Statistica Sinica Preprint No: SS-2016-0230	
Title	Hierarchical Models for Spatial Data with Errors that are Correlated with the Latent Process
Manuscript ID	SS-2016-0230
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202016.0230
Complete List of Authors	Jonathan R. Bradley Christopher K. Wikle and Scott H. Holan
Corresponding Author	Jonathan R. Bradley
E-mail	bradley@stat.fsu.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0230

Hierarchical Models for Spatial Data with Errors that are Correlated with the Latent Process

Jonathan R. Bradley[1], Christopher K. Wikle[2], and Scott H. Holan[2,3]

[1]*Florida State University,* [2]*University of Missouri,* [3]*U.S. Census Bureau*

Abstract

Predictions of spatial processes using large data sets have become an important area of research. Current solutions often include placing strong assumptions on the error process associated with the data. Specifically, it is typically assumed that the data are equal to the spatial process of principal interest plus a mutually independent error process, which avoids modeling confounded cross-covariances between the signal and the noise within an additive model. In this study, we consider an alternative latent-process model, in which we assume that the error process is spatially correlated *and* correlated with the latent process of interest. We show that such error-process dependencies allow us to obtain precise predictions and prevent confounded error covariances within an expression of the marginal distribution of the data. We refer to these covariances as "nonconfounded discrepancy error covariances." In addition, we develop a "process augmentation" technique as a computation aid. The proposed method is demonstrated using simulated examples and an analysis of a large data set from the U.S. Census Bureau's American Community Survey.

Keywords: Bayesian, Low rank, Machine learning, Mixed effects model, Nonresponse, Parsimony.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0230

1 Introduction

We introduce a general class of additive spatial models, where it is assumed that the error process is spatially correlated *and* correlated with the latent process of interest. Adopting these discrepancy error covariances is important in a variety of applications. For example, in federal/official statistics, the assumption of an independent "survey error" is standard and is a key component of the ubiquitous Fay−Herriot model (Fay and Herriot, 1979). However, in many settings, it is well known that these errors are dependent. For example, disseminated estimates are sometimes modified/suppressed based on the value of the latent process owing to disclosure limitations, which may induce correlations between the survey error and the latent process (e.g., see Quick et al. (2015) for a review in a spatial setting). The error induced by nonresponse may be due to the value of the latent process, and consequently, we expect cross-correlations between the error process and the latent process (Groves et al., 2001). In addition, sampling designs are often motivated by the spatial structure of the latent process, which may result in correlations between the latent process and the survey error (Wikle and Royle, 2005; Holan and Wikle, 2012). Thus, we investigate whether such dependencies exist among American Community Survey (ACS) period estimates (e.g., see Torrieri, 2007).

Although it is reasonable to expect that such discrepancy error covariances exist, they are often ignored. For example, standard spatial statistics textbooks focus almost exclusively on the case of a spatially correlated error term that is independent of a mutually independent error term (Cressie, 1993; Cressie and Wikle, 2011; Banerjee et al., 2015). This is partially because there is a problem of confounding between the covariances of the signal and those of the noise (Cressie, 1993). Henceforth, we use the terms "signal" and "latent process" interchangeably. Confounding between *fixed effects* and *spatial random effects* has become an important research topic in the spatial statistics literature (e.g., Clayton et al., 1993;

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0230

Reich et al., 2006; Hodges and Reich, 2011). Recently, spatial basis functions (e.g., Moran's I basis functions) have been developed to account for confounding within a spatial setting (Griffith, 2000, 2002, 2004; Hughes and Haran, 2013; Bradley et al., 2015a), and are related to the classical Moran's I statistic (Moran, 1950). However, to the best of our knowledge, no studies have examined confounding between the covariances of the signal and those of the discrepancy error.

Thus, the goal of this study is to provide a way to leverage error-process dependencies in a manner that is computationally feasible and that accounts for confounding in the marginal covariance matrix associated with the data. We achieve this goal by introducing non-negligible discrepancy error covariances that are not present in the marginal distribution of the data. We refer to this class of measurement-error covariance matrices as *nonconfounded discrepancy-error covariances*. Thus, we provide a general form of nonconfounded discrepancy-error covariance matrices and provide several parameterizations that can be used in practice. In particular, we show that the standard uncorrelated discrepancy-error assumption is a special case of the nonconfounded discrepancy-error covariances. Then, we provide parameterizations that represent slight departures from the standard assumption of uncorrelated discrepancy errors. To aid researchers in assessing the appropriateness of these assumptions, we develop a covariance penalized error (Efron, 2004) as a measure of the out-of-sample error.

To date, there are no competing spatial methodologies that capitalize on the correlations between the error and the latent process in such a computationally efficient manner. These dependencies lead to improvements in the predictions of the latent process. However, the general approach of leveraging the dependence between an error process and a latent process has been exploited in settings other than those of spatial statistics. For example, the time-series literature documents a methodological approach referred to as "leverage effects," applied in stochastic volatility models (e.g., see Black, 1976, for an early reference),

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0230

where volatility is assumed to be correlated with the latent process. Similar relationships are employed in "feedback models" (Zeger and Liang, 1991).

To aid our computations we use a type of data-augmentation approach (e.g., see Tanner and Wong, 1987; Albert and Chib, 1993; Wakefield and Walker, 1999; Wolpert and Ickstadt, 1998, among others). In our implementation, we augment the process and not the data. Hence, we refer to this strategy as "process augmentation." The implementation of our process-augmentation approach involves two steps. First we fit any well-defined Bayesian spatial model (e.g., see Banerjee et al., 2008; Cressie and Johannesson, 2008; Lindgren et al., 2011; Datta et al., 2014; Nychka et al., 2015; Katzfuss, 2017). The second step involves a posterior predictive simulation. Thus, our proposed model can be viewed as a diagnostic tool; that is, after fitting a Bayesian statistical model, our method checks whether the predictions can be improved by incorporating discrepancy-error covariances. Additionally, this two-step procedure shows that the estimation of the covariances associated with the data and the regression parameters are unaffected by incorporating nonconfounded discrepancy-error covariances.

The remainder of the paper is organized as follows. In Section 2, we introduce the nonconfounded discrepancy-error covariances and describe several special cases. In addition, we provide the kriging predictor (Cressie, 1993) to aid in the interpretation of these covariances. Then, in Section 3 we describe how to address out-of-sample performance and robustness when we depart from the model assumptions. In Section 4, we describe the implementation using a process-augmentation approach. In Section 5, we use simulation studies to illustrate the high predictive performance of our method, and demonstrate the method using a large data set of ACS estimates defined on census tracts. Section 6 concludes the paper. For convenience of exposition, the proofs of the technical results, model selection, and model fitting are provided as Supplementary Material.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0230

2 Methodology

We start our exposition with the motivating difficulty of incorporating discrepancy error covariances, namely, the presence of confounded cross-covariances. Then, we introduce the nonconfounded discrepancy-error covariance matrix (Section 2.1). Next, we provide several special cases and properties of the matrix in Sections 2.2−2.5. To aid in interpreting the matrix, we discuss these special cases in the context of kriging in Section 2.6.

2.1 Nonconfounded Discrepancy-Error Covariances

Suppose we observe data at a finite number of locations, denoted by $\mathbf{s}_1, \ldots, \mathbf{s}_m \in D \subset \mathbb{R}^d$, where D represents the spatial domain of interest. The observed data are denoted by $\{Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_m)\}$ and, in general, are realized through a spatial process. Here, $Z(\mathbf{s})$ is defined at unobserved locations $\mathbf{s} \notin \{\mathbf{s}_1, \ldots, \mathbf{s}_m\}$. We are interested in predictions at the set of locations $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$, which is not necessarily equal to the set $\{\mathbf{s}_1, \ldots, \mathbf{s}_m\}$. We follow the hierarchical modeling approach commonly used in the spatial statistics literature (e.g., see standard textbooks, Cressie and Wikle, 2011; Banerjee et al., 2015) and assume $Z(\mathbf{s}) = Y(\mathbf{s}) + \delta(\mathbf{s})$; $\mathbf{s} \in D$, where $Y(\cdot)$ represents the latent process of interest. The data process represents a corrupted version of the latent process, where the corruption is represented additively with the error process $\delta(\cdot)$. The spatial Gaussian process modeling literature assumes that both $Y(\mathbf{s})$ and $\delta(\mathbf{s})$ are Gaussian for any $\mathbf{s} \in D$.

Let $\mu(\cdot)$ be the mean function for both $Y(\cdot)$ and $Z(\cdot)$, and let the vectors $\mathbf{z} = \{Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_m)\}^\top$, $\mathbf{y} = \{Y(\mathbf{u}_1), \ldots, Y(\mathbf{u}_n)\}^\top$, $\boldsymbol{\delta} \equiv \{\delta(\mathbf{u}_1), \ldots, \delta(\mathbf{u}_n)\}^\top$, and $\boldsymbol{\mu} \equiv \{\mu(\mathbf{s}_1), \ldots, \mu(\mathbf{s}_m)\}^\top$. In addition, define the $n \times n$ matrices $\boldsymbol{\Sigma}_Y = \text{cov}(\mathbf{y})$, $\boldsymbol{\Sigma}_\delta = \text{cov}(\boldsymbol{\delta})$, and

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0230

$\boldsymbol{\Sigma}_{Y,\delta} = \text{cov}(\mathbf{y}, \boldsymbol{\delta})$. Then, the probability density function of $\mathbf{z}$ is

$$\begin{aligned} &f(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}_Y = \mathbf{C}_Y, \boldsymbol{\Sigma}_\delta = \mathbf{C}_\delta, \boldsymbol{\Sigma}_{Y,\delta} = \mathbf{C}_{Y,\delta}) \\ &\propto |\mathbf{C}_Y + \mathbf{C}_\delta + 2\mathbf{C}_{Y,\delta}|^{-1/2} \text{exp}\left\{-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu})^\top (\mathbf{C}_Y + \mathbf{C}_\delta + 2\mathbf{C}_{Y,\delta})^{-1} (\mathbf{z}-\boldsymbol{\mu})\right\}, \end{aligned} \tag{1}$$

where $\mathbf{C}_Y$, $\mathbf{C}_\delta$, and $\mathbf{C}_{Y,\delta}$ are distinct values in the parameter spaces of $\boldsymbol{\Sigma}_Y$, $\boldsymbol{\Sigma}_\delta$, and $\boldsymbol{\Sigma}_{Y,\delta}$, respectively. Note that we set $\{\mathbf{s}_1, \ldots, \mathbf{s}_m\} = \{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$. Although this is not always true, the method generalizes easily to this case. The confounding problem is immediately apparent in (1). For example, the commutative property of the matrices yields,

$$\begin{aligned} f(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}_Y = \mathbf{C}_Y, \boldsymbol{\Sigma}_\delta = \mathbf{C}_\delta, \boldsymbol{\Sigma}_{Y,\delta} = \mathbf{C}_{Y,\delta}) &= f(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}_Y = \mathbf{C}_\delta, \boldsymbol{\Sigma}_\delta = \mathbf{C}_Y, \boldsymbol{\Sigma}_{Y,\delta} = \mathbf{C}_{Y,\delta}) \\ &= f(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}_Y = \mathbf{C}_Y, \boldsymbol{\Sigma}_\delta = 2\mathbf{C}_{Y,\delta}, \boldsymbol{\Sigma}_{Y,\delta} = \frac{1}{2}\mathbf{C}_\delta). \end{aligned}$$

To mitigate these confounding issues, it is often assumed that $\boldsymbol{\Sigma}_{Y,\delta}$ is an $n \times n$ matrix of zeros (denoted by $\mathbf{0}_{n,n}$) and $\boldsymbol{\Sigma}_\delta = \sigma^2\mathbf{I}_n$, where $\sigma^2 > 0$ and $\mathbf{I}_n$ is an $n \times n$ identity matrix. (e.g., see Banerjee et al., 2008; Cressie and Johannesson, 2008; Finley et al., 2009; Lindgren et al., 2011; Sang and Huang, 2012; Nychka et al., 2015, among others). This assumption gives

$$\begin{aligned} &f(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}_Y = \boldsymbol{\Sigma}_w, \boldsymbol{\Sigma}_\delta = \sigma^2\mathbf{I}_n, \boldsymbol{\Sigma}_{Y,\delta} = \mathbf{0}_{n,n}) \\ &\propto |\boldsymbol{\Sigma}_w + \sigma^2\mathbf{I}_n|^{-1/2} \text{exp}\left\{-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu})^\top \left(\boldsymbol{\Sigma}_w + \sigma^2\mathbf{I}_n\right)^{-1} (\mathbf{z}-\boldsymbol{\mu})\right\}, \end{aligned} \tag{2}$$

where $\boldsymbol{\Sigma}_w$ is a generic positive semi-definite matrix in the parameter space of $\boldsymbol{\Sigma}_Y$. Here, confounded cross-covariances are not present in the likelihood for $\mathbf{z}$. However, we are no longer able to capitalize on the dependence between $Y(\cdot)$ and $\delta(\cdot)$ and the covariances between $\delta(\cdot)$ at different locations. Therefore, we introduce a nonzero structure to the cross-covariance

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0230

parameter $\mathbf{\Sigma}_{Y,\delta}$ *and* introduce a compatible (possibly nondiagonal) $\mathbf{\Sigma}_\delta$ to obtain the likelihood in (2). This leads to what we call the "General Assumption."

