Statistica Sinica **29** (2019), 1711-1737
doi:https://doi.org/10.5705/ss.202017.0008

SMOOTHED FULL-SCALE APPROXIMATION OF GAUSSIAN PROCESS MODELS FOR COMPUTATION OF LARGE SPATIAL DATA SETS

Bohai Zhang[1], Huiyan Sang[2] and Jianhua Z. Huang[2]

[1]Nankai University and [2]Texas A&M University

Abstract: Gaussian process (GP) models encounter computational difficulties with large spatial data sets, because the models' computational complexity grows cubically with the sample size n. Although a full-scale approximation (FSA) using a block modulating function provides an effective way to approximate GP models, it has several shortcomings. These include a less smooth prediction surface on block boundaries and sensitivity to the knot set under small-scale data dependence. To address these issues, we propose a smoothed full-scale approximation (SFSA) method for analyzing large spatial data sets. The SFSA leads to a class of scalable GP models, with covariance functions that consist of two parts: a reduced-rank covariance function that captures large-scale spatial dependence, and a covariance that adjusts the local covariance approximation errors of the reduced-rank part, both within blocks and between neighboring blocks. This method reduces the prediction errors on block boundaries, and leads to inference and prediction results that are more robust under different dependence scales owing to the better approximation of the residual covariance. The proposed method provides a unified view of approximation methods for GP models, grouping several existing computational methods for large spatial data sets into one common framework. These methods include the predictive process, FSA, and nearest neighboring block GP methods, allowing efficient algorithms that provide robust and accurate model inferences and predictions for large spatial data sets within a unified framework. We illustrate the effectiveness of the SFSA approach using simulation studies and a total column ozone data set.

Key words and phrases: Conditional likelihood, full-scale approximation, markov chain monte carlo, spatial covariance functions.

1. Introduction

Spatial data sets arising from ecology, climatology, and other disciplines are of considerable interest to scientists. With the advent of remote sensing and geographic information system (GIS) techniques, the quantity of spatial data available has increased dramatically. As a result, statisticians often need to process a large number of observations on variables of interest. This growth in the volume of data has imposed computational challenges on classical geostatistical models (Stein (1999); Banerjee, Gelfand and Carlin (2014)), resulting in the development of innovative computational methods capable of handling extremely large data sets (e.g., Sun, Li and Genton (2012)).

One of the most popular models used for spatial data sets is the Gaussian process (GP) model, which assumes that finite observations are jointly Gaussian. Although GP models enjoy mathematical tractability for model fitting and prediction, their computational complexity grows cubically with the sample size n, in general, owing to their expensive matrix operations. Specifically, calculations of the inverse and the determinant of an $n \times n$ covariance matrix of a GP typically require $\mathcal{O}(n^3)$ floating point operations per second (flops). Thus, fitting a GP model becomes computationally prohibitive for very large n.

Recently, Sang and Huang (2012) proposed the so-called full-scale approximation (FSA) approach to approximate the original covariance function of GP models for large spatial data sets. By combining the ideas of both low-rank models and sparse models, the FSA approach can approximate the data covariance matrix well for both large- and small-scale dependence structures. Popular low-rank models (e.g., Higdon (2002); Banerjee et al. (2008); Cressie and Johannesson (2008); Katzfuss and Cressie (2011); Nguyen, Cressie and Braverman (2012); Nguyen et al. (2014)) seek to approximate the original spatial process using a smoother process based on a reduced number of basis functions. Although low-rank models can enjoy computational complexity linear with n, they may fail to capture local variations well when using a limited number of basis functions (Finley et al. (2009); Stein (2014)). Sparse approximation techniques either shrink the covariance of distant pairs of spatial locations to zero to yield a sparse covariance matrix (Furrer, Genton and Nychka (2006); Kaufman, Schervish and Nychka (2008)), or assume a Gaussian–Markov property of the spatial random field to yield a sparse precision matrix (Rue and Tjelmeland (2002); Lindgren, Rue and Lindström (2011)). Another method used to induce sparsity of the precision matrix is to use conditional likelihoods (e.g., Vecchia (1988)). Here, Datta et al. (2016) proposed a nearest-neighbor GP (NNGP). A new permutation and grouping method that improves the performance of the NNGP method can be found in Guinness (2018). Recent “hybrid” methods that extend low-rank models include the works of Nychka et al. (2015), Katzfuss (2017), and Ma and Kang (2017). Modern versions of local GP models (e.g., (Gramacy and Apley (2015); Gramacy and Haaland (2016); Zhang, Lin and Ranjan (2018); Park and Apley (2018))) can also be applied effectively to model large or massive spatial data. Lastly, the divide-and-conquer-based approaches have been proposed to model large and nonstationary spatial data sets. See, for example, the treed GP (e.g., Gramacy and Lee (2008); Konomi et al. (2014)) and the spatial meta kriging (Guhaniyogi and Banerjee (2017)) methods.

Let $\mathcal{C}(\cdot, \cdot; \boldsymbol{\theta})$ be the original covariance function of a GP model. To give a more accurate approximation to $\mathcal{C}(\cdot, \cdot; \boldsymbol{\theta})$, the FSA approach first approximates the original covariance function using the covariance function of a Gaussian predictive process model (Banerjee et al. (2008)), denoted by $\mathcal{C}_l(\cdot, \cdot; \boldsymbol{\theta})$. Then the "residual" covariance, defined as $\mathcal{C}_s \equiv \mathcal{C} - \mathcal{C}_l$, is approximated by a sparse positive semidefinite function. The covariance function of the FSA, denoted by $\mathcal{C}^{\dagger}(\cdot, \cdot; \boldsymbol{\theta})$, can be written as $\mathcal{C}_l(\cdot, \cdot; \boldsymbol{\theta}) + \mathcal{C}_s(\cdot, \cdot; \boldsymbol{\theta})\mathcal{K}(\cdot, \cdot)$, where the function $\mathcal{K}$, referred to as a modulating function, is positive semidefinite and has a large number of zeros evaluated on observed spatial locations. If we choose $\mathcal{K}(\cdot, \cdot)$ as compactly supported covariance functions (Gneiting (2002)), the resulting approximation is referred to as *FSA-Taper*; if $\mathcal{K}(\cdot, \cdot) = 1$ when two locations belong to the same data block and $\mathcal{K}(\cdot, \cdot) = 0$ otherwise, then the resulting approximation is referred to as *FSA-Block*. Empirical results have shown that FSA-Block outperforms FSA-Taper (Sang, Jun and Huang (2011)), possibly because FSA-Block is an unbiased approximation of the covariance within each data block and uses parallel computation. Zhang, Sang and Huang (2015) extends the FSA-Block approximation to GP models for large spatio-temporal data sets.

Although the FSA-Block approach can lead to an effective and scalable approximation to the covariance function of a GP model, it has several shortcomings. First, its predictions around the boundaries of two adjacent blocks are less smooth than those of other regions, mainly because of its assumption of independent blocks for the residual covariance function, $\mathcal{C}_s$. Thus, mismatches between the predictions on block boundaries can result in large prediction errors for locations close to block boundaries. Second, the overall performance of the FSA-Block approach is more robust to the choice of knots and blocks, respectively, than that of the predictive process and independent block estimations. However, the approximation error for the residual covariance information across blocks can be severe when the predictive-process part does not perform well (e.g., when the underlying spatial process is less smooth or the number of knots is insufficient), leaving room for further improvement.

In this study, we develop a new covariance approximation for spatial GP models. We first extend the nearest-neighbor GP models developed by Datta et al. (2016) to construct a nearest neighboring block GP model. We then apply our model to approximate the residual covariance, which we combine with a reduced-rank predictive process. By doing so, we relax the independent-blocks assumption of FSA-Block to further account for the dependence between each block and its neighboring blocks in the residual covariance matrix. The proposed method alleviates the discontinuities of predictions on boundary locations for the FSA-Block approach. We call the proposed method the smoothed full-scale approximation (SFSA) method.

We further show that the SFSA approach defines a class of valid GP models that is scalable to large data sets. Therefore, the SFSA approach can perform both parameter estimations and predictions under a unified framework owing to the existence of a closed-form covariance function. The establishment of the SFSA GP also allows it to be flexibly embedded into hierarchical spatial models, which facilitates computation, while maintaining model richness.

The SFSA provides a unified view of approximations for spatial GP models, grouping together various existing popular approximation methods, including the predictive process, FSA, conditional composite likelihood, independent blocks method, and nearest-neighbor GP approximation. This unified modeling framework enables direct comparisons and reveals the relations between various computational methods for large spatial data sets.

The rest of the paper is organized as follows. Section 2 reviews the FSA-Block approach and formulates the proposed SFSA approach. Section 3 discusses the computational complexity of the SFSA and gives the algorithm used to evaluate its log-likelihood. Section 4 describes the parameter estimation and prediction procedures of the SFSA. Section 5 defines the valid GP constructed from the SFSA. Then, we compare the SFSA with other current methods using simulation studies in Section 6.1 and a total column ozone data set in Section 6.2. Finally Section 7 concludes the paper with a brief summary and a discussion of potential extensions to our work. The proofs of the theorems and additional numerical results are given in the Supplementary Material.

2. Methodology

2.1. The spatial regression model

Let $y(\mathbf{s})$ be a response variable observed at a spatial location $\mathbf{s} \in \mathcal{S} \subseteq \mathbb{R}^d$, where $\mathcal{S}$ is the spatial domain and $d = 1, 2, 3$. We model $y(\mathbf{s})$ through the following spatial regression model:

$$y(\mathbf{s}) = x(\mathbf{s})^T \boldsymbol{\beta} + w(\mathbf{s}) + \epsilon(\mathbf{s}), \tag{2.1}$$

where $x(\mathbf{s})$ is a p-dimensional vector of covariates, $\boldsymbol{\beta}$ is a vector of regression coefficients, $w(\mathbf{s})$ is a latent mean-zero GP, and $\epsilon(\mathbf{s})$ is a Gaussian white-noise process with a constant variance τ^2, independent of $w(\mathbf{s})$. The variance τ^2 is often referred to as the "nugget," accounting for the measurement-error effect. The dependence structure of $w(\mathbf{s})$ is specified by a valid covariance function, $\mathcal{C}(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta}) \equiv \text{cov}(w(\mathbf{s}), w(\mathbf{s}'))$. For example, the Matérn covariance function (e.g., see Stein (1999)) is widely used in spatial statistics owing to its flexibility in modeling the smoothness of a spatial process:

$$\mathcal{C}(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta}) = \frac{\sigma^2}{\Gamma(\nu)} 2^{1-\nu} \left(\frac{h}{\phi}\right)^{\nu} K_\nu \left(\frac{h}{\phi}\right), \tag{2.2}$$

where $\sigma^2 > 0$ is the variance parameter, $\phi > 0$ is the dependence range parameter, $\nu > 0$ is the smoothness parameter, $\Gamma(\cdot)$ is the gamma function, and $K_\nu(\cdot)$ is the modified Bessel function of the second kind of order ν. The Gaussian covariance function, $\mathcal{C}(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta}) = \sigma^2 \exp(-h^2/\phi)$, and the exponential covariance function, $\mathcal{C}(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta}) = \sigma^2 \exp(-h/\phi)$, are two special cases of (2.2), with $\nu \to \infty$ and $\nu = 0.5$, respectively.

