Statistica Sinica **30** (2020), 2203-2226
doi:https://doi.org/10.5705/ss.202018.0285

A CLASS OF MULTI-RESOLUTION APPROXIMATIONS FOR LARGE SPATIAL DATASETS

Matthias Katzfuss and Wenlong Gong

Texas A&M University

Abstract: Gaussian processes are popular and flexible models for spatial, temporal, and functional data, but they are computationally infeasible for large data sets. We discuss Gaussian-process approximations that use basis functions at multiple resolutions to achieve fast inference and that can (approximately) represent any spatial covariance structure. We consider two special cases of this multi-resolution approximation framework, namely a taper version and a domain-partitioning (block) version. We describe theoretical properties and inference procedures, and study the computational complexity of the methods. Numerical comparisons and an application to satellite data are also provided.

Key words and phrases: Basis functions, Gaussian process, kriging, predictive process, satellite data, sparsity.

1. Introduction

Gaussian processes (GPs) are popular models for spatial data, time series, and functions. They are flexible and allow natural uncertainty quantification, but their computational complexity is cubic in the data size. This prohibits GPs from being used directly for analyses of many modern data sets consisting of a large number of observations, such as satellite remote-sensing data.

As a result, numerous approximations and assumptions have been proposed that allow the application of GPs to large spatial data sets. Some of these approaches are most appropriate for capturing fine-scale structure (e.g., Furrer, Genton and Nychka (2006); Kaufman, Schervish and Nychka (2008)), while others are more suitable for capturing large-scale structure (e.g., Higdon (1998); Mardia et al. (1998); Wikle and Cressie (1999); Cressie and Johannesson (2008); Katzfuss and Cressie (2009, 2011, 2012)). Lindgren, Rue and Lindström (2011) proposed an approximation based on viewing a GP with a Matérn covariance as the solution to the corresponding stochastic partial differential equation. Vecchia's method and its extensions (e.g., Vecchia (1988); Stein, Chi and Welty (2004); Datta et al. (2016); Katzfuss and Guinness (2017)) are discontinuous and assume the so-called screening effect holds; in other words, they assume that any given observation is conditionally independent from other observations, given a small subset of (typically nearby) observations.

We propose a class of multi-resolution approximations (M-RAs) for GPs that allow us to capture spatial structure at all scales. The M-RA framework is based on an orthogonal decomposition of the GP of interest into processes at multiple resolutions by iteratively applying the predictive process (Quiñonero-Candela and Rasmussen (2005); Banerjee et al. (2008)). The process at each resolution has an equivalent representation as a weighted sum of spatial basis functions. As the resolution increases, the number of functions increases, and their scale decreases. Unlike other multi-resolution models or wavelets (e.g., Chui (1992); Nychka, Wikle and Royle (2002); Johannesson, Cressie and Huang (2007); Cressie and Johannesson (2008); Nychka et al. (2015)), our M-RA automatically specifies the basis functions (and the prior distributions of their weights) to adapt to the given covariance function of interest, without requiring restrictions on this covariance function. Thus, in contrast to other approaches, it is clear which covariance structure is approximated by the sum of the basis functions in the M-RA.

To achieve computational feasibility within the proposed framework, we require an approximation of the "remainder process" at each resolution, using so-called modulating functions. We consider two special cases. For the M-RA-taper, the modulating functions are taken as tapering functions (i.e., compactly supported correlation functions). For an increasing resolution, the remainder process is approximated with increasingly restrictive tapering functions, leading to increasingly sparse matrices. In contrast, the M-RA-block iteratively splits each region at each resolution into a set of subregions, with the remainder process assumed to be independent between these subregions. This can lead to discontinuities at the region boundaries. A special case of the M-RA-block (Katzfuss (2017)) performed very well in a recent comparison of methods for large spatial data (Heaton et al. (2019)). A further special case of the M-RA, with only one resolution, is given by the full-scale approximation (Snelson and Ghahramani (2007); Sang, Jun and Huang (2011); Sang and Huang (2012)).

The M-RA is suitable for inference based on large numbers of observations from a GP, which may be irregularly spaced. We will describe inference procedures that rely on operations on sparse matrices for computational feasibility. The M-RA-block can deal with massive data sets with tens of millions of observations or more, because it is amenable to parallel computations on modern distributed computing systems. It can be viewed as a Vecchia-type approximation (Katzfuss and Guinness (2017)), it can be extended to a Kalman-filter-type analysis of spatio-temporal data (Jurek and Katzfuss (2018)), and the approximated covariance matrix is a so-called hierarchical off-diagonal low-rank matrix (e.g., Ambikasaran et al. (2016)). The M-RA-taper leads to more general sparse matrices, and thus requires careful algorithms to fully exploit the sparsity structure. However, it has the advantage of not introducing artificial discontinuities.

