter-)dependence structures in a spatial setting and flexibly accommodating for the heterogeneities across multiple data sources.

Categorical spatial data commonly are encountered in research projects in, for example, geosciences, environmental science, natural resource management, decision support systems and planning. Typical examples of such data include land use classes, vegetation species, or socioeconomic census data, such as gender and ethnicity groups. A successful spatial prediction of categorical variables can benefit many areas of research, such as spatial data classification and change detection (Tso and Mather 2009; Atkinson and Lewis 2000; Foody 2002; Atkinson 2012), spatial data mining (Miller and Han 2003) and spatial uncertainty modeling (Zhang and Goodchild 2002; Goodchild et al. 2009; Yoo and Trgovac 2011; Li et al. 2012). With the availability of auxiliary spatial information, a key task in the spatial prediction of categorical variables is to estimate the posterior probability of class occurrences at a target location (where the actual class is unknown) jointly conditioned on all observed class labels and the observations of auxiliary spatial variables. The discrete nature of categorical spatial variables, such as sharp boundaries and complex geometrical characteristics, limits applications of standard statistical methods that have been developed for continuous variables. Considerable efforts have been devoted from different disciplines to improve the spatial prediction of categorical variables by incorporating auxiliary information and (inter-)dependence structures in a spatial setting. As efforts of adapting kriging family of geostatistical methods for categorical variables, indicator kriging (IK) is perhaps the most frequent method for estimating the posterior (conditional) probability of class occurrence at any target location (Journel 1983). Based on IK, several variants have been developed to improve the prediction accuracy of primary categorical variables. Indicator co-kriging (ICK), for example, is a natural extension of IK for multivariate cases (Journel and Alabert 1989; Goovaerts 1997) in which auxiliary variables are incorporated into the predictive process via (cross-)covariance functions of primary categorical variables and auxiliary variables. Practical applications of ICK, however, are cumbersome owing to a number of (cross-)covariance functions [often through the linear model of coregionalization (Goulard and Voltz 1992)] to be jointly fitted. When auxiliary variables are linearly related to the class occurrence of primary categorical variables, they can be incorporated into IK system as deterministic linear functions (non-stationary mean). This is referred to as indicator kriging with *external drift* (IKED) (Goovaerts 1997), whose implementation in practice is challenging since it is often problematic to simultaneously estimate the parameters of external drift and the covariance function of the stochastic component. As a hybrid method of kriging and multiple regression models, regression-kriging (RK) (Hengl et al. 2004, 2007) has been developed to combine a regression of the dependent variables on auxiliary variables with kriging of the regression residuals. An indicator variation of RK, regression-kriging of indicators (RKI), has been proposed for categorical variable and this method has evolved into regression-kriging of memberships (RKfM) by substituting crisp indicator values with a continuous membership values (Hengl et al. 2007). Most of these IK-based methods, however, share the inherent problems that the original IK suffers from: the probabilities of occurrence are not guaranteed to be between 0 and 1 (e.g., IK, ICK, RKI), the sum of the predicted probabilities may not be equal to 1 (e.g., IK, ICK, RKI and RKfM) and the outcome values of conditional cumulative distribution function may not be monotonic. A posterior correction of the resulting conditional probabilities is often necessary either through a Gaussian transformation or via a logistic regression model (Pardo-Igúzquiza et al. 2005).

Alternatively, a Bayesian maximum entropy (BME) approach (Christakos 1990), originally developed for statistical modeling of generic spatial variables, has been applied for modeling categorical spatial data (Bogaert 2002). This BME approach is based on a joint multivariate multinomial assumption of the categorical fields. The desirable joint probability is then estimated via a non-saturated log-linear model of main effects and interaction effects by maximizing the entropy under certain marginal constraints. Built upon the formal theory of entropy, this BME-based approach is free of the aforementioned inherent problems of IK-based methods. Most recently, the idea of BME has been applied to integrating categorical and continuous variables (Wibrin et al. 2006) through a mixed (multivariate) random field specified by (cross-)covariance functions across multiple (categorical and continuous) spatial variables. Within a more general paradigm of Bayesian statistics, Bogaert and Fasbender (2007) proposed a theoretical framework of data fusion in the context of spatial prediction while accounting for spatial dependence and heterogeneities. Similar with other variants of the BME approach (e.g., Bogaert 2002), inference of parameters is usually computationally intimidating particularly when non-Gaussian spatial variables are involved and number of sample size increases.

The use of spatially correlated latent variables is another statistical venue to model geo-referenced non-Gaussian responses. Most methods are developed within the context of exponential family distributions, which can be easily augmented with latent variables (often assumed multivariate Gaussian) within the framework of generalized linear mixed models (GLMMs) (Breslow and Clayton 1993). In such a spirit, Diggle et al. (1998) proposed GLMM-based methods for spatial count variables (with a log-linear link) and binary variables (with a logit link) and coined the term *model-based geostatistics*. Given that the posterior probability of introduced latent variables is not available in a closed form owing to the non-Gaussian response variables, Markov chain Monte Carlo (MCMC) sampling is often used for the inference of latent variables, while it has been criticized for convergence and computational burden issues (Rue et al. 2009). Alternatively, a spatial multinomial logistic mixed model (MLMM) (Cao et al. 2011) was proposed for spatially correlated categorical variables with multiple categorical outcomes. Instead of sampling the posterior probability of latent variables under the MCMC framework (Zhang 2002; Christensen 2004), or by using quasi-likelihood based generalized estimating equations (GEE) (Liang and Zeger 1986; Gotway and Stroup 1997), this spatial MLMM model approximates class occurrence probability as a multinomial logistic function of spatial covariances between the target and source locations within a reproducing kernel Hilbert space (RKHS) (Kimeldorf and Wahba 1970). Such RKHS-based methods (Wahba 1990) have proven to be remarkably successful in various disciplines including machine learning (e.g., support vector machines) (Schölkopf and Smola 2002), biostatistics (Schoölkopf et al. 2004), as well as geostatistics (Goovaerts 1998).

A spatial covariance function, or more generally a kernel function, is specified as a distance decay function controlled by a set of parameters. Such function measures attribute similarity between pairs of spatial locations and thus quantifies the implicit relationship (spatial dependency or similarity) in correlated (dependent) data, which renders model inference more intuitively and easily. From another perspective, spatial covariance functions or kernel functions actually project input spatial variables with inconsistent characteristics into a unified space of kernel (RKHS), and thus provide a straightforward venue for integrating heterogeneous spatial data. By taking advantage of this unified representation, Lanckriet et al. (2004) presented a statistical framework for genomic data fusion within the paradigm of support vector machines (SVMs). In this paper, we extend the approach of Cao et al. (2011) to account for spatial heterogeneities and dependencies in auxiliary variables by representing each of them as spatial covariance functions and combing them in a multinomial logistic fashion to estimate the occurrence probability of class labels. There are three immediate advantages in this extension. Firstly, spatial dependence information, as well as a wide range of heterogeneities, such as inconsistent spatial scales, attribute types (categorical and numerical), missing values (or spatially misaligned data), are easily accommodated via the representation of spatial covariance functions. Secondly, compared with SVM-based methods, this method offers an estimation of the occurrence probability for each class label, and can be naturally generalized to categorical variables with multiple outcomes. Thirdly, a recently proposed group least absolute shrinkage and selection operator (LASSO) (Yuan and Lin 2006) is applied for parameter inference to avoid the so-called *overfitting* issue, and at the same time, selects the optimal subset of variables that are most related to the primary categorical variable by shrinking the coefficients of the other variables.

The remainder of this paper is organized as follows: Section 2 presents the proposed statistical data fusion framework and discusses the associated inference problems, such as parameter estimation and choice of the spatial covariance functions. Case studies are provided in Sect. 3, followed by conclusions and discussion of future work in Sect. 4.

2 Methodology

2.1 Data fusion for prediction of categorical spatial data

Consider a spatially distributed categorical random variable (RV) $C(\boldsymbol{x})(\boldsymbol{x} \in \boldsymbol{R}^d)$ that may take one of several discrete outcomes $\{c_1, \ldots, c_K\}$, which we index 1, …, K. Each RV is associated with a location, which is denoted by coordinate vector $\boldsymbol{x}$. Let $\pi_k(\boldsymbol{x})$ denote the probability that the outcome of the RV $C(\boldsymbol{x})$ falls in the k-th class (category), i.e., $\pi_k(\boldsymbol{x}) = \boldsymbol{P}\{\boldsymbol{C}(\boldsymbol{x}) = \boldsymbol{c}_k\}$. Assuming that K classes are mutually exclusive and collectively exhaustive, the sum of marginal probabilities across all categories equals 1, i.e., $\sum_{k=1}^{K} \pi_k(\boldsymbol{x}) = \boldsymbol{1}$. The probability distribution of the categorical RV $C(\boldsymbol{x})$ is given by the multinomial distribution as

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

where $Mu(\cdot, \cdot)$ indicates the multinomial distribution, and a vector of marginal probabilities for K categories at $\boldsymbol{x}$ is represented by $\boldsymbol{\pi}(\boldsymbol{x}) = [\boldsymbol{\pi_1}(\boldsymbol{x}), \ldots, \boldsymbol{\pi_K}(\boldsymbol{x})]^{\boldsymbol{T}}$. The superscript T denotes a transposition of vector or matrix.