Now, suppose $y(\mathbf{s})$ is observed at n spatial locations in $S \equiv \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$. Let $\mathbf{y} = (y(\mathbf{s}_1), \ldots, y(\mathbf{s}_n))^T$ be the observed response vector and $\mathbf{x} = (x(\mathbf{s}_1), \ldots, x(\mathbf{s}_n))^T$ be the $n \times p$ design matrix. The log-likelihood function is:

$$\ell(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\beta}) = -\frac{1}{2}(\mathbf{y} - \mathbf{x}\boldsymbol{\beta})^T C_{\mathbf{y}}^{-1}(\mathbf{y} - \mathbf{x}\boldsymbol{\beta}) - \frac{1}{2}|C_{\mathbf{y}}| - \frac{n}{2}\log(2\pi), \tag{2.3}$$

where $C_{\mathbf{y}} \equiv \text{var}(\mathbf{y})$ is the data covariance matrix. In general, evaluating (2.3) requires $\mathcal{O}(n^3)$ flops to calculate $|C_{\mathbf{y}}|$ and $C_{\mathbf{y}}^{-1}$. Thus, the computational cost can be very high (and even prohibitive) when n is very large.

2.2. The FSA-Block approach

In this subsection, we briefly review the FSA-Block approach. The approach (Sang, Jun and Huang (2011); Sang and Huang (2012)) is motivated by the decomposition of the latent spatial process $w(\mathbf{s})$:

$$w(\mathbf{s}) = w_l(\mathbf{s}) + w_s(\mathbf{s}), \tag{2.4}$$

where $w_l(\mathbf{s})$ is the Gaussian predictive process (Banerjee et al. (2008)), and $w_s(\mathbf{s})$ is referred to as the residual process of $w(\mathbf{s})$ that is independent of $w_l(\mathbf{s})$. Approximating $w(\mathbf{s})$ using only $w_l(\mathbf{s})$ results in a loss of residual covariance information in $w_s(\mathbf{s})$, which could lead to bias in the parameter estimations and inaccuracy in the spatial predictions (e.g., see Finley et al. (2009); Stein (2014)).

Let $S^* \equiv \{\mathbf{s}_1^*, \ldots, \mathbf{s}_m^*\}$ be a (pre-specified) set of locations in $\mathcal{S}$, referred to as the knot set. In the following, we use the generic notation $\mathcal{C}(A, B) \equiv [\mathcal{C}(\mathbf{s}_i, \mathbf{s}_j)]_{\mathbf{s}_i \in A, \mathbf{s}_j \in B}$ to denote the covariance matrix for two location sets, A and B. The covariance function of $w_l(\mathbf{s})$ is given by

$$\mathcal{C}_l(\mathbf{s}, \mathbf{s}') = \mathcal{C}(\mathbf{s}, S^*)\mathcal{C}(S^*, S^*)^{-1}\mathcal{C}(\mathbf{s}', S^*)^T. \tag{2.5}$$

It follows that the covariance function of $w_s(\mathbf{s})$ takes the form

$$\mathcal{C}_s(\mathbf{s}, \mathbf{s}') = \mathcal{C}(\mathbf{s}, \mathbf{s}') - \mathcal{C}_l(\mathbf{s}, \mathbf{s}'). \tag{2.6}$$

Let $C_{w_l} \equiv \mathcal{C}_l(S, S)$ be the covariance matrix of the predictive-process component and $C_{w_s} \equiv \mathcal{C}_s(S, S)$ be the residual covariance matrix. By the Schur complement property of linear algebra, C_{w_s} is positive definite when $S \cap S^* = \emptyset$, and positive semidefinite otherwise. In general, C_{w_s} has a dependence structure of a smaller scale than that of the full covariance; however, it is still a dense matrix. Sang, Jun and Huang (2011) proposed approximating C_{w_s} using a block-diagonal matrix to reduce the number of computations, while preserving the residual covariance entries within blocks. Specifically, let $\mathcal{P}$ be a partition rule that partitions the observed data vector $\mathbf{y}$ into K disjoint subvectors $\mathbf{y}_k$ of length n_k, for $k = 1, \ldots, K$. If we group the observations according to blocks, then the likelihood approximated by the FSA-Block approach follows the Gaussian distribution, $\mathcal{N}(\mathbf{x}\boldsymbol{\beta}, (C_{w_l} + C_{w_s} \circ \mathcal{T}_B + \tau^2 I_n))$, where $\mathcal{T}_B$ is a block-diagonal matrix with $\mathbf{1}_{n_k}\mathbf{1}_{n_k}^T$ as its k-th block, $\mathbf{1}_{n_k}$ is an $n_k \times 1$ vector of ones, I_n is an identity matrix of size n, and $\circ$ is the Schur product (entrywise product) of two matrices. Compared with C_{w_l} from the predictive process model, the FSA-Block approach incorporates an additional block-diagonal residual covariance matrix to correct the approximation errors within each data block. Because $(C_{w_s} \circ \mathcal{T}_B + \tau^2 I_n)$ is block-diagonal, it takes $\mathcal{O}(n)$ order flops to compute its inverse and determinant. It can be shown that the computational complexity of the FSA-Block approach is linear with n (Sang, Jun and Huang (2011)).

However, the independent-blocks approximation of C_{w_s} ignores the residual dependence across blocks. The loss of dependence information can be severe when $w_l(\mathbf{s})$ does not provide a good approximation for $w(\mathbf{s})$, such that the entries across the blocks of the residual covariance matrix are not negligible (e.g., the knots are not placed properly or the number of knots is insufficient). More importantly, because the approximation errors of the covariance matrix by FSA-Block are zero for data within the same block and nonzero for data across blocks, there exist jumps of approximation errors between each data block and its neighboring blocks. This discontinuity of approximation errors can harm the prediction performance, particularly around the block boundaries (see Section 6.1). To address these issues, we seek a new method that partially preserves the entries of C_{w_s} across data blocks, while maintaining computational efficiency.

2.3. The SFSA approach

Let $\mathbf{w}^* = (w(\mathbf{s}_1^*), \ldots, w(\mathbf{s}_m^*))^T$ denote the vector of $w(\cdot)$ evaluated on the knot set. To motivate the new method, we write the data likelihood as

$$p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\theta}) = \int p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{w}^*) \cdot p(\mathbf{w}^*|\boldsymbol{\theta}) d\mathbf{w}^*,$$

where $p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{w}^*)$ follows $\mathcal{N}(\mathbf{x}\boldsymbol{\beta} + \mathcal{C}(S, S^*)\mathcal{C}(S^*, S^*)^{-1}\mathbf{w}^*, C_{w_s} + \tau^2 I_n)$. The computational bottleneck lies in evaluating $p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{w}^*)$, because C_{w_s} is a dense matrix, in general. We propose replacing $p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{w}^*)$ with some Gaussian density that has less expensive computations. Then, after integrating out $\mathbf{w}^*$, we obtain an approximated Gaussian likelihood with a reduced computational cost. Note that, compared with the original data covariance matrix, the covariance matrix in $p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{w}^*)$ has entries closer to zero. Therefore, data located in distant blocks are more likely to be independent, conditional on $\mathbf{w}^*$. This observation motivates us to use the conditional block composite likelihood (CBCL) approach of Stein, Chi and Welty (2004) to approximate $p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{w}^*)$.

Specifically, let $\mathcal{P}$ be a partition rule leading to the partition $S = \cup_{k=1}^K S_k$, with a corresponding partition of observations $\mathbf{y} = \cup_{k=1}^K \mathbf{y}_k$, where S_k and $\mathbf{y}_k$ have size n_k, and $\sum_{k=1}^K n_k = n$. Let $\mathbf{y}_{(k-1)} = (\mathbf{y}_1^T, \ldots, \mathbf{y}_{k-1}^T)^T$ for $k \geq 2$ and $\mathbf{y}_{(0)} = \emptyset$. By the chain rule,

$$p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{w}^*) = \prod_{k=1}^K p(\mathbf{y}_k|\mathbf{y}_{(k-1)}, \boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{w}^*).$$

When n is very large, it is computationally expensive to evaluate the full conditional density, $p(\mathbf{y}_k|\mathbf{y}_{(k-1)}, \boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{w}^*)$, for large k, because $\mathbf{y}_{(k-1)}$ is high-dimensional. Thus, following Stein, Chi and Welty (2004), we choose the conditional set as a subvector of $\mathbf{y}_{(k-1)}$ for the k-th block:

$$\tilde{p}(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{w}^*) = \prod_{k=1}^K p(\mathbf{y}_k|\mathbf{y}_{N(k)}, \boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{w}^*), \tag{2.7}$$

where $\mathbf{y}_{N(k)}$ is an $n_{N(k)}$-dimensional subvector of $\mathbf{y}_{(k-1)}$ with the location set $S_{N(k)}$ (i.e., the neighboring observations of $\mathbf{y}_k$ in $\mathbf{y}_{(k-1)}$). Here, we use the convention that $S_{N(1)} = \emptyset$. For notational simplicity, we focus on the special case in which $S_{N(k)}$ contains all locations in the q nearest neighboring blocks of the k-th block (e.g., "closeness" is measured by the Euclidean distances between the block centers). Specifically, $S_{N(k)}$ is defined as

$$S_{N(k)} = \begin{cases} \emptyset, & \text{if } k = 1; \\ \{S_1, S_2, \dots, S_{k-1}\}, & \text{if } k \le q; \\ q \text{ nearest blocks in } \{S_1, S_2, \dots, S_{k-1}\}, & \text{if } k > q. \end{cases}$$

In practice, we choose K to partition the data such that each data block contains only a few hundred observations, for computation efficiency. By choosing $q \ll K$, evaluating $\tilde{p}(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{w}^*)$ is computationally efficient. The FSA-Block approach is a special case of the proposed method when we use $\emptyset$ as the conditional set for every $\mathbf{y}_k$. Usually, we choose $q \ge 1$ to ease the discontinuity issue of approximation errors across data blocks. We later show that the prediction errors around block boundaries can be reduced by applying the proposed approach. This yields the proposed SFSA approach.

