Statistica Sinica Preprint No: SS-2017-0008

Title	Smoothed Full-Scale Approximation of Gaussian Process Models for Computation of Large Spatial Datasets
Manuscript ID	SS-2017-0008
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202017.0008
Complete List of Authors	Bohai Zhang Huiyan Sang and Jianhua Z. Huang
Corresponding Author	Bohai Zhang
E-mail	bohai@uow.edu.au

Notice: Accepted version subject to English editing.

Smoothed Full-Scale Approximation of Gaussian Process Models for Computation of Large Spatial Datasets

Bohai Zhang, Huiyan Sang and Jianhua Z. Huang

University of Wollongong and Texas A&M University

Abstract: Gaussian process (GP) models encounter computational difficulties with large spatial datasets since its computational complexity grows cubically with sample size n. Although the Full-Scale Approximation (FSA) using a block modulating function provides an effective way for approximating GP models, it has several shortcomings such as the less smooth prediction surface on block boundaries and sensitiveness to the knot set under small-scale data dependence. To address these issues, we propose a Smoothed Full-Scale Approximation (SFSA) method for the analysis of large spatial dataset. The SFSA leads to a class of scalable GP models, whose covariance functions consist of two parts: A reduced-rank covariance function capturing large-scale spatial dependence and a covariance adjusting local covariance approximation errors of the reduced-rank part both within blocks and between neighboring blocks. This method can alleviate the prediction errors on block boundaries; it also leads to more robust inference and prediction results under different dependence scales due to better approximation of the residual covariance. The proposed method provides a unified view of approximation methods for GP models, encompassing several existing computational methods for large spatial datasets into one common framework, including the predictive process, the FSA, and the nearest neighboring block Gaussian process methods, allowing efficient algorithms for more robust and accurate model inference and prediction for large spatial datasets in a unified framework. We illustrate the effectiveness of the SFSA approach through simulation studies and a total column ozone dataset.

Key words and phrases: Conditional likelihood, full-scale approximation, Markov chain Monte Carlo, spatial covariance functions.

1. Introduction

Spatial datasets arising from ecology, climatology, and other disciplines have generated considerable interests for scientists. With the advent of remote sensing and Geographic Information System (GIS) techniques, the spatial data collection capacity increases dramatically, and statisticians nowadays are often facing a large number of observations on variables of interest. The growth in data volume imposes computational challenges to classical geostatistical models (Stein (1999); Banerjee et al. (2014)) and has driven the innovations of computational methods scalable to handle large datasets (e.g., Sun et al. (2012)).

One of the most popular models for spatial datasets is the Gaussian process (GP) model, assuming that finite observations are jointly Gaussian. Although GP models enjoy mathematical tractability for model fitting and prediction, its computational complexity generally grows cubically with sample size n, due to 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), making model fitting of a GP model computationally prohibitive for very large n.

Recently, Sang and Huang (2012) proposed a so-called Full-Scale Approximation (FSA) approach to approximate the original covariance function of GP models for large spatial datasets. By combining the ideas of both low-rank models and sparse models, it can approximate the data covariance matrix well under 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 et al. (2012, 2014)) seek to approximating the original spatial process by 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 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 for yielding a sparse covariance matrix (Furrer et al. (2006); Kaufman et al. (2008)), or assume Gaussian-Markov property of the spatial random field for yielding a sparse precision matrix (Rue and Tjelmeland (2002); Lindgren et al. (2011)). Another way for inducing sparsity of the precision matrix is to use conditional likelihoods (e.g., Vecchia (1988)), and most recently, Datta et al. (2016) proposed a nearest-neighbor GP (NNGP); a new permutation and grouping method for improving the performance of NNGP can be found in Guinness (2016). Most recent "hybrid" methods extending the low-rank models include Nychka et al. (2015); Katzfuss (2017); Ma and Kang (2017). The modern-version local GP models (e.g., Gramacy and Apley, 2015; Gramacy and Haaland, 2016; Zhang et al., 2016; Park and Apley, 2017) can also be applied very effectively to modeling large or massive spatial data. Lastly, the divide and conquer based approaches have also been proposed to model large and nonstationary spatial datasets. See, e.g., the treed GP (e.g., Gramacy and Lee, 2008; Konomi et al., 2014) and the spatial meta kriging (Guhaniyogi and Banerjee, 2017).

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 semi-definite function. The covariance function of 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 semi-definite and has a large number of zeros evaluated on observed spatial locations. If we choose $\mathcal{K}(\cdot,\cdot)$ to be 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*. It turns out that FSA-Block outperforms FSA-Taper empirically (Sang et al. (2011)), possibly due to its nature of being an unbiased approximation of covariance within each data block, and the convenience of using parallel computation. Zhang et al. (2015) extends the FSA-Block approximation to GP models for large spatio-temporal datasets.