Relative to the M-RA-block in Katzfuss (2017), the contributions of our study are as follows. We introduce a general framework for M-RAs that provides a new, intuitive perspective on this approach. This allows an extension of the M-RA-block of Katzfuss (2017) that removes the requirement that knots at the finest resolution correspond to the observed locations. Furthermore, it enables us to introduce a novel M-RA-taper approach that extends the ideas of Sang and Huang (2012) to more than one resolution. We provide more insights about the theoretical and computational properties of both versions of the M-RA. We also include further implementation details and numerical comparisons.

The remainder of this paper is organized as follows. In Section 2, we first describe an exact orthogonal multi-resolution decomposition of a GP, which leads to the M-RA framework and the two special cases described above after applying the appropriate modulating functions. We also study their theoretical properties. In Section 3, we discuss the algorithms necessary for statistical inference using the M-RA, and we provide details of their computational complexity. Numerical comparisons on simulated and real data are given in Sections 4 and 5, respectively. We conclude in Section 6. The online Supplementary Material contains all proofs and additional simulation results. All code will be provided upon publication.

2. M-RAs

2.1. The true GP

Let $\{y_0(\mathbf{s}) : \mathbf{s} \in \mathcal{D}\}$, or $y_0(\cdot)$, be the true spatial field or process of interest on a continuous (non-gridded) domain $\mathcal{D} \subset \mathbb{R}^d$, for $d \in \mathbb{N}^+$. We assume that $y_0(\cdot) \sim GP(0, C_0)$ is a zero-mean GP with covariance function C_0. We place no restrictions on C_0, other than assuming it is a positive-definite function that is known up to a vector of parameters, $\boldsymbol{\theta}$. In real applications, $y_0(\cdot)$ will often not have a zero mean; however, estimating and subtracting the mean is usually not computationally difficult. Once $y_0(\cdot)$ has been observed at a set of n spatial locations $\mathcal{S}$, the primary goal of spatial statistics is to make (likelihood-based) inference on the parameters $\boldsymbol{\theta}$, and to obtain spatial predictions of $y_0(\cdot)$ at a set of locations $\mathcal{S}^P$ (i.e., to obtain the posterior distribution of $\mathbf{y}_0(\mathcal{S}^P)$). Direct computation using the Cholesky decomposition of the resulting covariance matrix requires $\mathcal{O}(n^3)$ time and $\mathcal{O}(n^2)$ memory complexity, which is computationally infeasible when $n \gg 10^4$.

2.2. Preliminaries

A multi-resolution approximation with M resolutions (M-RA) requires two main inputs: knots and modulating functions. The multi-resolutional set of knots, $\mathcal{Q} := \{\mathcal{Q}_0, \ldots, \mathcal{Q}_M\}$, is chosen such that, for all $m = 0, 1, \ldots, M$, $\mathcal{Q}_m = \{\mathbf{q}_{m,1}, \ldots, \mathbf{q}_{m,r_m}\}$ is a set of r_m knots, with $\mathbf{q}_{m,i} \in \mathcal{D}$. We assume that the number of knots increases with the resolution (i.e., $r_0 < r_1 < \cdots < r_M$). For example, we could choose each $\mathcal{Q}_m$ to be a regular grid over $\mathcal{D}$, as illustrated for a simple toy example in Figure 1. Alternatively, the knot set could be a partition of the set of observed spatial locations: $\mathcal{S} = \dot{\cup}_{m=0}^{M} \mathcal{Q}_m$.

The second imput is a set of modulating functions (Sang, Jun and Huang (2011)), $T := \{T_0, T_1, \ldots, T_M\}$, where $T_m : \mathcal{D} \times \mathcal{D} \to [0, 1]$ is a symmetric, nonnegative-definite function. In Section 2.5 we consider two specific examples; for now, we merely require that $T_m(\mathbf{s}_1, \mathbf{s}_2)$ is equal to one when $\mathbf{s}_1 = \mathbf{s}_2$, and is (exactly) equal to zero when $\mathbf{s}_1$ and $\mathbf{s}_2$ are far apart. Here, the meaning of "far" depends on the resolution m, in that, with increasing m, the modulating function should be equal to zero for increasingly large sets of pairs of locations in $\mathcal{D}$.