General Assumption: Let $\mathbf{\Sigma}_\delta = \mathbf{\Sigma}_Y + \mathbf{\Sigma}_w - 2\mathbf{\Sigma}_{Y,w} + \sigma^2\mathbf{I}_n$ and $\mathbf{\Sigma}_{Y,\delta} = \mathbf{\Sigma}_{Y,w} - \mathbf{\Sigma}_Y$, where $\mathbf{\Sigma}_w$ is a positive semi-definite matrix, $\mathbf{\Sigma}_{Y,w}$ is an $n \times n$ real matrix, and

$$\text{cov}\left\{ \begin{pmatrix} \mathbf{y} \\ \boldsymbol{\delta} \end{pmatrix} | \mathbf{\Sigma}_Y, \mathbf{\Sigma}_w, \mathbf{\Sigma}_{Y,w}, \sigma^2 \right\} = \begin{pmatrix} \mathbf{\Sigma}_Y & \mathbf{\Sigma}_{Y,w} - \mathbf{\Sigma}_Y \\ \mathbf{\Sigma}_{Y,w}^\top - \mathbf{\Sigma}_Y^\top & \mathbf{\Sigma}_Y + \mathbf{\Sigma}_w - 2\mathbf{\Sigma}_{Y,w} + \sigma^2\mathbf{I}_n \end{pmatrix} \tag{3}$$

is positive semi-definite.

Now, substituting the General Assumption into (1), we obtain

$$\begin{aligned} & f(\mathbf{z}|\boldsymbol{\mu}, \mathbf{\Sigma}_Y, \mathbf{\Sigma}_\delta = \mathbf{\Sigma}_Y + \mathbf{\Sigma}_w - 2\mathbf{\Sigma}_{Y,w} + \sigma^2\mathbf{I}_n, \mathbf{\Sigma}_{Y,\delta} = \mathbf{\Sigma}_{Y,w} - \mathbf{\Sigma}_Y) \\ & \propto |\mathbf{\Sigma}_w + \sigma^2\mathbf{I}_n|^{-1/2} \text{exp}\left\{ -\frac{1}{2}(\mathbf{z} - \boldsymbol{\mu})^\top \left(\mathbf{\Sigma}_w + \sigma^2\mathbf{I}_n\right)^{-1} (\mathbf{z} - \boldsymbol{\mu}) \right\}, \end{aligned} \tag{4}$$

because,

$$\begin{aligned} & \text{cov}(\mathbf{z}|\mathbf{\Sigma}_Y, \mathbf{\Sigma}_w, \mathbf{\Sigma}_{Y,w}, \sigma^2) \\ & = \text{cov}(\mathbf{y}|\mathbf{\Sigma}_Y, \mathbf{\Sigma}_w, \mathbf{\Sigma}_{Y,w}, \sigma^2) + \text{cov}(\boldsymbol{\delta}|\mathbf{\Sigma}_Y, \mathbf{\Sigma}_w, \mathbf{\Sigma}_{Y,w}, \sigma^2) + 2\text{cov}(\mathbf{y}, \boldsymbol{\delta}|\mathbf{\Sigma}_Y, \mathbf{\Sigma}_w, \mathbf{\Sigma}_{Y,w}, \sigma^2) \\ & = \mathbf{\Sigma}_Y + \mathbf{\Sigma}_Y + \mathbf{\Sigma}_w - 2\mathbf{\Sigma}_{Y,w} + \sigma^2\mathbf{I} + 2\mathbf{\Sigma}_{Y,w} - 2\mathbf{\Sigma}_Y \\ & = \mathbf{\Sigma}_w + \sigma^2\mathbf{I}_n. \end{aligned}$$

The likelihood in (4) is of the same form as the likelihood in Equation (2), which did not have confounded cross-covariances. The difference in our approach is that we *also* have

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0230

correlations between $Y(\cdot)$ and $\delta(\cdot)$. That is,

$$\text{cov}\left(\mathbf{y}, \boldsymbol{\delta}\right) = \boldsymbol{\Sigma}_{Y,w} - \boldsymbol{\Sigma}_{Y}.$$

Thus, the General Assumption gives us a Gaussian likelihood in (2) with nonconfounded cross-covariances, while simultaneously allowing for cross-correlations between the latent process and the error process.

A special case of the General Assumption is the more conventionally used uncorrelated signal and noise, with no cross-spatial dependence between the discrepancy errors. In particular, let $\boldsymbol{\Sigma}_w = \boldsymbol{\Sigma}_{Y,w} = \boldsymbol{\Sigma}_Y$, such that,

$$\text{cov}\left\{\begin{pmatrix} \mathbf{y} \\ \boldsymbol{\delta} \end{pmatrix} | \boldsymbol{\Sigma}_Y, \boldsymbol{\Sigma}_w, \boldsymbol{\Sigma}_{Y,w}, \sigma^2 \right\} = \begin{pmatrix} \boldsymbol{\Sigma}_Y & \mathbf{0}_{n,n} \\ \mathbf{0}_{n,n} & \sigma^2\mathbf{I}_n \end{pmatrix}, \tag{5}$$

which is positive definite. Henceforth, we refer to $\boldsymbol{\Sigma}_w = \boldsymbol{\Sigma}_{Y,w} = \boldsymbol{\Sigma}_Y$ as the "Standard Assumption." This helps to interpret the new matrix-valued parameters $\boldsymbol{\Sigma}_w$ and $\boldsymbol{\Sigma}_{Y,w}$. That is, as the difference between $\boldsymbol{\Sigma}_{Y,w}$ and $\boldsymbol{\Sigma}_Y$ increases, we obtain a larger cross dependence between the signal $Y(\cdot)$ and the noise $\delta(\cdot)$. The moment properties of this model also help to interpret each matrix-valued parameter.

Because we are primarily interested in developing nonconfounded additive error covariances, for illustration purposes, we use standard "off-the-shelf" covariance functions to define $\boldsymbol{\Sigma}_Y$ and $\boldsymbol{\Sigma}_w$. In particular, when $\mathbf{s}_1, \ldots, \mathbf{s}_n$ are continuous, we use the Matérn covariogram (Matérn, 1960). When $\mathbf{s}_1, \ldots, \mathbf{s}_n$ define a lattice, we use a conditional autoregressive (CAR) model (Besag, 1974). In general, our framework can be implemented using any well-defined covariance function.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0230

2.2 Moment Properties of the General Assumption

This section discusses basic moment results, which lead to illuminating interpretations of $\mathbf{\Sigma}_w$ and $\mathbf{\Sigma}_{Y,w}$ defined in the General Assumption.

Proposition 1 (Moment Properties): Let the data vector $\boldsymbol{z}$ have a probability density function defined as in (1). Suppose the General Assumption from Section 2.1 holds. Then, we have the following moment results:

a. $\text{cov}(\boldsymbol{z}|\mathbf{\Sigma}_w, \sigma^2) = \mathbf{\Sigma}_w + \sigma^2 \boldsymbol{I}_n$;

b. $E(\boldsymbol{z}|\mathbf{\Sigma}_w, \sigma^2) = \boldsymbol{\mu}$;

c. $\text{cov}(\boldsymbol{z}, \boldsymbol{y}|\mathbf{\Sigma}_w, \sigma^2) = \mathbf{\Sigma}_{Y,w}^\top$;

d. $\text{cov}(\boldsymbol{z}|\mathbf{y}, \mathbf{\Sigma}_w, \sigma^2) = \mathbf{\Sigma}_w + \sigma^2 \boldsymbol{I}_n - \mathbf{\Sigma}_{Y,w}^\top \mathbf{\Sigma}_Y^{-1} \mathbf{\Sigma}_{Y,w}$;

e. $E(\boldsymbol{z}|\mathbf{y}, \mathbf{\Sigma}_w, \sigma^2) = \boldsymbol{\mu} + \mathbf{\Sigma}_{Y,w}^\top \mathbf{\Sigma}_Y^{-1}(\mathbf{y} - \boldsymbol{\mu})$.

Proof: The proof follows immediately from the General Assumption and the rules for the conditional and marginal distributions of Gaussian random vectors (Ravishanker and Dey, 2002).

Proposition 1(a,b) are the motivating features discussed in Section 2.1 (i.e., the marginal density of $\mathbf{z}$ does not contain confounded cross-correlations). However, they also show that the off-diagonals of $\mathbf{\Sigma}_w$ represent the cross-spatial correlations of the data, and that σ^2 represents the extra variability not accounted for in $\mathbf{\Sigma}_w$. These play a similar role to the nugget in classical spatial statistics (see Cressie, 1993; Cressie and Wikle, 2011;and Banerjee et al., 2015, for standard references). In addition, Proposition 1(c) shows that $\mathbf{\Sigma}_{Y,w}^\top$ represents the cross-covariance between $\mathbf{z}$ and $\mathbf{y}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0230

Proposition 1(d,e) imply that the data are *not* conditionally unbiased and are *not* conditionally uncorrelated, given the latent process $Y(\cdot)$. In the ACS example (Section 5.3), $Z(\cdot)$ represents the *disseminated* (log transform) median income of individuals in a particular census tract, whereas $Y(\cdot)$ represents the actual (log transform) median income of individuals in the census tract. The difference between $Y(\cdot)$ and $Z(\cdot)$ may be due to a combination of the sampling design, a nonresponse bias, modifications due to disclosure avoidance concerns, and many other sources of error. Thus, it may be reasonable to assume that the disseminated ACS data contain some bias and/or unaccounted for covariability between the survey errors.

2.3 Special Case 1

In this section, we consider a slight departure from the Standard Assumption that $\boldsymbol{\Sigma}_w = \boldsymbol{\Sigma}_{Y,w} = \boldsymbol{\Sigma}_Y$. Specifically, we assume $\boldsymbol{\Sigma}_w = \boldsymbol{\Sigma}_{Y,w} \neq \boldsymbol{\Sigma}_Y$. In this case, the expression in (3) is given by

$$\text{cov}\left\{\begin{pmatrix}\mathbf{y}\\ \boldsymbol{\delta}\end{pmatrix}\right\} = \begin{pmatrix}\boldsymbol{\Sigma}_Y & \boldsymbol{\Sigma}_{Y,w} - \boldsymbol{\Sigma}_Y \\ \boldsymbol{\Sigma}_{Y,w}^\top - \boldsymbol{\Sigma}_Y^\top & \boldsymbol{\Sigma}_Y + \boldsymbol{\Sigma}_w - 2\boldsymbol{\Sigma}_{Y,w} + \sigma^2\mathbf{I}_n\end{pmatrix} = \begin{pmatrix}\boldsymbol{\Sigma}_Y & -\boldsymbol{\Sigma}_1 \\ -\boldsymbol{\Sigma}_1 & \boldsymbol{\Sigma}_1 + \sigma^2\mathbf{I}_n\end{pmatrix}, \tag{6}$$

where $\boldsymbol{\Sigma}_1 \equiv \boldsymbol{\Sigma}_Y - \boldsymbol{\Sigma}_w$. In Appendix A of the Supplementary Material, we show that (6) is positive definite when $\boldsymbol{\Sigma}_w$ and $\boldsymbol{\Sigma}_1$ are positive definite. We refer to (6) as "Special Case 1." Here, we see an indirect relationship between the signal-to-noise cross-covariance and the marginal covariance of the discrepancy error. In addition, Special Case 1 leads to conditionally unbiased data (substitute $\boldsymbol{\Sigma}_w = \boldsymbol{\Sigma}_{Y,w} \neq \boldsymbol{\Sigma}_Y$ into Proposition 1(e)). Some might consider Special Case 1 to be more realistic in an official statistics setting. As discussed in the Introduction, federal agencies expend much effort to produce highly accurate estimates; however, there is no guarantee that discrepancy error correlations are not present in these estimates.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0230

2.4 Special Case 2

Consider the assumption that $\boldsymbol{\Sigma}_w \neq \boldsymbol{\Sigma}_{Y,w} = \boldsymbol{\Sigma}_Y$. In this case, the expression in (3) is given by

$$\text{cov}\left\{\begin{pmatrix} \mathbf{y} \\ \boldsymbol{\delta} \end{pmatrix}\right\} = \begin{pmatrix} \boldsymbol{\Sigma}_Y & \boldsymbol{\Sigma}_{Y,w} - \boldsymbol{\Sigma}_Y \\ \boldsymbol{\Sigma}_{Y,w}^\top - \boldsymbol{\Sigma}_Y^\top & \boldsymbol{\Sigma}_Y + \boldsymbol{\Sigma}_w - 2\boldsymbol{\Sigma}_{Y,w} + \sigma^2\mathbf{I}_n \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_Y & \mathbf{0}_{n,n} \\ \mathbf{0}_{n,n} & \boldsymbol{\Sigma}_2 + \sigma^2\mathbf{I}_n \end{pmatrix}, \tag{7}$$

where $\boldsymbol{\Sigma}_2 \equiv \boldsymbol{\Sigma}_w - \boldsymbol{\Sigma}_Y$. It follows immediately that (7) is positive definite, provided that $\boldsymbol{\Sigma}_2$ and $\boldsymbol{\Sigma}_Y$ are positive definite. We refer to the assumption that $\boldsymbol{\Sigma}_w \neq \boldsymbol{\Sigma}_{Y,w} = \boldsymbol{\Sigma}_Y$ as "Special Case 2." In (7), we see that covariances are present within the discrepancy errors, but that cross-covariances between the signal and the noise are not present. In addition, Proposition 1(e) shows that Special Case 2 implies that the data are conditionally unbiased for the latent process.