Although the FSA-Block approach can lead to effective and scalable approximation to the covariance function of GP models, it has several shortcomings. First, the predictions around boundaries of two adjacent blocks by the FSA-Block approach are less smooth than the rest of the regions, mainly due to the independent-blocks assumption for the residual covariance function, $\mathcal{C}_s$; the mismatches of predictions on block boundaries can result in large prediction errors on locations close to block boundaries. Second, although the overall performance of the FSA-Block is more robust than that of the predictive process and independent block estimations to the choice of knots and blocks, 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 knot number is insufficient), leaving room for further improvement.

In this paper 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 propose to apply it to approximate the residual covariance and combine this approximated residual covariance with a reduced-rank predictive process. By doing this, we relax the independent-blocks assumption in FSA-Block to further account for the dependence between each block and its neighboring blocks in the residual covariance matrix. The proposed method can alleviate the discontinuities of predictions on boundary locations for the FSA-Block approach. We name the new proposed method the Smoothed Full-Scale Approximation (SFSA).

We further show that the SFSA approach defines a class of valid Gaussian process models scalable to large datasets. Therefore, both parameter estimation and prediction by the SFSA approach can be readily done under a unified framework due to the existence of a closed-form covariance function. The establishment of the SFSA Gaussian process also allows it being flexibly embedded into hierarchical spatial models to facilitate computation while maintaining model richness.

The SFSA also provides a unifying view of approximations for spatial GP models, putting various existing popular approximation methods under one umbrella, including the predictive process, the FSA, the conditional composite likelihood, the independent blocks method, and the nearest-neighbor GP approximations. This unified modeling framework enables direct comparison and reveals the relation among various computational methods for large spatial datasets.

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 SFSA and gives the algorithm for evaluating its log-likelihood. Section 4 gives details on the parameter-estimation and prediction procedures of SFSA. Section 5 defines the valid GP constructed from the SFSA. We compare SFSA with other state-of-the-art methods through simulation studies in Section 6.1 and a total column ozone dataset in Section 6.2. Finally Section 7 concludes the paper with a brief summary and discussions of some potential extensions. The proof of theorems and additional numerical results are given in the supplementary materials.

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 Gaussian process, 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, $\mathcal{C}(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta}) \equiv \mathrm{cov}(w(\mathbf{s}), w(\mathbf{s}'))$. For example, the Matérn covariance function (e.g., see Stein (1999)) is widely used in spatial statistics due to its flexibility for modeling different smoothness of a spatial process:

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

where $\sigma^2 > 0$ is the variance parameter, $\phi > 0$ is the dependence range parameter, and $\nu > 0$ 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. Since evaluating (2.3) in general requires $\mathcal{O}(n^3)$ flops for calculating $|C_{\mathbf{y}}|$ and $C_{\mathbf{y}}^{-1}$, the computational cost can be very expensive or even prohibitive when n is very large.

2.2 The FSA-Block approach

In this subsection, we will briefly review the FSA-Block approach. The FSA-Block approach (Sang et al. (2011); Sang and Huang (2012)) is motivated from 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 $w(\mathbf{s})$ by only using $w_l(\mathbf{s})$ will result in loss of residual covariance information in $w_s(\mathbf{s})$, which could subsequently lead to bias in parameter estimations and inaccuracy in 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 semi-definite otherwise. In general, C_{w_s} has a dependence structure of a smaller scale than the full covariance but it is still a dense matrix. Sang et al. (2011) proposed to approximate C_{w_s} by a block-diagonal matrix to reduce 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 one groups observations according to blocks, then the approximated likelihood 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} by the predictive process model, an additional block-diagonal residual covariance matrix is incorporated to correct the approximation errors within each data block. Since $(C_{w_s} \circ \mathcal{T}_B + \tau^2 I_n)$ is block-diagonal, it takes $\mathcal{O}(n)$ order flops for computing its inverse and determinant. It can be shown that the computational complexity of the FSA-Block approach is linear with n (Sang and Huang (2012)).

However, the independent-blocks approximation of C_{w_s} ignores the residual dependence across different blocks. The loss of dependence information can be severe when $w_l(\mathbf{s})$ does not provide a good approximation for $w(\mathbf{s})$ so that the entries across blocks of the residual covariance matrix are not negligible (e.g., the knots are not placed properly or the knot number is insufficient). More importantly, since 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, in particular around block boundary locations (see Section 6.1). To address these issues, we seek a new method that can partially preserve the entries of C_{w_s} across data blocks, while still maintaining the 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 $p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{w}^*)$, since C_{w_s} is a dense matrix in general. We propose to replace $p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{w}^*)$ with some Gaussian density whose computations are less expensive; then after integrating out $\mathbf{w}^*$, an approximated Gaussian likelihood with a reduced computational cost can be readily obtained. 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 in Stein et al. (2004) for approximating $p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{w}^*)$.

