A Geostatistical Framework for Heterogeneous Spatial Data Fusion

Guofeng Cao

guofeng.cao@ttu.edu
Department of Geosciences
Texas Tech University

Abstract

This paper proposes a geostatistical framework of spatial data fusion for spatial prediction and uncertainty modeling, while accounting for varieties of spatial heterogeneities and complex spatial dependencies. Within this proposed framework, spatial variables are characterized via spatial covariance functions, which measure the spatial dependencies of pair-wise locations and project heterogeneous spatial data into a unified space of similarity (or kernel space). The representation of spatial covariance thus provides a convenient venue to integrate heterogeneous data source while taking into account spatial dependencies. We show that the probability distribution at target locations given the neighboring observations can be represented as a generalized linear combination of spatial covariance functions between the target and observed locations. For parameter estimation, a recently proposed group LASSO (Least Absolute Shrinkage and Selection Operator)) approach is adopted to prevent the model from over-fitting. Case studies are conducted to showcase the advantages of the proposed framework.

1 Introduction

With the continuing advancement of spatial data acquisition and dissemination technology in the past few decades, the availability of geospatial data has been increasing dramatically for many geographical or environmental research problems. In the case of soil mapping, for example, measurements of external environmental conditions, such as climate, vegetation, are often available in addition to the measurements of soil properties (e.g., moisture and nutrition). These external environmental conditions are known to be related to the distribution of soil types. It would be ideal to fuse these diverse information efficiently to achieve a comprehensive perspective. These geospatial data, however, often demonstrate incompatible heterogeneities with each other in terms of data nature (continuous or categorical), spatial support (areal or point-reference data), spatial scales, and sample locations (missing values). 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.

Recently, in the context of spatial prediction of categorical variables, Cao et al. (2014) proposed a statistical approach to integrate diverse spatial data sources. In this approach, each heterogeneous spatial variable is characterized via spatial covariance functions, and the desirable probability mass function at targeted (unsampled) locations can be represented as a multinomial logistic combination of spatial covariance values between targeted and source (sampled) locations. As similarity (dependence) measures of spatially correlated variables between pairs of locations, spatial covariance functions essentially project the heterogeneous spatial variables into a unified space of kernel (reproducing kernel Hilbert space), and thus provide a straightforward venue for integrating the heterogeneous spatial data while accounting for complex spatial dependencies and varieties of spatial heterogeneities. In this paper, we propose to extend the approach of Cao et al. (2014) into a data fusion framework in a general context of spatial prediction and spatial uncertainty modeling.

2 Methodology

Consider a spatially indexed random variable (RV) $Z(\boldsymbol{u})$ ($\boldsymbol{u} \in R^d$), and that for an arbitrary location $\boldsymbol{u}$, $Z(\boldsymbol{u})$ is assumed to follow a distribution of exponential family. There exist N sampled values $\boldsymbol{z} = \{z(\boldsymbol{u}_1), \ldots, z(\boldsymbol{u}_N)\}$ corresponding to sampled locations $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N$. In addition to the primary RV $Z(\boldsymbol{u})$, there is a vector of P auxiliary variables $X(\boldsymbol{u}) = \{X_1(\boldsymbol{u}), \ldots, X_p(\boldsymbol{u})\}$, and for each of them, we have N number of observations, denoted as $\boldsymbol{x}_p = \{x_p(\boldsymbol{u}_1), \ldots, x_p(\boldsymbol{u}_N)\}, p = 1, \ldots, P$. Given a target (unsampled) location $\boldsymbol{u}^*$, the objective in spatial prediction context is to estimate the conditional probability $P\{Z(\boldsymbol{u}^*)|\boldsymbol{u}^*, \boldsymbol{z}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_P\}$, or $P\{Z(\boldsymbol{u}^*)|\mathcal{D}\}$ for notation simplicity. Please note that $Z(\boldsymbol{u})$ and $X_p(\boldsymbol{u})$ don't have to be Gaussian distributed, and observations $\boldsymbol{z}$ and $\boldsymbol{x}_p$ don't have to be collocated with each other (e.g., missing values in $\boldsymbol{x}_p(\boldsymbol{u})$).