In general, any valid covariance function can be used to represent $\boldsymbol{\Sigma}_2$. For example, when $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ consists of point-referenced locations, we define the (i,j)-th element of $\boldsymbol{\Sigma}_2$ as being formed by the Matérn covariogram with a correlation parameter $\tau > 0$ and a variance parameter $\sigma_Y^2 > 0$ (Matérn, 1960). When $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ consists of areal locations, we can define $\boldsymbol{\Sigma}_2$ as the covariance from a CAR model with a correlation parameter $\tau > 0$ and a variance parameter $\sigma_Y^2 > 0$ (Besag, 1974). See Appendix B of the Supplementary Material for more details.

Special Case 2 is equivalent to the Standard Assumption after a transformation. Specifically, let $\mathbf{z}^* = (\frac{1}{\sigma^2}\boldsymbol{\Sigma}_2 + \mathbf{I}_n)^{-1}\mathbf{z}$. In a similar manner, define $\mathbf{y}^* = (\frac{1}{\sigma^2}\boldsymbol{\Sigma}_2 + \mathbf{I}_n)^{-1/2}\mathbf{y}$ and $\boldsymbol{\delta}^* = (\frac{1}{\sigma^2}\boldsymbol{\Sigma}_2 + \mathbf{I}_n)^{-1/2}\boldsymbol{\delta}$. Then, it follows that

$$\text{cov}\left\{\begin{pmatrix} \mathbf{y}^* \\ \boldsymbol{\delta}^* \end{pmatrix}\right\} = \begin{pmatrix} (\frac{1}{\sigma^2}\boldsymbol{\Sigma}_2 + \mathbf{I}_n)^{-1/2}\boldsymbol{\Sigma}_Y(\frac{1}{\sigma^2}\boldsymbol{\Sigma}_2 + \mathbf{I}_n)^{-1/2} & \mathbf{0}_{n,n} \\ \mathbf{0}_{n,n} & \sigma^2\mathbf{I}_n \end{pmatrix},$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0230

which has the same form as (5). Consequently, in Section 5 we focus less on Special Case 2 than we do on other choices because implementing Special Case 2 is identical (after a transformation) to implementing a model using the Standard Assumption.

2.5 Special Case 3

In this section, we consider a slight departure from the Standard Assumption that $\boldsymbol{\Sigma}_w = \boldsymbol{\Sigma}_{Y,w} = \boldsymbol{\Sigma}_Y$. Specifically, we assume $\boldsymbol{\Sigma}_{Y,w} \neq \boldsymbol{\Sigma}_w = \boldsymbol{\Sigma}_Y$. In this case, the expression in (3) is given by

$$\text{cov}\left\{\begin{pmatrix}\mathbf{y}\\ \boldsymbol{\delta}\end{pmatrix}\right\} = \begin{pmatrix}\boldsymbol{\Sigma}_Y & \boldsymbol{\Sigma}_{Y,w} - \boldsymbol{\Sigma}_Y \\ \boldsymbol{\Sigma}_{Y,w}^\top - \boldsymbol{\Sigma}_Y^\top & \sigma^2\mathbf{I}_n - 2(\boldsymbol{\Sigma}_{Y,w} - \boldsymbol{\Sigma}_Y)\end{pmatrix} = \begin{pmatrix}\boldsymbol{\Sigma}_Y & \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma}^\top & \sigma^2\mathbf{I}_n - 2\boldsymbol{\Sigma}\end{pmatrix}, \quad (8)$$

where $\boldsymbol{\Sigma} \equiv \boldsymbol{\Sigma}_{Y,w} - \boldsymbol{\Sigma}_Y$. In Appendix A of the Supplementary Material, we show that (8) is positive semi-definite, provided that $\boldsymbol{\Sigma}_Y$ and $\boldsymbol{\Sigma}_Y - \boldsymbol{\Sigma}_{Y,w}^\top \boldsymbol{\Sigma}_Y^{-1} \boldsymbol{\Sigma}_{Y,w}$ are both positive semi-definite. We refer to the assumption that $\boldsymbol{\Sigma}_{Y,w} \neq \boldsymbol{\Sigma}_w = \boldsymbol{\Sigma}_Y$ as "Special Case 3."

2.6 Special Case 4

Assume that $\boldsymbol{\Sigma}_w \approx \boldsymbol{\Sigma}_Y$ and $\boldsymbol{\Sigma}_{Y,w} \approx \boldsymbol{\Sigma}_Y$, where $\boldsymbol{\Sigma}_w$ and $\boldsymbol{\Sigma}_{Y,w}$ are defined as projections onto a reduced dimensional space. Specifically, let $\boldsymbol{\Psi} \in \mathbb{R}^n \times \mathrm{R}^r$ be an $n \times r$ $(r \leq n)$ matrix consisting of spatial basis functions evaluated at $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$. Several examples of $\boldsymbol{\Psi}$ are provided in the Appendix C of the Supplementary Material. Then, we assume that $\boldsymbol{\Sigma}_w$ is the left-and-right-projection of $\boldsymbol{\Sigma}_Y$ onto the column space of $\boldsymbol{\Psi}$, and that $\boldsymbol{\Sigma}_{Y,w}$ is the right

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0230

projection of $\boldsymbol{\Sigma}_Y$ onto the column space of $\boldsymbol{\Psi}$. That is,

$$\boldsymbol{\Sigma}_w = \boldsymbol{\Psi}\left(\underset{\mathbf{K}\in\mathbb{R}^r\times\mathbb{R}^r}{\arg\min}||\boldsymbol{\Psi}\mathbf{K}\boldsymbol{\Psi}^\top - \boldsymbol{\Sigma}_Y||_F^2\right)\boldsymbol{\Psi}^\top = \boldsymbol{\Psi}(\boldsymbol{\Psi}^\top\boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}^\top\boldsymbol{\Sigma}_Y\boldsymbol{\Psi}(\boldsymbol{\Psi}^\top\boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}^\top \tag{9}$$

$$\boldsymbol{\Sigma}_{Y,w} = \underset{\mathbf{C}\in\mathbb{R}^n\times\mathbb{R}^r}{\arg\min}||\mathbf{C}\boldsymbol{\Psi}^\top - \boldsymbol{\Sigma}_Y||_F^2 = \boldsymbol{\Sigma}_Y\boldsymbol{\Psi}(\boldsymbol{\Psi}^\top\boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}^\top, \tag{10}$$

respectively. The operator $||\cdot||_F^2$ is known as the Frobenius norm, and for any square real-valued matrix $\mathbf{M}$, we have that $||\mathbf{M}||_F^2 = \text{trace}(\mathbf{M}^\top\mathbf{M})$. The expressions on the far right-hand side of (9) and (10) are an immediate consequence of a result in Cressie and Johannesson (2008). For notational convenience, denote the hat matrix $\mathbf{P} = \boldsymbol{\Psi}(\boldsymbol{\Psi}^\top\boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}^\top$.

In this case, the expression in (3) is given by

$$\text{cov}\left\{\begin{pmatrix}\mathbf{y}\\ \boldsymbol{\delta}\end{pmatrix}\right\} = \begin{pmatrix}\boldsymbol{\Sigma}_Y & -\boldsymbol{\Sigma}_Y(\mathbf{I}_n - \mathbf{P})\\ -(\mathbf{I}_n - \mathbf{P})\boldsymbol{\Sigma}_Y & \boldsymbol{\Sigma}_Y(\mathbf{I}_n - \mathbf{P}) - (\mathbf{I}_n - \mathbf{P})\boldsymbol{\Sigma}_Y\mathbf{P} + \sigma^2\mathbf{I}_n\end{pmatrix}. \tag{11}$$

In Appendix A of the Supplementary Material, we show that (11) is positive semi-definite, provided that $\boldsymbol{\Sigma}_Y$ is positive semi-definite. The cross-covariance term between the signal and the noise is determined by the covariance of $\mathbf{y}$ and the basis functions evaluated at all locations in $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$. Thus, it is especially important that the basis functions are chosen carefully. In Appendix C of the Supplementary Material, we describe an algorithm for selecting the basis functions using an out-of-sample measure of error (see Section 3).

There is an interesting relationship between (11) and the Standard Assumption. That is, if we set $\boldsymbol{\Sigma}_Y = \mathbf{PHP}$ for some $n \times n$ positive definite matrix $\mathbf{H}$, we obtain,

$$\text{cov}\left\{\begin{pmatrix}\mathbf{y}\\ \boldsymbol{\delta}\end{pmatrix}\right\} = \begin{pmatrix}\mathbf{PHP} & \mathbf{0}_{n,n}\\ \mathbf{0}_{n,n} & \sigma^2\mathbf{I}_n\end{pmatrix}.$$

Thus, to capture discrepancy-error dependencies, we require that the columns of $\boldsymbol{\Sigma}_Y$ fall

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0230

within the orthogonal column space of $\mathbf{P}$. In our empirical results, we check whether all eigenvectors of $\boldsymbol{\Sigma}_Y$ are in the orthogonal complement space of $\mathbf{P}$.

2.7 The Kriging Predictor With Nonconfounded discrepancy-error Covariances

The traditional kriging predictor (e.g., see Matheron, 1963; Cressie, 1990, among others) is a standard optimal predictor (in terms of minimizing the mean squared prediction error (MSPE)) in spatial statistics, and should be discussed under the General Assumption. In Appendix A of the Supplementary Material, we show that the kriging predictor, assuming $\boldsymbol{\Sigma}_w \neq \boldsymbol{\Sigma}_{Y,w} \neq \boldsymbol{\Sigma}_Y$, is given by

$$E(\mathbf{y}|\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Sigma}_w, \boldsymbol{\Sigma}_{Y,w}, \boldsymbol{\Sigma}_Y, \sigma^2) = \boldsymbol{\mu} + \boldsymbol{\Sigma}_{Y,w}^\top \left(\boldsymbol{\Sigma}_w + \sigma^2 \mathbf{I}_n\right)^{-1} (\mathbf{z} - \boldsymbol{\mu}). \tag{12}$$

This kriging predictor has the following covariance (see Appendix A of the Supplementary Material):

$$\text{cov}(\mathbf{y}|\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Sigma}_w, \boldsymbol{\Sigma}_{Y,w}, \boldsymbol{\Sigma}_Y, \sigma^2) = \boldsymbol{\Sigma}_Y - \boldsymbol{\Sigma}_{Y,w}^\top \left(\boldsymbol{\Sigma}_w + \sigma^2 \mathbf{I}_n\right)^{-1} \boldsymbol{\Sigma}_{Y,w}. \tag{13}$$

The special cases discussed in Sections 2.1−2.5 lead to illuminating special cases of the kriging predictor.