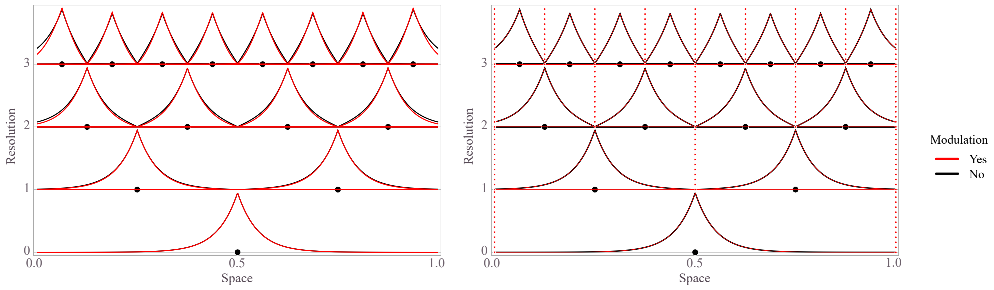
Based on these two inputs, we can now provide two definitions.

Definition 1 (Predictive process)**.** For a generic GP $x(\cdot) \sim GP(0, C)$, define $x^{(m)}(\cdot)$ as the predictive-process approximation (Quiñonero-Candela and Rasmussen (2005); Banerjee et al. (2008)) of $x(\cdot)$, based on the knots $\mathcal{Q}_m$:

$$x^{(m)}(\mathbf{s}) := E\big(x(\mathbf{s})|\mathbf{x}(\mathcal{Q}_m)\big) = \mathbf{b}(\mathbf{s})'\boldsymbol{\eta}, \ \mathbf{s} \in \mathcal{D},$$

where $\mathbf{b}(\mathbf{s})' = C(\mathbf{s}, \mathcal{Q}_m)$ and $\boldsymbol{\eta} \sim \mathcal{N}_{r_m}(\mathbf{0}, \boldsymbol{\Lambda}^{-1})$, with $\boldsymbol{\Lambda} = C(\mathcal{Q}_m, \mathcal{Q}_m)$.

That is, the predictive process is a conditional expectation, and hence is a smooth, low-rank approximation of $y(\cdot)$, which can also be written as a linear combination of basis functions (cf., Katzfuss (2013)). Furthermore, the remainder $x(\cdot) - x^{(m)}(\cdot) \sim GP(0, C_R)$ is independent of $x(\cdot)$, with positive-definite covariance function $C_R(\mathbf{s}_1, \mathbf{s}_2) = C(\mathbf{s}_1, \mathbf{s}_2) - \mathbf{b}(\mathbf{s}_1)'\boldsymbol{\Lambda}^{-1}\mathbf{b}(\mathbf{s}_2)$ (Sang and Huang (2012)).



(a) M-RA-taper with $d_m = 1.75/2^m$

(b) M-RA-block

Figure 1. For $y_0(\cdot) \sim GP(0, C_0)$ with exponential covariance function C_0 on $\mathcal{D} = [0, 1]$, a set of multi-resolution knots (black dots) and their corresponding basis functions, based on the orthogonal decomposition in (2.1) (black lines) and on two versions of the M-RA (red lines), with $r_0 = 1$, $J = 2$, and $M = 3$. The M-RA-block is exact in this setting (see Proposition 6) and, hence, the red and black lines overlap.

Definition 2 (Modulated process)**.** For a GP $x(\cdot) \sim GP(0, C)$, define $[x]_{[m]}(\cdot)$ to be the "modulated" process corresponding to $x(\cdot)$:

$$[x]_{[m]}(\cdot) \sim GP(0, [C]_{[m]}), \text{ where } [C]_{[m]}(\mathbf{s}_1, \mathbf{s}_2) = C(\mathbf{s}_1, \mathbf{s}_2) \cdot T_m(\mathbf{s}_1, \mathbf{s}_2), \ \mathbf{s}_1, \mathbf{s}_2 \in \mathcal{D}.$$

We find that $x(\cdot)$ and $[x]_{[m]}(\cdot)$ have the same variance structure (because $T_m(\mathbf{s}, \mathbf{s}) = 1$), but that $[x]_{[m]}(\cdot)$ has a compactly supported covariance function that is an increasingly bad approximation of C as m and the distance between $\mathbf{s}_1$ and $\mathbf{s}_2$ increase.

2.3. Exact multi-resolution decompositions of GPs

For any GP $y_0(\cdot) \sim GP(0, C_0)$ (as specified in Section 2.1), using Definition 1, we can write $y_0(\cdot) = \tau_0(\cdot) + \delta_1(\cdot)$, where $\tau_0(\cdot) := y_0^{(0)}(\cdot)$ is the predictive process of $y_0(\cdot)$ based on the knots $\mathcal{Q}_0$, and $\delta_1(\cdot) := y_0(\cdot) - \tau_0(\cdot) \sim GP(0, w_1)$ is independent of τ_0, and is itself a GP with a (positive-definite) covariance function w_1. Thus, we can again apply the predictive process to $\delta_1(\cdot)$ (this time based on the knots $\mathcal{Q}_1$) to obtain the decomposition $\delta_1(\cdot) = \tau_1(\cdot) + \delta_2(\cdot)$, and so forth, up to some resolution $M \in \mathbb{N}$.