Next, we show that the SFSA approach generates a Gaussian likelihood with a closed-form expression for its covariance matrix. For residual covariance matrices of $(w_s(\cdot) + \epsilon(\cdot))$, we use the generic notation $\Sigma_{A,B} \equiv \operatorname{cov}(w_s(S_A) + \epsilon(S_A), w_s(S_B) + \epsilon(S_B))$ and $\Sigma_A \equiv \operatorname{var}(w_s(S_A) + \epsilon(S_A))$, where S_A and S_B are two sets of spatial locations. Now, for $k, l = 1, \dots, K$, define

$$B_{k,l} = \begin{cases} I_{n_k}, & \text{if } l = k; \\ \left[-\Sigma_{k,N(k)}\Sigma_{N(k)}^{-1}\right](\cdot, n_{(l-1)} + 1 : n_{(l)}), & \text{if } l \in N(k); \\ \mathbf{0}, & \text{otherwise}, \end{cases} \tag{2.8}$$

where $n_{(l)} = \sum_{1 \le i \le l, i \in N(k)} n_i$. Here, $B_{k,l}$ is an $n_k \times n_l$ matrix that encodes the conditional dependence information between the k-th block and the l-th block. Let $B_k^* = (B_{k,1}, \dots, B_{k,K})$. Then, it can be shown that (see the Supplementary Material, Section S1.1) the conditional density, $p(\mathbf{y}_k|\mathbf{y}_{N(k)}, \boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{w}^*)$, is proportional to

$$|\Sigma_{k|N(k)}|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{y} - \mathbf{x}\boldsymbol{\beta} - U\mathbf{w}^*)^T B_k^{*^T} \Sigma_{k|N(k)}^{-1} B_k^*(\mathbf{y} - \mathbf{x}\boldsymbol{\beta} - U\mathbf{w}^*)\right\},$$

where $\Sigma_{k|N(k)} = \Sigma_k - \Sigma_{k,N(k)}\Sigma_{N(k)}^{-1}\Sigma_{k,N(k)}^T$ is the residual covariance of the k-th block, conditional on its neighboring blocks, and $U = \mathcal{C}(S, S^*)\mathcal{C}(S^*, S^*)^{-1}$. The SFSA approach yields the following likelihood:

$$\tilde{p}(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\theta}) = \int_{\mathbf{w}^*} \prod_{k=1}^{K} p(\mathbf{y}_k|\mathbf{y}_{N(k)}, \boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{w}^*) \cdot p(\mathbf{w}^*|\boldsymbol{\theta}) d\mathbf{w}^*.$$

The following theorem shows that this approximated likelihood corresponds to a Gaussian density with a closed-form covariance matrix.

Theorem 1. *Let* $\mathbf{y} \sim \mathcal{N}(\mathbf{x}\boldsymbol{\beta}, C_{\mathbf{y}})$. *Then, the approximated likelihood by the SFSA approach,* $\tilde{p}(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\theta})$, *follows* $\mathcal{N}(\mathbf{x}\boldsymbol{\beta}, C_{\mathbf{y}}^{\dagger})$, *where*

$$C_{\mathbf{y}}^{\dagger} = B^{-1}\Sigma_{con}B^{T^{-1}} + \mathcal{C}(S, S^*)\mathcal{C}(S^*, S^*)^{-1}\mathcal{C}(S, S^*)^T,$$

where Σ_{con} *is a block-diagonal matrix, with* $\Sigma_{k|N(k)}$ *as its k-th block, and* $B = (B_1^{*^T}, \ldots, B_K^{*^T})^T \in \mathbb{R}^{n\times n}$.

The proof is given in the Supplementary Material, Section S1.1.

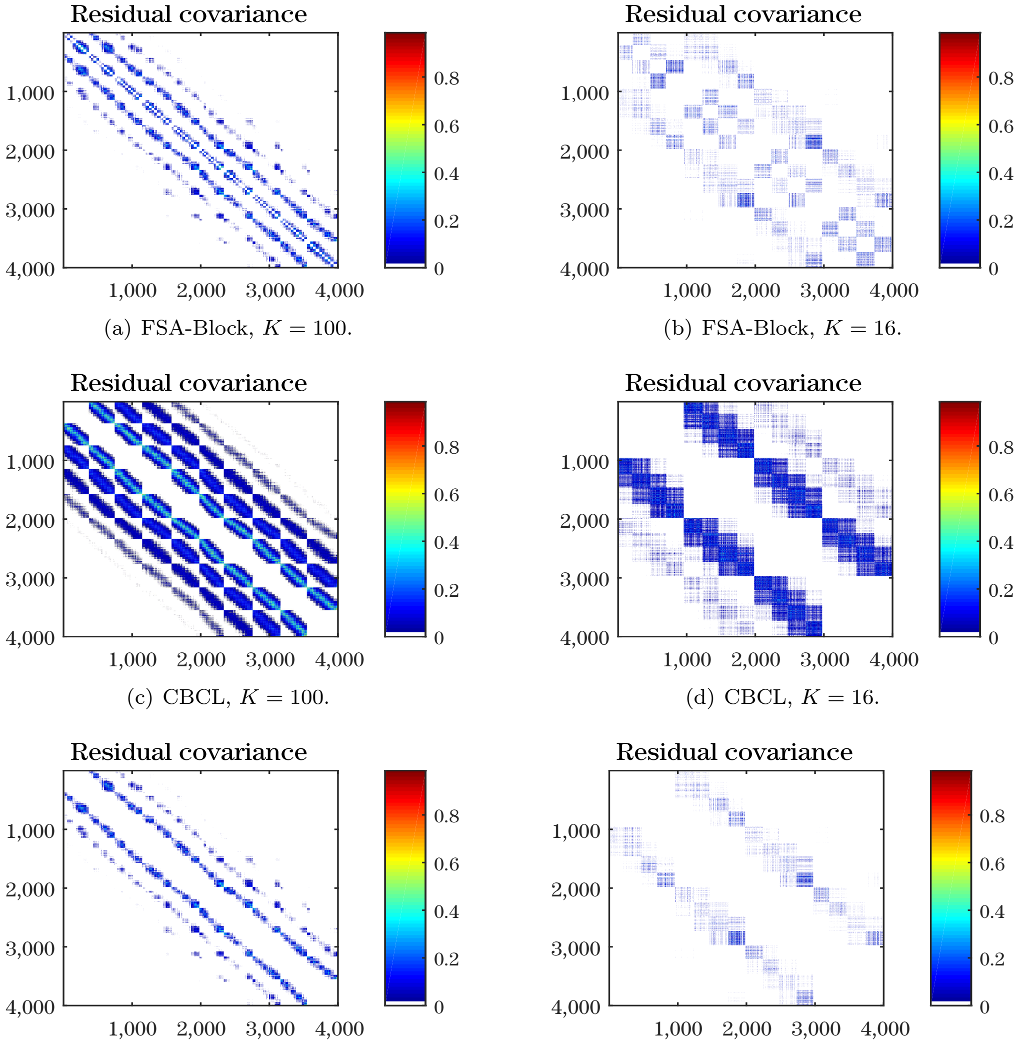
2.4. A new unifying view

The proposed method offers a unified approximation method for spatial GP models. Evidently, the method is a direct generalization of the FSA-Block approach (SFSA with $q = 0$) and the conditional block composite likelihood approach (SFSA with $m = 0$). Hence, it includes both as special cases. Thus, the SFSA provides a unified approximation framework for spatial GP models that allows us to compare different methods directly.

Below, we compare the performance of each method in terms of their covariance matrix approximations. Figure 1 shows the absolute differences of entries between the approximated data covariance matrix and the original data covariance matrix for each of the approaches. Specifically, 4,000 locations were randomly generated in a square domain $[0, 10] \times [0, 10]$. Then, we used the exponential covariance function $\mathcal{C}(\mathbf{s}, \mathbf{s}') = \exp(-\|\mathbf{s} - \mathbf{s}'\|)$ with a nugget effect of 0.01 to generate the covariance matrix on these locations. For all three approaches, equally spaced blocks are generated and the block index takes an increasing order from northwest to southeast; locations within the same block are grouped together. For the SFSA and FSA-Block approaches, $m = 50$ knots were uniformly selected in the square domain; for the SFSA and CBCL, the neighboring block set is the nearest neighboring block. We observe that for locations within a certain band, the approximation errors by the SFSA are much smaller than those by the FSA-Block approach, owing to the corrections of the residual covariance between neighboring blocks. Compared with the CBCL approximation, both approaches provide good approximations for covariance entries within a certain location band. However, the SFSA approach leads to smaller approximation errors for the residual covariance entries off the location band, owing to the inclusion of the low-rank predictive-process component.

2.5. Choices of tuning parameters for the SFSA approach

The SFSA approach requires specifications of several tuning parameters, in-



(a) FSA-Block, $K = 100$.

(b) FSA-Block, $K = 16$.

(c) CBCL, $K = 100$.

(d) CBCL, $K = 16$.

(e) SFSA, $K = 100$.

(f) SFSA, $K = 16$.

Figure 1. Plots of the absolute differences between the approximated data covariance matrix and the original data covariance matrix for three methods. SFSA: the smoothed full-scale approximation; CBCL: the conditional block composite likelihood approximation.

cluding the knot set, block partition scheme, ordering of the data blocks, and number of neighboring blocks q. For the knot set, random sampling, Latin hypercube sampling (McKay, Conover and Beckman (1979)), or a spatial grid can be applied to place knots with a good space coverage. Alternatively, we can treat the knots as unknown parameters and model them stochastically (Guhaniyogi et al. (2011); Katzfuss (2013); Zhang, Sang and Huang (2015)). For the block partition, Eidsvik et al. (2014) recommend using the empirical variogram to determine the block width. The K-means clustering algorithm based on the Euclidean distances of locations is a simple choice for creating blocks; alternatively, we can apply a clustering algorithm based on the estimated covariance matrix from a pilot study to account for nonstationarity. For uniformly spaced spatial locations, we recommend using regular rectangular blocks (e.g., see Eidsvik et al. (2014); Katzfuss (2017)), which empirically work very well. We adopt this method here. For highly nonuniformly distributed data, Delauny triangulation (e.g., see Lee and Schachter (1980)) might be more effective to create meshes for the proposed method.

After creating the blocks, it is necessary to order the blocks to construct the residual likelihood of the SFSA. Following Guinness (2018), we compared the model-fitting performance of the SFSA for a few ordering methods, including the sorted-coordinate (SC) ordering, random ordering, maximum-minimum-distance (MMD) ordering, and center-out (CO) ordering (see the Supplementary Material, Section S2.1). Based on our simulation results, we recommend using the SC ordering for uniformly spaced data and the CO ordering for nonuniformly spaced data.