Proposition 2 (Kriging Predictors): Let the data vector $\boldsymbol{z}$ have a probability density function defined as in (1). Suppose the General Assumption from Section 2.1 holds. Then, we have the following expressions for the kriging predictor and kriging covariances:

a. *Standard Assumption ($\boldsymbol{\Sigma}_w = \boldsymbol{\Sigma}_{Y,w} = \boldsymbol{\Sigma}_Y$):*

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0230

$$E(\mathbf{y}|\mathbf{z}) = \boldsymbol{\mu} + \boldsymbol{\Sigma}_w \left(\boldsymbol{\Sigma}_w + \sigma^2 \mathbf{I}_n\right)^{-1} (\mathbf{z} - \boldsymbol{\mu});$$

$$\text{cov}(\mathbf{y}|\mathbf{z}) = \boldsymbol{\Sigma}_w - \boldsymbol{\Sigma}_w \left(\boldsymbol{\Sigma}_w + \sigma^2 \mathbf{I}_n\right)^{-1} \boldsymbol{\Sigma}_w;$$

b. *Special Case 1 ($\boldsymbol{\Sigma}_w = \boldsymbol{\Sigma}_{Y,w} \neq \boldsymbol{\Sigma}_Y$):*

$$E(\mathbf{y}|\mathbf{z}) = \boldsymbol{\mu} + \boldsymbol{\Sigma}_w \left(\boldsymbol{\Sigma}_w + \sigma^2 \mathbf{I}_n\right)^{-1} (\mathbf{z} - \boldsymbol{\mu});$$

$$\text{cov}(\mathbf{y}|\mathbf{z}) = \boldsymbol{\Sigma}_Y - \boldsymbol{\Sigma}_w \left(\boldsymbol{\Sigma}_w + \sigma^2 \mathbf{I}_n\right)^{-1} \boldsymbol{\Sigma}_w;$$

c. *Special Case 2 ($\boldsymbol{\Sigma}_w \neq \boldsymbol{\Sigma}_{Y,w} = \boldsymbol{\Sigma}_Y$):*

$$E(\mathbf{y}|\mathbf{z}) = \boldsymbol{\mu} + \boldsymbol{\Sigma}_Y \left(\boldsymbol{\Sigma}_w + \sigma^2 \mathbf{I}_n\right)^{-1} (\mathbf{z} - \boldsymbol{\mu});$$

$$\text{cov}(\mathbf{y}|\mathbf{z}) = \boldsymbol{\Sigma}_Y - \boldsymbol{\Sigma}_Y \left(\boldsymbol{\Sigma}_w + \sigma^2 \mathbf{I}_n\right)^{-1} \boldsymbol{\Sigma}_Y;$$

d. *Special Case 3 ($\boldsymbol{\Sigma}_{Y,w} \neq \boldsymbol{\Sigma}_w = \boldsymbol{\Sigma}_Y$):*

$$E(\mathbf{y}|\mathbf{z}) = \boldsymbol{\mu} + \boldsymbol{\Sigma}_{Y,w}^\top \left(\boldsymbol{\Sigma}_w + \sigma^2 \mathbf{I}_n\right)^{-1} (\mathbf{z} - \boldsymbol{\mu});$$

$$\text{cov}(\mathbf{y}|\mathbf{z}) = \boldsymbol{\Sigma}_w - \boldsymbol{\Sigma}_{Y,w}^\top \left(\boldsymbol{\Sigma}_w + \sigma^2 \mathbf{I}_n\right)^{-1} \boldsymbol{\Sigma}_{Y,w};$$

e. *Special Case 4 ($\boldsymbol{\Sigma}_w \approx \boldsymbol{\Sigma}_{Y,w} \approx \boldsymbol{\Sigma}_Y$):*

$$E(\mathbf{y}|\mathbf{z}) = \boldsymbol{\mu} + \boldsymbol{\Sigma}_Y \mathbf{P} \left(\boldsymbol{\Sigma}_w + \sigma^2 \mathbf{I}_n\right)^{-1} (\mathbf{z} - \boldsymbol{\mu});$$

$$\text{cov}(\mathbf{y}|\mathbf{z}) = \boldsymbol{\Sigma}_Y - \boldsymbol{\Sigma}_Y \mathbf{P} \left(\boldsymbol{\Sigma}_w + \sigma^2 \mathbf{I}_n\right)^{-1} \mathbf{P} \boldsymbol{\Sigma}_Y.$$

Proof: The proof follows immediately from the General Assumption and the rules for the conditional and marginal distributions of Gaussian random vectors (Ravishanker and Dey, 2002).

In Proposition 2 (a,b), the kriging predictors are the same. However, the kriging variances are larger in Special Case 1 than are those under the Standard Assumption. In Proposition 2 (c,d,e), both the kriging predictor and kriging covariance differ from the expressions developed under the Standard Assumption. In Proposition 2(d), if we set $\boldsymbol{\Sigma}_{Y,w} = \boldsymbol{\Sigma}_Y \mathbf{P}$ and $\boldsymbol{\Sigma}_w = \mathbf{P} \boldsymbol{\Sigma}_Y \mathbf{P}$, we obtain the same kriging predictor as in Special Case 4. Henceforth, we use

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0230

$\boldsymbol{\Sigma}_{Y,w} = \boldsymbol{\Sigma}_Y \mathbf{P}$ when applying Special Case 3.

3 An Empirical Measure of an Out-of-Sample Error

The strategies presented in Sections 2.3−2.5 are all based on making a small change to the Standard Assumption that $\boldsymbol{\Sigma}_w = \boldsymbol{\Sigma}_{Y,w} = \boldsymbol{\Sigma}_Y$. By providing flexible modeling assumptions, it is incumbent on us to provide an empirical measure to assess the appropriateness of these new assumptions in practice. We propose the following criterion:

$$\begin{aligned}
&\sum_{\mathbf{s}} \left\{ Z(\mathbf{s}) - \widehat{Y}(\mathbf{s}) \right\}^2 + 2 \sum_{\mathbf{s}} \text{cov} \left\{ Z(\mathbf{s}) - Y(\mathbf{s}) \right\} \widehat{Y}(\mathbf{s}) \\
&= \sum_{\mathbf{s}} \left\{ Y(\mathbf{s}) - \widehat{Y}(\mathbf{s}) \right\}^2 + \sum_{\mathbf{s}} \left\{ Z(\mathbf{s}) - Y(\mathbf{s}) \right\}^2 + 2 \sum \left\{ Z(\mathbf{s}) - Y(\mathbf{s}) \right\} Y(\mathbf{s}), \qquad (14)
\end{aligned}$$

where $\widehat{Y}(\mathbf{s})$ is a generic real-valued function of $\mathbf{z}$, which represents a prediction at $\mathbf{s} \in D$. Note that the left-hand side of Equation (14) is the so-called covariance penalized error introduced in Efron (1983), and is a measure of the out-of-sample error. This out-of-sample criterion is preferable to others because it is general enough to capture nonconfounded discrepancy-error covariances. In practice, we use the posterior expected value of the cross-product between $Z(\mathbf{s}) - Y(\mathbf{s})$ and $\widehat{Y}(\mathbf{s})$ to estimate the covariance term on the left-hand side of (14). See Appendix B of the Supplementary Material for more details on the implementation.

In Special Cases 1 and 3, we use the estimated covariance penalized error to select the parameters that define $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$. Recall that we consider Matérn and CAR model specifications of the matrices. Similarly, in Special Case 4, the choice of basis functions evaluated at pre-specified locations partially defines the discrepancy-error covariances. Consequently, we provide a stepwise algorithm that uses the covariance penalized error in (14) (see Appendix C of the Supplementary Material).

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0230

4 Bayesian Implementation: Process Augmentation

4.1 Process Augmentation

It is often useful to introduce an artificial latent random variable such that, upon marginalization, we obtain the original joint probability function (e.g., see Tanner and Wong, 1987; Albert and Chib, 1993; Wakefield and Walker, 1999; Wolpert and Ickstadt, 1998, among others). We extend this strategy to our setting by developing a similar "process-augmentation" approach. Specifically, let

$$Z(\mathbf{s}) = Y(\mathbf{s}) + \delta(\mathbf{s}) \tag{15}$$

$$\delta(\mathbf{s}) = w(\mathbf{s}) - Y(\mathbf{s}) + \epsilon(\mathbf{s});\ \ \mathbf{s} \in D, \tag{16}$$

where $w(\mathbf{s}) - Y(\mathbf{s}) \neq 0$ for at least one location $\mathbf{s} \in D$, $w(\cdot)$ is a Gaussian process with mean function $\mu(\cdot)$, and $\epsilon(\cdot)$ is a mutually independent error term with mean zero and variance σ^2.

Let $\mathbf{z}_n = \{Z(\mathbf{u}_1), \ldots, Z(\mathbf{u}_n)\}^\top$, $\mathbf{w} = (w(\mathbf{u}_1), \ldots, w(\mathbf{u}_n))^\top$, and $\boldsymbol{\epsilon} = (\epsilon(\mathbf{u}_1), \ldots, \epsilon(\mathbf{u}_n))^\top$, such that (15) becomes

$$\mathbf{z}_n = \mathbf{y} + \boldsymbol{\delta} \tag{17}$$

$$\boldsymbol{\delta} = \mathbf{w} - \mathbf{y} + \boldsymbol{\epsilon}. \tag{18}$$

Let $\text{cov}(\mathbf{y}, \mathbf{w}) \equiv \boldsymbol{\Sigma}_{Y,w}$ and $\text{cov}(\mathbf{w}) \equiv \boldsymbol{\Sigma}_w$. Then, from (17) and (18), it follows that

$$\boldsymbol{\Sigma}_{Y,\delta} = \boldsymbol{\Sigma}_{Y,w} - \boldsymbol{\Sigma}_Y$$

$$\boldsymbol{\Sigma}_\delta = \boldsymbol{\Sigma}_Y + \boldsymbol{\Sigma}_w - 2\boldsymbol{\Sigma}_{Y,w} + \sigma^2 \mathbf{I}_n,$$

yielding the General Assumption. Many of the special cases arise from Equations (17) and

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0230

(18). For example, Special Case 3 occurs when we define $\text{cov}(\mathbf{w}) = \text{cov}(\mathbf{y})$. The remaining special cases are organized into Proposition 3.

Proposition 3 (Process Augmentation, Special Cases): Assume the model described in (15) and (16); complete regularity conditions are provided in Appendix D of the Supplementary Material. Let $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\} \in D$ *be a generic collection of points in* D*. Suppose the General Assumption from Section 2.1 holds. Then, we have the following:*

a. *The Standard Assumption (*$\boldsymbol{\Sigma}_w = \boldsymbol{\Sigma}_{Y,w} = \boldsymbol{\Sigma}_Y$*) is obtained by assuming* $w(\cdot) \equiv Y(\cdot)$*.*

b. *Special Case 1 (*$\boldsymbol{\Sigma}_w = \boldsymbol{\Sigma}_{Y,w} \neq \boldsymbol{\Sigma}_Y$*) is obtained by assuming* $Y(\cdot) \equiv w(\cdot) + \epsilon_Y(\cdot)$*, where* $\epsilon_Y(\cdot)$ *is independent of* $w(\cdot)$ *and is a Gaussian process.*

c. *Special Case 2 (*$\boldsymbol{\Sigma}_w \neq \boldsymbol{\Sigma}_{Y,w} = \boldsymbol{\Sigma}_Y$*) is obtained by assuming* $w(\cdot) \equiv Y(\cdot) + \epsilon_w(\cdot)$*, where* $\epsilon_w(\cdot)$ *is independent of* $Y(\cdot)$ *and is a Gaussian process with mean zero.*

d. *Special Case 3 (*$\boldsymbol{\Sigma}_w = \boldsymbol{\Sigma}_Y \neq \boldsymbol{\Sigma}_{Y,w}$*) is obtained by assuming* $w(\cdot)$ *and* $Y(\cdot)$ *are identically distributed, where* $w(\cdot)$ *and* $Y(\cdot)$ *are dependent Gaussian processes.*

e. *Special Case 4 (*$\boldsymbol{\Sigma}_w \approx \boldsymbol{\Sigma}_{Y,w} \approx \boldsymbol{\Sigma}_Y$*) is obtained by assuming* $\boldsymbol{w} \equiv \boldsymbol{\mu} + \boldsymbol{\Psi}\boldsymbol{\eta} + \boldsymbol{\xi}$*, where* $\boldsymbol{\Psi}$ *is defined in Section 2.6, and* $\boldsymbol{\eta}$ *is a mean-zero Gaussian random vector such that* $\text{cov}(\boldsymbol{\eta}) = (\boldsymbol{\Psi}^\top\boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}^\top\boldsymbol{\Sigma}_Y\boldsymbol{\Psi}(\boldsymbol{\Psi}^\top\boldsymbol{\Psi})^{-1}$ *and* $\text{cov}(\boldsymbol{y}, \boldsymbol{\eta}) = \boldsymbol{\Sigma}_Y\boldsymbol{\Psi}(\boldsymbol{\Psi}^\top\boldsymbol{\Psi})^{-1}$*, and* $\boldsymbol{\xi}$ *is an independent* n*-dimensional Gaussian random vector.*

Proof: The proof follows immediately from the General Assumption and the rules for the conditional and marginal distributions of Gaussian random vectors (Ravishanker and Dey, 2002).