Generalized linear models with mixed effects (Breslow & Clayton 1993) provide a flexible framework for investigating the relationship between non-Gaussian response variables and covariates while accounting for clustering effects. This framework has been specifically adopted for modeling non-Gaussian spatial variables (e.g., Diggle et al. 1998). In such models, one usually first specify the probability distribution for the response variable $Z(\boldsymbol{u}^*)$, and then model the connection between the expected response $E\{Z(\boldsymbol{u}^*)\}$ and a linear combination of covariates (a linear predictor), via a specified link function $g(\cdot)$, i.e.:

$$g(E\{Z(\boldsymbol{u}^*)\}) = \sum_{p=1}^{P} \beta_p x_p(\boldsymbol{u}^*) + S(\boldsymbol{u}^*)$$

where $S(\boldsymbol{u}^*)$ is a Gaussian random field with zero mean $E[S(\boldsymbol{u}^*)] = 0$ for all $\boldsymbol{u}$ and covariance functions $\sigma(S(\boldsymbol{u}^*), S(\boldsymbol{u}'); \boldsymbol{\theta})$ for all $\boldsymbol{u}^*$ and $\boldsymbol{u}'$, where the spatial covariance function $\sigma(\cdot, \cdot)$ depends on a vector parameter $\boldsymbol{\theta}$ under the stationarity assumption. Specifically if $Z(\boldsymbol{u})$ is a categorical variable $Z(\boldsymbol{u}) \sim \{1, \ldots, K\}$, based on Representer Theorem in reproducing kernel Hilbert space (Schölkopf & Smola 2002), the desirable (posterior) class occurrence probability $P\{Z(\boldsymbol{u}^*) = k|\mathcal{D}\}$, or link function $g(\cdot)$, could be represented as a weighted combination of spatial covariance function, and measurements of collocated $X(\boldsymbol{u})$ if there is any (Cao et al. 2014). This conclusion for response variables with multinomial distribution could actually be extended into other distributions in exponential family (e.g., Poisson), i.e.:

$$g(E\{Z(\boldsymbol{u}^*)\}) = \beta_0 + \sum_{p=1}^{P} \sum_{i=1}^{N} \beta_{i,p} \sigma_p(\boldsymbol{u}^*, \boldsymbol{u}_i; \boldsymbol{\theta}_p) \tag{1}$$

where $\sigma_p(\boldsymbol{u}^*, \boldsymbol{u}_i; \boldsymbol{\theta}_p)$ is spatial covariance function for variable $X_p(\boldsymbol{u})$ and β_0 is the intercept. If observations of auxiliary variable $x_p(\boldsymbol{u})$ are available (collocated with $Z(\boldsymbol{u})$) at targeted location

$\boldsymbol{u}^*$, it could also be incorporated into the link function, i.e.:

$$g(E\{Z(\boldsymbol{u}^*)\}) = \beta_0 + \sum_{p=0}^{P} \alpha_p x_p(\boldsymbol{u}^*) + \sum_{p=1}^{P} \sum_{i=1}^{N} \beta_{i,p} \sigma_p(\boldsymbol{u}^*, \boldsymbol{u}_i; \boldsymbol{\theta}_p) \quad (2)$$

Under assumption of stationarity, the spatial covariance function could be simplified as a covariogram $\sigma_p(\boldsymbol{h}; \boldsymbol{\theta}_p) = \sigma_p(\boldsymbol{u}^*, \boldsymbol{u}_i; \boldsymbol{\theta}_p)$ where $\boldsymbol{h} = \boldsymbol{u}^* - \boldsymbol{u}_i$ is a seperation vector. Similar as in kriging family of methods, $\sigma_p(\boldsymbol{h}; \boldsymbol{\theta}_p)$ could be fit by scanning sampled values.