Specifically, let $\mathcal{P}$ be a partition rule leading to a partition $S = \cup_{k=1}^K S_k$ with the corresponding partition of observations $\mathbf{y} = \cup_{k=1}^K \mathbf{y}_k$, where S_k and $\mathbf{y}_k$ have a size of 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 a large k, because $\mathbf{y}_{(k-1)}$ is high dimensional. Thus, following Stein et al. (2004), we choose the conditional set to be 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 notation simplicity, in this paper, we focus on the special case that $S_{N(k)}$ contains all locations in the q-nearest neighboring blocks of the k-th block (e.g., "closeness" may be measured by the Euclidean distances of block centers). In specific, $S_{N(k)}$ is defined as

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

In practice, one chooses K to partition the data such that each data block contains only a few hundreds 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 one uses $\emptyset$ as the conditional set for every $\mathbf{y}_k$. Usually we choose $q \geq 1$ 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. We name the proposed approach the Smoothed FSA (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 notations $\Sigma_{A,B} \equiv \operatorname{cov}(w_s(S_A) + \epsilon(S_A), w_s(S_B) + \epsilon(S_B))$ 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 $k, l = 1, \ldots, K$, define

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

where $n_{(l)} = \sum_{1 \leq i \leq l, i \in N(k)} n_i$. Here $B_{k,l}$ is an n_k by n_l matrix encoding the conditional dependence information between the k-th block and the l-th block. Let $B^*_k = (B_{k,1}, \ldots, B_{k,K})$, then it can be shown that (in 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

$$|\Sigma_{k|N(k)}|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\mathbf{y} - \mathbf{x}\boldsymbol{\beta} - U\mathbf{w}^*)^T B^{*^T}_k \Sigma^{-1}_{k|N(k)} B^*_k(\mathbf{y} - \mathbf{x}\boldsymbol{\beta} - U\mathbf{w}^*)\},$$

where $\Sigma_{k|N(k)} = \Sigma_k - \Sigma_{k,N(k)}\Sigma^{-1}_{N(k)}\Sigma^T_{k,N(k)}$ 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* $\mathcal{N}(\mathbf{x}\boldsymbol{\beta}, C^{\dagger}_{\mathbf{y}})$*, where*

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

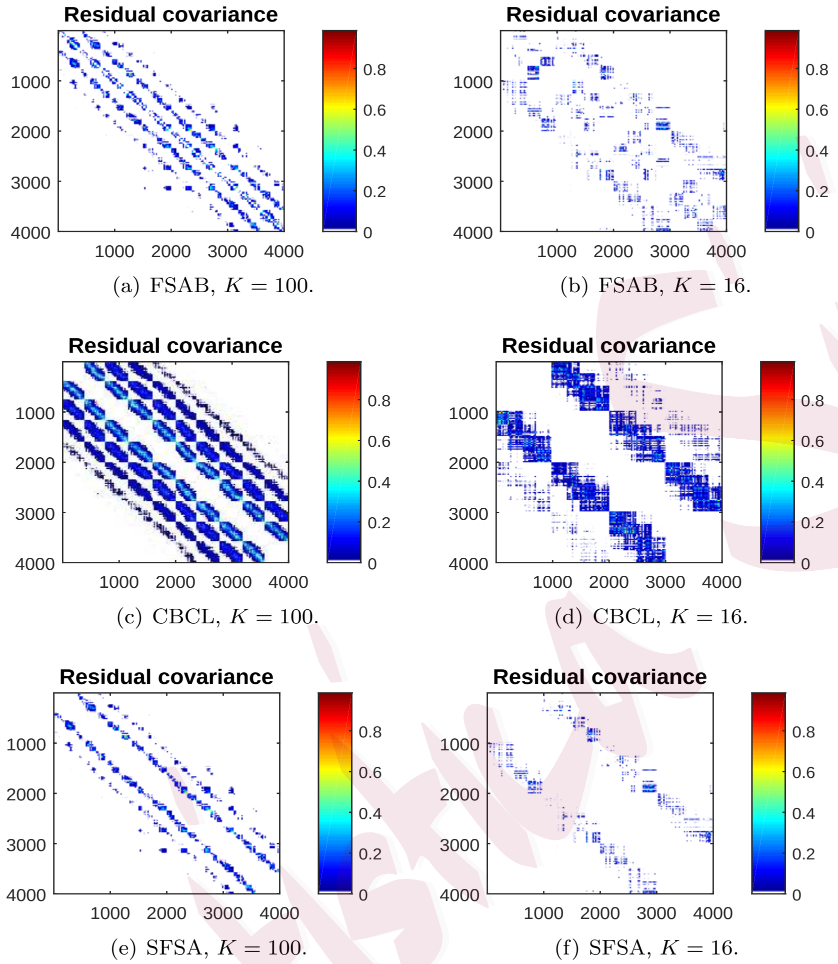
where Σ_{con} *is a block-diagonal matrix with* $\Sigma_{k|N(k)}$ *as its k-th block, and* $B = (B_1^{*^T}, \ldots, B_K^{*^T})^T \in \mathbb{R}^{n \times n}$.