In our exposition, the process $w(\cdot)$ is interpreted as an artificial quantity, which aids in the implementation (see the discussion in Tanner and Wong, 1987; Albert and Chib, 1993;

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0230

Wakefield and Walker, 1999; Wolpert and Ickstadt, 1998, among others). However, several current methods interpret $w(\cdot)$ as an approximation of $Y(\cdot)$ (Banerjee et al., 2008; Cressie and Johannesson, 2008; Sang and Huang, 2012), but still paradoxically assume the Standard Assumption that $w(\cdot) \equiv Y(\cdot)$. We assert that it is more realistic to assume that $w(\cdot)$ is separate from $Y(\cdot)$ when $w(\cdot)$ represents an approximation of $Y(\cdot)$.

Consider Special Case 4, which, from Proposition 3, is equivalent to assuming $w(\mathbf{s}) = \sum_{i=1}^{r} \psi_i(\mathbf{s})\eta_i$, where $\{\eta_i\}$ are random effects and $\{\psi_i\}$ are real-valued spatial basis functions. Consider the Karhunen−Loéve representation of a spatial random process $Y(\cdot)$ (Karhunen, 1947; Loéve, 1978):

$$Y(\mathbf{s}) = \sum_{i=1}^{\infty} \psi_i(\mathbf{s})\alpha_i, \tag{19}$$

where $\{\psi_i(\cdot)\}$ are orthonormal and the random variables $\{\alpha_i\}$ are uncorrelated, with mean zero and variance $\{\lambda_i\}$. Model $w(\cdot)$ using the *truncated* Karhunen−Loéve expansion,

$$w(\mathbf{s}) = \sum_{i=1}^{r} \psi_i(\mathbf{s})\alpha_i,$$

where, for "large" r, $w(\cdot)$ approximates $Y(\cdot)$. Now, setting $w(\cdot) \equiv Y(\cdot)$ under the Standard Assumption is the same as claiming

$$\sum_{i=r+1}^{\infty} \psi_i(\mathbf{s})\alpha_i = 0, \tag{20}$$

$$Y(\mathbf{s}) = \sum_{i=1}^{r} \psi_i(\mathbf{s})\alpha_i \tag{21}$$

$$Z(\mathbf{s}) = \sum_{i=1}^{r} \psi_i(\mathbf{s})\alpha_i + \epsilon(\mathbf{s}), \tag{22}$$

for *every* $\mathbf{s} \in D$. Stein (2014) provides inferential problems with the KL-divergence measure when making the assumptions given in (20), (21), and (22). Some have tried to adjust for

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0230

this by using the Standard Assumption and an artificial model (i.e., tapered covariance, or white noise) for $\sum_{i=r+1}^{\infty} \psi_i(\mathbf{s})\alpha_i$, typically chosen for computational reasons (Finley et al., 2009; Sang and Huang, 2012). Instead of using an artificial model for $\sum_{i=r+1}^{\infty} \psi_i(\mathbf{s})\alpha_i$, we choose to model $Y(\cdot)$ in (19) directly through Special Case 4.

For this heuristic, Special Case 4 can be viewed as the following assumption:

$$Y(\mathbf{s}) = \sum_{i=1}^{\infty} \psi_i(\mathbf{s})\alpha_i$$
$$Z(\mathbf{s}) = \sum_{i=1}^{r} \psi_i(\mathbf{s})\alpha_i + \epsilon(\mathbf{s}).$$

Thus, if the Standard Assumption is correct, we can use our modeling approach to make correct assumptions for $Y(\cdot)$ and $\sum_{i=r+1}^{\infty} \psi_i(\mathbf{s})\alpha_i$, but use an approximation to model $Z(\cdot)$. In this case, the discrepancy-error variances are induced through a misspecified/approximated model for $Z(\cdot)$.

In Section 5, we compare the predictor under Special Case 4 with several methods that assume an approximated process $w(\cdot)$ that is equivalent to the exact process $Y(\cdot)$. Specifically, we compare it to the full-scale approximation (FSA), modified predictive processes (MPP), and Bayesian fixed-rank kriging (FRK) (Cressie and Johannesson, 2008; Finley et al., 2009; Sang and Huang, 2012). See Appendix E of the Supplementary Material for a review of these methods.

4.2 Posterior Predictive Distributions

The presence of $w(\cdot)$ can be used to obtain computationally efficient predictions. The following technical results demonstrate this for the Bayesian setting by showing a useful conditional independence property of the posterior predictive distribution of $\mathbf{y}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0230

Theorem 1: Let $\mathcal{S} \subset D \subset \mathbb{R}^d$ *be an open set. For each* $k \in \mathbb{N} = \{1, 2, 3, \ldots\}$ *and finite collection of locations* $\mathbf{s}_1, \ldots, \mathbf{s}_k \in \mathcal{S}$*, define the* n*-dimensional random vectors* $\boldsymbol{w} = \{w(\mathbf{s}_{k+1}), \ldots, w(\mathbf{s}_{k+n})\}^\top$*,* $\mathbf{z} = \{Z(\mathbf{s}_{k+1}), \ldots, Z(\mathbf{s}_{k+n})\}^\top$*, and* $\boldsymbol{\epsilon} = \{\epsilon(\mathbf{s}_{k+1}), \ldots, \epsilon(\mathbf{s}_{k+n})\}^\top$*, for* $\{\mathbf{s}_{k+1}, \ldots, \mathbf{s}_{k+n}\} \in D$*. Suppose all marginal and conditional densities are proper (see Regularity Conditions in Appendix D of the Supplementary Material). Furthermore, let* $f(Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_k)|\boldsymbol{w}, \boldsymbol{\theta})$ *be Kolmogorov consistent. Then, there exists a probability space (with sample space* Ω*, sigma-algebra* $\mathcal{A}$*, and probability measure* $\mathbb{P}$*) and stochastic process* $Y : \mathcal{S} \times \Omega \to \mathbb{R}$*, such that*

$$\begin{aligned}\mathbb{P}\left\{Y(\mathbf{s}_1) \in A_1, \ldots, Y(\mathbf{s}_k) \in A_k\right\} &= \int_{A_1} \ldots \int_{A_k} f(Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_k)|\boldsymbol{w}, \boldsymbol{\theta}, \mathbf{z}) dY(\mathbf{s}_1) \ldots dY(\mathbf{s}_k) \\ &= \int_{A_1} \ldots \int_{A_k} f(Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_k)|\boldsymbol{w}, \boldsymbol{\theta}) dY(\mathbf{s}_1) \ldots dY(\mathbf{s}_k),\end{aligned} \tag{23}$$

for all $\mathbf{s}_1, \ldots, \mathbf{s}_k \in \mathcal{S}$*,* $k \in \mathbb{N}$*, and measurable sets* $A_i \subset \mathbb{R}$*;* $i = 1, \ldots, k$*.*

Proof: See Appendix F of the Supplementary Material.

The conditional independence property in Equation (23) has been referred to as "Bayesianly unidentified" (e.g., see Banerjee et al., 2015, page 157). However, this differs from no Bayesian learning, which implies that $f(Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_k)|\mathbf{w}, \boldsymbol{\theta}, \mathbf{z}) = f(Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_k))$. This independence property is not present in our modeling framework. To clarify the model structure, we present a pictorial representation of the augmented joint probability density function in Figure 1.

In practice, we do not observe the entire spatial field D, nor do we predict over the (possibly) uncountably infinite spatial domain. Thus, the following corollary is important

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0230

for practical purposes.

Corollary 1: Define the n-dimensional random vector $\boldsymbol{y} = \{Y(\boldsymbol{u}_1), \ldots, Y(\boldsymbol{u}_n)\}^\top$ *and the m-dimensional random vectors* $\boldsymbol{w} = \{w(\mathbf{s}_1), \ldots, w(\mathbf{s}_m)\}^\top$, $\mathbf{z} = \{Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_m)\}^\top$, *and* $\boldsymbol{\epsilon} = \{\epsilon(\mathbf{s}_1), \ldots, \mathbf{s}_m\}^\top$, *for* $\{\mathbf{s}_1, \ldots, \mathbf{s}_m\} \in D$ *and* $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\} \subset \mathbb{R}^d \in D$. *Let* $(\Omega, \mathcal{A}, \mathbb{P})$ *denote the probability space for* $\boldsymbol{y}$, $\boldsymbol{w}$, $\mathbf{z}$, *and* $\boldsymbol{\epsilon}$. *Assume, for every* $\omega \in \Omega$ *and* $\boldsymbol{h} \in \mathbb{R}^m$, *the set* $\{\omega : \mathbf{z}(\omega) = \boldsymbol{h}\} = \{\omega : \boldsymbol{\epsilon}(\omega) = \boldsymbol{h} - \mathbf{y}_Z(\omega)\}$ *(i.e., the additive model holds, almost surely). In addition, assume that* $\mathbf{y}$ *and* $\boldsymbol{\epsilon}$ *are mutually independent. Let the probability density functions for* $\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{\theta}, \mathbf{z}$ *and* $\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{\theta}$ *exist, and denote them as* $f(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{\theta}, \mathbf{z})$ *and* $f(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{\theta})$, *respectively. Let* $\boldsymbol{\theta}$ *be a generic k-dimensional real-valued parameter vector. Then,* $f(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{\theta}, \mathbf{z}) = f(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{\theta})$.

Proof: The proof follows immediately from Theorem 1 by setting $k = n$.

In Corollary 1, the probability density function $f(\mathbf{y}|\mathbf{w}, \boldsymbol{\theta}, \mathbf{z})$ is the so-called predictive distribution for $\mathbf{y}$ (Berger, 1985). That is, if $\mathbf{w}^{[0]}$ and $\boldsymbol{\theta}^{[0]}$ are generated from their respective posterior distributions $f(\mathbf{w}, \boldsymbol{\theta}|\mathbf{z})$, then a random vector with probability density function $f(\mathbf{y}|\mathbf{z})$ can be simulated using $f(\mathbf{y}|\mathbf{w}^{[0]}, \boldsymbol{\theta}^{[0]}, \mathbf{z})$. This leads to a composition sampling approach for the model implementation, where we first simulate $\mathbf{w}^{[0]}$ and $\boldsymbol{\theta}^{[0]}$ from $f(\mathbf{w}, \boldsymbol{\theta}|\mathbf{z})$, and then simulate them from $f(\mathbf{y}|\mathbf{w}^{[0]}, \boldsymbol{\theta}^{[0]})$.

This implies that two steps are required to obtain posterior replicates of $\mathbf{y}$: (1) simulating from the posterior distribution of $\mathbf{w}$ and $\boldsymbol{\theta}$; and (2) simulating from $f(\mathbf{y}|\mathbf{w}, \boldsymbol{\theta})$.

4.3 Bayesian Inference of the Latent Process

For each special case, we assume $\mathbf{w} = \boldsymbol{\Psi}\boldsymbol{\eta}$, where $\boldsymbol{\eta}$ is a mean-zero Gaussian random vector such that $\text{cov}(\boldsymbol{\eta}|\boldsymbol{\theta}) = (\boldsymbol{\Psi}^\top\boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}^\top\mathbf{H}(\boldsymbol{\theta})\boldsymbol{\Psi}(\boldsymbol{\Psi}^\top\boldsymbol{\Psi})^{-1}$. For point-referenced data, the (i, j)-th