In the proposed methods (Equation 1 and Equation 2), each spatially distributed variable is characterized as spatial covariance functions, and the posterior probability, or link function, for a target (unknown) location is obtained by a function of the data-to-unknown covariance values for each spatial variable (Equation 1), and collocated attribute values of each spatial variable at the target location, if there is any (Equation 2). 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. Similar representation has been used in the dual form of kriging family of methods (Goovaerts 1997), but this dual form only works best in the Gaussian cases. Through these spatial covariance functions, incompatible spatial variables can be combined in a straightforward manner while accounting for spatial (inter-) dependencies across these variables. For parameter estimation, group LASSO (Least Absolute Shrinkage and Selection Operator) (Yuan & Lin 2006), is adopted to prevent the model from *over-fitting* and simultaneously select an optimal subset of important information (variables).

3 Preliminary Case Study

In this preliminary case study, we consider a primary variable as a categorical variable with three class labels, and three continuous auxiliary variables. Maps of these three auxiliary variables (100×100, see Figures 1(b-d)) were stochastically simulated based on three independent Gaussian random fields (GRFs) with different specification of covariance functions. Based on a multinomial linear combination of these three auxiliary variables, a categorical map with three class labels was generated and considered as a reference map of the primary categorical variable, as displayed in Figure 1(a). To demonstrate the performance of the proposed methods, we sampled the reference map (Figure 1(a)) at a set of randomly selected locations (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 (Figure 1(a) using the sampled class labels (Figure 2(a)) with an aid of the observed three spatial auxiliary variables (Figures 1(b)-(d)). Indicator variants of kriging family of methods (Goovaerts 1997) could be applied to this case, but they tend to lead to problematic posterior probabilities.

To apply the proposed framework, the spatial covariance functions of the primary categorical variable and the auxiliary spatial variables were first modeled. Since in this case, the observations of auxiliary variables are collocated with the primary variable, Equation 2 were adopted by including spatial covariance models of all the spatial variables, and measurement values of auxiliary variables. Group LASSO was then applied to estimate model parameters, and the sought-after conditional class occurrence probability at each unknown location was obtained according to Equation 2. Last, the class label with maximum occurrence probability was assigned to the unknown locations. Figure

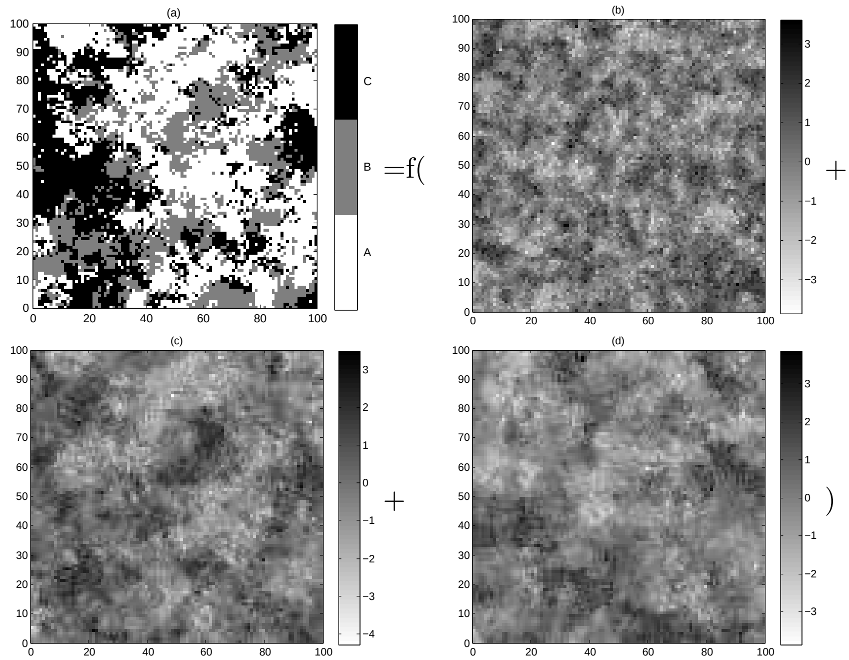

Figure 1: Reference map of categorical data with three classes (a) generated from three realizations of Gaussian random fields (b-d).