Sample data have been measured at N locations, which consist of the observations of the primary categorical variable $C(\boldsymbol{x})$, denoted by a $(N \times 1)$ vector $\boldsymbol{c} = [\boldsymbol{c}(\boldsymbol{x_1}), \ldots, c(\boldsymbol{x_N})]^{\boldsymbol{T}}$, and the measurements of P auxiliary variables $\{Z_p(\boldsymbol{x}), \boldsymbol{p} = \boldsymbol{1}, \ldots, \boldsymbol{P}\}$, denoted by $z_{\boldsymbol{p}} = [z_{\boldsymbol{p}}(\boldsymbol{x_1}), \ldots, z_p(\boldsymbol{x_N})]^{\boldsymbol{T}}$ for the p-th auxiliary variable. For further notational simplicity, we combine the observed class labels with a collection of P auxiliary data as $\mathcal{D} = \{\boldsymbol{c}, \boldsymbol{Z}\}$, where $\boldsymbol{Z} = [z_{\boldsymbol{1}}, \ldots, z_{\boldsymbol{P}}]$ denotes a $(N \times P)$ matrix of the measurements of the P auxiliary variables at N sample locations.

As discussed in the previous section, our goal is to predict the class occurrence probability $P\{C(\boldsymbol{x}^*)|z(\boldsymbol{x}^*), \mathcal{D}\}$ for a given target location $\boldsymbol{x}^*$ using $\mathcal{D}$, and the auxiliary information $z(\boldsymbol{x}^*) = [z_{\boldsymbol{1}}(\boldsymbol{x}^*), \ldots, z_{\boldsymbol{P}}(\boldsymbol{x}^*)]^{\boldsymbol{T}}$ at location $\boldsymbol{x}^*$ if there is any, while accounting for both spatial dependencies and spatial heterogeneities in $C(\boldsymbol{x})$ and $Z(\boldsymbol{x})$.

Within a general paradigm of GLMM, Cao et al. (2011) constructed a two-stage model for the spatial prediction of categorical variables by introducing Gaussian distributed K intermediate latent variables $\boldsymbol{u}(\boldsymbol{x}) = \{\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{1}), \ldots, \boldsymbol{u}(\boldsymbol{x}, \boldsymbol{K})\}$. In this paper, we follow the modeling framework proposed by Cao et al. (2011), but allow each latent variable $\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{k}) = [\boldsymbol{u}(\boldsymbol{x_1}, \boldsymbol{k}), \ldots, \boldsymbol{u}(\boldsymbol{x_N}, \boldsymbol{k})]^{\boldsymbol{T}}$ for $k = 1, \ldots, K$ to be a multivariate Gaussian Random Field (GRF). A multivariate GRF is specified by a mean $\boldsymbol{\mu_k}$ and a positive definite covariance function $\sigma_k(\boldsymbol{u}(z_i), \boldsymbol{u}(z_j); \boldsymbol{\theta})$ informed by the sample data $\mathcal{D}$. The probability distribution of a latent variable is defined as:

$$P(\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{k})|\boldsymbol{Z}) = \mathbb{N}(\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}; \boldsymbol{\theta}), \tag{2}$$

where $\boldsymbol{\Sigma_k}$ is the covariance matrix (a Gram matrix with a element $\boldsymbol{\Sigma_{kij}} = [\sigma_k(\boldsymbol{u}(\boldsymbol{x_i}), \boldsymbol{u}(\boldsymbol{x_j}))]$) and $\boldsymbol{\theta}$ is the hyperparameter vector for the mean $\boldsymbol{\mu_k}$ and covariance function $\sigma_k(\cdot, \cdot)$. Without losing generality, we use a zero mean $\boldsymbol{\mu_k} \equiv \boldsymbol{0}$ hereafter. We assume that the K latent RFs are independent of each other, $\sigma(\boldsymbol{u}(\boldsymbol{x_i}, \boldsymbol{k}), \boldsymbol{u}(\boldsymbol{x_j}, \boldsymbol{k'})) = \boldsymbol{0}$ for $k \neq k'$ and $\sigma(\boldsymbol{u}(\boldsymbol{x_i}, \boldsymbol{k}), \boldsymbol{u}(\boldsymbol{x_j}, \boldsymbol{k})) = \boldsymbol{\sigma_k}(\boldsymbol{u}(\boldsymbol{x_i}), \boldsymbol{u}(\boldsymbol{x_j}))$, otherwise. Under the assumption of second-order stationarity, the covariance function can be simplified as $\sigma_k(\boldsymbol{u}(\boldsymbol{x_i}), \boldsymbol{u}(\boldsymbol{x_j})) = \boldsymbol{\sigma_k}(\boldsymbol{x_i} - \boldsymbol{x_j})$.

We further assume that the covariance matrix for the k-th latent variable $\boldsymbol{\Sigma_k}$ represents the spatial variation and dependence information of the k-th GRF implied in the observations $\mathcal{D}$. The mixture covariance matrix is estimated by combining the individual K latent variable covariance matrices $\boldsymbol{\Sigma_{k,p}}$ with restriction of resulting positive definite covariance matrix $\boldsymbol{\Sigma_k}$. Many statistical methods based on multivariate GRFs, such as coKriging family of methods (Wackernagel 1998), construct the multivariate covariance matrix by modeling all spatial interactions across different variables via auto- and cross-covariance functions. Not surprisingly, these approaches tend to dramatically increase the size of the covariance matrix as the number of auxiliary variables or sample locations increases. Despite the intimidating complications, it does not always guarantee improved performance. Furthermore, the construction of eligible multivariate covariance matrix is often difficult, while it is possible to define such a covariance matrix through a linear model of coregionalization (Goulard and Voltz 1992). It is unclear how the multivariate covariance matrix should be defined, although they may be built independently for each class in practice. In this paper, we approximate the covariance matrix for the k-th latent variable $\boldsymbol{\Sigma_k}$ as a linear combination of covariance matrices of each variables $\boldsymbol{\Sigma_{k,p}}, \boldsymbol{p} = \boldsymbol{1}, \ldots, \boldsymbol{P}$, as:

$$\boldsymbol{\Sigma_k} = \sum_{\boldsymbol{p=1}}^{\boldsymbol{P}} \boldsymbol{\tau_p \Sigma_{k,p}}, \tag{3}$$

whose (i, j)-th element can be expressed as a form of covariance functions,

$$\sigma_k(\boldsymbol{x_i}, \boldsymbol{x_j}) = \sum_{\boldsymbol{p=1}}^{\boldsymbol{P}} \boldsymbol{\tau_p \sigma_{k,p}}(\boldsymbol{x_i}, \boldsymbol{x_j}; \boldsymbol{\theta_p}), \tag{4}$$

where $\tau_p \geq 0$ and $\sigma_{k,p}(\cdot, \cdot)$, $p \in \{1, \ldots, P\}$ represents k-th covariance function for the p-th auxiliary variable Z_p. The positive definiteness of each $\sigma_{k,p}(\boldsymbol{x_i}, \boldsymbol{x_j}; \boldsymbol{\theta_p})$ guarantees that their linear combination $\sigma_k(\boldsymbol{x_i}, \boldsymbol{x_j})$ is also a positive definite function.

The ultimate goal is to estimate the class occurrence probability $P\{C(\boldsymbol{x}^*)|z(\boldsymbol{x}^*), \mathcal{D}\}$ at a target location $\boldsymbol{x}^*$ using all available source data. By introducing latent variables within a Bayesian approach, the predictive function can be given as:

$$P\{C(\boldsymbol{x}^*)|z(\boldsymbol{x}^*), \mathcal{D}\} = \int \boldsymbol{P}\{\boldsymbol{C}(\boldsymbol{x}^*)|z(\boldsymbol{x}^*), \boldsymbol{u})\boldsymbol{P}(\boldsymbol{u}|\mathcal{D}\}\boldsymbol{du} \tag{5}$$

where $P\{\boldsymbol{u}|\mathcal{D}\}$ is the posterior probability of the latent variable $\boldsymbol{u}$ given $\mathcal{D}$, which can be further written as: $P\{\boldsymbol{u}|\mathcal{D}\} \propto \boldsymbol{P}\{\boldsymbol{c}|\boldsymbol{u}\}\boldsymbol{P}\{\boldsymbol{u}|\boldsymbol{Z}\}$. Thus, one needs to integrate out all $N \times K$ multivariate latent variables $\boldsymbol{u}(\boldsymbol{x_i}, \boldsymbol{k})$, which is computationally intractable. A common approximation, the so-called *Laplace approximation* (Williams and Barber