Lastly, the selection of the number of neighboring blocks (q) is a trade-off between the computational time and the statistical efficiency. Here, a larger number of neighboring blocks yields a more accurate approximation by the SFSA. Based on our simulation results (see the Supplementary Material, Section S2.3), a small number of neighboring blocks, such as $q = 3$ or 4 (with a few hundred observations), can already lead to statistically efficient parameter-estimation results. Here, because we focus on using regular rectangular blocks, the Euclidean distance between the block centers becomes a natural choice to determine the q nearest neighboring blocks. Alternatively, we can define the “closeness” of two blocks using a distance metric between the residual correlations of observations in two blocks. However, such an approach requires estimating residual correlation, which increases the computational cost. Further details on finding the nearest neighboring blocks using residual correlations are provided in the Supplementary Material (Section S3).

3. Computational Aspects of the SFSA Approach

We first determine the computational complexity of evaluating the log-like-

Table 1. Notation for the SFSA.

Sample size:	n
Knot number:	m
Block size:	n_b
Number of blocks:	K
Number of neighbors:	q

lihood of the SFSA (Table 1 gives the notations for the SFSA). For simplicity, suppose all data blocks have an equal block size n_b, such that $n = Kn_b$, and each data block has at most q neighbors. The log-likelihood function of the SFSA, up to a constant, is (see equation (S1.1))

$$\begin{aligned}\log \tilde{p}(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\theta}) = &-\frac{1}{2}(\mathbf{y}-\mathbf{x}\boldsymbol{\beta})^T B^T(\Sigma_{con}^{-1} - \Sigma_{con}^{-1} BU\Sigma_{\mathbf{w}^*}U^T B^T \Sigma_{con}^{-1})B(\mathbf{y}-\mathbf{x}\boldsymbol{\beta}) \\ &-\frac{1}{2}|U^T B^T \Sigma_{con}^{-1} BU| - \frac{1}{2}|\Sigma_{con}| - \frac{1}{2}|C_*|, \qquad (3.1)\end{aligned}$$

where $\Sigma_{\mathbf{w}^*} = (U^T B^T \Sigma_{con}^{-1} BU + C_*^{-1})^{-1} \in \mathbb{R}^{m\times m}$ and $C_* \equiv \mathcal{C}(S^*, S^*)$.

Evaluating the determinant of the SFSA likelihood is computationally efficient, because we need only calculate the determinants of two $m\times m$ matrices and a block-diagonal matrix. When evaluating $|U^T B^T \Sigma_{con}^{-1} BU|$, we need to obtain BU and Σ_{con} first. Note that B is a sparse matrix with at most (qn_b+1) nonzero entries per row. Hence, calculating BU is computationally inexpensive, with complexity $\mathcal{O}(nmqn_b)$. To obtain each diagonal block of the block-diagonal matrix Σ_{con}, we need to invert a $(qn_b \times qn_b)$ residual covariance matrix for neighboring observations, which has computational complexity $\mathcal{O}(q^3 n_b^3)$. Hence, obtaining Σ_{con} has the order $\mathcal{O}(Kq^3 n_b^3) = \mathcal{O}(nq^3 n_b^2)$.

Now, suppose Σ_{con} has been obtained. To evaluate the quadratic term in (3.1), the required quantities are $(\mathbf{y}-\mathbf{x}\boldsymbol{\beta})^T B^T \Sigma_{con}^{-1} B(\mathbf{y}-\mathbf{x}\boldsymbol{\beta})$, $U^T B^T \Sigma_{con}^{-1} BU$, and $U^T B^T \Sigma_{con}^{-1} B(\mathbf{y}-\mathbf{x}\boldsymbol{\beta})$. Recall that Σ_{con} is block-diagonal and that its inverse takes $\mathcal{O}(Kn_b^3) = \mathcal{O}(nn_b^2)$ flops. Furthermore, BU has computational complexity $\mathcal{O}(nmqn_b)$, because B is a lower-triangular matrix with at most (qn_b+1) nonzero entries per row, and U is an $n\times m$ matrix. Similarly, evaluating $B(\mathbf{y}-\mathbf{x}\boldsymbol{\beta})$ needs $\mathcal{O}(nqn_b)$ flops. After Σ_{con}^{-1}, BU, and $B(\mathbf{y}-\mathbf{x}\boldsymbol{\beta})$ are calculated, $(\mathbf{y}-\mathbf{x}\boldsymbol{\beta})^T B^T \Sigma_{con}^{-1} B(\mathbf{y}-\mathbf{x}\boldsymbol{\beta})$ needs $\mathcal{O}(nn_b+n)$ flops, $U^T B^T \Sigma_{con}^{-1} BU$ needs $\mathcal{O}(nm^2 + nmn_b)$ flops, and $U^T B^T \Sigma_{con}^{-1} B(\mathbf{y}-\mathbf{x}\boldsymbol{\beta})$ needs $\mathcal{O}(nn_b+nm)$ flops.

Therefore, the computational complexity of the SFSA approach has the order $\mathcal{O}(nq^3 n_b^2 + nmqn_b + nm^2)$. In practice, the data are partitioned into K blocks, such that each block contains a few hundred observations. If we choose the knot size m to be a few hundred as well, and set $q \ll K$, the SFSA approach then has computational complexity linear with n.

Algorithm 1 : Evaluating the log-likelihood function of the SFSA.

```
Compute $C_* = \mathcal{C}(S^*, S^*)$ and $U = \mathcal{C}(S, S^*)\mathcal{C}(S^*, S^*)^{-1}$. Factorize $C_* = Q_*^T Q_*$.
foreach $k = 1$ to $K$
    Compute $\Sigma_k$, $\Sigma_{k,N(k)}$, and $\Sigma_{N(k)}$. Then, compute $B_k^* = (B_{k,1}, \ldots, B_{k,K})$ according to (2.8).
    Compute $\Sigma_{k|N(k)} = \Sigma_k - \Sigma_{k,N(k)}\Sigma_{N(k)}^{-1}\Sigma_{k,N(k)}^T$. Factorize $\Sigma_{k|N(k)} = Q_{k|N(k)}^T Q_{k|N(k)}$.
    Compute the quantities $(\mathbf{y}-\mathbf{x}\boldsymbol{\beta})^T {B_k^*}^T \Sigma_{k|N(k)}^{-1} B_k^*(\mathbf{y}-\mathbf{x}\boldsymbol{\beta})$, $U^T {B_k^*}^T \Sigma_{k|N(k)}^{-1} B_k^* U$, and $U^T {B_k^*}^T \Sigma_{k|N(k)}^{-1} B_k^*(\mathbf{y} - \mathbf{x}\boldsymbol{\beta})$.
end foreach
Sum the quantities for each block to obtain $(\mathbf{y} - \mathbf{x}\boldsymbol{\beta})^T B^T \Sigma_{con}^{-1} B(\mathbf{y} - \mathbf{x}\boldsymbol{\beta})$, $U^T B^T \Sigma_{con}^{-1} BU$, and $U^T B^T \Sigma_{con}^{-1} B(\mathbf{y} - \mathbf{x}\boldsymbol{\beta})$.
Compute the quadratic term in (3.1) and $\Sigma_{\mathbf{w}^*} = U^T B^T \Sigma_{con}^{-1} BU + C_*^{-1}$. Factorize $\Sigma_{\mathbf{w}^*} = Q_{\mathbf{w}^*}^T Q_{\mathbf{w}^*}$.
Compute the log of determinants: $\log|\Sigma_{\mathbf{w}^*}| = 2\log|Q_{\mathbf{w}^*}|$, $\log|\Sigma_{con}| = 2\sum_{k=1}^{K} \log|Q_{k|N(k)}|$ and $\log|C_*| = 2\log|Q_*|$.
Evaluate the log-likelihood function in (3.1).
```

Parallel computation is possible for evaluating the SFSA's likelihood. Recall that $B = ({B_1^*}^T, \ldots, {B_K^*}^T)^T$ is a lower-triangular matrix, where $B_k^* = (B_{k,1}, \ldots, B_{k,K})$ encodes the residual conditional dependence information between the k-th block and each of the individual blocks, for $k = 1, \ldots, K$. Because Σ_{con} is a block-diagonal matrix with $\Sigma_{k|N(k)}$ as its k-th block, we have

$$(\mathbf{y} - \mathbf{x}\boldsymbol{\beta})^T B^T \Sigma_{con}^{-1} B(\mathbf{y} - \mathbf{x}\boldsymbol{\beta}) = \sum_{k=1}^{K} (\mathbf{y} - \mathbf{x}\boldsymbol{\beta})^T {B_k^*}^T \Sigma_{k|N(k)}^{-1} B_k^*(\mathbf{y} - \mathbf{x}\boldsymbol{\beta}).$$

Similarly,

$$U^T B^T \Sigma_{con}^{-1} BU = \sum_{k=1}^{K} U^T {B_k^*}^T \Sigma_{k|N(k)}^{-1} B_k^* U$$

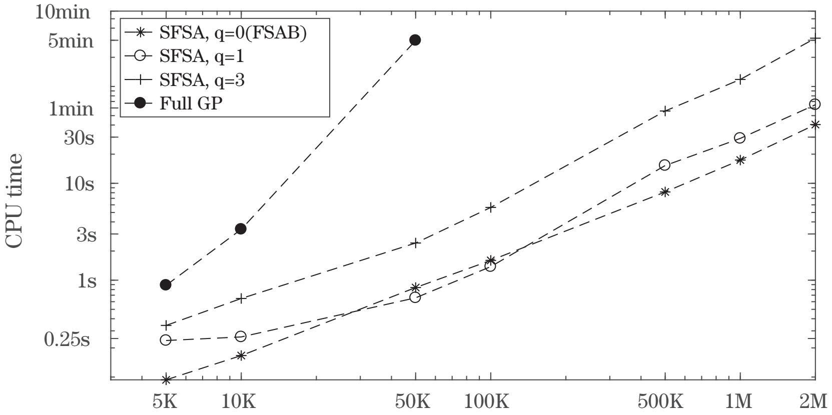


Figure 2. Computational time (on a log scale) per likelihood evaluation versus sample size (on a log scale). For the SFSA and FSAB, $m = n_b = 200$; the results were obtained using 16 CPU cores.

and

$$U^T B^T \Sigma_{con}^{-1} B(\mathbf{y} - \mathbf{x}\boldsymbol{\beta}) = \sum_{k=1}^{K} U^T B_k^{*T} \Sigma_{k|N(k)}^{-1} B_k^* (\mathbf{y} - \mathbf{x}\boldsymbol{\beta}).$$

Algorithm 1 describes how to evaluate $\log \tilde{p}(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\theta})$. A parallel-computing technique can be applied to obtain the required quantities for each block simultaneously and, hence, avoid the loop in algorithm 1. Using K cores, the SFSA has computational complexity $\mathcal{O}(q^3 n_b^3 + mqn_b^2 + n_b m^2)$. In addition, because $U \in \mathbb{R}^{n\times m}$, it will require a significant amount of memory to store U for very large n. Note that because of the sparsity of B_k^*, only $U_k = \mathcal{C}(S_k, S^*)\mathcal{C}(S^*, S^*)^{-1}$ and $U_{N(k)} = \mathcal{C}(S_{N(k)}, S^*)\mathcal{C}(S^*, S^*)^{-1}$ are required to calculate $B_k^* U$ when evaluating the likelihood of the SFSA.