This procedure enables us to exactly decompose any $y_0(\cdot) \sim GP(0, C_0)$ into orthogonal (i.e., independent) components at multiple resolutions:

$$y_0(\cdot) \overset{d}{=} \tau_0(\cdot) + \cdots + \tau_{M-1}(\cdot) + \delta_M(\cdot), \tag{2.1}$$

where $\tau_m(\cdot) := \delta_m^{(m)}(\cdot)$ is the predictive process of $\delta_m(\cdot)$ based on knots $\mathcal{Q}_m$,

$\delta_0(\cdot) := y_0(\cdot)$, and $\delta_m(\cdot) := \delta_{m-1}(\cdot) - \tau_{m-1}(\cdot) \sim GP(0, w_m)$, for $m = 1, \ldots, M$. Furthermore, using the basis-function representation from Definition 1, we can write each component of the decomposition as $\tau_m(\cdot) = \mathbf{a}_m(\cdot)'\boldsymbol{\gamma}_m$, where $\boldsymbol{\gamma}_m \overset{ind.}{\sim} \mathcal{N}_{r_m}(\mathbf{0}, \mathbf{\Omega}^{-1})$, and starting with $w_0 = C_0$, we have for $m = 1, \ldots, M-1$:

$$\begin{aligned} \mathbf{a}_m(\mathbf{s})' &:= w_m(\mathbf{s}, \mathcal{Q}_m),\ \mathbf{s} \in \mathcal{D}, \\ \mathbf{\Omega}_m &:= w_m(\mathcal{Q}_m, \mathcal{Q}_m), \\ w_{m+1}(\mathbf{s}_1, \mathbf{s}_2) &:= w_m(\mathbf{s}_1, \mathbf{s}_2) - \mathbf{a}_m(\mathbf{s}_1)'\mathbf{\Omega}_m^{-1}\mathbf{a}_m(\mathbf{s}_2),\ \mathbf{s}_1, \mathbf{s}_2 \in \mathcal{D}. \end{aligned} \tag{2.2}$$

An important feature of this decomposition is that components $\tau_m(\cdot)$ with low resolution m capture mostly smooth, long-range dependence, whereas high-resolution components capture mostly the fine-scale, local structure. This is because the predictive process at each resolution m is an approximation of the first r_m terms in the Karhunen Loéve expansion of $\delta_m(\cdot)$ (Sang and Huang (2012)). Figure 1 illustrates the resulting basis functions in our toy example.

It is straightforward to show that the decomposition of the process $y_0(\cdot) \sim GP(0, C_0)$ in (2.1) also implies an equivalent decomposition of the covariance function C_0:

$$C_0(\mathbf{s}_1, \mathbf{s}_2) = \sum_{m=0}^{M-1} w_m(\mathbf{s}_1, \mathcal{Q}_m) w_m(\mathcal{Q}_m, \mathcal{Q}_m)^{-1} w_m(\mathcal{Q}_m, \mathbf{s}_2) + w_M(\mathbf{s}_1, \mathbf{s}_2),\ \mathbf{s}_1, \mathbf{s}_2 \in \mathcal{D}. \tag{2.3}$$

2.4. The M-RA

The M-RA is a "modulated" version of the exact decomposition in (2.1), which at each resolution m, modulates the remainder using the function T_m from Section 2.2. The key idea is that the predictive processes at low resolutions capture the low-frequency variation in $y_0(\cdot)$, resulting in remainder terms that exhibit variability on increasingly smaller scales as m increases. Thus, approximating the remainder using increasingly restrictive modulating functions causes little approximation error.

Definition 3 (M-RA)**.** For a given $M \in \mathbb{N}$, the M-RA of a process $y_0(\cdot) \sim GP(0, C_0)$, based on a set of knots $\mathcal{Q} = \{\mathcal{Q}_0, \ldots, \mathcal{Q}_M\}$ and a set of modulating functions $T = \{T_0, \ldots, T_M\}$, is given by

$$y_M(\cdot) = \sum_{m=0}^{M} \widetilde{\tau}_m(\cdot) = \sum_{m=0}^{M} \mathbf{b}_m(\mathbf{s})'\boldsymbol{\eta}_m, \tag{2.4}$$