2002), is to replace the integral by the value of the integrand at the mode of the posterior distribution where $P\{\boldsymbol{u}|\mathcal{D}\}$ is maximal, i.e. maximum a posteriori (MAP) estimation of $\boldsymbol{u}$. With this approximation, Eq. (5) can be written as:

$$P\{C(\boldsymbol{x}^*)|z(\boldsymbol{x}^*),\mathcal{D}\} \approx \boldsymbol{P\{C(x^*)}|z(\boldsymbol{x}^*),\boldsymbol{u_{MAP}},\mathcal{D}\}, \tag{6}$$

where $\boldsymbol{u_{MAP}} = \underset{\boldsymbol{u}}{\boldsymbol{argmaxP}}\{\boldsymbol{u}|\mathcal{D}\}$. Based on a conditional independence assumption of $\boldsymbol{c}$ given latent variables $\boldsymbol{u}$, the posterior distribution of $\boldsymbol{u}$ can be obtained by:

$$P\{\boldsymbol{u}|\mathcal{D}\} \propto \boldsymbol{P}\{\boldsymbol{u}|\boldsymbol{Z}\}\prod_{\boldsymbol{i=1}}^{\boldsymbol{N}}\boldsymbol{p}\{\boldsymbol{c}(\boldsymbol{x_i})|\boldsymbol{u}(\boldsymbol{x_i})\} \tag{7}$$

To find $\boldsymbol{u_{MAP}}$, one can combine the multivariate GRFs prior over $\boldsymbol{u}$ and take the logarithm of the posterior density $P\{\boldsymbol{u}|\mathcal{D}\}$ as:

$$\log P\{\boldsymbol{u}|\mathcal{D}\} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}}\boldsymbol{p}\{\boldsymbol{c}(\boldsymbol{x_i})|\boldsymbol{u}(\boldsymbol{x_i})\} - \frac{\boldsymbol{\lambda}}{\boldsymbol{2}}\boldsymbol{u^T\Sigma^{-1}u} + \rho, \tag{8}$$

where ρ is a constant that accounts for the normalized information but does not influence the search of $\boldsymbol{u}$ maximizing Eq. (8), and is, therefore, dropped for notational simplicity. A multinomial logistic function (or soft-max function) is used to model $p\{c(\boldsymbol{x_i})|\boldsymbol{u}(\boldsymbol{x_i})\}$, that links an observed class $c(\boldsymbol{x_i})$ at the i-th sample point to the latent variable $\boldsymbol{u}(\boldsymbol{x_i},\boldsymbol{k})$ as:

$$P\{c(\boldsymbol{x_i}) = \boldsymbol{k}|\boldsymbol{u}(\boldsymbol{x_i})\} \propto \exp\{\boldsymbol{\beta_0^k} + \boldsymbol{u}(\boldsymbol{x_i},\boldsymbol{k})\} \tag{9}$$

Cao et al. (2011) showed that Eq. (8) takes a form of the following by applying the Representer Theorem (Kimeldorf and Wahba 1970; Schölkopf and Smola 2002) to the maximizer $\boldsymbol{u_{MAP}}$ **:**

$$\boldsymbol{u}(\boldsymbol{x_i},\boldsymbol{k})_{\boldsymbol{MAP}} = \sum_{\boldsymbol{j=1}}^{\boldsymbol{N}}\boldsymbol{w_j^k\sigma_k}(\boldsymbol{x_i},\boldsymbol{x_j};\,\boldsymbol{\theta}) \tag{10}$$

We combine Eqs. (4) and (10), and apply them into Eq. (9). The desirable class occurrence probability at a target location $\boldsymbol{x}^*$ is re-written as:

$$\hat{P}\{C(\boldsymbol{x}^*) = \boldsymbol{k}|\mathcal{D}\} \propto \exp\{\boldsymbol{\beta_0^k} + \sum_{\boldsymbol{p=1}}^{\boldsymbol{P}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}}\boldsymbol{\beta_{i,p}^k\sigma_{k,p}}(\boldsymbol{x}^*,\boldsymbol{x_i};\,\boldsymbol{\theta_{k,p}})\} \tag{11}$$

In Eq. (11), we can easily see that the estimated class occurrence probability for location $\boldsymbol{x}^*$ only includes spatial covariance functions and does not explicitly rely on the auxiliary variables $Z_p(\boldsymbol{x}^*)$. This indicates that, similar with coKriging, Eq. (11) allows for missing values (or spatially misaligned data) in the measurements and does not require each auxiliary variables collocated with each other as long as the spatial covariance function $\sigma_{k,p}(\cdot,\cdot)$ could successfully capture the spatial variabilities of the p-th auxiliary variable.

2.2 Incorporating multiple collocated auxiliary information

In practice, measurements of auxiliary variables are oftentimes collocated with each other, and analog to the regression kriging and kriging with external drift, we may know that a specific parametric and deterministic (non-stationary mean and trend) function of this set of collocated variables is a part of the solution. It would be unwise not to incorporate these additional information. The extension of the Representer Theorem, namely semi-parametric Representer Theorem (1), provides a convenient venue to take into account this parametric and deterministic functions (Schölkopf and Smola 2002; Schölkopf et al. 2001).

Theorem 1 (Semi-parametric Representer Theorem) *Let $\mathcal{H}$ be a reproducing kernel Hilbert space with a kernel $\delta : \mathcal{X}\times\mathcal{X}\rightarrow\mathcal{R}$, and a set of M real-valued functions $\{\psi_p\}_{p=1}^{M} : \mathcal{X}\rightarrow\mathcal{R}$, with the property that the $m\times M$ matrix $(\psi_p(x_i))_{ip}$ has rank M. For any function $G : \mathcal{R}^n \rightarrow \mathcal{R}\bigcup\{\infty\}$ and $\tilde{f} := f + h$ with $f\in\mathcal{H}$ and $h\in span\{\psi_p\}$, and any non-decreasing function $\Omega : [0,\infty)\rightarrow\mathcal{R}$, if the optimization problem can be well-defined as:*

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

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

As a special case, suppose we know the primary categorical variable at location $\boldsymbol{x}^*$ is related to a weighted combination of covariates $\sum_{p=0}^{P}\alpha_p^kZ_p(\boldsymbol{x}^*)$. Equation (11) can be re-written as below by applying Theorem (1):

$$\begin{aligned}\hat{P}\{C(\boldsymbol{x}^*) = k|\mathcal{D}\} \propto \exp\{&\boldsymbol{\beta_0^k} + \sum_{\boldsymbol{p=0}}^{\boldsymbol{P}}\boldsymbol{\alpha_p^kZ_p}(\boldsymbol{x}^*)\\ &+ \sum_{p=1}^{P}\sum_{i=1}^{N}\boldsymbol{\beta_{i,p}^k\sigma_p}(\boldsymbol{x}^*,\boldsymbol{x_i};\,\boldsymbol{\theta_p})\}\end{aligned} \tag{12}$$

2.3 Model inference

Under the assumption of stationarity, the covariance function $\sigma_{k,p}(\boldsymbol{x_i},\boldsymbol{x_j};\,\boldsymbol{\theta_{k,p}})$ can be written as $\sigma_{k,p}(\boldsymbol{x_i} - \boldsymbol{x_j};\,\boldsymbol{\theta_{k,p}}) = \boldsymbol{\sigma_{k,p}}(\boldsymbol{h};\,\boldsymbol{\theta_{k,p}})$, where $\boldsymbol{h} = \boldsymbol{x_i} - \boldsymbol{x_j}$ is a vector of separation, and the *covariogram* $\sigma_{k,p}(\boldsymbol{h};\,\boldsymbol{\theta_{k,p}})$ is a monotonically decreasing and positive definite function representing spatial variabilities of the p-th auxiliary variable Z_p. Behaviors of covariograms are often assumed to be controlled by a set of parameters $\boldsymbol{\theta} = \{\upsilon,\boldsymbol{a}\}$, where υ is the variance or scale parameter, and a is the *range* to represent the influence of this covariance function. A valid covariogram includes Gaussian, exponential and spherical covariograms whose properties have been extensively studied (Chiles and Delfiner 1999). For each covariogram $\sigma_{k,p}(\boldsymbol{h;\theta_{k,p}})$, we follow the covariogram fitting procedure that is routinely used in geostatistics; we initially compute the empirical covariances based on observed data, and estimate the covariance function parameters $\boldsymbol{\theta_{k,p}}$ through least squares methods. Alternatively, coefficient parameters, i.e., $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, can be estimated by maximizing the likelihood or minimizing the loss. Consider a $(K \times 1)$ indicator vector defined at the i-th sample location as $\boldsymbol{j}(\boldsymbol{x_i}) = [\boldsymbol{j_k}(\boldsymbol{x_i}),\ \boldsymbol{k=1,\ldots,K}]^{\boldsymbol{T}}$, where $j_k(\boldsymbol{x_i}) = \boldsymbol{1}$ if the observed class belongs to the k-th class $c(\boldsymbol{x_i}) = \boldsymbol{c_k}$, 0 otherwise. Based on the simplified representation of $\boldsymbol{u}$ in Eq. (10), the loss function $\mathcal{L}(\boldsymbol{\beta})$ based on Eq. (11) can be rewritten as:

$$\mathcal{L}(\boldsymbol{\beta}) = -\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}}\{\boldsymbol{j}(\boldsymbol{x_i})^{\boldsymbol{T}}(\boldsymbol{\Sigma}(\boldsymbol{x_i},\cdot)\boldsymbol{\beta})^{\boldsymbol{T}} - \log\sum_{\boldsymbol{k'=1}}^{\boldsymbol{K}}\exp\{\boldsymbol{\Sigma}(\boldsymbol{x_i},\cdot)(\boldsymbol{\beta^{k'}})\}\}, \tag{13}$$