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 $m = 0$), and hence includes both as special cases. Thus, 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 covariance matrix approximation. Figure 1 shows the absolute differences of entries between the approximated data covariance matrix and the original data covariance matrix, for each of these three approaches. Specifically, 4000 locations were randomly generated in a square domain $[0, 10] \times [0, 10]$, and the exponential covariance function, $\mathcal{C}(\mathbf{s}, \mathbf{s}') = \exp(-\|\mathbf{s} - \mathbf{s}'\|)$ with a nugget effect 0.01, was used to generate the covariance matrix on those locations. For all three approaches, equally spaced blocks were generated and the block index takes an increasing order from northwest to southeast; the locations within the same block were grouped together. For SFSA and FSA-Block approaches, $m = 50$ knots were uniformly selected in the square domain; and for SFSA and CBCL, the neighboring block set was the nearest neighboring block. We observe that for locations within a certain band, the approximation errors by SFSA are much smaller than those by the FSA-Block approach, due to the corrections of residual covariance between neighboring blocks. Compared with the CBCL approximation, while both approaches provide good approximations for covariance entries within a certain location band, the SFSA approach leads to smaller approximation errors for the residual covariance entries off the location band, due to the inclusion of the low-



(a) FSAB, $K = 100$.

(b) FSAB, $K = 16$.

(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; FSAB: the FSA-Block approximation; CBCL: the Conditional Block Composite Likelihood approximation.

rank predictive process component.

2.5 Choices of tuning parameters for SFSA

The SFSA approach requires specifications of several tuning parameters, including a knot set, a scheme of block partition, the ordering of data blocks, and the number of neighboring blocks q. For the knot set, random sampling, Latin Hypercube Sampling (McKay et al. (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 et al. (2015)). For the block partition, Eidsvik et al. (2014) provided some guidance on blocking strategy, and recommended to use the empirical variogram to determine the block width. The K-means clustering algorithm based on Euclidean distances of locations is a simple choice for creating blocks; alternatively, one may 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 the regular rectangle blocks (e.g., see Eidsvik et al. (2014); Katzfuss (2017)) which empirically work very well. In this paper we have focused on using regular rectangle blocks for implementation of our method. For highly non-uniformly distributed data, Delauny triangulation (e.g., see Lee and Schachter (1980)) might be a more effective partition method to create meshes for the proposed method.

After creating the blocks, it is necessary to order the blocks for constructing the residual likelihood of SFSA. Following Guinness (2016), we compared the model-fitting performances of SFSA for a few ordering methods, including the sorted-coordinate (SC) ordering, the random ordering, the maximum-minimum-distance (MMD) ordering, and the center-out (CO) ordering (see the supplementary material, Section S2.1). Based on the simulation-study results, we recommend using the SC ordering for uniformly spaced data and the CO ordering for non-uniformly spaced data.

Lastly, the selection of the number of neighboring blocks (q) is a trade-off between the computational time and the statistical efficiency. Apparently, the larger the number of neighboring blocks, the more accurate approximation the SFSA leads to. Based on the simulation results (see the supplementary material, Section S2.3), a small number of neighboring blocks such as $q = 3$ or 4 (with a few hundreds observations) can already lead to parameter-estimation results of good statistical efficiency. In this paper since we have focused on using regular rectangle blocks, the Euclidean distance between block centers becomes a natural choice to determine the q-nearest neighboring blocks. Alternatively, one may define the “closeness” of two blocks by a distance metric between the residual correlations of observations in two blocks. But apparently such an approach requires the estimation of residual correlation and will increase computational cost.

Table 1: Notations for SFSA.

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

More detailed discussion on finding the nearest neighboring blocks based on the residual correlations is provided in the supplementary material (Section S3).

3. Computational Aspects of the SFSA Approach

We first determine the computational complexity for evaluating the log-likelihood of 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 by SFSA, up to a constant, is (see equation (S1.1))

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

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, since we only need to 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. We remark that B is a sparse matrix with at most $(qn_b + 1)$ nonzero entries per row, and hence calculating BU is cheap with computational 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 the computational complexity of order $\mathcal{O}(q^3 n_b^3)$. Hence obtaining Σ_{con} has the order $\mathcal{O}(Kq^3n_b^3) = \mathcal{O}(nq^3n_b^2)$.