where $\widetilde{\tau}_m(\cdot) := \widetilde{\delta}_m^{(m)}(\cdot)$ and $\boldsymbol{\eta}_m \overset{ind.}{\sim} \mathcal{N}_{r_m}(\mathbf{0}, \boldsymbol{\Lambda}_m^{-1})$, for $m = 0, \dots, M$; $\widetilde{\delta}_0(\cdot) := [y_0]_{[0]}(\cdot) \sim GP(0, v_0)$, with $v_0 = [C_0]_{[0]}$; $\widetilde{\delta}_m(\cdot) = [\widetilde{\delta}_{m-1} - \widetilde{\tau}_{m-1}]_{[m]}(\cdot) \sim GP(0, v_m)$, for $m = 1, \dots, M$; and

$$\begin{aligned} \mathbf{b}_m(\mathbf{s})' &:= v_m(\mathbf{s}, \mathcal{Q}_m),\ \mathbf{s} \in \mathcal{D},\ m = 0, \dots, M, \\ \boldsymbol{\Lambda}_m &:= v_m(\mathcal{Q}_m, \mathcal{Q}_m),\ m = 0, \dots, M, \\ v_{m+1}(\mathbf{s}_1, \mathbf{s}_2) &:= \big(v_m(\mathbf{s}_1, \mathbf{s}_2) - \mathbf{b}_m(\mathbf{s}_1)' \boldsymbol{\Lambda}_m^{-1} \mathbf{b}_m(\mathbf{s}_2)\big) \cdot T_{m+1}(\mathbf{s}_1, \mathbf{s}_2),\ \mathbf{s}_1, \mathbf{s}_2 \in \mathcal{D}, \\ & \quad m = 0, \dots, M-1. \end{aligned} \tag{2.5}$$

Figure 1 shows the M-RA basis functions for our toy example. As shown, the M-RA is similar to a wavelet model, in that for increasing resolution m, we have an increasing number of basis functions with increasingly compact support. However, in contrast to wavelets, the basis functions $\mathbf{b}(\cdot)$ and the precision matrix $\boldsymbol{\Lambda}$ of the corresponding weights in the M-RA adapt to the covariance function C_0. Defining the basis functions recursively allows the M-RA to approximate C_0. In other approaches (e.g., wavelets, or that of Nychka et al. (2015)) with explicit expressions for the basis functions, the resulting covariance is less clear.

For ease of notation, we often stack the basis functions as $\mathbf{b}(\cdot) := \big(\mathbf{b}_0(\cdot)', \dots, \mathbf{b}_M(\cdot)'\big)'$, and the corresponding coefficients as $\boldsymbol{\eta} := \big(\boldsymbol{\eta}_0', \dots, \boldsymbol{\eta}_M'\big)'$, such that

$$y_M(\cdot) = \mathbf{b}(\cdot)'\boldsymbol{\eta}, \text{ where } \boldsymbol{\eta} \sim \mathcal{N}_r(\mathbf{0}, \boldsymbol{\Lambda}^{-1}), \tag{2.6}$$

with $\boldsymbol{\Lambda} := blockdiag(\boldsymbol{\Lambda}_0, \dots, \boldsymbol{\Lambda}_M)$ and $r = \sum_{m=0}^{M} r_m$.

2.5. Examples

As described in Section 2.2, the M-RA requires two inputs: knots and modulating functions. In light of the computational complexities discussed in Sections 3.2 – 3.3 below, we introduce a factor J, which is often set equal to 2 or 4. Then, starting with some (small) number of knots r_0 at resolution $m = 0$, we henceforth assume $r_m = J r_{m-1}$, for $m = 1, \dots, M$.

For the modulating functions, the two choices discussed next lead to important versions of the M-RA.

2.5.1. M-RA-block

To define the M-RA-block, we need a recursive partitioning of the spatial domain $\mathcal{D}$, in which each of J regions, $\mathcal{D}_1, \dots, \mathcal{D}_J$, is again divided into J smaller subregions, and so forth, up to level M:

$$\mathcal{D}_{j_1,\ldots,j_{m-1}} = \dot{\bigcup}_{j_m=1,\ldots,J} \mathcal{D}_{j_1,\ldots,j_m},\ j_1,\ldots,j_{m-1} = 1,\ldots,J;\ m = 1,\ldots,M.$$

We then assume, for each resolution m, that the modulated remainder $\delta_m(\cdot)$ is independent across partitions at the mth resolution. That is, the modulating function is defined as

$$T_m(\mathbf{s}_i,\mathbf{s}_j) = \begin{cases} 1, & (i_1,\ldots,i_m) = (j_1,\ldots,j_m), \\ 0, & \text{otherwise}, \end{cases} \quad \mathbf{s}_i \in \mathcal{D}_{i_1,\ldots,i_m},\ \mathbf{s}_j \in \mathcal{D}_{j_1,\ldots,j_m}. \tag{2.7}$$