where $\boldsymbol{\beta} = [\boldsymbol{\beta^1},\ldots,\boldsymbol{\beta^K}]$ and each of $\boldsymbol{\beta^k}$ is a $(NP \times 1)$ vector of weights for the observed indicator data for the class k and $\boldsymbol{\Sigma}(\boldsymbol{x_i},\cdot)$ indicates the i-th row of the covariance matrix $\boldsymbol{\Sigma}$. Due to the large number of $\beta_{i,p}$, a direct minimization of the loss function [see Eq. (13)] would cause the *over-fitting* problem. To address this problem, we adopt an inference method based on *group* l_1-regularization (group LASSO) (Meier et al. 2008; Obozinski et al. 2007; Yuan and Lin 2006). Specifically, the $\boldsymbol{\beta}$s in Eq. (11) are grouped according to the associated covariates and each group is penalized by applying a regularization parameter. With group l_1-regularization, we update the loss function $\mathcal{L}(\boldsymbol{\beta})$ as $\mathcal{L}(\boldsymbol{\beta})_{\boldsymbol{p}}$ as:

$$\mathcal{L}(\boldsymbol{\beta})_{\boldsymbol{p}} = \mathcal{L}(\boldsymbol{\beta}) + \sum_{\boldsymbol{p=1}}^{\boldsymbol{P}}\boldsymbol{\lambda_p}||\boldsymbol{\beta}_{\mathcal{I}_p}||_2 \tag{14}$$

where $\lambda_p \geq 0$ is an adjustable regularization parameter and $\mathcal{I}_p$ is the index set that belongs to the p-th group of covariates, $p = 1, \ldots, P$.

The second component on the right side of Eq. (14) denotes the regularization term in block l_1 norm, which can be viewed as an intermediate between l_1-norm and l_2-norm. In the context of LASSO (Tibshirani 1996), the l_1-norm tends to produce sparse group solution by penalizing the regression coefficients of groups to zero (i.e., a process of variable selection), while the l_2-norm tends to yield soft penalization on the coefficients within a group. By balancing these two regularization terms, group LASSO applies the l_2-norm to the parameters within each group and the l_1-norm applies to each group. The solution of the optimal $\boldsymbol{\beta}$ in Eq. (11) is obtained by minimizing $\mathcal{L}(\boldsymbol{\beta})_{\boldsymbol{p}}$ with $\lambda_p \geq 0$, which can be transformed into a constrained convex optimization problem. Commonly used Barzilai–Borwein approximation methods (Barzilai and Borwein 1988), such as the spectral projected gradient (SPG) method (Birgin et al. 2000), can be used to solve the optimizing problem. These methods, however, tend to suffer from performance issues when the objective function becomes complex and costly to evaluate. A limited-memory projected quasi-Newton (PQN) algorithm (Schmidt et al. 2009) was recently proposed to address the performance issues of high-dimensional constrained optimization problems. This method could be taken as an extension of the commonly used L-BFGS method (Nocedal 1980) and it is particularly efficient when the number of parameters to be estimated is large, evaluation of the objective function is computationally expensive and the parameters have constraints (Schmidt 2010), which makes it very suitable for finding the optimal values of $\boldsymbol{\beta}$ in Eq. (14). Same procedure can be easily applied for Eq. (12) by taking each Z_p as an extra group and extending parameters from $\boldsymbol{\beta}$ to $[\boldsymbol{\beta^T},\boldsymbol{\alpha^T}]^{\boldsymbol{T}}$, where $\boldsymbol{\alpha} = [\boldsymbol{\alpha^1},\ldots,\boldsymbol{\alpha^K}]$ and each of $\boldsymbol{\alpha^k}$ is a $(p \times 1)$ vector of weights for observations of Z_p.

2.4 Summary

An efficient statistical framework is proposed to combine multiple spatial variables for the prediction of categorical spatial variables. In the proposed framework, each spatially distributed variable is represented as spatial covariance functions, and the class occurrence probability for a target (unknown) location is obtained by a multinomial logistic function of the data-to-unknown covariance values for each spatial variable [Eq. (11)], and collocated attribute values of each spatial variable at the target location, if there is any [Eq. (12)]. The described framework enjoys several appealing features over existing methods. Firstly, the spatial covariance functions quantify the similarity or dependency in spatially distributed variables and provide a unified representation for heterogeneous types of spatial variables (e.g., categorical vs. continuous). It should be noted here that multiple spatial covariance functions can be defined for each spatial variable for a better representation of spatial variations. Through these spatial covariance functions, incompatible spatial variables can be combined in a straightforward manner while accounting for spatial (inter-) dependencies across these variables. Secondly, a LASSO-based method, namely group LASSO, was adopted for model inference. By maximizing the likelihood adjusted by a regularization term, group LASSO simultaneously

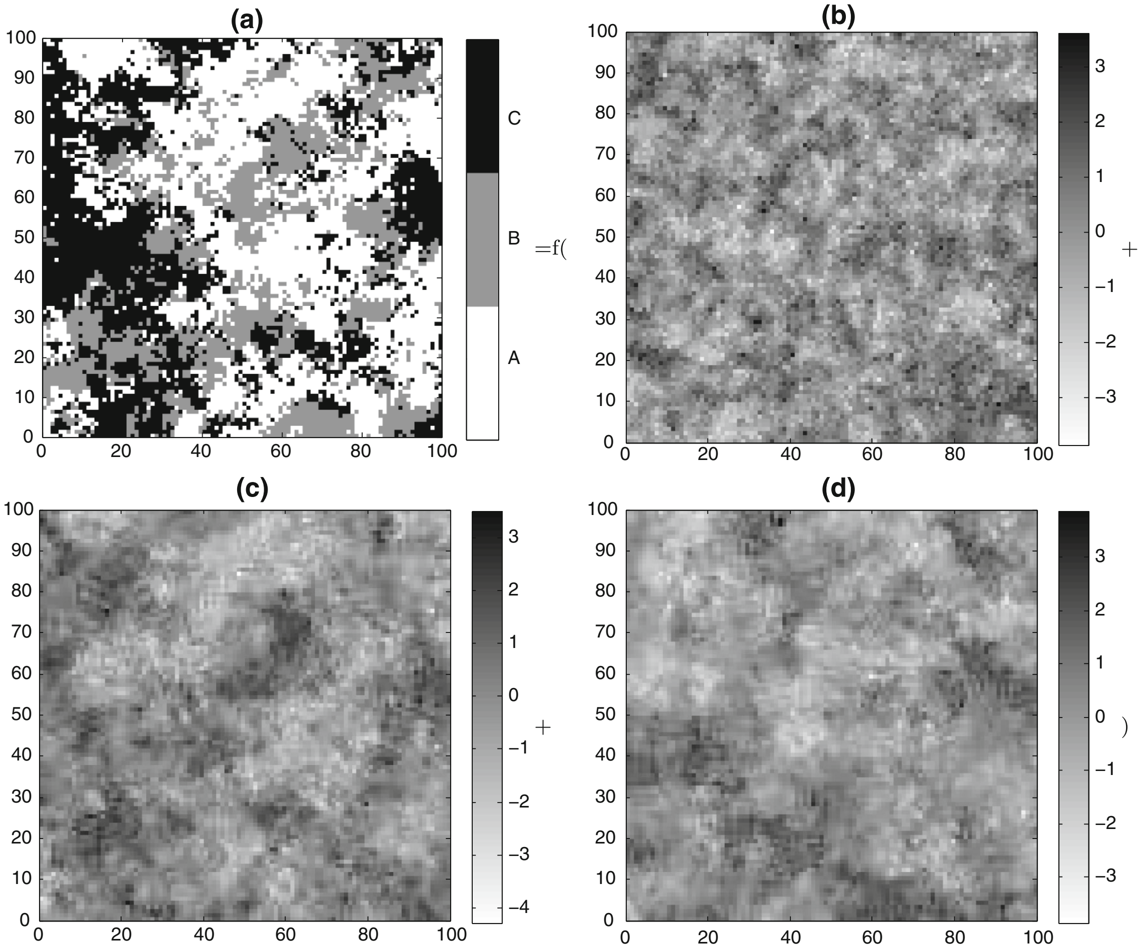


Fig. 1 Map of categorical data with three classes (**a**) generated from three realizations of Gaussian random fields (**b–d**)

estimates the coefficients and selects an optimal subset of variables in the model. Thirdly, compared with other methods such as indicator kriging family of methods and SVM-based classification methods, the proposed method produced a clear probabilistic interpretation by outputting class occurrence probability for each class label. Although the derivations of class occurrence probability in Eqs. (11) and (12) were based on the initial assumption of latent GRFs, the link between Bayesian estimation and reproducing kernels-based methods (Schölkopf et al. 2001) allows the described framework extensible to general cases.

3 Case study

The proposed framework has been implemented within the computing environment of Matlab, and a software toolbox is publicly available at http://www.cigi.uiuc.edu/guofeng/software.html. In order to illustrate the concepts and to demonstrate the performance of the proposed statistical data fusion framework, two case studies are presented in this section with one case using synthetic data and the other using real data. Due to the limitations of space, not all of the concepts introduced above could be illustrated in this paper. The synthetic case study showcases the performance of the described framework in incorporating collocated spatial information by following Eq. (12), whereas the real case study demonstrates the capability of the proposed statistical framework in integrating heterogeneous categorical and continuous spatial variables by following Eq. (11).