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 its inverse takes $\mathcal{O}(Kn_b^3) = \mathcal{O}(nn_b^2)$ flops; 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 by 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 $\mathcal{O}(nq^3n_b^2+nmqn_b+nm^2)$. In practice, the data is partitioned into K blocks such that each block has a block size of a few hundreds. If we also choose the knot size m to be a few hundreds and set $q \ll K$, the SFSA approach then has the computational complexity linear with n.

Parallel computation is possible for evaluating 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 $k = 1, \ldots, K$. Since Σ_{con} is a block-diagonal matrix with $\Sigma_{k|N(k)}$ as its k-th block, then

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

Similarly,

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

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 shows how to evaluate $\log \tilde{p}(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\theta})$ step by step. Parallel-computing technique can be applied to obtain the required quantities for each block simultaneously and hence avoid the loop in algorithm 1. By using K cores, the computational complexity of SFSA has the order $\mathcal{O}(q^3n_b^3 + mqn_b^2 + n_bm^2)$. Besides, since $U \in \mathbb{R}^{n\times m}$, it will require large memory to store U for very large n. We remark that due to 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 SFSA.

Algorithm 1 Evaluating the log-likelihood function of SFSA.

```
Compute $C_* = \mathcal{C}(S^*, S^*)$ and $U = \mathcal{C}(S, S^*)\mathcal{C}(S^*, S^*)^{-1}$. Factorize $C_* = Q_*^T Q_*$.
for $k = 1$ to $K$ do
    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 for
Sum up 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).
```

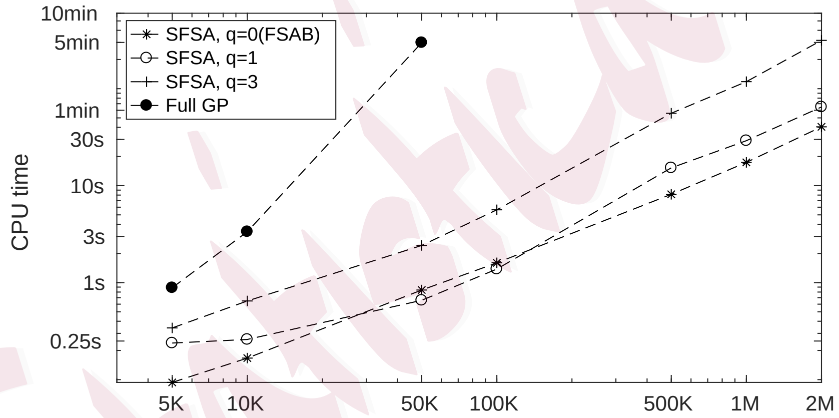


Figure 2: Computational times (on a log scale) per likelihood evaluation versus different sample sizes (on a log scale). For SFSA and FSAB, $m = n_b = 200$, and the results were obtained by using 16 CPU cores.

Figure 2 shows the computational time of SFSA for different sample sizes. We can see that for the data size of two millions, evaluating the likelihood of SFSA with $q = 1$ can still be done in a short time.

4. Parameter Estimation and Prediction

4.1 Maximum likelihood estimation

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

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

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

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

Calculation of $U^TB^T\Sigma_{con}^{-1}BU$ involves the multipulication of an $n \times n$ matrix B and an $n \times m$ matrix U. Recall that B is a sparse matrix with at most $(qn_b + 1)$ nonzero entries per row, and hence calculating BU is cheap with computational complexity $\mathcal{O}(nqn_bm)$. Then efficient computations are achieved, since calculating $C_{\mathbf{y}}^{\dagger}$ only involves computing the determinants of two $m \times m$ matrices and one block-diagonal matrix.

4.2 Bayesian inference on model parameters

The Bayesian inference starts from 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 a 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, 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; and for smoothness parameter ν, usually a uniform prior at $(0, 3]$ is used, since high values of smoothness can hardly be identified from real datasets.

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 $\Sigma_{\mathbf{y}} = C_{\mathbf{y}} + \mathbf{x}\Sigma_0\mathbf{x}^T$. Since 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 $p(\mathbf{y}|\boldsymbol{\theta})$ by using Metropolis-Hastings algorithm (Gelman et al. (2014)). Since $p(\boldsymbol{\beta}|\mathbf{y}) = \int_{\boldsymbol{\theta}} p(\boldsymbol{\beta}|\boldsymbol{\theta}, \mathbf{y})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$ and $p(\boldsymbol{\beta}|\boldsymbol{\theta}, \mathbf{y})$ is Gaussian, 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 $p(\mathbf{w}|\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\beta})$, for drawing 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 $S_{p,k}$, $k = 1, \ldots, K$. 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, since $p(\mathbf{y}_p|\mathbf{y}, \mathbf{w}^*, \boldsymbol{\beta}, \boldsymbol{\theta})$ 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,