Figure 2 shows the computational time of the SFSA for different sample sizes. As the figure shows, evaluating the likelihood of the SFSA with $q = 1$ can still be done in a short time, even for two million observations.

4. Parameter Estimation and Prediction

4.1. Maximum likelihood estimation

The maximum likelihood estimators maximize the log-likelihood function in (2.3). To facilitate the computations, we replace the full covariance matrix $C_{\mathbf{y}}$ with the approximated covariance matrix $C_{\mathbf{y}}^{\dagger}$ in Theorem 1. The log-likelihood approximated by the SFSA is:

$$\log \tilde{p}(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\theta}) = -\frac{1}{2}(\mathbf{y}-\mathbf{x}\boldsymbol{\beta})^T C_{\mathbf{y}}^{\dagger^{-1}}(\mathbf{y}-\mathbf{x}\boldsymbol{\beta}) - \frac{1}{2}|C_{\mathbf{y}}^{\dagger}| - \frac{n}{2}\log(2\pi).$$

We calculate the inverse covariance matrix as (see equation (S1.2))

$$C_{\mathbf{y}}^{\dagger^{-1}} = \Sigma_{con}^{-1} - \Sigma_{con}^{-1} BU\Sigma_{\mathbf{w}^*}U^T B^T \Sigma_{con}^{-1}.$$

Then, we can evaluate the quadratic term in the log-likelihood efficiently using Algorithm 1. For $|C_{\mathbf{y}}^{\dagger}|$ (see equation (S1.3)),

$$|C_{\mathbf{y}}^{\dagger}| = |U^T B^T \Sigma_{con}^{-1} BU + C_*^{-1}| \cdot |\Sigma_{con}| \cdot |C_*|.$$

Calculating $U^T B^T \Sigma_{con}^{-1} BU$ involves multiplying an $n \times n$ matrix B by an $n \times m$ matrix U. Recall that B is a sparse matrix with at most $(qn_b + 1)$ nonzero entries per row. Hence, calculating BU is computationally inexpensive with complexity $\mathcal{O}(nqn_b m)$. Then, efficient computations are achieved, because calculating $C_{\mathbf{y}}^{\dagger}$ requires computing the determinants of two $m \times m$ matrices and one block-diagonal matrix only.

4.2. Bayesian inference on model parameters

The Bayesian inference starts by specifying the prior distributions of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. The conjugate Gaussian prior $\pi(\boldsymbol{\beta}) \sim \mathcal{N}(\boldsymbol{\mu}_0, \Sigma_0)$ can be assigned to $\boldsymbol{\beta}$. The prior of $\boldsymbol{\theta}$ depends on the form of the covariance function. Taking the Matérn covariance function in (2.2) as an example, the inverse gamma prior $IG(a,b)$ can be assigned to the variance parameter σ^2 and the nugget τ^2, where hyperparameters a and b are chosen to assign vague priors, or to reflect reasonable guesses for the mean and variance. For the dependence range parameter ϕ, a uniform prior with a reasonable support of practical dependence ranges can be used. For the smoothness parameter ν, a uniform prior at $(0,3]$ is commonly used, because high smoothness values are rarely identified from real data sets.

The marginalized likelihood that integrates out both $\boldsymbol{\beta}$ and $\boldsymbol{w}$ is

$$p(\mathbf{y}|\boldsymbol{\theta}) = \int_{\boldsymbol{\beta}} p(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\theta})\pi(\boldsymbol{\beta})d\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{y}}, \Sigma_{\mathbf{y}}),$$

where $\boldsymbol{\mu}_{\mathbf{y}} = \Sigma_{\mathbf{y}} C_{\mathbf{y}}^{-1}\mathbf{x}(\mathbf{x}^T C_{\mathbf{y}}^{-1}\mathbf{x} + \Sigma_0^{-1})^{-1}\Sigma_0^{-1}\boldsymbol{\mu}_0$ and $\Sigma_{\mathbf{y}} = C_{\mathbf{y}} + \mathbf{x}\Sigma_0\mathbf{x}^T$. Because the posterior distribution of $\boldsymbol{\theta}$ does not have a closed form, we first draw posterior samples of $\boldsymbol{\theta}$ based on the marginalized likelihood $p(\mathbf{y}|\boldsymbol{\theta})$ using the Metropolis–Hastings algorithm (Gelman et al. (2014)). Since $p(\boldsymbol{\beta}|\boldsymbol{\theta},\mathbf{y})$ is Gaussian and $p(\boldsymbol{\beta}|\mathbf{y}) = \int_{\boldsymbol{\theta}} p(\boldsymbol{\beta}|\boldsymbol{\theta},\mathbf{y})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$, the posterior samples of $\boldsymbol{\beta}$ can be drawn from $p(\boldsymbol{\beta}|\mathbf{y})$ using the method of composition. Similarly, the posterior samples of $\mathbf{w}$ can be recovered by sampling from

$$p(\mathbf{w}|\mathbf{y}) = \int_{\boldsymbol{\theta}} \int_{\boldsymbol{\beta}} p(\mathbf{w}|\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\beta}) p(\boldsymbol{\beta}|\mathbf{y}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\beta} d\boldsymbol{\theta}.$$

When n is large, we replace $C_{\mathbf{y}}$ with $C_{\mathbf{y}}^{\dagger}$ (see Theorem 1) in $p(\mathbf{y}|\boldsymbol{\theta})$, $p(\boldsymbol{\beta}|\boldsymbol{\theta}, \mathbf{y})$, and $p(\mathbf{w}|\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\beta})$ in order to draw posterior samples efficiently.

4.3. Prediction

Let $S_p \equiv \{\mathbf{s}_1, \ldots, \mathbf{s}_{n_p}\}$ be a set of predictive spatial locations such that $S_p \cap S = \emptyset$, with $\mathbf{y}_p = (y(\mathbf{s}_1), \ldots, y(\mathbf{s}_{n_p}))^T$ as the corresponding response vector. Using the same partition rule $\mathcal{P}$ that partitions S into K disjoint blocks, suppose S_p is partitioned into K disjoint location blocks $S_{p,k}$ ($S_{p,k}$ may be empty), with $\mathbf{y}_{p,k}$ as the response vector of $y(\cdot)$ evaluated on $S_{p,k}$, for $k = 1, \ldots, K$. We start from the joint density of $\mathbf{y}_p$ and $\mathbf{y}$,

$$p(\mathbf{y}_p, \mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\theta}) = \int p(\mathbf{y}_p|\mathbf{y}, \mathbf{w}^*, \boldsymbol{\beta}, \boldsymbol{\theta}) \cdot p(\mathbf{y}|\mathbf{w}^*, \boldsymbol{\beta}, \boldsymbol{\theta}) \cdot p(\mathbf{w}^*|\boldsymbol{\theta}) d\mathbf{w}^*.$$

When n is very large, because $p(\mathbf{y}_p|\mathbf{y}, \mathbf{w}^*, \boldsymbol{\beta}, \boldsymbol{\theta})$ and $p(\mathbf{y}|\mathbf{w}^*, \boldsymbol{\beta}, \boldsymbol{\theta})$ are high-dimensional, their exact computations may not be feasible. Thus, we define the following approximated conditional density:

$$\tilde{p}(\mathbf{y}_p|\mathbf{y}, \mathbf{w}^*, \boldsymbol{\beta}, \boldsymbol{\theta}) = \prod_{k=1}^{K} p(\mathbf{y}_{p,k}|\mathbf{y}_k, \mathbf{y}_{N(k)}, \mathbf{w}^*, \boldsymbol{\beta}, \boldsymbol{\theta}),$$

where we set $p(\mathbf{y}_{p,k}|\mathbf{y}_k, \mathbf{y}_{N(k)}, \mathbf{w}^*, \boldsymbol{\beta}, \boldsymbol{\theta}) = 1$ if $\mathbf{y}_{p,k} = \emptyset$. This definition assumes that $\mathbf{y}_{p,k}$ is independent of the other predictive responses (conditional on $\mathbf{w}^*$), the observations in the same block $\mathbf{y}_k$, and the observations in the neighboring blocks $\mathbf{y}_{N(k)}$. Note that for the predictive response vector $\mathbf{y}_{p,k}$, its neighboring location set is $S_{p,N(k)} \equiv \{S_{N(k)}, S_k\}$, for $k = 1, \ldots, K$, where $S_{N(k)}$ is the neighboring location set for the observed response vector $\mathbf{y}_k$.

Then, an approximated marginal joint density with computational efficiency can be obtained as

$$\tilde{p}(\mathbf{y}_p, \mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\theta}) = \int \tilde{p}(\mathbf{y}_p|\mathbf{y}, \mathbf{w}^*, \boldsymbol{\beta}, \boldsymbol{\theta}) \cdot \tilde{p}(\mathbf{y}|\mathbf{w}^*, \boldsymbol{\beta}, \boldsymbol{\theta}) \cdot p(\mathbf{w}^*|\boldsymbol{\theta}) d\mathbf{w}^*, \qquad (4.1)$$

where $\tilde{p}(\mathbf{y}|\mathbf{w}^*, \boldsymbol{\beta}, \boldsymbol{\theta})$ is the Gaussian density given in (2.7). The (approximated) predictive distribution, $\tilde{p}(\mathbf{y}_p|\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\theta})$, can be readily obtained from $\tilde{p}(\mathbf{y}_p, \mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\theta})$.

Let $\mathbf{x}_{p,k}$ be the design matrix for $\mathbf{y}_{p,k}$, $U_{p,k} = \mathcal{C}(S_{p,k}, S^*) C_*^{-1}$, and $\Sigma_{p,k|N(k)}$ be the residual conditional variance of $\mathbf{y}_{p,k}$, given $\mathbf{y}_k$ and $\mathbf{y}_{N(k)}$. Then, define $B_{p,k} = (B_{p,k,1}, \ldots, B_{p,k,K})$, where $B_{p,k,l}$ is defined similarly to $B_{k,l}$ in (2.8). This encodes the residual conditional dependence information of $\mathbf{y}_{p,k}$ given its neighbors $y(S_{p,N(k)})$ for the l-th block, for $l = 1, \ldots, K$. Let $\mathbf{x}_p = (\mathbf{x}_{p,1}^T, \ldots, \mathbf{x}_{p,K}^T)^T$,

$U_p = (U_{p,1}^T, \ldots, U_{p,K}^T)^T$, $B_p = (B_{p,1}^T, \ldots, B_{p,K}^T)^T$, and $\Sigma_{p,con}$ be a block-diagonal matrix with $\Sigma_{p,k|N(k)}$ as its k-th block. The following proposition shows that $\tilde{p}(\mathbf{y}_p|\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\theta})$ follows a Gaussian distribution.