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0230

Model	$\boldsymbol{\Sigma}_w$	$\boldsymbol{\Sigma}_Y$	$\boldsymbol{\Sigma}_{Y,w}$	$\text{cov}(\mathbf{y}, \boldsymbol{\eta})$
Standard Assumption	Let $\boldsymbol{\Sigma}_w = \mathbf{PHP}$, where for point referenced data the (i,j)-th element of $\mathbf{H}$ is $C_M(\|\mathbf{s}_i - \mathbf{s}_j\|)$, and for areal data $\mathbf{H}$ is the covariance from a CAR model.	$\boldsymbol{\Sigma}_Y = \boldsymbol{\Sigma}_w$	$\boldsymbol{\Sigma}_{Y,w} = \boldsymbol{\Sigma}_w$	$\text{cov}(\mathbf{y}, \boldsymbol{\eta}) = \mathbf{PH\Psi}(\boldsymbol{\Psi}^\top\boldsymbol{\Psi})^{-1}$
Special Case 1	Let $\boldsymbol{\Sigma}_w = \mathbf{PHP}$, where for point referenced data the (i,j)-th element of $\mathbf{H}$ is $C_M(\|\mathbf{s}_i - \mathbf{s}_j\|)$, and for areal data $\mathbf{H}$ is the covariance from a CAR model.	Let $\boldsymbol{\Sigma}_Y = \mathbf{PHP} + \boldsymbol{\Sigma}_1$. For point referenced data the (i,j)-th element of $\boldsymbol{\Sigma}_1$ is $C_M(\|\mathbf{s}_i - \mathbf{s}_j\|)$, and for areal data $\boldsymbol{\Sigma}_1$ is the covariance from a CAR model. Recall from Section 4, the parameters of $\boldsymbol{\Sigma}_1$ are chosen using the covariance penalized error.	$\boldsymbol{\Sigma}_{Y,w} = \boldsymbol{\Sigma}_w$	$\text{cov}(\mathbf{y}, \boldsymbol{\eta}) = \mathbf{PH\Psi}(\boldsymbol{\Psi}^\top\boldsymbol{\Psi})^{-1}$
Special Case 3	Let $\boldsymbol{\Sigma}_w = \mathbf{PHP}$, where for point referenced data the (i,j)-th element of $\mathbf{H}$ is $C_M(\|\mathbf{s}_i - \mathbf{s}_j\|)$, and for areal data $\mathbf{H}$ is the covariance from a CAR model.	$\boldsymbol{\Sigma}_w = \boldsymbol{\Sigma}_Y$	Following the discussion at the end of Section 2.7, we set $\boldsymbol{\Sigma}_{Y,w} = \boldsymbol{\Sigma}_Y\mathbf{P}$.	$\text{cov}(\mathbf{y}, \boldsymbol{\eta}) = \mathbf{H\Psi}(\boldsymbol{\Psi}^\top\boldsymbol{\Psi})^{-1}$
Special Case 4	$\boldsymbol{\Sigma}_w = \mathbf{P}\boldsymbol{\Sigma}_Y\mathbf{P}$	For point referenced data the (i,j)-th element of $\boldsymbol{\Sigma}_Y$ is $C_M(\|\mathbf{s}_i - \mathbf{s}_j\|)$ and for areal data $\boldsymbol{\Sigma}_Y$ is the covariance from a CAR model.	$\boldsymbol{\Sigma}_{Y,w} = \boldsymbol{\Sigma}_Y\mathbf{P}$	$\text{cov}(\mathbf{y}, \boldsymbol{\eta}) = \boldsymbol{\Sigma}_Y\boldsymbol{\Psi}(\boldsymbol{\Psi}^\top\boldsymbol{\Psi})^{-1}$

Table 1: Assumptions made for the Bayesian inference in Section 4.3. The choice of $\boldsymbol{\Sigma}_w = \mathbf{PHP}$ stays the same. Thus, to obtain $\boldsymbol{\Sigma}_w$, we set $\mathbf{w} = \boldsymbol{\Psi}\boldsymbol{\eta}$, where $\boldsymbol{\eta}$ is a mean-zero Gaussian random vector, such that $\text{cov}(\boldsymbol{\eta}) = (\boldsymbol{\Psi}^\top\boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}^\top\mathbf{H}\boldsymbol{\Psi}(\boldsymbol{\Psi}^\top\boldsymbol{\Psi})^{-1}$. In the fifth column we give the expression of $\text{cov}(\mathbf{y}, \boldsymbol{\eta})$ that produces $\boldsymbol{\Sigma}_{Y,w}$. We do not include Special Case 2, which can be implemented using the Standard Assumption after a transformation.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0230

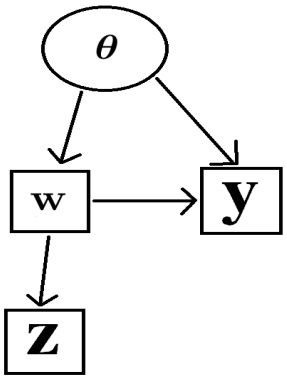

Figure 1: A pictorial representation of the process augmentation. Circles represent the parameters, and squares represent the random vectors that stack the data, augmented process, and latent process over different spatial locations. These vectors are defined in Corollary 1. The n-dimensional vector $\mathbf{z}$ is the observed data vector, $\mathbf{w}$ is the associated n-dimensional vector of the augmented values, and $\mathbf{y}$ is an N-dimensional vector of the values of the latent process.

element of $\mathbf{H}(\boldsymbol{\theta})$ is $C_M(||\mathbf{s}_i - \mathbf{s}_j||; \boldsymbol{\theta})$, and for areal data, $\mathbf{H}(\boldsymbol{\theta})$ is the covariance from a CAR model, where $\boldsymbol{\theta} = (\sigma^2_Y, \tau)'$. The assumptions made for the implementation are outlined in Table 1.

In each setting, the procedure for the posterior inference on $Y(\cdot)$ begins by obtaining B posterior replicates of $\boldsymbol{\beta}$, $\boldsymbol{\eta}$, $\boldsymbol{\xi}$, and $\boldsymbol{\theta}$, which we denote as $\boldsymbol{\beta}^{[b]}$, $\boldsymbol{\eta}^{[b]}$, $\boldsymbol{\xi}^{[b]}$, and $\boldsymbol{\theta}^{[b]}$, respectively for $b = 1, \ldots, B$. Let the N-dimensional vector $\mathbf{w}^{[b]} = \left(w^{[b]}(\mathbf{s}_1), \ldots, w^{[b]}(\mathbf{s}_m)\right)^\top$ and $w^{[b]}(\mathbf{s}) = \mathbf{x}(\mathbf{s})^\top\boldsymbol{\beta} + \boldsymbol{\psi}(\mathbf{s})^\top\boldsymbol{\eta}^{[b]} + \xi^{[b]}(\mathbf{s})$ for each b and $\mathbf{s}$. In general, the full-conditional distributions for $\boldsymbol{\beta}$, $\boldsymbol{\eta}$, $\boldsymbol{\xi}$, and $\boldsymbol{\theta}$ are well known (e.g., see Cressie and Wikle, 2011, Ch. 7), and straightforward to compute. For ease of exposition, we outline the final statistical model and the corresponding full conditional distributions in Appendix B of the Supplementary Material.

We point out that the computationally intensive likelihood associated with $Y(\cdot)$ is not needed to obtain the posterior replicates of $\boldsymbol{\beta}$, $\boldsymbol{\eta}$, $\boldsymbol{\xi}$, and $\boldsymbol{\theta}$. Instead, we need only to use the

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0230

probability density functions $[\mathbf{z}|\boldsymbol{\beta},\boldsymbol{\eta},\boldsymbol{\xi},\boldsymbol{\theta}]$, $[\boldsymbol{\eta}|\boldsymbol{\theta}]$, $[\boldsymbol{\xi}|\boldsymbol{\theta}]$, $[\boldsymbol{\beta}]$, and $[\boldsymbol{\theta}]$. Thus, obtaining $\boldsymbol{\beta}$, $\boldsymbol{\eta}$, $\boldsymbol{\xi}$, and $\boldsymbol{\theta}$ requires Br^3 computations, in general. In addition, confounding between $Y(\cdot)$ and $\delta(\cdot)$ is avoided when estimating these parameters, because obtaining $\boldsymbol{\beta}^{[b]}$, $\boldsymbol{\eta}^{[b]}$, $\boldsymbol{\xi}^{[b]}$, and $\boldsymbol{\theta}^{[b]}$ does not require the joint modeling of $Y(\cdot)$ and $\delta(\cdot)$.

Given the MCMC replicates of the random effects and the parameters, we can now use the posterior predictive distribution of $\mathbf{y}$ to obtain samples from $[\mathbf{y}|\mathbf{z}]$. Using Corollary 1 and standard results for the Gaussian distribution, the predictive distribution for $Y(\mathbf{s})$ is given by

$$Y(\mathbf{s})|\boldsymbol{\eta},\boldsymbol{\xi},\boldsymbol{\theta},\mathbf{z} \sim \text{Gau}\left(\mathbf{e}(\mathbf{s})^\top E(\mathbf{y}|\boldsymbol{\eta},\boldsymbol{\xi},\boldsymbol{\theta}), \mathbf{e}(\mathbf{s})^\top \mathbf{K}(\boldsymbol{\theta})\mathbf{e}(\mathbf{s})\right), \tag{24}$$

where $\mathbf{K}(\boldsymbol{\theta}) = \boldsymbol{\Sigma}_Y - \text{cov}\left(\mathbf{y},\boldsymbol{\eta}|\boldsymbol{\theta}\right)(\boldsymbol{\Psi}^\top\boldsymbol{\Psi})(\boldsymbol{\Psi}^\top\mathbf{H}(\boldsymbol{\theta})\boldsymbol{\Psi})^{-1}(\boldsymbol{\Psi}^\top\boldsymbol{\Psi})\text{cov}\left(\boldsymbol{\eta},\mathbf{y}|\boldsymbol{\theta}\right)$, the elemental vector $\mathbf{e}(\mathbf{s}^*) \equiv (I(\mathbf{s}^* = \mathbf{s}) : \mathbf{s} \in \{\mathbf{u}_1,\ldots,\mathbf{u}_n\})^\top$, and $I(\cdot)$ is the indicator function. The choice of $\mathbf{K}$ depends on which special case we are considering. These choices are outlined in Table 1, where we explicitly provide $\boldsymbol{\Sigma}_Y$ and $\text{cov}(\mathbf{y},\boldsymbol{\eta}|\boldsymbol{\theta})$, which is needed to compute $\mathbf{K}$.

For each b and $\mathbf{s}$,

$$\begin{aligned} Y^{[b]}(\mathbf{s}) &= \mathbf{e}(\mathbf{s})^\top E\left(\mathbf{y}|\boldsymbol{\eta}^{[b]},\boldsymbol{\xi}^{[b]},\boldsymbol{\theta}^{[b]}\right) + \{\mathbf{e}(\mathbf{s})^\top\mathbf{K}(\boldsymbol{\theta}^{[b]})\mathbf{e}(\mathbf{s})\}^{1/2}\phi \\ &= \mathbf{x}(\mathbf{s})^\top\boldsymbol{\beta}^{[b]} + \mathbf{e}(\mathbf{s})^\top\text{cov}\left(\mathbf{y},\boldsymbol{\eta}|\boldsymbol{\theta}^{[b]}\right)\mathbf{K}(\boldsymbol{\theta}^{[b]})^{-1}\boldsymbol{\eta}^{[b]} + \mathbf{e}(\mathbf{s})^\top\boldsymbol{\xi}^{[b]} + \{\mathbf{e}(\mathbf{s})^\top\mathbf{K}(\boldsymbol{\theta}^{[b]})\mathbf{e}(\mathbf{s})\}^{1/2}\phi, \end{aligned}$$

where ϕ is a draw from a standard normal random distribution. Then, the posterior predictions and prediction variances of $Y(\cdot)$ can be estimated as follows:

$$\begin{aligned} \widehat{E}(Y(\mathbf{s})|\mathbf{z}) &= \frac{1}{B}\sum_{b=1}^{B} Y^{[b]}(\mathbf{s}) \\ \widehat{\text{var}}(Y(\mathbf{s})|\mathbf{z}) &= \frac{1}{B}\sum_{b=1}^{B}\left\{Y^{[b]}(\mathbf{s}) - \widehat{E}(Y(\mathbf{s})|\mathbf{z})\right\}^2, \end{aligned} \tag{25}$$

where we let $\widehat{\mathbf{y}} \equiv \left\{\widehat{E}(Y(\mathbf{u}_i)|\mathbf{z}) : i = 1,\ldots,n\right\}^\top$. Note that the computations needed to ob-

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0230

tain the predictions of $Y(\cdot)$ are of order N (i.e., we simulate from Equation (24) B times). In addition, we do not need to store $\mathbf{K}$, but need only to store the n values $\{\mathbf{e}(\mathbf{s})^\top \mathbf{K}(\boldsymbol{\theta}^{[b]})\mathbf{e}(\mathbf{s})\}^{1/2}$ and the $n \times r$ matrix $\boldsymbol{\Psi}$, because r is presumed to be small in this setting. In addition, we do not compute and store $\mathbf{H}$, nor do we compute and store $\mathbf{K}$ before we obtain $\{\mathbf{e}(\mathbf{s})^\top \mathbf{K}(\boldsymbol{\theta}^{[b]})\mathbf{e}(\mathbf{s})\}^{1/2}$. The order of the matrix multiplications is chosen carefully to avoid storing an $n \times n$ dense matrix. For example, to obtain $\mathbf{K}$ we compute the first column of $\mathbf{H}$ and pre-multiply the column by the n-dimensional vector $(\boldsymbol{\Psi}^\top \boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}^\top$, which gives an r-dimensional vector. The remaining steps are described in Appendix B of the Supplementary Material.

In principle, we could use *any* prior on $\boldsymbol{\theta}$, but not every prior on $\boldsymbol{\theta}$ is computationally feasible. Thus, we suggest using a discrete uniform prior. Suppose the discrete uniform prior on $\boldsymbol{\theta}$ takes on M values. Then, we need only store the n values in $\{\mathbf{e}(\mathbf{s})^\top \mathbf{K}(\boldsymbol{\theta})\mathbf{e}(\mathbf{s})\}^{1/2}$ for each of the M values of the support of $\boldsymbol{\theta}$. However, if the prior distribution has continuous support, then the n values in $\{\mathbf{e}(\mathbf{s})^\top \mathbf{K}(\boldsymbol{\theta})\mathbf{e}(\mathbf{s})\}^{1/2}$ need to be computed each time $\boldsymbol{\theta}$ is updated within a Gibbs sampler, which is not computationally feasible. Of course, in low dimensions, a continuous support for $\boldsymbol{\theta}$ would be straightforward to implement.