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

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 rest of predictive responses, conditional on $\mathbf{w}^*$, the observations in the same block $\mathbf{y}_k$, and the observations in neighboring blocks $\mathbf{y}_{N(k)}$. Notice 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 recall that $S_{N(k)}$ 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 $B_{p,k,l}$ has the similar definition to $B_{k,l}$ in (2.8), encoding the residual conditional dependence information of $\mathbf{y}_{p,k}$ given its neighbors $y(S_{p,N(k)})$ for the l-th block, $l = 1, \ldots, K$. 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: There are different ways to approximate $p(\mathbf{y}_p, \mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\theta})$ for prediction. For example, consider the case in which the predictive response vector $\mathbf{y}_p$ belongs to some block l, then we can have an approximation of the augmented data likelihood, $\tilde{p}(\mathbf{y}_p, \mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\theta})$, defined as

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

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

5. The Smoothed FSA 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 parameter estimation and prediction of SFSA can be performed in a unified GP framework. Recall 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 SFSA approximates the process $\tilde{w}_s(\mathbf{s})$ by using the nearest neighboring block GP that extends the nearest neighbor GP developed in Datta et al. (2016), and hence the approximated process 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 on 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 $S_{p,k}$, $k = 1, \ldots, K$. 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

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

We remark that assumptions (5.1) and (5.2) are the block versions of 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 finite dimensional distribution defined in* (5.3)*. Then $\tilde{w}_s^{\dagger}(\mathbf{s})$ is a valid Gaussian process with 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 similarly defined as B_p in Section 4.3, under the special scenario that the predictive location set $S_p = \{\mathbf{s}\}$ or *$\{\mathbf{s}'\}$; $\Sigma_{\mathbf{y}}^{\dagger} \equiv B^{-1}\Sigma_{con}B^{T^{-1}}$ 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 $\tilde{w}_s(S)$; $\Sigma_{\mathbf{y}}^{\dagger}(S_1, S_2)$ 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 giving in Theorem 2, we can recover the conditional distribution expression given in Proposition 1 by 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 $(U_p C_* U^T - B_p \Sigma_{\mathbf{y}}^{\dagger})$; the usual kriging formula yields the conditional mean and the conditional variance of $\mathbf{y}_p$ given $\mathbf{y}$ presented in Proposition 1.

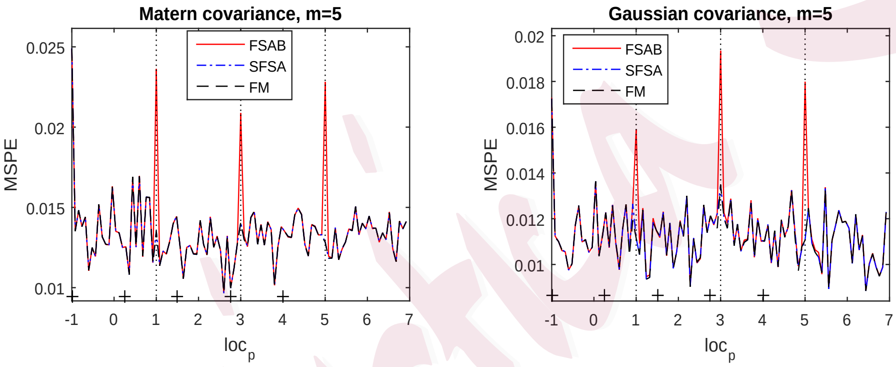
6. Numerical Examples

In this section, we illustrate the effectiveness of our method through simulation studies. The implementations of the NNGP, SFSA, and SFSA's variants (FSAB and CBCL) were written in Matlab; we used the R package "laGP" to obtain results of the local Gaussian process method with adaptive local designs (Gramacy and Apley (2015)). All methods ran on a AMD Opteron (tm) processor with 2.3 GHz CPUs and 32 GB memory. For log-likelihood function optimization, we used the matlab function, `fminunc`, which implements a Broyden-Fletcher-Goldfarb-Shanno (BFGS) based Quasi-Newton method. We used `parfor` command in Matlab for parallel computations.

6.1. Simulation Studies

We use the following example to show that compared to FSAB, the SFSA approach with $q \geq 1$ can alleviate the prediction errors around block boundaries.

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



(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 datasets.

Figure 3 shows the plots of Mean Squared Prediction Errors (MSPE) against the predictive locations. For the Matérn-covariance case (left panel), the MSPEs by the FSAB approach are particularly very large around block boundaries ($s = 1, 3, 5$). In contrast, the SFSA approach can reduce prediction errors around block-boundary points by borrowing dependence information from neighboring blocks. It can be seen that the MSPEs of 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, since it can significantly alleviate discontinuities of predictions around block boundaries.