Proposition 1. *The approximated conditional density $\tilde{p}(\mathbf{y}_p|\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\theta})$ based on* (4.1) *follows a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_p, \Sigma_p)$, where*

$$\begin{aligned}
\boldsymbol{\mu}_p &= \mathbf{x}_p\boldsymbol{\beta} + F_p C_{\mathbf{y}}^{\dagger^{-1}}(\mathbf{y} - \mathbf{x}\boldsymbol{\beta}),\\
\Sigma_p &= \Sigma_{p,con} + B_p B^{-1}\Sigma_{con} B^{T^{-1}} B_p^T + U_p C_* U_p^T - F_p C_{\mathbf{y}}^{\dagger^{-1}} F_p^T,\\
F_p &= (-B_p B^{-1}\Sigma_{con} B^{T^{-1}} + U_p C_* U^T).
\end{aligned}$$

The proof of Proposition 1 is given in the Supplementary Material, Section S1.2. The conditional mean $\boldsymbol{\mu}_p$ is the kriging formula for spatial predictions under the SFSA approximation. In fact, in Section 5, we prove that the SFSA approach can induce a valid GP with a closed-form covariance function.

Remark 1. There are different ways to approximate $p(\mathbf{y}_p, \mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\theta})$ for predictions. For example, consider the case in which the predictive response vector $\mathbf{y}_p$ belongs to some block l. Then, we can approximate the augmented data likelihood, $\tilde{p}(\mathbf{y}_p, \mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\theta})$, defined as

$$\int \prod_{1\le k\le K, k\ne l} p(\mathbf{y}_k|\mathbf{y}_{N(k)}^{aug}, \mathbf{w}^*, \boldsymbol{\beta}, \boldsymbol{\theta}) \cdot p(\mathbf{y}_p, \mathbf{y}_l|\mathbf{y}_{N(l)}, \mathbf{w}^*, \boldsymbol{\beta}, \boldsymbol{\theta}) \cdot p(\mathbf{w}^*|\boldsymbol{\theta}) d\mathbf{w}^*,$$

where $\mathbf{y}_{N(k)}^{aug} = (\mathbf{y}_p^T, \mathbf{y}_{N(k)}^T)^T$ if block l is a neighbor of the k-th block, and $\mathbf{y}_{N(k)}^{aug} = \mathbf{y}_{N(k)}$ otherwise. However, the prediction obtained by this approximation cannot yield a valid GP, because integrating out $\mathbf{y}_p$ cannot, in general, lead to $\tilde{p}(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\theta})$ in Theorem 1 (except for the special case where $l = K$, such that $\mathbf{y}_{N(k)}^{aug} = \mathbf{y}_{N(k)}$ for $k = 1, \ldots, K$, which corresponds to the proposed prediction method).

5. The SFSA Spatial Process

In this section, we show that the SFSA approach equipped with the prediction method in Section 4 yields a valid spatial GP with a closed-form covariance function. Therefore, both the parameter estimation and the prediction of the SFSA can be performed in a unified GP framework. Recall that in Section 2.2 we showed that the underlying spatial process $w(\mathbf{s})$ can be decomposed into two independent processes $w_l(\mathbf{s})$ and $w_s(\mathbf{s})$, where $w_l(\mathbf{s})$ is the predictive process with covariance function $\mathcal{C}_l(\cdot,\cdot)$, and $w_s(\mathbf{s})$ is the exact residual process with covariance function $\mathcal{C}(\cdot,\cdot) - \mathcal{C}_l(\cdot,\cdot)$. Let $\tilde{w}_s(\mathbf{s}) = w_s(\mathbf{s}) + \epsilon(\mathbf{s})$ be the new residual process that incorporates the measurement-error term. Then, the data process is

$$y(\mathbf{s}) = x^T(\mathbf{s})\boldsymbol{\beta} + w_l(\mathbf{s}) + \tilde{w}_s(\mathbf{s}).$$

In the following, we show that the SFSA approximates the process $\tilde{w}_s(\mathbf{s})$ using the nearest neighboring block GP, thus extending the nearest-neighbor GP developed in Datta et al. (2016). Hence, the process approximated by the SFSA approach is a valid GP.

Given a partition rule $\mathcal{P}$ leading to $S = \cup_{k=1}^K S_k$, the key assumption when deriving the likelihood of the SFSA approach is

$$\tilde{p}(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{w}^*) = \prod_{k=1}^K p(\mathbf{y}_k|\mathbf{y}_{N(k)}, \boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{w}^*),$$

which is equivalent to

$$\tilde{p}(\tilde{w}_s(S)|\boldsymbol{\theta}) = \prod_{k=1}^K p(\tilde{w}_s(S_k)|\tilde{w}_s(S_{N(k)}), \boldsymbol{\theta}). \tag{5.1}$$

Let $\mathcal{P}$ also partition a set of predictive locations S_p into K disjoint blocks $S_{p,k}$, for $k = 1, \ldots, K$. The assumption in Section 4,

$$\tilde{p}(\mathbf{y}_p|\mathbf{y}, \mathbf{w}^*, \boldsymbol{\theta}) = \prod_{k=1}^K p(\mathbf{y}_{p,k}|\mathbf{y}_k, \mathbf{y}_{N(k)}, \mathbf{w}^*, \boldsymbol{\theta}),$$

is equivalent to

$$\tilde{p}(\tilde{w}_s(S_p)|\tilde{w}_s(S), \boldsymbol{\theta}) = \prod_{k=1}^K p(\tilde{w}_s(S_{p,k})|\tilde{w}_s(S_k), \tilde{w}_s(S_{N(k)}), \boldsymbol{\theta}). \tag{5.2}$$

Note that assumptions (5.1) and (5.2) are the block versions of the key assumptions for the nearest-neighbor GP defined on $\tilde{w}_s(\mathbf{s})$.

Consider an arbitrary set of locations $S_v \subset \mathcal{S}$. Let $S_p = S_v \setminus S$ be the subset of S_v that is outside of S (predictive locations). We define

$$\tilde{p}(\tilde{w}_s(S_v)|\boldsymbol{\theta}) = \int \tilde{p}(\tilde{w}_s(S_p)|\tilde{w}_s(S), \boldsymbol{\theta})\tilde{p}(\tilde{w}_s(S)|\boldsymbol{\theta}) \prod_{\mathbf{s}_i \in S \setminus S_v} d\tilde{w}_s(\mathbf{s}_i), \tag{5.3}$$

where $\tilde{p}(\tilde{w}_s(S_p)|\tilde{w}_s(S), \boldsymbol{\theta})$ has the expression in (5.2) and $\tilde{p}(\tilde{w}_s(S)|\boldsymbol{\theta})$ has the expression in (5.1). The following theorem shows that the approximated process with finite-dimensional densities defined in (5.3) is a valid GP.

Theorem 2. *Let $\tilde{w}_s^\dagger(\mathbf{s})$ be the constructed process with the finite-dimensional distribution defined in* (5.3)*. Then, $\tilde{w}_s^\dagger(\mathbf{s})$ is a valid GP with a covariance function defined as*

$$
\tilde{\mathcal{C}}_s^{\dagger}(\mathbf{s},\mathbf{s}') = \begin{cases} \Sigma_{\mathbf{y}}^{\dagger}(\mathbf{s},\mathbf{s}'), & \text{if } \mathbf{s},\mathbf{s}' \in S; \\ -B_{\mathbf{s}}\Sigma_{\mathbf{y}}^{\dagger}(S,\mathbf{s}'), & \text{if } \mathbf{s} \notin S,\ \mathbf{s}' \in S; \\ B_{\mathbf{s}}\Sigma_{\mathbf{y}}^{\dagger}B_{\mathbf{s}'}^T, & \text{if } \mathbf{s},\mathbf{s}' \notin S, \text{and } \mathbf{s},\mathbf{s}' \\ & \text{belong to different blocks;} \\ B_{\mathbf{s}}\Sigma_{\mathbf{y}}^{\dagger}B_{\mathbf{s}'}^T + \Sigma_{p,k|N(k)}(\mathbf{s},\mathbf{s}'), & \text{if } \mathbf{s},\mathbf{s}' \notin S, \text{and } \mathbf{s},\mathbf{s}' \text{ belong} \\ & \text{to the same block } k, \end{cases} \tag{5.4}
$$

where $B_{\mathbf{s}}$ and $B_{\mathbf{s}'}$ are defined similarly to B_p in Section 4.3*, under the special scenario that the predictive location set $S_p = \{\mathbf{s}\}$ or $\{\mathbf{s}'\}$. In addition, $\Sigma_{\mathbf{y}}^{\dagger} \equiv B^{-1}\Sigma_{con}B^{T^{-1}}$ is the approximated residual covariance matrix in Theorem* 1*; $\Sigma_{p,k|N(k)}$ is the residual variance of $\tilde{w}_s(S_{p,k})$, conditional on its neighbors in $\tilde{w}_s(S)$; and $\Sigma_{\mathbf{y}}^{\dagger}(S_1,S_2)$ and $\Sigma_{p,k|N(k)}(S_1,S_2)$ denote the sub-matrices of $\Sigma_{\mathbf{y}}^{\dagger}$ and $\Sigma_{p,k|N(k)}$ for corresponding location sets S_1 and S_2, respectively.*

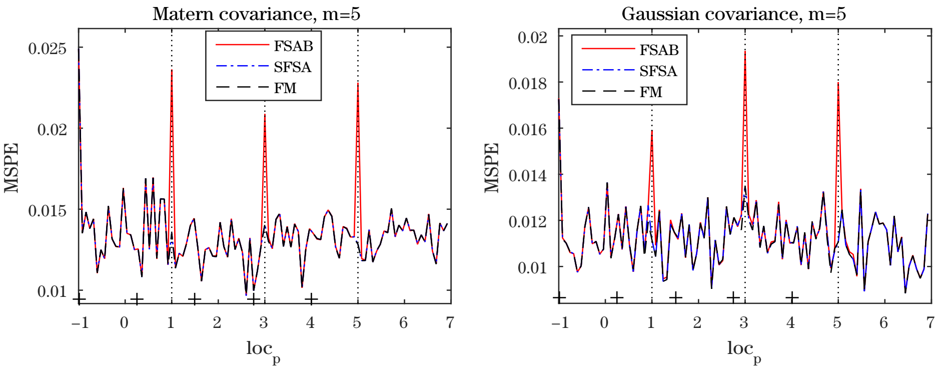
The proof of Theorem 2 basically follows Datta et al. (2016) (see the Supplementary Material, Section S1.3). Now, adding the predictive process covariance-function part, the covariance function of the SFSA GP is

$$
\mathcal{C}^{\dagger}(\mathbf{s},\mathbf{s}') = \mathcal{C}_l(\mathbf{s},\mathbf{s}') + \tilde{\mathcal{C}}_s^{\dagger}(\mathbf{s},\mathbf{s}'). \tag{5.5}
$$

Utilizing the finite-dimensional distribution given in Theorem 2, we can recover the conditional distribution expression given in Proposition 1 using the properties of multivariate Gaussian distributions. Specifically, following the results in (5.4) and (5.5), the approximated cross-covariance between the prediction set S_p and the training set S is $(U_pC_*U^T - B_p\Sigma_{\mathbf{y}}^{\dagger})$; the usual kriging formula yields the conditional mean and the conditional variance of $\mathbf{y}_p$, given $\mathbf{y}$, presented in Proposition 1.