To choose the support of the discrete uniform distribution, we suggest considering several discrete fixed intervals for the range parameter of a Matérn or CAR model, and using the deviance information criterion of Spiegelhalter et al. (2002) to select from among the candidate supports. Our independent simulation studies suggest that the results are robust to this specification, provided that the discrete uniform support is not too coarse (i.e., M is small). In Appendix B of the Supplementary Material, we outline the statistical model in full and provide the full-conditional distributions for the Gibbs sampler.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0230

5 Empirical Results

In this section, we provide several analyses of simulation studies, as well as an analysis of median household income using a large data set of ACS five-year period estimates at the census tract level. We do not provide empirical results for Special Case 2 because it is closely related to the Standard Assumption (see Section 2.4). In addition, following the discussion at the end of Section 2.7, we only investigate predictions using Special Case 4, and not Special Case 3 (our parameterization of Special Case 3 produces the same kriging predictor as that from Special Case 4). A simulation study of Special Case 1 is provided in Appendix G of the Supplementary Material.

5.1 Simulation Study of Special Case 4: Robust to Departures from the Standard Assumption

In this simulation study, we compare a Bayesian prediction under Special Case 4 to an MPP approach, Bayesian FRK, and an FSA (Finley et al., 2009; Kang and Cressie, 2011; Sang and Huang, 2012). The Gibbs sampler for the Bayesian FRK is outlined in Appendix B of the Supplementary Material, the spBayes R-package is used to compute the MPP predictor (Finley et al., 2015), and Matlab code is used to compute an empirical Bayes implementation of the FSA (Castillo and Tajbakhsh, 2015). The same equally spaced knot locations are used for all methods that share the same rank r. We consider several choices of r for each method. The goal of this analysis is to show that Special Case 4 is robust to departures from the Standard Assumption by comparing it to other reasonable choices for spatial predictions used in the literature.

We generate a random process on a 40×40 grid $D \equiv \{(s_1, s_2)^\top : s_1, s_2 = 0, 0.025, \ldots, 1\}$, and let $\{Y(\mathbf{s}) : \mathbf{s} \in D\}$ be generated from a Matérn process with an unknown mean $\mathbf{x}(\mathbf{s})^\top \boldsymbol{\beta} = 2 + u_1 + 7u_2$ for $\mathbf{s} = (u_1, u_2) \in D$, a smoothing parameter of 0.5 (i.e., an exponential

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0230

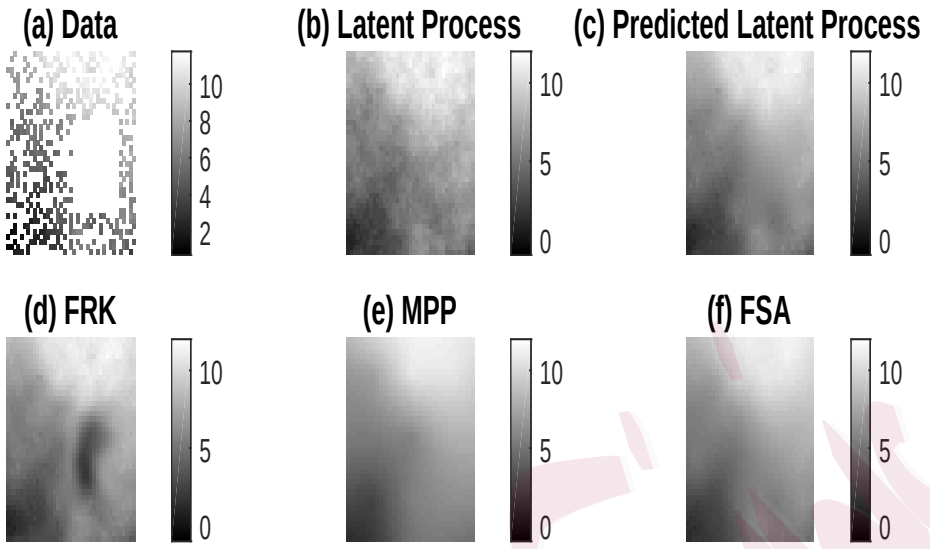

Figure 2: In Panel (a), we display the simulated data $\{Z(\cdot)\}$ over a collection of observed locations $D_O \subset D$, which are generated by randomly selecting points outside of a rectangular region in D. Here, D is a 40×40 grid $D \equiv \{(s_1, s_2)^\top : s_1, s_2 = 0, 0.025, \ldots, 1\}$. White areas indicate missing observations. Panel (b) represents a simulation of the latent process with a Matérn covariance function, as specified in the last row of Table 1. In Panels (c), (d), (e), and (f), we present the posterior expected values of $Y(\cdot)$ from MD = SC4, FRK, MPP, and FSA defined in Table 1, respectively.

covariagram), and a unit variance. The range parameter of the Matérn process is $\tau = 1/12$, such that the spatial range is moderate at $1/4$. Let $\text{var}\{\epsilon(\cdot)\} \equiv 0.5$, resulting in a large signal-to-noise ratio (≈ 10).

To obtain the simulated data, we add independent error,

$$Z(\mathbf{s}) = Y(\mathbf{s}) + \epsilon(\mathbf{s}); \;\; \mathbf{s} \in D_O,$$

where $\{\epsilon(\mathbf{s}_i) : i = 1, \ldots, m\}$ consists of independent and identically distributed Gaussian random variables with mean zero and variance 0.5. Note that $Z(\cdot)$ is not generated according

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0230

to the General Assumption. In Figure 3(a), we plot Z over our choice of observed data locations, with observations missing at random at locations outside of a large square region in D. Note that our model makes an incorrect assumption about the data process, but correctly specifies the latent process.

The bi-square radial basis functions of Cressie and Johannesson (2008) are used. These functions depend on a collection of r knots. The location and number of these knots are found using an algorithm that compares estimates of the latent process to estimates of $w(\cdot)$. In Appendix C of the Supplementary Material, we outline how we select the basis functions. For the predictions in Figure 3, we set $r = 100$.

The predictions in Figure 3 are based on a single realization of $Z(\cdot)$, and are given in Figure 3(a). Let the total MSPE be defined as

$$MSPE(\text{MD}) \equiv \frac{1}{|D|} \sum \left[Y(\mathbf{s}) - E_{MD} \left\{ Y(\mathbf{s}) | \mathbf{z} \right\} \right]^2 ; \text{MD} = \text{SC4}, \text{FRK}, \text{MPP}, \text{FSA},$$

where "MD" represents the model from which the posterior expected value is taken, and "SC4" denotes Special Case 4.

For the example shown in Figure 3, the total MSPEs are as follows: 0.08 for the SC4 approach, 0.75 for FRK, 0.23 for MPP, and 0.12 for FSA. Thus, it is clear that the MSPE is considerably smaller using the SC4 approach. For the Bayesian FRK and MPP, we observe a strange circular artifact in the large 'square' missing region in D, which is a well-known consequence of low rank modeling (Datta et al., 2014). This deficiency in low-rank modeling is not present when using our model, nor is it present for the full-scale approximation of Sang and Huang (2012). These behaviors are consistent over multiple replicates. That is, after generating 100 different sets of $\{Z(\mathbf{s})\}$ and $\{Y(\mathbf{s})\}$, we obtain the results presented in

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0230

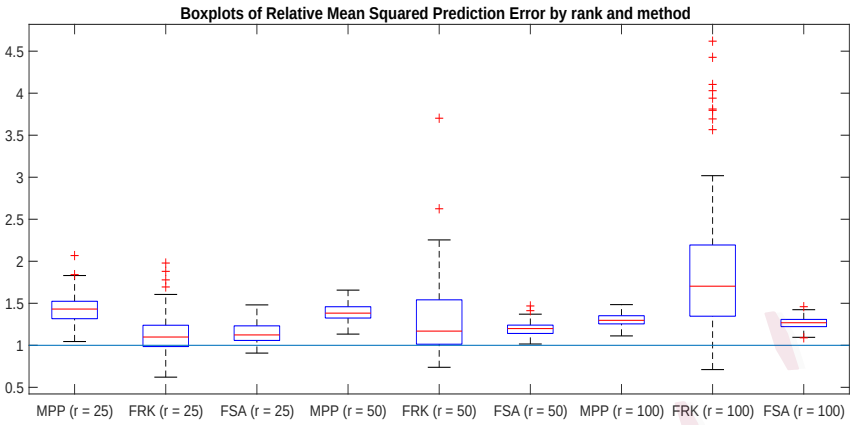

Figure 3: Boxplots of the rMSPE value defined in (26). The x-axis indicates the method and the choice of rank r. The y-axis is relative to the MSPE of SC4. When the rMSPE is greater than one this indicates that the MSPE of SC4 is smaller than that of the model labeled on the x-axis.

Figure 4. Here, we plot the relative MSPE (rMSPE),

$$rMSPE(MD) = \frac{MSPE(\text{MD})}{MSPE(\text{SC4})}; \text{ MD} = \text{SC4}, \text{BFRK}, \text{MPP}, \text{FSA}. \tag{26}$$

For each choice of r, we tend to have values of rMSPE greater than one, which indicates that our approach (i.e., Model 1) outperforms each of its competitors. Furthermore, for each method, as we increase the rank, the range of the rMSPE-values tends to move further away from one, which indicates increasingly better performances as a result of using the SC4 approach.

The data are simulated in a manner such that the optimal kriging predictor should outperform SC4. The MSPE of the optimal kriging predictor tends (over the replicate simulations) to be around 0.02. In addition, the rMSPE associated with the traditional kriging predictor tends to be around 0.91. This suggests that SC4 produces predictions with

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0230

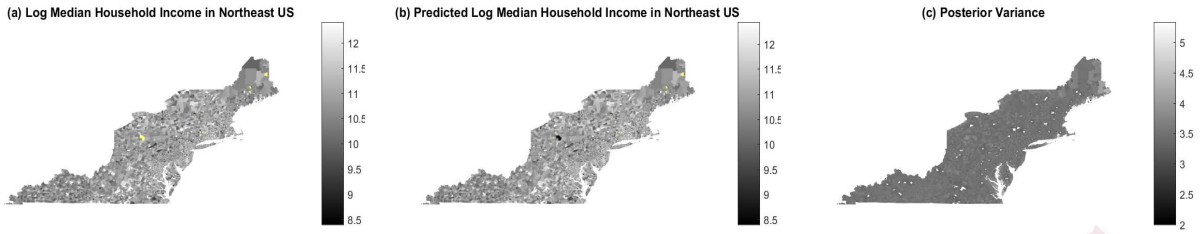

Figure 4: In Panel (a), we plot the log of the ACS 2013 five-year period estimates of median household income over census tracts in the northeast United States. Panel (b) contains the predicted log median income, and Panel (c) displays the corresponding posterior variances. Data, predictions, and prediction variances are available for the whole of the United States, however, we only plot the northeast United States in Panels (a), (b), and (c) for brevity.

an MSPE that is close to (albeit greater than) that of the optimal predictor.

5.2 Census-Tract Level ACS Five-Year Period Estimates of Median Household Income

The U.S. Census Bureau's ACS is a key data source for U.S. demographics. ACS estimates have a unique structure, where the estimates are reported for different periods. Specifically, the ACS currently produces one-year and five-year period estimates for various U.S. demographic variables. We consider a subset of the ACS data and provide a spatial analysis of median household income for the period 2009−2013. We consider a fairly large data set of 72,361 observations, which includes of estimates of median household income over all census tracts in the contiguous United States. A subset of the data is presented in Figure 4(a).

This particular example is especially interesting from the point of view of a spatial analysis, because ACS period estimates of median household income have been modeled using Moran's I (MI) basis functions (e.g., see Bradley et al., 2015b, 2016). The literature suggests

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0230

that MI basis functions require a large r to obtain a reasonable fit when using a low-rank approach; for example, in Bradley et al. (2015b) $r = 4,750$. Hence, one of our goals is to determine whether our approach allows us to model w using fewer basis functions than is common in the literature. Furthermore, federal data sets, such as the public-use ACS estimates, are typically modeled under the assumption of independent "survey errors" (i.e., δ) (Fay and Herriot, 1979). Thus, evidence suggesting that the survey errors are dependent may have important implications for the modeling of federal data.