Essentially, we have $T_m(\mathbf{s}_1,\mathbf{s}_2) = 1$ if $\mathbf{s}_1$ and $\mathbf{s}_2$ are in the same region $\mathcal{D}_{j_1,\ldots,j_m}$, and $T_m(\mathbf{s}_1,\mathbf{s}_2) = 0$ otherwise. At resolution m, $\mathcal{D}$ is split into J^m subregions. Typically, we assume that the knots at each resolution are roughly equally spread throughout the domain; as a result, there are roughly the same number $r_m/J^m = r_0$ of knots in every such region.

The M-RA-block and the corresponding domain partitions are illustrated in the toy example shown in Figure 1b. The M-RA-block was first proposed in Katzfuss (2017), with the restriction that $\mathcal{Q}_M = \mathcal{S}$. Another special case for $M = 1$ is the block-full-scale approximation (Snelson and Ghahramani (2007); Sang, Jun and Huang (2011)). Further discussion on the knot choice and partitioning schemes for the M-RA-block can be found in Katzfuss (2017, Sec. 2.5).

2.5.2. M-RA-taper

We can also specify the modulating functions as compactly supported correlation functions, often refered to as tapering functions. For simplicity, we assume here that the modulating functions are of the form

$$T_m(\mathbf{s}_1,\mathbf{s}_2) = T_*\left(\frac{\|\mathbf{s}_1 - \mathbf{s}_2\|}{d_m}\right),$$

with $d_{m+1} = d_m/J^{1/d}$, where d is the dimension of $\mathcal{D}$, $\|\cdot\|$ is some norm on $\mathcal{D}$, and T_* is a compactly supported correlation function, scaled such that $T^*(x) = 0$, for all $x \geq 1$. For simplicity, we use Kanter's function (Kanter (1997)) in all data examples:

$$T_*(x) := \begin{cases} 1, & x = 0, \\ (1-x)\frac{\sin(2\pi x)}{2\pi x} + \frac{1-\cos(2\pi x)}{2\pi^2 x}, & x \in (0,1), \\ 0, & x \geq 1. \end{cases}$$

For other possible choices of tapering functions, see Gneiting (2002). The M-RA-taper is illustrated in Figure 1a. A special case of the M-RA-taper for $M = 1$ is the taper-full-scale approximation (Sang and Huang (2012); Katzfuss (2013)).

2.6. Properties of the M-RA process

Throughout this subsection, let $y_M(\cdot)$ be the M-RA (as described in Definition 3) of $y_0(\cdot) \sim GP(0, C_0)$ on domain $\mathcal{D}$, based on knots $\mathcal{Q} = \{\mathcal{Q}_0, \ldots, \mathcal{Q}_M\}$ and modulating functions $T = \{T_0, \ldots, T_M\}$. All proofs are given in the Supplementary Material.

Proposition 1 (Distribution of the M-RA). *The M-RA is a GP, $y_M(\cdot) \sim GP(0, C_M)$, with covariance function*

$$C_M(\mathbf{s}_1, \mathbf{s}_2) = \sum_{m=0}^{M} v_m(\mathbf{s}_1, \mathcal{Q}_m) v_m(\mathcal{Q}_m, \mathcal{Q}_m)^{-1} v_m(\mathcal{Q}_m, \mathbf{s}_2), \ \mathbf{s}_1, \mathbf{s}_2 \in \mathcal{D}, \tag{2.8}$$

where v_m is defined in (2.5)*. We call C_M the M-RA of the covariance function C_0.*

Proposition 2 (Duplication of knots). *If $\mathbf{q} \in \mathcal{Q}_m$, then $v_{m+l}(\mathbf{q}, \mathbf{s}) = 0$, for any $\mathbf{s} \in \mathcal{D}$ and $l \geq 1$.*

This proposition implies that there is no benefit to designating the same locations as knots at different resolutions; that is, the knot locations in $\mathcal{Q}$ should not be too close together. In addition, although C_M in (2.8) is not a strictly positive function, we can ensure that the matrix $C_M(\mathcal{S}, \mathcal{S})$ is positive definite, for any set of unique locations $\mathcal{S}$, by setting $\mathcal{Q} = \mathcal{S}$.

Proposition 3 (Exact variance). *If $\mathbf{s} \in \mathcal{Q}$, then the M-RA variance at location $\mathbf{s}$ is exact; that is, $C_M(\mathbf{s}, \mathbf{s}) = C_0(\mathbf{s}, \mathbf{s})$.*

This proposition implies that, in contrast to other recent basis-function approaches (e.g., Lindgren, Rue and Lindström (2011); Nychka et al. (2015)), no variance or “edge” correction is needed for the M-RA if we place a knot location at each observed and prediction location.