6. Numerical Examples

In this section, we illustrate the effectiveness of our method by means of simulations. The implementations of the NNGP, SFSA, and FSA-Block were written in MATLAB. We used the R package “laGP” to obtain the results of the local GP method with adaptive local designs (Gramacy and Apley (2015)). All methods were run on an AMD Opteron (tm) processor with 2.3 GHz CPUs and 32 GB memory. For the log-likelihood function optimization, we used the matlab function, fminunc, which implements a Broyden–Fletcher–Goldfarb–Shanno (BFGS)-based quasi-Newton method. We used the parfor command in MATLAB for parallel computations.



(a) Matérn covariance

(b) Gaussian covariance

Figure 3. MSPEs versus the predictive locations. Crosses denote the locations of knots and the dotted lines indicate the block boundaries. The results were obtained based on 200 simulated data sets.

6.1. Simulation studies

We use the following example to show that, compared with the FSA-Block approach, the SFSA approach with $q \geq 1$ can alleviate the prediction errors around block boundaries. We generated 500 data observations from a GP with mean zero and Matérn covariance function in (2.2) on an equally spaced grid in the domain $\mathcal{S} \equiv [-1, 7]$. Then, predictions were performed at 100 equally spaced locations in $\mathcal{S}$, and the rest of data were used for training. For both the FSA-Block and SFSA approaches, we partitioned $[-1, 7]$ equally to create $K = 4$ blocks, with five knots placed equally on $[-1, 4]$. Therefore, the block boundaries are $s = 1, 3, 5$, and there are no knots close to the boundary $s = 5$. For the SFSA, we set $q = 1$ and $N(k) = \{k - 1\}$, for $k = 2, 3, 4$. We experimented with both the Matérn covariance function, with $\sigma^2 = 1, \nu = 1.5, \phi = 0.2$, and $\tau^2 = 0.01$, and the Gaussian covariance function, with $\sigma^2 = 1, \phi = 0.1$, and $\tau^2 = 0.01$. The parameter settings correspond to smooth GP processes with relatively small dependence ranges.

Figure 3 plots the mean squared prediction errors (MSPE) against the predictive locations. For the Matérn-covariance case (left panel), the MSPEs by the FSA-Block approach are particularly large around the block boundaries ($s = 1, 3, 5$). In contrast, the SFSA approach reduces the prediction errors around the block-boundary points by borrowing dependence information from neighboring blocks. Thus, the MSPEs of the SFSA are almost indistinguishable from those of the full model. Similar conclusions hold for the Gaussian-covariance case (right panel). Thus, for a smooth spatial process with a relatively small dependence range, the SFSA approach with $q \geq 1$ is preferred, because it significantly alleviates discontinuities of predictions around block boundaries.

6.2. Application to a total column ozone data set

In this section, we analyze the total column ozone (TCO) level 2 data set collected on October 1, 1988 (previous analyses of this data set can be found in Cressie and Johannesson (2008); Eidsvik et al. (2014)). This TCO level 2 data set has $n = 173,405$ observations. We partitioned the data into a training set and a prediction set under two prediction scenarios: 1) predictions on $25,000$ randomly selected locations (MAR); and 2) predictions on locations in a hold-out 15 degree × 15 degree rectangular region (MBD), which consists of around 600 predictive locations. For both prediction scenarios, we randomly generated three sets to evaluate the prediction performance of the various methods.

Following the analysis in Eidsvik et al. (2014), we used a fixed mean parameter and a Cauchy covariance function $\mathcal{C}(\mathbf{s}, \mathbf{s}') = \sigma^2(1+\|\mathbf{s}-\mathbf{s}'\|/\phi)^{-3}$ with a nugget effect to model the TCO data set, where $\sigma^2 > 0$ and $\phi > 0$ are the variance and range parameters, respectively. We also considered a Matérn covariance function (see (2.2)) with the smoothness ν fixed at one, as suggested by a pilot study using the full covariance model on $10,000$ randomly selected observations. The constant mean was removed before estimating the covariance function parameters. We compare the SFSA with the FSA-Block, NNGP, and LaGP methods in terms of their prediction performance, considering the MSPE and the mean continuous rank probability score (CRPS) (e.g., see Gneiting and Raftery (2007)). For the SFSA and FSA-Block, we used 24×24 regular blocks and 225 regular-grid knots such that both the block size n_b and the knot size m are around 200. For the SFSA, we applied the sorted-coordinates (SC) ordering for the MAR scenario and the center-out (CO) ordering for the MBD scenario. We specified the number of neighboring blocks as $q = 1$. For the NNGP and LaGP, 50 neighbors were used for both the parameter estimation and the prediction. For the LaGP, the "mspe" heuristic was considered.

Table 2 shows the prediction results for the various methods. We focus on the results of the Matérn covariance, because, in general, it leads to better MSPE results than the Cauchy covariance does, except for the SFSA under the MBD scenario. For the MAR scenario testing the small-range predictions, the NNGP performs best, with slightly smaller MSPE and CRPS values than those of the SFSA. However, for the MBD scenario, the SFSA method results in the best

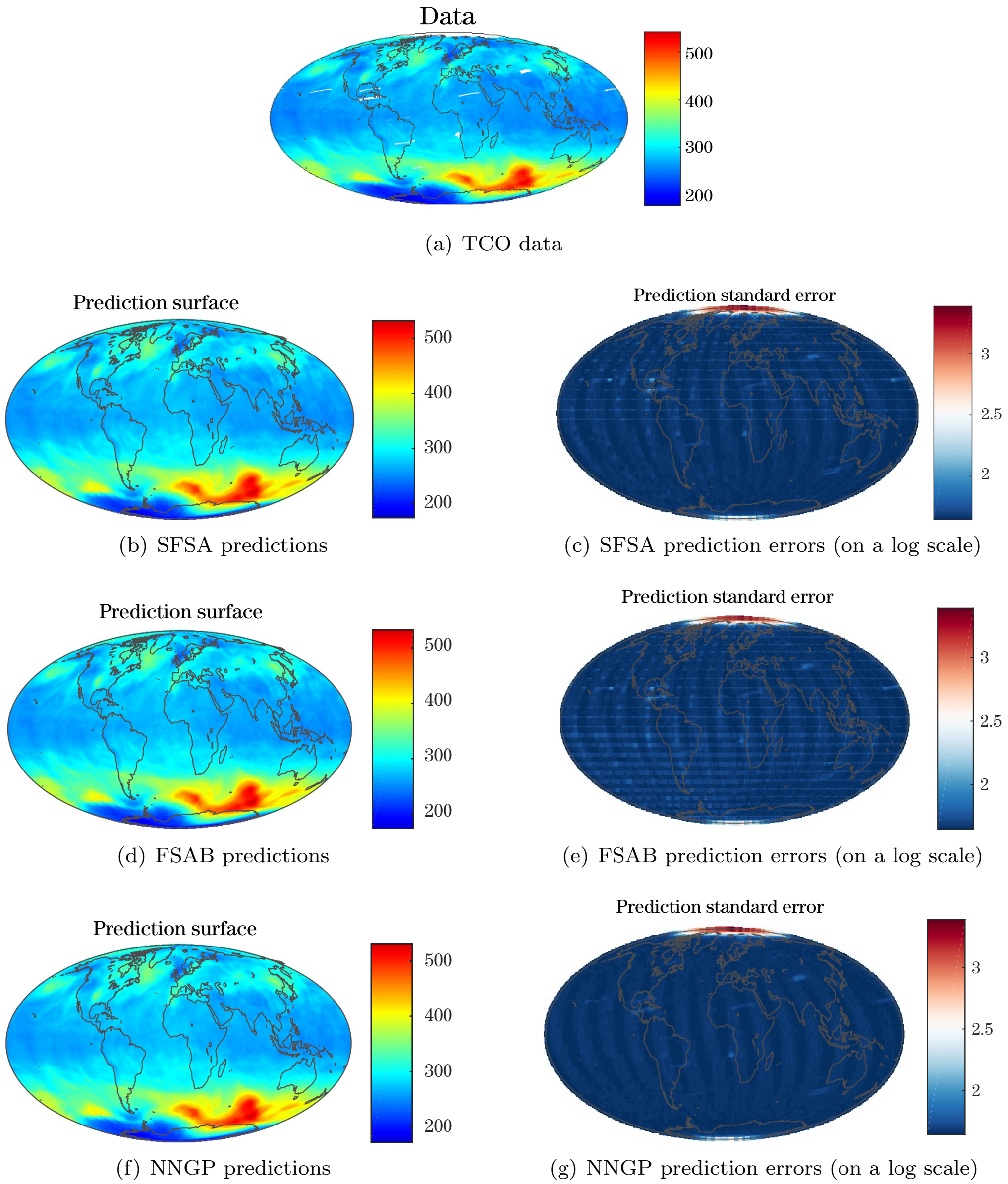
Table 2. Prediction performance of the SFSA, FSA-Block, NNGP, and LaGP for the TCO data. The results were obtained based on three prediction sets for each prediction scenario.

Scenarios		SFSA		FSA-Block		NNGP		LaGP-mspe
		Matérn	Cauchy	Matérn	Cauchy	Matérn	Cauchy	
MAR	MSPE	27.06	27.77	27.24	27.98	**26.67**	27.43	38.03
	CRPS	2.51	2.53	2.53	2.55	**2.50**	2.52	3.78
	Time (min)	58	33	47	31	57	27	121
MBD	MSPE	16.73	**16.32**	21.46	24.26	21.77	23.88	23.47
	CRPS	2.75	**2.54**	2.91	2.90	2.88	2.85	3.31
	Time (min)	79	27	72	31	100	38	4

prediction results. The NNGP results in larger MSPE and CRPS values than those of the SFSA for the MBD scenario, which may be because the correlations of the TCO data have a relatively large scale. As a result, borrowing information from non-neighboring locations helps to improve the prediction accuracy. Compared with the other methods, the LaGP leads to much larger prediction errors (especially for the MAR scenario), which may be because its methodology is developed based on the Gaussian covariance, which is too smooth for modeling the TCO data set. Using other covariance functions (e.g., the Cauchy or Matérn covariance functions) may improve its prediction performance significantly, but the current LaGP package does not support this.