3.1 Synthetic case

In this synthetic case study, three models of GRFs were considered and each of them is characterized by a *zero* mean and an exponential covariogram with unit sill, 0.1 % nugget effects, and a range value of 10, 20, 30 *units*, respectively. For each of the GRF models, stochastic simulations were conducted over a regular grid (100 × 100) with a *unit* spacing. Out of the realization maps of each simulation, one map was randomly chosen and taken as a map of an auxiliary spatial variable in the subsequent analysis (see Fig. 1b–d). Based on a

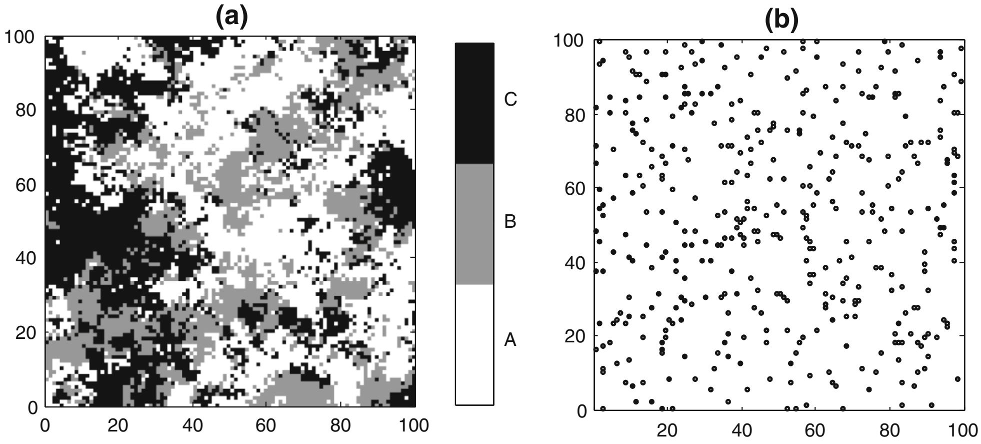


Fig. 2 Reference categorical map (**a**), and locations of 400 sample class labels (**b**)

multinomial linear combination of the three auxiliary variables, denoted as #1, #2 and #3 respectively, a categorical map with three class labels, namely #*A*, #*B* and #*C*, was generated and considered as a reference map of the primary categorical variable, as displayed in Fig. 1a.

To demonstrate the performance of the proposed data fusion method, we sampled the reference map at a set of randomly selected locations. Figure 2a and b present the reference map (same as Fig. 1a), and locations of a set of 400 samples which amount to 4 % of total locations in the reference map. The goal is to reconstruct the reference map of the primary categorical variable (Fig. 1a or Fig. 2a) using the sampled class labels (Fig. 2b) with an aid of the observed three spatial auxiliary variables (Fig. 1b–d).

To apply the proposed framework, we first represented the primary categorical variables and the auxiliary spatial variables as spatial covariance functions. Both the empirical and fitted covariance functions are presented in Fig. 3. The full model in Eq. (12) calls for the spatial covariance models of all the spatial variables including categories #*A*, #*B*, #*C*, and auxiliary variables #1, #2, #3. Group LASSO was then applied to estimate model parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}$, and the sought-after conditional class occurrence probability at each unknown location was obtained according to Eq. (12). Last, the class label with maximum occurrence probability was assigned to the unknown locations. Figure 5l shows the resulting prediction map with the corrected estimation rate of 75.6 %. Recall that LASSO-based methods (Tibshirani 1996), including group LASSO (Yuan and Lin 2006), estimate coefficients, while simultaneously selecting the most important variables by shrinking the coefficients of the others. The estimated coefficients are presented in Fig. 4. One can clearly see that the coefficients for spatial covariance values (spatial effects) of category #*A* and #*B* of the primary categorical variable, auxiliary variable #2 and #3 are nearly zeros, which suggest that the contribution of spatial covariances of these variables is not substantial as other variables are for the occurrence of the categorical variable of interest. We dropped these four variables from the model, repeated the process and we obtained almost identical prediction map as that of the full model (Fig. 5l).

Two other methods—the multinomial GLM and the spatial multinomial linear mixed model (MLMM) (Cao et al. 2011)—were also applied to this synthetic case study. The former tends to ignore the spatial dependence information in a spatial setting, and the latter doesn't account for auxiliary information. The resulting prediction maps are displayed in Figure 5j, k, respectively. One can clearly see that the proposed method (Fig. 5l) better reproduces the reference map (Fig. 2a). The corrected estimation rates of these two methods were 64.4 and 65.7 %, respectively, both inferior to 75.7 % of the proposed method. Figure 5 also displays the estimated probability maps of the three different methods for category #*A*, #*B*, and #*C*. One can see that spatial MLMM (Fig. 5b, e, h), with no consideration of auxiliary variables, tends to yield continuous results, while multinomial GLM (Fig. 5a, d, g), ignoring spatial dependence information, tends to be unsure (most of the estimated probability values in a range of 0.5–0.6) at target (unknown) locations. This suggests that both auxiliary variables and the spatial dependence information play important roles in the prediction of class occurrence of the primary categorical variable in this case. This result should be expected considering that reference categorical map was generated from a linear combination of realizations of three auxiliary variables with strong spatial dependency. We repeated the above process for different numbers of sampled sizes. Table 1 lists the correct estimation rates for the three different methods. Apparently, the proposed method tends to yield substantially better correct estimation rates in every case than the other two methods.

3.2 Real case

Public land survey (PLS) data of the general land office (GLO) have been widely used in landscape studies of the forest and woodlands in pre- and early-European settled Midwestern and Western US. Forest vegetation distribution

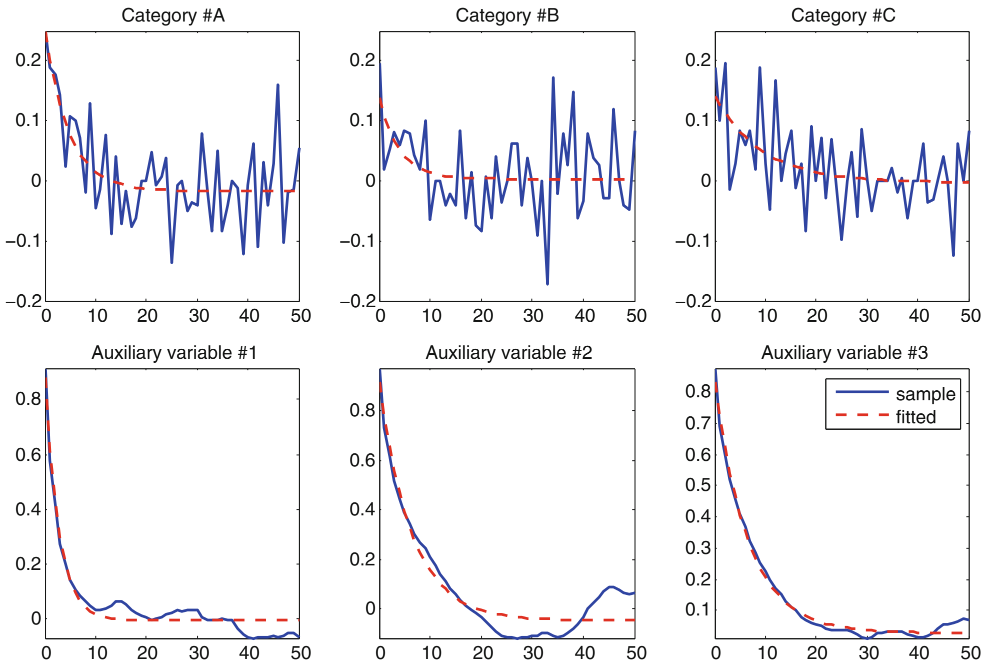
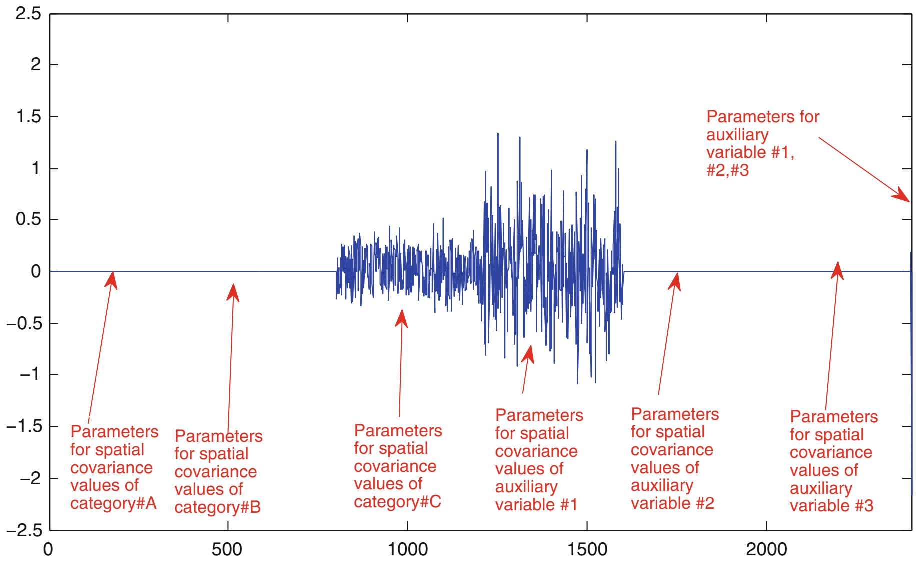


Fig. 3 Sample and fitted covariograms for category #*A*, #*B* and #*C*, and auxiliary variable #1, #2 and #3

Fig. 4 Plot of estimated coefficients of group LASSO method for covariance values of category #*A*, #*B*, and #*C*, covariance and observed values of spatial auxiliary variables #1, #2, and #3. One can clearly see that coefficients for covariances of category #*A*, and auxiliary variable #1 and #2 are penalized as nearly zeros



maps at a finer spatial resolution than available is oftentimes needed. In this case study, we aim to reconstruct the spatial distribution of the three most abundant tree species (*post oak*, *black oak* and *elm*) from PLS data in the Arbuckle Mountains of south-central Oklahoma, with availability of information from multiple environmental covariates, including elevation (continuous type), geological and soil types (categorical type). To demonstrate the advantages of the proposed method in incorporating heterogeneous auxiliary information, we compared the prediction result of the proposed method with the result from the spatial MLMM model (Cao et al. 2011), where the auxiliary environmental covariate information is not taken into account.