6.2. Application to a Total Column Ozone Dataset

In this section, we analyze the total column ozone (TCO) level 2 dataset collected on October, 1st, 1988 (previous analysis of this dataset can be found in Cressie and Johannesson (2008); Eidsvik et al. (2014)). This TCO level 2 dataset has $n = 173,405$ observations, and we partitioned the data into a training set and a prediction set under two prediction scenarios: 1) Prediction on $25,000$ randomly selected locations (MAR); and 2) prediction on locations in a hold-out 15 degree × 15 degree rectangle region (MBD) that consists of around 600 predictive locations. For both prediction scenarios, we randomly generated three sets for evaluating the prediction performance of all comparison 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 for modeling the TCO dataset, where $\sigma^2 > 0, \phi > 0$ are the variance and range parameters, respectively. We also considered a Matérn covariance function (see (2.2)), with the smoothness ν fixed at 1 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 SFSA with the FSAB, NNGP, and LaGP methods in terms of prediction performance, considering the mean squared prediction error (MSPE) and the mean continuous rank probability score (CRPS) (e.g., see Gneiting and Raftery (2007)). For SFSA and FSAB, we used 24×24 regular blocks and 225 regular-grid knots so that both the block size n_b and the knot size m are around 200. For SFSA, we applied the sorted-coordinates (SC) ordering for the MAR scenario and the center-out (CO) ordering for the MBD scenario; then the number of neighboring blocks $q = 1$ was specified. For NNGP and LaGP, 50 neighbors were used for both parameter estimation and prediction. For LaGP, the "mspe" heuristic was considered.

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

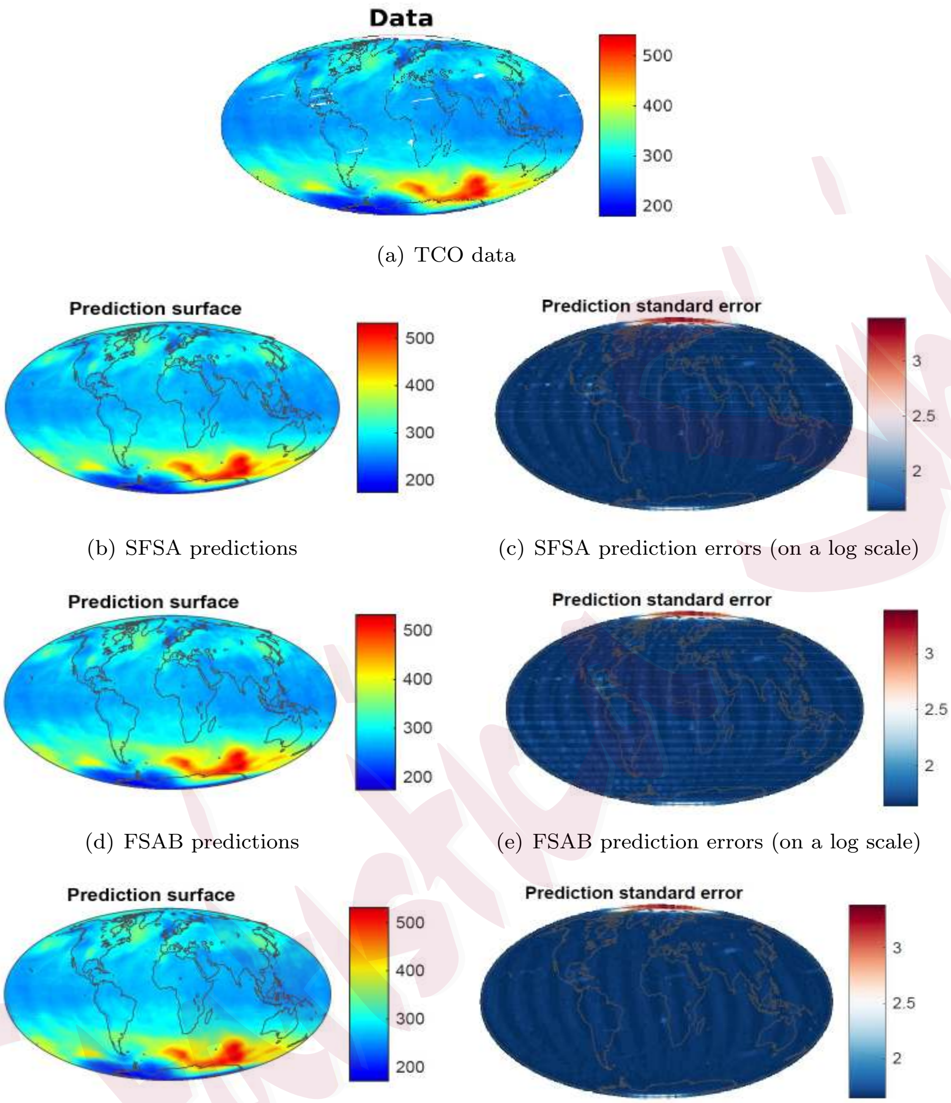
Table 2: Prediction performances of SFSA, FSAB, NNGP, and LaGP for the TCO data. The results were obtained based on 3 prediction sets for each prediction scenario.