Smoothness (i.e., differentiability) is an important concept in spatial statistics, and has led to the popularity of the Matérn covariance class with a parameter that flexibly regulates differentiability (e.g., Stein (1999)). The following proposition shows that any desired smoothness can be preserved when applying the M-RA.

Proposition 4 (Smoothness)**.** *If* $\mathbf{y}_0(\cdot)$ *is exactly* p*-times (mean-square) differentiable at* $\mathbf{s} \in \mathcal{Q}$*, where* $p \in \mathbb{Z}_{\geq 0}$*, then* $\mathbf{y}_M(\cdot)$ *is also exactly* p*-times differentiable at* $\mathbf{s}$*, provided that* $C_0(\cdot, \mathbf{q})$ *and* $T_m(\cdot, \mathbf{q})$ *are at least* $2p$*-times differentiable at* $\mathbf{s}$*, for any* $\mathbf{q} \in \mathcal{Q}$ *and* $m = 1, \ldots, M$*.*

Many commonly used covariance functions (e.g., Matérn) are infinitely differentiable away from the origin. If C_0 is such a function, the M-RA-block will have the same smoothness as the original process $\mathbf{y}_0(\cdot)$ at any $\mathbf{s}$ not located on the boundary between subregions, at any resolution (cf., Katzfuss (2017)). Tapering functions are often smooth away from the origin, except at the distance at which they become exactly zero. Thus, the M-RA-taper will typically have the same smoothness at $\mathbf{s}$ as $\mathbf{y}_0(\cdot)$ if T is at least $2p$-times differentiable at the origin, and $\mathbf{s}$ is not exactly at distance d_m from any $\mathbf{q} \in \mathcal{Q}_m$, for all $m = 1, \ldots, M$. Note that this result does not require that the smoothness of y_0 be the same at all locations $\mathbf{s}$; if the smoothness (or other local characteristics) of the covariance function C_0 varies over space, the M-RA will automatically adapt to this nonstationarity, and vary over space accordingly.

There is, however, an issue with the continuity of the M-RA-block process at the region boundaries, which can be highly undesirable in prediction maps.

Proposition 5 (Continuity)**.** *Assume that* C_0 *is a continuous function. Then, for the* M*-RA-taper, the realizations of the corresponding process* $y_M(\cdot)$ *and the posterior mean (i.e., kriging prediction) surface* $\mu_M(\mathbf{s}) := E(y_M(\mathbf{s})|\mathbf{z})$ *based on observations* $\mathbf{z}$*, as in* (3.1)*, are both continuous, assuming that* T_m *is continuous for all* $m = 0, 1, \ldots, M$*. In contrast, for the* M*-RA-block,* $y_M(\cdot)$ *and* $\mu_M(\cdot)$ *are both discontinuous, in general, at any* $\mathbf{s}$ *on the boundary between any two subregions.*

Proposition 6 (Exactness of M-RA-block)**.** *Let* C_0 *be a (stationary) exponential covariance function on the real line,* $\mathcal{D} = \mathbb{R}$*. In addition, let* C_M *be the covariance function of the corresponding* M*-RA-block (see Section* 2.5.1*), with* $r_m = (J-1)J^m$ *knots, for* $m = 0, \ldots, M-1$*, placed such that at each resolution* m*, a knot is located on each boundary between two subregions at resolution* $m+1$*. Then, the* M*-RA is exact at every knot location; that is,* $C_M(\mathbf{s}_1, \mathbf{s}_2) = C_0(\mathbf{s}_1, \mathbf{s}_2)$*, for any* $\mathbf{s}_1, \mathbf{s}_2 \in \mathcal{Q}$*.*

This proposition is illustrated in Figure 1b. As discussed in Section 3.2, this result allows us to exactly decompose an $n \times n$ exponential covariance matrix into a sparse matrix with n rows, but with only about $\log_2 n$ nonzero elements per row, where $r_0 = 1$ and $J = 2$. This leads to tremendous computational savings (e.g., $\log_2(n) < 30$ for $n = 1$ billion).

The exact result in Proposition 6 relies on the Markov property and the exact screening effect of the exponential covariance function (which is a Matérn covariance with smoothness parameter $\nu = 0.5$). However, similar, albeit approximate results are expected to hold for larger smoothness parameters in one dimension. Specifically, Stein (2011) shows that an asymptotic screening effect holds for $\nu = 1.5$ when using conditioning sets of size 2. He conjectures that an asymptotic screening effect holds for any ν when using conditioning sets of size greater than ν. This conjecture is also explored numerically in Katzfuss and Guinness (2017). To exploit this screening effect for the M-RA-block, we can simply place $c > \nu$ knots near every subregion boundary (i.e., $r_0 = c(J-1)$).