Figure 6(a) shows the locations of a total of 2,561 witness trees obtained from the 1870s survey. We focused on the three most abundant species: *post oak* (48.0 %), *black oak* (20.2 %), and *elm* (12.8 %), and re-categorized the rest species as *other-type* (19 %). As evidenced in Fig. 6, *post oak* is the most abundant tree species with a strong concentration in the southern portion of the study area, while *black oak* is more evenly distributed with a few clusters in

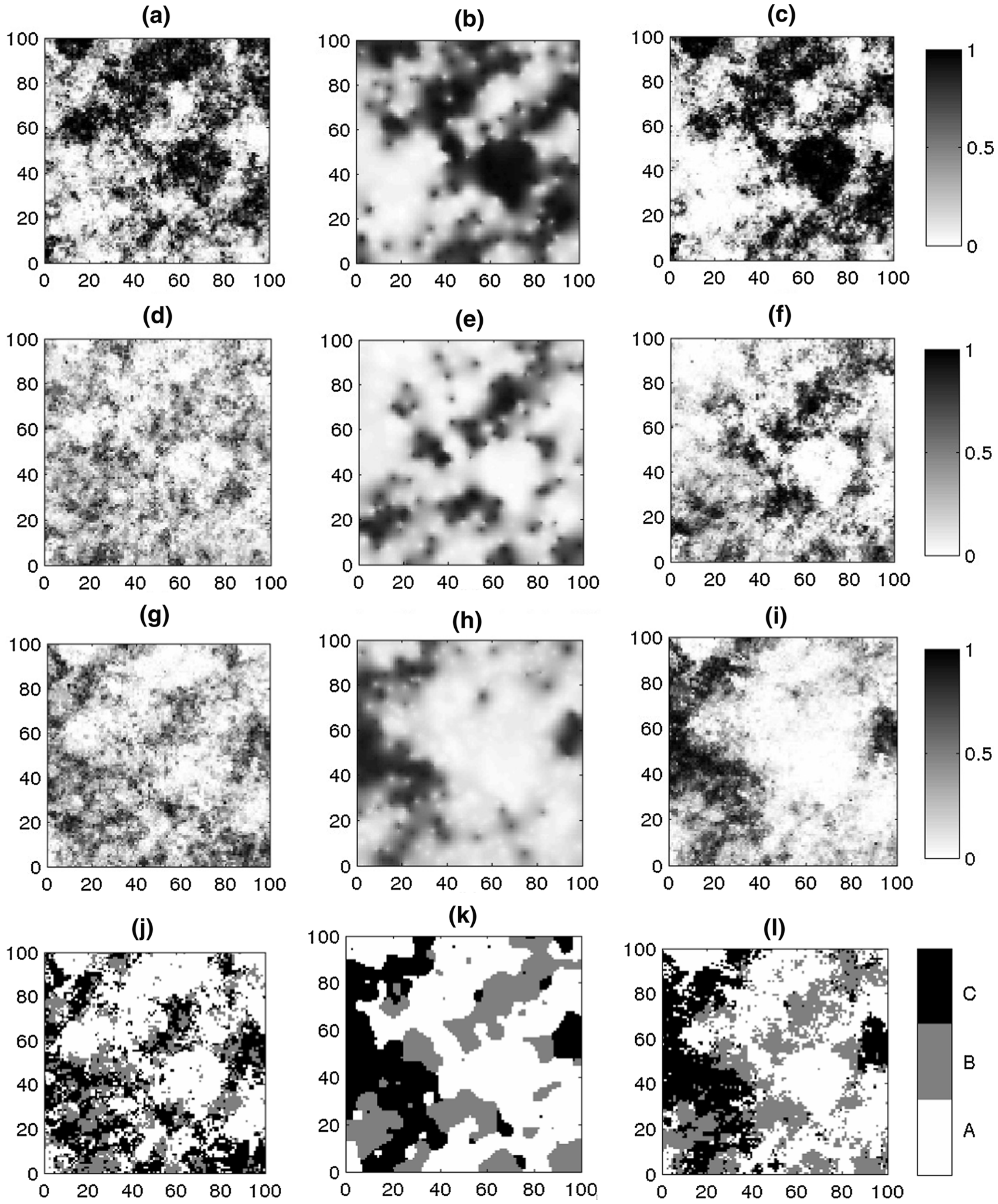


Fig. 5 **a–c** Represent the estimated probability map for class label #*A* via multinomial GLM, spatial MLMM, and Eq. 12 respectively, **d–f** for class label #*B* and **g–i** for class label #*C*; **j–l** represent prediction results of these three methods

the central and east regions of the study area. In contrast to these two oaks, *elm* appears more often in the north east side of the study area. All three trees species show the presence of spatial clusters, but with different intensities (Yoo et al. 2013). These witness tree data was collected at 0.8 km (quarter mile) intervals, typical in public land survey system (He et al. 2000), and only a small fraction of the tree observations (approximately 0.07 % of total tree pairs) are less than 0.4 km apart. The objective of this case study was to model and reconstruct the distribution of tree species distribution. It is expected that, without further auxiliary information, using this witness tree dataset alone might lead to unreliable results. Earlier efforts (Fagin and Hoagland 2011; He et al. 2007; Yoo et al. 2013) have

Table 1 Correct estimation rates of three different methods (GLM, spatial MLMM, and the proposed data fusion method)

Percentage of sample size (%)	GLM (%)	Spatial MLMM (%)	The proposed method (%)
1	63.9	54.2	68.1
2	61.8	59.4	69.9
3	64.5	61.5	72.8
4	64.4	65.7	75.6
5	65.1	66.1	76.1
6	64.6	67.8	77.5
7	64.9	68.5	78.2
8	65.2	69.2	78.5
9	65.2	71.5	79.2
9.99	65.2	71.6	80.0

confirmed that environmental conditions to which tree species respond play an important role in the reconstruction of the spatial distribution of forest vegetation. We selected three predictors that have varying degrees of influence on each tree species based on literature and preliminary data analyses. Table 2 provides a brief description of covariates, whose spatial distribution is shown in Fig. 6b–d (Yoo et al. 2013).

As mentioned above, the auxiliary data consist of a continuous (elevation) and two categorical (geological and soil types) variables, which are typically challenging to combine using conventional statistical methods. We reconciled this problem using the proposed method, that is, by applying Eq. (11) to these categorical and continuous data. Similar to the synthetic case study, the predicted probabilities for each tree species at each prediction location were first estimated, and then a class with the maximum posterior probability was identified. As a result, we can recover the tree species that is most likely to have been present at the prediction location (see Fig. 7). Figure 7a shows the resulting map of species occurrence based on spatial MLMM without covariates, while Fig. 7b presents the results based on the proposed model [Eq. (11)] with covariates.