Scenarios		SFSA		FSAB		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

for the MBD scenario, the SFSA method results in the best prediction results. NNGP results in larger MSPE and CRPS values than those of SFSA for the MBD scenario, which may be because the correlations of the TCO data have a relatively large scale so that borrowing information from non-neighboring locations is helpful for improving the prediction accuracy. Compared with 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 that is too smooth for modeling the TCO dataset; 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 computational times (including both the parameter-estimation and prediction steps), SFSA, FSAB and NNGP have comparable computational speeds. Compared with other methods, LaGP has much longer computational time for the MAR scenario but much shorter time for the MBD scenario. The reason is that its computational time mainly depends on the total number of the predictive locations. In contrast, for SFSA, FSAB, and NNGP, the computational bottleneck lies in the parameter-estimation step rather than the prediction step, since the high-dimensional likelihood of all 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 three methods, SFSA, FSAB, and NNGP produce very similar prediction surfaces, due to their comparable capability on the short-range predictions. Their associated predic-



(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 the prediction standard errors (on a log scale) for SFSA, FSAB, and NNGP.

tion standard errors (on a log scale) are shown in the right column, and we can observe that the prediction standard errors are particularly large for regions without any observations. For SFSA and FSAB, relatively larger prediction standard errors are observed around block boundaries; and this artifact is alleviated for

SFSA compared with FSAB.

7. Discussion

We propose a Smoothed Full-Scale Approximation approach (SFSA) that extends the FSA-Block approach by correcting the approximation errors of covariance between each block and its neighboring blocks. We prove that the SFSA approach yields a class of valid Gaussian process models so that both parameter estimation and prediction of SFSA can be performed in a unified framework. The proposed method incorporates the FSA-Block approach and the block conditional composite likelihood approach as special cases, and hence it can achieve better statistical efficiency. Compared with the FSA-Block approach, the SFSA approach can reduce prediction errors at locations around block boundaries, which can help produce a smoother prediction surface.

A natural extension of the proposed method is the spatio-temporal setting (Katzfuss and Cressie (2011); Bevilacqua et al. (2012); Zhang et al. (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 of spatio-temporal locations may not be a good measure for finding neighbors. We will explore using other measures to define the block partition and neighboring blocks that minimize the residual covariance for non-neighboring blocks to improve the approximation accuracy.

For modeling non-Gaussian observations from exponential family of distributions, SFSA can be embedded in the hierarchical spatial generalized linear models (GLM) (e.g., Diggle et al. (1998); Banerjee et al. (2014)) to speed up computations. The spatial GLM proposed by Diggle et al. (1998) for modeling non-Gaussian spatially dependent observations involves two stages. In the first stage, the data conditional on a latent spatial process are i.i.d exponential family random variables; and 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 $w(\mathbf{s})$ with the process induced by 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 and 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 model parameters for the spatial GLM (e.g., see Sengupta and Cressie (2013)).

Supplementary Materials

The supplementary material contains the proofs of Theorems, and additional numerical results for comparing SFSA with other methods.

Acknowledgements

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.

References

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

Bevilacqua, M., C. Gaetan, J. Mateu, and E. Porcu (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 G. Johannesson (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., S. Banerjee, A. O. Finley, and A. E. Gelfand (2016). Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association 111*, 800–812.

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

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

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

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

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (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 A. E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association 102*, 359–378.

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

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

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

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

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

Guinness, J. (2016). Permutation methods for sharpening Gaussian process approximations. *arXiv preprint arXiv:1609.05372*.

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

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

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

Lindgren, F., H. Rue, and J. Lindström (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 E. L. Kang (2017). Fused gaussian process for very large spatial data. *arXiv preprint arXiv:1702.08797*.

McKay, M. D., W. J. Conover, and R. J. Beckman (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., N. Cressie, and A. Braverman (2012). Spatial statistical data fusion for remote sensing applications. *Journal of the American Statistical Association 107*, 1004–1018.

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

Nychka, D., S. Bandyopadhyay, D. Hammerling, F. Lindgren, and S. Sain (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 D. Apley (2017). Patchwork kriging for large-scale Gaussian process regression. *arXiv preprint arXiv:1701.06655*.

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

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

Sun, Y., B. Li, and M. G. Genton (2012). Geostatistics for Large Datasets. In *Advances and challenges in space-time modelling of natural events*, pp. 55–77. Springer.

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

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

Zhang, R., C. D. Lin, and P. Ranjan (2016). Local Gaussian process model for large-scale dynamic computer experiments. *arXiv preprint arXiv:1611.09488*.

National Institute for Applied Statistics Research Australia (NIASRA), University of Wollongong, Wollongong, NSW 2522, Australia.

E-mail: bohai@uow.edu.au

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