For this example, let $\{R(\mathbf{s}) : \mathbf{s} \in D\}$ represent the median household income over the census tracts in the contiguous United States (denoted by D). Histograms of the logarithm of R appear roughly symmetric, indicating that normality of $Z(\cdot) = \log\{R(\cdot)\}$ is reasonable. Thus, for this example, we assume that Z (i.e., the logarithm of the data) is Gaussian. In addition, the survey variances are converted to a log scale using the delta method (Oehlert, 1992). We use an intercept-only model and set $\mathbf{X}$ equal to a vector of ones. We consider two different methods to analyze areal data sets: a version of Bayesian kriging, as in Hughes and Haran (2013, among others); and the CAR model (Besag, 1974). In Appendix H of the Supplementary Material, we clearly explain the model assumptions, as well as the implementations for Special Case 1, Bayesian FRK, and the CAR model.

All Markov chains in this section use a burn-in of 1,000 replications and generate $B = 10,000$ posterior replications. Convergence is assessed visually using trace plots of the sample chain, with no lack of convergence detected. The rank r is specified according to Appendix C of the Supplementary Material. Here, we find that $r = 30$ is reasonable, and a comparison with historical choices of r when using MI basis functions shows that we obtain a significant dimension reduction. For this data set $\sigma^2_Y = 9$ and $\tau = 0.91$ minimize the covariance penalized error in (14). In Figure 4(b,c) we display subsets of the posterior mean and variances. Here, we see that the overall pattern of the predictions are similar to, but smoother than, the spatial patterns of the data. In addition, the estimates of Efron (2004) covariance

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0230

penalized error (see (14)) are -0.9874 for our model, -0.1321 for the FRK, and 0.9153 for the CAR model. Thus, the proposed model outperforms competing methods for spatial prediction.

6 Discussion

In this paper, we have introduced a hierarchical model that leverages spatial dependencies within the error process associated with the data, as well as leveraging cross-dependencies between the error process and the latent process of principal interest. This is done by introducing nondegenerate discrepancy error covariances that are not confounded within the marginal distribution of the data. The "Standard Assumption" of uncorrelated discrepancy errors is an important special case of our general parameterization, which occurs when three matrix-valued parameters are equal to each other. We consider four additional special cases in which we allow the matrix-valued parameters to differ.

Our parameterization between the error process and the latent process leads to a computationally useful process-augmentation approach. The first step in our implementation is to fit a model for the process that does not assume a dependent error process (see Appendix B of the Supplementary Material). Then, the second step produces predictions of a latent process that is dependent on the error process. The latter predictions are based on the output from the first step of the algorithm (see Section 5.2 and Appendix B of the Supplementary Material). This feature of our model greatly increases the applicability of the approach because any statistical model that is based on the assumption of mutually independent errors can be used in the first step of the algorithm.

Our empirical results suggest that our approach is robust to departures from our model assumptions. We illustrated this by implementing Special Case 3, where the data are generated using a full-rank Matérn specification. Low-rank statistical models are known to be

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0230

sensitive to the setting when the data are very sparse over the spatial domain. Hence, sparse data were generated. In this setting, our predictor outperformed many of the current methods used for predictions. These result suggest that if the assumption of a dependent error process is not correct, then nonconfounded discrepancy-error covariances may still be useful.

The nonconfounded discrepancy error covariance can also be scaled to large data sets, which we demonstrated using a data set consisting of ACS estimates of median household income defined on census tracts.Note that we have produced precise predictions that have a full-rank specification, as well as computational gains for a reduced-rank model. In a sense, the proposed method benefits from a full-rank specification and low-rank specification. Furthermore, our example demonstrated that there appears to be dependent error in a popular survey data set. This is a setting in which it is typically assumed that the error process is independent of the latent process (i.e., the Fay−Herriot model used in federal statistics). This could have important implications for small area estimation because the Fay−Herriot model is a ubiquitous choice for modeling area-level data in the official statistics literature.

There are many opportunities for future research. For example, the selection of a basis function (e.g., see Huang et al., 2006; Bradley et al., 2011) is an important and recurring inferential problem. We specified knot locations so that the augmented process was close to the latent process. This type of strategy can also be used to choose the class of spatial basis functions. In general, there is great potential to use the proposed modeling paradigm to specify parsimonious models in an informed manner.

Supplementary Materials

The Supplementary Material provides proofs of the technical results, a review of current methods in spatial statistics, an additional simulation, and additional details on the model specification and implementation.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0230

Acknowledgments

We would like to express our gratitude to the editor, associate editor, and referees for their very helpful comments. This research was partially supported by the U.S. National Science Foundation (NSF) and the U.S. Census Bureau under NSF grant SES-1132031, funded by the NSF-Census Research Network (NCRN) program. This paper is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the authors and not necessarily those of the NSF or the U.S. Census Bureau.

References

Albert, J. H. and Chib, S. (1993). "Bayesian Analysis of Binary and Polychotomous Response Data." *Journal of the American Statistical Association*, 88, 669–679.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2015). *Hierarchical Modeling and Analysis for Spatial Data*. London, UK: Chapman and Hall.

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008). "Gaussian predictive process models for large spatial data sets." *Journal of the Royal Statistical Society Series B*, 70, 825–848.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. New York, NY: Springer-Verlag.

Besag, J. E. (1974). "Spatial Interaction and the statistical analysis of lattice systems (with discussion)." *Journal of the Royal Statistical Society, Series B*, 36, 192–236.

Black, B. (1976). "Studies of Stock Price Volatility Changes." *Proceedings of the 1976*

Meetings of the American Statistical Association, Business and Economic Statistics, 177–181.

Bradley, J. R., Cressie, N., and Shi, T. (2011). "Selection of rank and basis functions in the Spatial Random Effects model." In *Proceedings of the 2011 Joint Statistical Meetings*, 3393–3406. Alexandria, VA: American Statistical Association.

Bradley, J. R., Holan, S. H., and Wikle, C. K. (2015a). "Multivariate spatio-temporal models for high-dimensional areal data with application to Longitudinal Employer-Household Dynamics." *The Annals of Applied Statistics*, 9, 1761–1791.

Bradley, J. R., Wikle, C. K., and Holan, S. H. (2015b). "Spatio-Temporal Change of Support with Application to American Community Survey Multi-Year Period Estimates." *Stat*, 4, 255–270.

— (2016). "Bayesian spatial change of support for count-valued survey data." *Journal of the American Statistical Association*, 111, 472 – 487.

Castillo, E. and Tajbakhsh, B. M. C. . S. D. (2015). "Geodesic Gaussian Processes for the Reconstruction of a Free-Form Surface." *Technometrics*, 57, 87–99.

Clayton, D., Bernardinelli, L., and Montomoli, C. (1993). "Spatial correlation in ecological analysis." *International Journal of Epidemiology*, 6, 1193–1202.

Cressie, N. (1990). "The origins of kriging." *Mathematical Geology*, 22, 239–252.

— (1993). *Statistics for Spatial Data,* rev. edn. New York, NY: Wiley.

Cressie, N. and Johannesson, G. (2008). "Fixed rank kriging for very large spatial data sets." *Journal of the Royal Statistical Society, Series B*, 70, 209–226.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0230

Cressie, N. and Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data*. Hoboken, NJ: Wiley.

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2014). "Hierarchical Nearest-Nerighbor Gaussian Process Models for Large Geostatistical Datasets." *arXiv preprint: 1406.7343*.

Efron, B. (1983). "Estimating the error rate of a prediction rule: Improvement on cross-validation." *Journal of the Americal Statistical Association*, 78, 316–331.

— (2004). "The estimation of prediction error: Covariance penalties and cross-validation." *Journal of the Americal Statistical Association*, 99, 619–642.

Fay, R. and Herriot, R. (1979). "Estimates of income for small places: an application of James-Stein procedures to census data." *Journal of the American Statistical Association*, 74, 269–277.

Finley, A. O., Banerjee, S., and Carlin, B. (2015). "Package 'spBayes'." http://cran.r-project.org/web/packages/spBayes/spBayes.pdf. Retrieved April, 2015.

Finley, A. O., Sang, H., Banerjee, S., and Gelfand, A. E. (2009). "Improving the performance of predictive process modeling for large datasets." *Computational Statistics and Data Analysis*, 53, 2873–2884.

Griffith, D. (2000). "A linear regression solution to the spatial autocorrelation problem." *Journal of Geographical Systems*, 2, 141–156.

— (2002). "A spatial filtering specification for the auto-Poisson model." *Statistics and Probability Letters*, 58, 245–251.

— (2004). "A spatial filtering specification for the auto-logistic model." *Environment and Planning A*, 36, 1791–1811.

Groves, R., Dillman, D. A., Eltinge, J. L., and Little, R. J. A. (2001). *Survey Nonresponse (Wiley Series in Survey Methodology)*. New York, NY: Wiley-Interscience.

Hodges, J. S. and Reich, B. J. (2011). "Adding Spatially-Correlated Errors Can Mess Up the Fixed Effect You Love." *The American Statistician*, 64, 325–334.

Holan, S. H. and Wikle, C. K. (2012). "Semiparametric Dynamic Design of Monitoring Networks for Non-Gaussian Spatio-Temporal Data." In *Spatio-Temporal Design Advances in Efficient Data Acquisition*, eds. J. Mateu and W. Muller. New York, NY: Wiley.

Huang, H. C., Hsu, N. J., Theobald, D., and Breidt, F. J. (2006). "Spatial LASSO with applications to GIS model selection."

Hughes, J. and Haran, M. (2013). "Dimension reduction and alleviation of confounding for spatial generalized linear mixed model." *Journal of the Royal Statistical Society, Series B*, 75, 139–159.

Kang, E. L. and Cressie, N. (2011). "Bayesian inference for the Spatial Random Effects model." *Journal of the American Statistical Association*, 106, 972 – 983.

Karhunen, K. (1947). "Uber lineare Methoden in der Wahrscheinlichkeitsrechnung." *Ann. Acad. Sci. Fennicae. Ser. A. I. Math.-Phys*, 37, 1–49.

Katzfuss, M. (2017). "A multi-resolution approximation for massive spatial datasets." *Journal of the American Statistical Association*, 112, 201–214.

Lindgren, F., Rue, H., and Lindström, J. (2011). "An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach." *Journal of the Royal Statistical Society, Series B*, 73, 423–498.

Loéve, M. (1978). "Probability theory, vol. ii." *Graduate texts in mathematics*, 46, 0–387.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0230

Matérn, B. (1960). "Spatial Variation." *Meddelanden fran Statens Skogsforskningsinstitut*, 49, 1–144.

Matheron, G. (1963). "Principles of geostatistics." *Economic Geology*, 58, 1246–1266.

Moran, P. A. P. (1950). "Notes on Continuous Stochastic Phenomena." *Biometrika*, 37, 17–23.

Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F., and Sain, S. (2015). "A multi-resolution Gaussian process model for the analysis of large spatial data sets." *Journal of Computational and Graphical Statistics*, 2, 579–599.

Oehlert, G. (1992). "A note on the delta method." *The American Statistician*, 46, 27 – 29.

Quick, H., Holan, S. H., Wikle, C. K., and Reiter, J. P. (2015). "Bayesian marked point process modeling for generating fully synthetic public use data with point-referenced geography." *Spatial Statistics*, 14, 439–451.

Ravishanker, N. and Dey, D. K. (2002). *A First Course in Linear Model Theory*. Boca Raton, FL: Chapman and Hall/CRC.

Reich, B. J., Hodges, J. S., and Zadnik, V. (2006). "Effects of Residual Smoothing on the Posterior of the Fixed Effects in Disease-Mapping Models." *Biometrics*, 62, 1197–1206.

Sang, H. and Huang, J. (2012). "A full-scale approximation of covariance functions for large spatial data sets." *Journal of the Royal Statistical Society: Series B*, 74, 111–132.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). "Bayesian measures of model complexity and fit." *Journal of the Royal Statistical Society, Series B*, 64, 583–616.

Stein, M. (2014). "Limitations on low rank approximations for covariance matrices of spatial data." *Spatial Statistics*, 8, 1–19.

Tanner, M. A. and Wong, W. H. (1987). "The Calculation of Posterior Distributions by Data Augmentation." *Journal of the American Statistical Association*, 82, 528–540.

Torrieri, N. (2007). "America is changing, and so is the census: The American Community Survey." *American Statistician*, 61, 16–21.

Wakefield, J. and Walker, S. (1999). "Gibbs Sampling for Bayesian Non-Conjugate and Hierarchical Models by Using Auxiliary Variables." *Journal of the Royal Statistical Society*, 82, 331–344.

Wikle, C. K. and Royle, J. A. (2005). "Dynamic design of ecological monitoring networks for nonGaussian spatio-temporal data." *Environmetrics*, 16, 507–522.

Wolpert, R. and Ickstadt, K. (1998). "Poisson/gamma random field models for spatial statistics." *Biometrika*, 85, 251–267.

Zeger, S. L. and Liang, K.-Y. (1991). "Feedback models for discrete and continuous time series." *Statistica Sinica*, 1, 51–64.