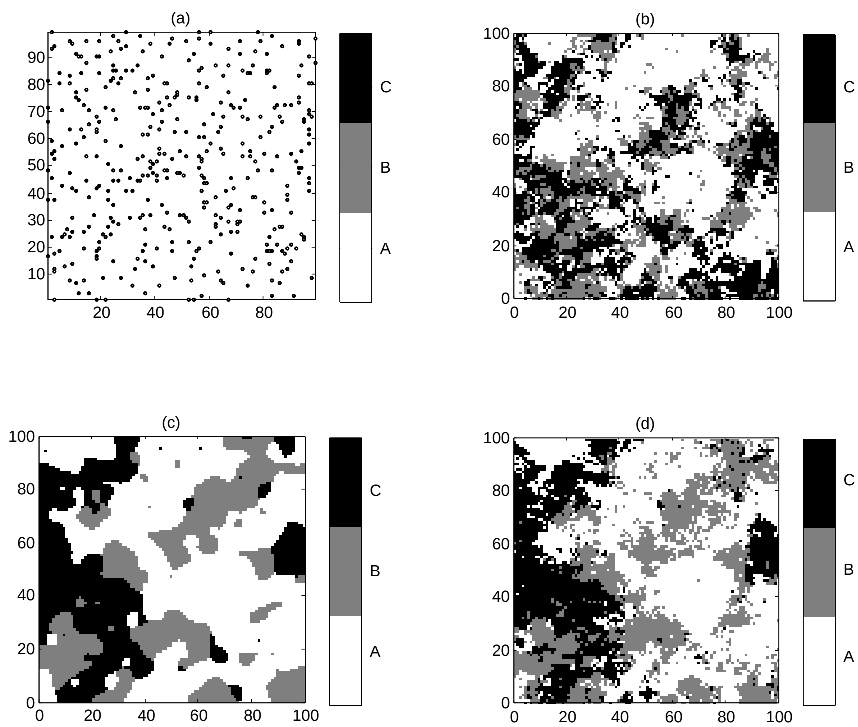

Figure 2: Locations of 400 sample class labels (a), and estimated maps via GLM (b), spatial MLMM (c) and the proposed method (d)

2(d) shows the resulting prediction map with the corrected estimation rate of 75.6%. Two other methods, the multinomial GLM and the spatial multinomial linear mixed model (MLMM) (Cao et al. 2011), were also applied to this preliminary 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 2(b) and Figure 2(c), respectively. The corrected estimation rates of these two methods were 64.4% and 65.7%, respectively, both inferior to 75.7% of the proposed method.

4 Conclusion

In this paper, we described a geostatistical framework of heterogeneous spatial data fusion, where each spatial variable is characterized via spatial covariance functions. The spatial covariance function essentially project 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. 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 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. Further investigations of such extensions and the performances in real cases are needed in the future work.

References

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

Cao, G., Kyriakidis, P. & Goodchild, M. (2011), 'A multinomial logistic mixed model for the prediction of categorical spatial data', *International Journal of Geographical Information Science* **25**(12), 2071–2086.

Cao, G., Yoo, E.-h. & Wang, S. (2014), 'A statistical framework of data fusion for spatial prediction of categorical variables', *Stochastic Environmental Research and Risk Assessment* pp. 1–15.

Diggle, P., Tawn, J. & Moyeed, R. (1998), 'Model-based geostatistics', *Applied Statistics* **47**(3), 299–350.

Goovaerts, P. (1997), *Geostatistics for Natural Resources Evaluation*, Oxford University Press, New York.

Schölkopf, B. & Smola, A. (2002), *Learning with kernels: Support vector machines, regularization, optimization, and beyond*, the MIT Press.

Yuan, M. & Lin, Y. (2006), 'Model selection and estimation in regression with grouped variables', *Journal of the Royal Statistical Society: Series B* **68**, 49–67.