We assess the predictive performance of the two models using cross-validation, where data are split into validation and training data. Validation data consist of a subset (10 %) of observed witness tree data, which are withheld in the model fitting process and later used to validate model outcomes. In other words, only the training data are used for model fitting. The size of validation data might not be sufficient for an effective assessment of the model accuracy and the sectional bias might be involved in the data split process. Therefore, we repeated the validation process

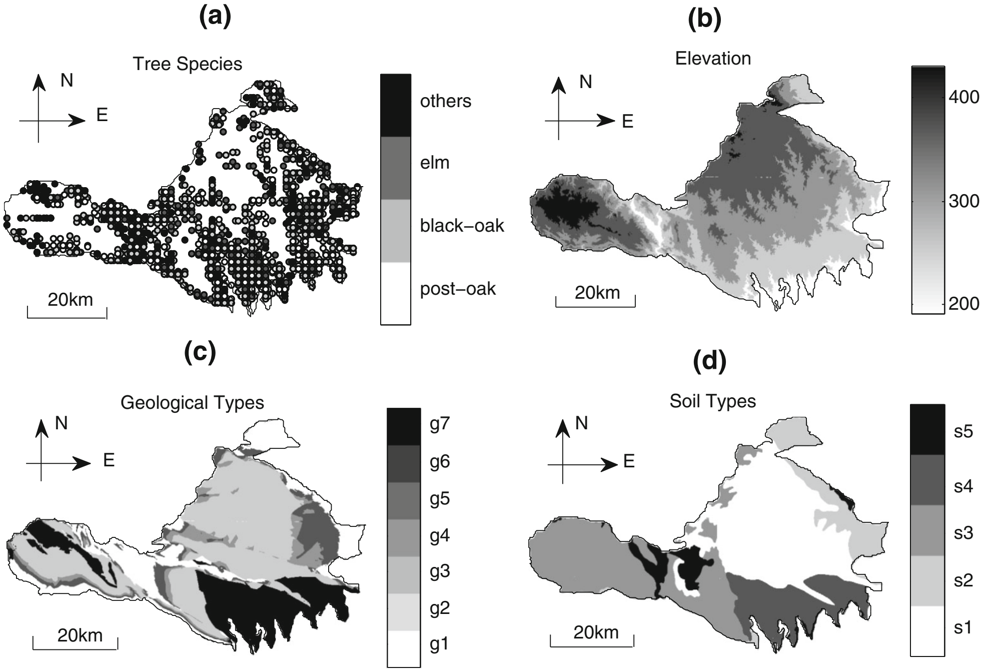


Fig. 6 Maps of the witness tree species data and three environmental covariates in the Arbuckle Mountains area (latitude ranging between 34.21694° and 34.71635° and longitude between −97.36979° and −96.43859°): **a** survey locations of the three most abundant tree species: post oak, black oak, and elm; **b** elevation; **c** geology type; **d** soil type. All the three covariates maps (**b**–**d**) are at 30 × 30 m spatial resolution. See Table 2 for the definition of each category

Table 2 Description of environmental covariates [modified from (Yoo et al. 2013)]

Type	Interval/categories	Source
Elevation		
Continuous	[191, 432] m	USGS
Geology		
Categorical	Shale/limestone (g1)	USGS
	Shale (g2)	
	Limestone (g3)	
	Oil creek sandstone (g4)	
	Bromide sandstone (g5)	
	Granite (g6)	
	Other-type (g7)	
Soil		
Categorical	Shidler-Scullin-Lula-Claremore-rock outcrop (s1)	NRCS STATSGO
	Shidler-rock outcrop (s2)	
	Kiti-rock outcrop (s3)	
	Chigley-Agan-rock outcrop (s4)	
	Reinach-McLain-Dale (s5)	

USGS United States Geological Survey, *NRCS* Natural Resources Conservation Service, *STATSGO* State Soil Geographic Database

iteratively (100×) where a new set of training and validation data are randomly selected at each time, and model accuracy is calculated based on new model fit and newly selected validation data. The performance of the proposed model, with average correct estimation rate 62.0 % is substantially better than that of the spatial MLMM 50.9 %. The output coefficients for the covariance values of elevation, soil types s_1, s_2 and s_3, are penalized to zeros (Fig. 8). This result suggests that spatial covariances of these variables do not explain to the occurrence of tree species, which may be due to the homogeneous physiographic features and the strong spatial association between soil types and geological composition in the study area. For example, the limestone and dolomite substrates constitute 69 % of the surface rocks in the study area, and shallow soils characterize areas where granite and rhyolite are common. The most extensive soil type in the Arbuckle Mountains is the Shidler-Scullin-Lula-Claremore-rock outcrop complex (s_1), a silty clay loam that covers the greatest areal extent (37.7 %), which occurs primarily on fractured limestone (g_3) (Bogard 1973; Burgess 1977), one of the most dominant geology type. The Kiti-rock outcrop complex (s_3) is a clay loam soil that occupies 27.4 % of the study area and consists of moderately alkaline loam of very shallow, well-drained soils (Bogard 1973; Burgess 1977). The elevation of the study area varies from 191 to 432 m, and this range is relatively too small to impact the distribution of tree species, and there is strong colinearity existing between the elevation, soil types s_1, s_2 and s_3, and the geological types in terms of spatial distribution (dependence information) (see Fig. 6). The spatial variability of tree species occurrence are captured by the spatial covariance values of geological types (with non-zero coefficients), and because of the colinearity, little valuable information can be further contributed by the spatial covariances of soil types (s_1, s_2, and s_3). Similarly with the synthetic cases, we dropped the spatial covariances of these variables in the model, repeated the process, and we obtained almost identical prediction result as shown in Fig. 7b.

4 Concluding discussion

As the amount of spatial data grows exponentially, diverse sources of spatial data have become increasingly available in geospatial research. These diverse sources of spatial data, however, tend to be heterogeneous and incompatible with each other, which calls for efficient spatial data fusion methods. It is challenging to reconcile these heterogeneous data sources particularly when considering the wide range of heterogeneities and complex (inter-)dependence structures in spatial settings. This paper describes a statistical framework of heterogeneous spatial data fusion for the

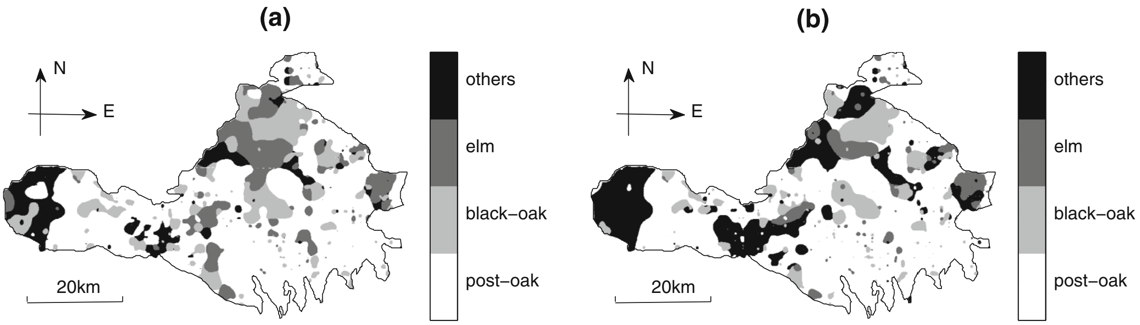


Fig. 7 Predicted map of species occurrences in the Arbuckle Mountains area obtained (same geographic extent as that of Fig. 6) from two models: **a** spatial multinomial logistic mixed model; **b** The proposed model [Eq. (11)]. Both of the two maps are at 30 × 30 m spatial resolution

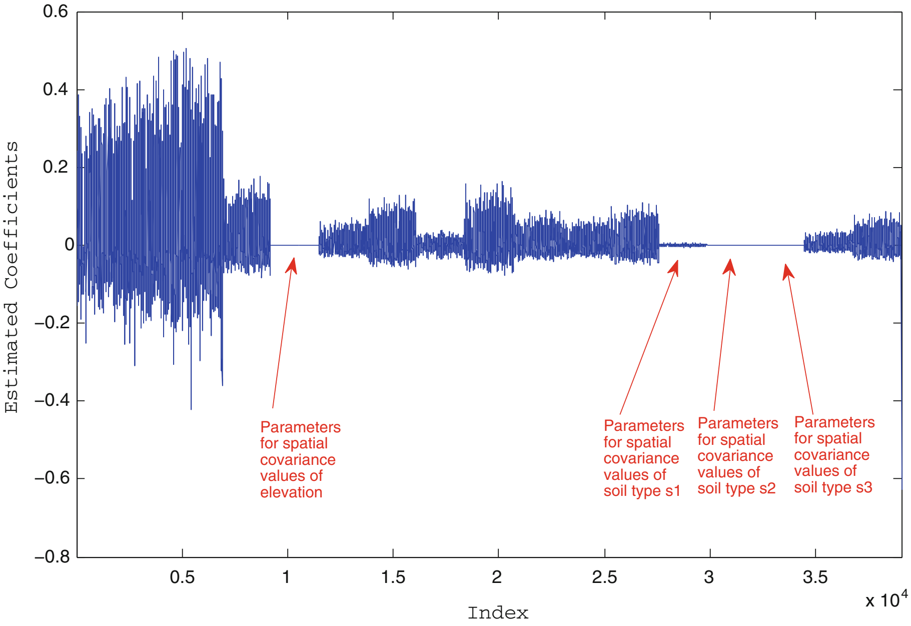


Fig. 8 Estimated parameters of group LASSO, and the coefficients for covariances of elevation dataset are penalized as near zeros

prediction of categorical spatial variables. In this framework, each spatial variable is represented via spatial covariance functions. This representation of spatial covariance functions has a number of virtues for statistical analysis of spatial data. A spatial covariance function defines the similarity (or dependency) of a spatial variable as functions of separating vectors. It should be noted that more than one covariance function could be defined for a spatial variable to better capture its spatial variation characteristics. From another perspective, a spatial covariance function essentially projects the heterogeneous input spatial variables into a unified reproducing kernel Hilbert space, and thus provide a unified representation for heterogeneous types of spatial data, independent of data nature and object complexity. Through spatial covariance functions, information implied in heterogeneous spatial data can then be combined in a straightforward fashion while accounting for the spatial (inter-)dependencies across these spatial variables. Although the discussion of this paper focused on the spatial prediction of categorical variables, this spatial covariance functions-based data fusion strategy could be extended into a general spatial prediction context. In addition to integrating spatial information with spatial support of points in two dimensional spaces, the described framework could be extended to account for more general types of spatial supports, such as areal units or volumes in higher dimensional spaces. Specific types of spatial covariance functions, however, need to be carefully designed to capture spatial variations of these variables. Areal-to-areal spatial covariance functions may be necessary for spatial variables represented by areal units, and areal-to-point covariance functions for similarities between points and areal spatial variables. Careful investigations of such extensions are warranted in future research.