For the computational times (including both the parameter-estimation and prediction steps), the SFSA, FSA-Block and NNGP have comparable speeds. Compared with the other methods, the LaGP has a much longer computational time for the MAR scenario, but a much shorter time for the MBD scenario. The reason is that its computational time depends mainly on the total number of the predictive locations. In contrast, for the SFSA, FSA-Block, and NNGP, the computational bottleneck lies in the parameter-estimation step, rather than the prediction step, because the high-dimensional likelihood of the training data needs to be evaluated repeatedly by the optimization function in the parameter-estimation step.

The prediction plots on a 288 × 180 longitude–latitude regular grid using the Matérn covariance are shown in the left column of Figure 4. The SFSA, FSA-Block, and NNGP produce very similar prediction surfaces, owing to their comparable capability for short-range predictions. Their associated prediction standard errors (on a log scale) are shown in the right column. We can observe that the prediction standard errors are particularly large for regions without



(a) TCO data

(b) SFSA predictions

(c) SFSA prediction errors (on a log scale)

(d) FSAB predictions

(e) FSAB prediction errors (on a log scale)

(f) NNGP predictions

(g) NNGP prediction errors (on a log scale)

Figure 4. Prediction surfaces and prediction standard errors (on a log scale) for the SFSA, FSA-Block, and NNGP.

any observations. For the SFSA and FSA-Block, relatively larger prediction standard errors are observed around the block boundaries. However, this effect is less evident for the SFSA than it is for the FSA-Block.

7. Discussion

We have proposed the SFSA approach that extends the FSA-Block approach by correcting the approximation errors of the covariance between each block and its neighboring blocks. We prove that the SFSA approach yields a class of valid GP models, such that both the parameter estimation and the prediction of the SFSA can be performed within a unified framework. The proposed method incorporates the FSA-Block approach and the block conditional composite likelihood approach as special cases. Hence, it achieves better statistical efficiency. Compared with the FSA-Block, the SFSA reduces prediction errors at locations around block boundaries, which helps to produce a smoother prediction surface.

A natural extension of the proposed method is to the spatio-temporal setting (Katzfuss and Cressie (2011); Bevilacqua et al. (2012); Zhang, Sang and Huang (2015)), where we consider a spatio-temporal partition of observations and define the neighboring blocks in space and time. In this case, the Euclidean distance between spatio-temporal locations may not be a good way to identify neighbors. We will explore using other measures to define the block partition and neighboring blocks that minimize the residual covariance for nonneighboring blocks in order to improve the approximation accuracy.

For modeling non-Gaussian observations from the exponential family of distributions, the SFSA can be embedded in hierarchical spatial generalized linear models (GLM) (e.g., Diggle, Tawn and Moyeed (1998); Banerjee, Gelfand and Carlin (2014)) to speed up computations. The spatial GLM proposed by Diggle, Tawn and Moyeed (1998) for modeling non-Gaussian spatially dependent observations involves two stages. In the first stage, the data conditional on a latent spatial process are independent and identically distributed exponential family random variables. In the second stage, the latent spatial process is modeled as a GP, with both fixed and random effects. For this modeling strategy, the SFSA approximation is applied to $\eta(\mathbf{s}) \equiv g(E(y(\mathbf{s})|\eta(\cdot))) = \mathbf{x}(\mathbf{s})^T\boldsymbol{\beta} + w(\mathbf{s}) + \epsilon(\mathbf{s})$ in the second stage, where $g(\cdot)$ is a link function, $y(\cdot)$ is the data process, and $\eta(\cdot)$ is the latent spatial process. Similarly to the Gaussian case, we approximate $w(\mathbf{s})$ using the process induced by the SFSA, denoted by $w^{\dagger}(\mathbf{s})$, to facilitate computations for evaluating the joint likelihood function. However, the marginalized likelihood that integrates out the latent spatial process $\eta(\cdot)$ does not have an analytical form. Hence, MCMC algorithms need to be employed to obtain posterior samples of model parameters $\boldsymbol{\theta}$, along with $\eta(\mathbf{s})$. Alternatively, the EM algorithm can be used to estimate the model parameters for the spatial GLM (e.g., see Sengupta and Cressie (2013)).

Supplementary Material

The supplementary material contains the proofs of the theorems, and additional numerical results used to compare the SFSA with other methods.

Acknowledgments

The research of Huiyan Sang is partially supported by NSF CNS-1343155, NSF DMS-1622433, and NSF DMS-1737885. The research of Jianhua Z. Huang is partially supported by NSF grant DMS-1208952. The research of Bohai Zhang was partially supported by an Australian Research Council Discovery Project, number DP150104576, and is partially supported by the Key Laboratory of Pure Mathematics and Combinatorics of Ministry of Education (LPMC) and the Key Laboratory for Medical Data Analysis and Statistical Research of Tianjin (KLMDASR).

References

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

Banerjee, S., Gelfand, A. E. and Carlin, B. P. (2014). *Hierarchical Modeling and Analysis for Spatial Data*. CRC Press.

Bevilacqua, M., Gaetan, C., Mateu, J. and Porcu, E. (2012). Estimating space and space-time covariance functions for large data sets: A weighted composite likelihood approach. *Journal of the American Statistical Association* **107**, 268–280.

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

Datta, A., Banerjee, S., Finley, A. O. and Gelfand, A. E. (2016). Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association* **111**, 800–812.

Diggle, P. J., Tawn, J. and Moyeed, R. (1998). Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **47**, 299–350.

Eidsvik, J., Shaby, B. A., Reich, B. J., Wheeler, M. and Niemi, J. (2014). Estimation and prediction in spatial models with block composite likelihoods. *Journal of Computational and Graphical Statistics* **23**, 295–315.

Finley, A., Sang, H., Banerjee, S. and Gelfand, A. (2009). Improving the performance of predictive process modeling for large datasets. *Comput. Statist. Data Anal.* **53**, 2873–2884.

Furrer, R., Genton, M. and Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics* **15**, 502–523.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2014). *Bayesian Data Analysis*, Volume 2. Chapman & Hall/CRC Boca Raton, FL, USA.

Gneiting, T. (2002). Compactly supported correlation functions. *Journal of Multivariate Analysis* **83**, 493–508.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102**, 359–378.

Gramacy, R. B. and Apley, D. W. (2015). Local Gaussian process approximation for large computer experiments. *Journal of Computational and Graphical Statistics* **24**, 561–578.

Gramacy, R. B. and Haaland, B. (2016). Speeding up neighborhood search in local gaussian process prediction. *Technometrics* **58**, 294–303.

Gramacy, R. B. and Lee, H. K. H. (2008). Bayesian treed Gaussian process models with an application to computer modeling. *Journal of the American Statistical Association* **103**, 1119–1130.

Guhaniyogi, R. and Banerjee, S. (2017). Meta-kriging: Scalable Bayesian modeling and inference for massive spatial datasets.

Guhaniyogi, R., Finley, A., Banerjee, S. and Gelfand, A. (2011). Adaptive gaussian predictive process models for large spatial datasets. *Environmetrics* **22**, 997–1007.

Guinness, J. (2018). Permutation methods for sharpening Gaussian process approximations. *Technometrics* To appear.

Higdon, D. (2002). Space and space-time modeling using process convolutions. In *Quantitative Methods for Current Environmental Issues*, pp. 37–56. Springer.

Katzfuss, M. (2013). Bayesian nonstationary spatial modeling for very large datasets. *Environmetrics* **24**, 189–200.

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

Katzfuss, M. and N. Cressie (2011). Spatio-temporal smoothing and em estimation for massive remote-sensing data sets. *Journal of Time Series Analysis* **32**, 430–446.

Kaufman, C. G., Schervish, M. J. and Nychka, D. W. (2008). Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association* **103**, 1545–1555.

Konomi, B., Karagiannis, G., Sarkar, A., Sun, X. and Lin, G. (2014). Bayesian treed multivariate Gaussian process with adaptive design: Application to a carbon capture unit. *Technometrics* **56**, 145–158.

Lee, D.-T. and Schachter, B. J. (1980). Two algorithms for constructing a delaunay triangulation. *International Journal of Computer & Information Sciences* **9**, 219–242.

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 (Statistical Methodology)* **73**, 423–498.

Ma, P. and Kang, E. L. (2017). Fused Gaussian process for very large spatial data. *arXiv Preprint ArXiv:1702.08797*

McKay, M. D., Conover, W. J. and Beckman, R. J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* **21**, 239–245.

Nguyen, H., Cressie, N. and Braverman, A. (2012). Spatial statistical data fusion for remote sensing applications. *Journal of the American Statistical Association* **107**, 1004–1018.

Nguyen, H., Katzfuss, M., Cressie, N. and Braverman, A. (2014). Spatio-temporal data fusion for very large remote sensing datasets. *Technometrics* **56**, 174–185.

Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F. and Sain, S. (2015). A multiresolution Gaussian process model for the analysis of large spatial datasets. *Journal of Computational and Graphical Statistics* **24**, 579–599.

Park, C. and Apley, D. (2018). Patchwork kriging for large-scale Gaussian process regression. *The Journal of Machine Learning Research* **19**, 269–311.

Rue, H. and Tjelmeland, H. (2002). Fitting Gaussian Markov random fields to Gaussian fields. *Scandinavian Journal of Statistics* **29**, 31–49.

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 (Statistical Methodology)* **74**, 111–132.

Sang, H., Jun, M. and Huang, J. (2011). Covariance approximation for large multivariate spatial data sets with an application to multiple climate model errors. *The Annals of Applied Statistics* **5**, 2519–2548.

Sengupta, A. and Cressie, N. (2013). Hierarchical statistical modeling of big spatial datasets using the exponential family of distributions. *Spatial Statistics* **4**, 14–44.

Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging.* Springer Science & Business Media, New York.

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

Stein, M. L., Chi, Z. and Welty, L. J. (2004). Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **66**, 275–296.

Sun, Y., Li, B. and Genton, M. G. (2012). Geostatistics for Large Datasets. In *Advances and Challenges in Space-time Modelling of Natural Events*, 55–77. Springer.

Vecchia, A. V. (1988). Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **50**, 297–312.

Zhang, B., Sang, H. and Huang, J. Z. (2015). Full-scale approximations of spatio-temporal covariance models for large datasets. *Statistica Sinica* **25**, 99–114.

Zhang, R., Lin, C. D. and Ranjan, P. (2018). Local Gaussian process model for large-scale dynamic computer experiments. *Journal of Computational and Graphical Statistics* **27**, 798–807.

School of Statistics and Data Science, Nankai University, Tianjin 300071, China.

E-mail: bohaizhang@nankai.edu.cn

Department of Statistics, Texas A&M University, College Station, TX 77840, USA.

E-mail: huiyan@stat.tamu.edu

Department of Statistics, Texas A&M University, College Station, TX 77840, USA.

E-mail: jianhua@stat.tamu.edu

(Received January 2017; accepted December 2017)