A recently proposed group LASSO was adopted in this paper for model inference to avoid the over-fitting issues. By penalizing parameters of less informative variables in the model (spatial covariance functions) into nearly zeros, group LASSO actually selects an subset of most relevant variables in the model. Advantages of this described framework in the spatial prediction of categorical variables have been discussed. In the setting of spatial analysis, only one observation can usually be made for a spatial variable at a certain location, and the observation typically is not repeatable. Therefore, as the complexity of the spatial analysis increases by incorporating more spatial variables and accounting for complex spatial interactions across these variables, one could end up with an underdetermined model with insufficient degree of freedom [more unknown parameters than observations, known as “large p, small n” paradigm (West 2003)] to estimate the model, which would be difficult for conventional methods to handle. LASSO-based methods, as the group LASSO adopted in this paper, demonstrates great potentials to address such problems by enforcing the parameters of irrelevant group of variables to zeros (group sparsity).

Acknowledgments We gratefully acknowledge the funding provided by the National Science Foundation under grant number OCI-1047916 to support this research. We would like to thank Professors Bruce W. Hoagland and Todd D. Fagin from the University of Oklahoma for valuable discussions and the datasets they kindly provided. We would also thank the anonymous reviewers for the constructive comments and suggestions, and thank Professor Jeff Lee from Texas Tech University for his proofreading which has profoundly improved the composition of this manuscript.

References

Atkinson P, Lewis P (2000) Geostatistical classification for remote sensing: an introduction. Comput Geosci 26(4):361–371

Atkinson PM (2012) Downscaling in remote sensing. Int J Appl Earth Obs Geoinf

Barzilai J, Borwein JM (1988) Two-point step size gradient methods. IMA J Numer Anal 8(1):141–148

Birgin E, Marttinez J, Raydan M (2000) Nonmonotone spectral projected gradient methods on convex sets. SISM SISM J Optim 10:1196–1211

Bogaert P (2002) Spatial prediction of categorical variables: the Bayesian maximum entropy approach. Stoch Environ Res Risk Assess 16(6):425–448

Bogaert P, Fasbender D (2007) Bayesian data fusion in a spatial prediction context: a general formulation. Stoch Environ Res Risk Assess 21:695–709

Bogard V (1973) Soil survey of Pontotoc County, Oklahoma, U.S. Soil Conservation Service

Breslow N, Clayton D (1993) Approximate inference in generalized linear mixed models. J Am Stat Assoc 88(421):9–25

Burgess D (1977) Soil survey of Johnston County, Oklahoma, National Cooperative Soil Survey

Cao G, Kyriakidis P, Goodchild M (2011) A multinomial logistic mixed model for the prediction of categorical spatial data. Int J Geogr Inf Sci 25(12):2071–2086

Chiles J, Delfiner P (1999) Geostatistics: modeling spatial uncertainty. Wiley, New York

Christakos G (1990) A Bayesian/maximum-entropy view to the spatial estimation problem. Math Geol 22(7):763–777

Christensen O (2004) Monte Carlo maximum likelihood in model-based geostatistics. J Comput Graph Stat 13(3):702–718

Diggle P, Tawn J, Moyeed R (1998) Model-based geostatistics. Appl Stat 47(3):299–350

Fagin T, Hoagland B (2011) Patterns from the past: modeling Public Land Survey witness tree distributions with weights-of-evidence. Plant Ecol 212:207–217

Foody GM (2002) Status of land cover classification accuracy assessment. Remote Sens Environ 80:185–201

Goodchild M, Zhang J, Kyriakidis P (2009) Discriminant models of uncertainty in nominal fields. Trans GIS 13(1):7–23

Goovaerts P (1997) Geostatistics for natural resources evaluation. Oxford University Press, New York

Goovaerts P (1998) Accounting for estimation optimality criteria in simulated annealing. Math Geol 30(5):511–534

Gotway CA, Stroup WW (1997) A generalized linear model approach to spatial data analysis and prediction. J Agric Biol Environ Stat 2(2):157

Goulard M, Voltz M (1992) Linear coregionalization model: tools for estimation and choice of cross-variogram matrix. Math Geol 24(3):269–286

He H, Dey D, Fan X, Hooten M, Kabrick J, Wikle C, Fan Z (2007) Mapping pre-European settlement vegetation at fine resolutions using a hierarchical Bayesian model and GIS. Plant Ecol 11:85–94

He H, Mladenoff D, Sickley T, Guntenspergen G (2000) GIS interpolations of witness tree records (1839–1866) for Northern Wisconsin at multiple scales. J Biogeogr 27:1131–1042

Hengl T, Heuvelink G, Rossiter D (2007) About regression-kriging: from equations to case studies. Comput Geosci 33(10):1301–1315

Hengl T, Heuvelink G, Stein A (2004) A generic framework for spatial prediction of soil variables based on regression-kriging. Geoderma 120(1):75–93

Hengl T, Toomanian N, Reuter H, Malakouti M (2007) Methods to interpolate soil categorical variables from profile observations: lessons from Iran. Geoderma 140:417–427

Journel AG (1983) Nonparametric estimation of spatial distributions. Math Geol 15(3):445–468

Journel AG, Alabert F (1989) No-Gaussian data expansion in the Earth Sciences. Terra Nova 1(1):123–134

Kimeldorf G, Wahba G (1970) A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. Ann Math Stat 41(2):495–502

Lanckriet GRG, De Bie T, Cristianini N, Jordan MI, Noble WS (2004) A statistical framework for genomic data fusion. Bioinformatics 20(16):2626–2635

Li D, Zhang J, Wu H (2012) Spatial data quality and beyond. Int J Geogr Inf Sci 26(12):2277–2290

Liang K, Zeger S (1986) Longitudinal data analysis using generalized linear models. Biometrika 73(1):13

Meier L, Geer SVD, Bühlmann P (2008) The group lasso for logistic regression. J R Stat Soc B 70:53–71

Miller HJ, Han J (2003) Geographic data mining and knowledge discovery. CRC Press, Boca Raton

Nocedal J (1980) Updating quasi-newton matrices with limited storage. Math Comput 35(151):773–782

Obozinski G, Taskar B, Jordan M (2007) Joint covariate selection for grouped classification, technical report, University of California, Berkeley

Pardo-Igúzquiza E, Dowd P, Pardoiguzquiza E (2005) Multiple indicator cokriging with application to optimal sampling for environmental monitoring. Comput Geosci 31(1):1–13

Rue H, Martino S, Chopin N (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J R Stat Soc B 71(2):319–392

Schmidt M (2010) Graphical model structure learning with $l1$-regularization. PhD thesis, University of British Columbia

Schmidt M, Berg EVD, Friedlander M, Murphy K (2009) Optimizing costly functions with simple constraints: a limited-memory projected quasi-newton algorithm. In: Proceedings of the 12th international conference on artificial intelligence and statistics (AISTATS), pp. 456–463

Schölkopf B, Herbrich R, Smola A (2001) A generalized representer theorem. In: Proceedings of the annual conference on computational learning theory, pp. 416–426

Schölkopf B, Smola A (2002) Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT Press, Cambridge

Schoölkopf B, Tsuda K, Vert J-P (2004) Kernel methods in computational biology. MIT Press, Cambridge

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc 58:267–288

Tso B, Mather P (2009) Classification methods for remotely sensed data. CRC Press, Boca Raton

Wackernagel H (1998) Multivariate geostatistics—an Introduction with applications, 2nd edn. Springer, New York

Wahba G (1990) Spline models for observational data, vol. 59. Society for Industrial and Applied Mathematics, Philadelphia

West M (2003) Bayesian factor regression models in the large p, small n paradigm. Bayesian Stat 7(2003):723–732

Wibrin M, Bogaert P, Fasbender D (2006) Combining categorical and continuous spatial information within the Bayesian Maximum Entropy paradigm. Stoch Environ Res Risk Assess 20:423–433

Williams C, Barber D (2002) Bayesian classification with Gaussian processes. Pattern Anal Mach Intell IEEE Trans 20(12):1342–1351

Yoo E-H, Hoagland BW, Cao G, Fagin T (2013) Spatial distribution of trees and landscapes of the past: a mixed spatially correlated multinomial logit model approach for the analysis of the public land survey data. Geogr Anal 45(4):419–440

Yoo E-H, Trgovac A (2011) Scale effects in uncertainty modeling of presettlement vegetation distribution. Int J Geogr Inf Sci 25(3):405–421

Yuan M, Lin Y (2006) Model selection and estimation in regression with grouped variables. J R Stat Soc B 68:49–67

Zhang H (2002) On estimation and prediction for spatial generalized linear mixed models. Biometrics 58(1):129–136

Zhang J, Goodchild M (2002) Uncertainty in geographic information. Taylor & Francis, London