3. Inference

In this section, we describe inference for the M-RA based on a set of n measurements at locations $\mathcal{S}$. We assume additive, independent measurement error, such that

$$\mathbf{z} = \mathbf{y}_M(\mathcal{S}) + \boldsymbol{\epsilon}, \ \boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{V}_\epsilon), \tag{3.1}$$

where $\mathbf{V}_\epsilon$ is a diagonal matrix. We assume that C_0 and $\mathbf{V}_\epsilon$ are fully determined by the parameter vector $\boldsymbol{\theta}$, which is assumed fixed at a particular value, unless noted otherwise. For the sparsity and complexity calculations, we assume $r_m = r_0 J^m$ and $n = \mathcal{O}(r_M)$.

3.1. General inference results

3.1.1. Prior matrices

For a given set of parameters, the covariance function C_0 and, hence, the basis functions $\mathbf{b}(\cdot)$ and the precision matrix $\boldsymbol{\Lambda}$ in (2.6), are fixed. The prerequisite for inference is to calculate the prior matrices $\boldsymbol{\Lambda}$ and $\mathbf{B} := [\mathbf{B}_0, \ldots, \mathbf{B}_M] := [\mathbf{b}_0(\mathcal{S}), \ldots, \mathbf{b}_M(\mathcal{S})]$. Define $\mathbf{W}^k_{m,l} := v_k(\mathcal{Q}_m, \mathcal{Q}_l)$ and $\mathbf{W}^k_{\mathcal{S},m} := v_k(\mathcal{S}, \mathcal{Q}_m)$, such that $\boldsymbol{\Lambda}_m = \mathbf{W}^m_{m,m}$ and $\mathbf{B}_m = \mathbf{W}^m_{\mathcal{S},m}$. For $m = 0, \ldots, M$, starting with $\mathbf{W}^0_{m,l} = v_0(\mathcal{Q}_m, \mathcal{Q}_l)$ and $\mathbf{W}^0_{\mathcal{S},m} = v_0(\mathcal{S}, \mathcal{Q}_m)$, it is straightforward to verify that

$$\mathbf{W}^{k+1}_{m,l} = \big(\mathbf{W}^k_{m,l} - \mathbf{W}^k_{m,k}\boldsymbol{\Lambda}_k^{-1}\mathbf{W}^k_{l,k}{}'\big) \circ T_{k+1}(\mathcal{Q}_m, \mathcal{Q}_l), \ k = 0, \ldots, l-1; \ l = 0, \ldots, m; \tag{3.2}$$

and

$$\mathbf{W}^{k+1}_{\mathcal{S},m} = \big(\mathbf{W}^k_{\mathcal{S},m} - \mathbf{W}^k_{\mathcal{S},k}\boldsymbol{\Lambda}_k^{-1}\mathbf{W}^k_{m,k}{}'\big) \circ T_{k+1}(\mathcal{S}, \mathcal{Q}_m), \ k = 0, \ldots, m-1. \tag{3.3}$$

Here, $\circ$ denotes the Hadamard or element-wise product. Note that $\boldsymbol{\Lambda}_m$ and $\mathbf{B}_m$ both grow in dimension and become increasingly sparse with increasing resolution m. We have $(\boldsymbol{\Lambda}_m)_{i,j} = 0$ if $T_m(\mathbf{q}_{m,i}, \mathbf{q}_{m,j}) = 0$, and $(\mathbf{B}_m)_{i,j} = 0$ if $T_m(\mathbf{s}_i, \mathbf{q}_{m,j}) = 0$.

3.1.2. Posterior inference

Once $\boldsymbol{\Lambda}$ and $\mathbf{B}$ have been obtained, the posterior distribution of the unknown weight vector, $\boldsymbol{\eta}$, is given by well-known formulae for conjugate normal-normal Bayesian models:

$$\boldsymbol{\eta} \,|\, \mathbf{z} \sim \mathcal{N}_r\left(\widetilde{\boldsymbol{\nu}}, \widetilde{\boldsymbol{\Lambda}}^{-1}\right), \tag{3.4}$$

where $\widetilde{\boldsymbol{\Lambda}} = \boldsymbol{\Lambda} + \mathbf{B}'\mathbf{V}_\epsilon^{-1}\mathbf{B}$, $\widetilde{\boldsymbol{\nu}} = \widetilde{\boldsymbol{\Lambda}}^{-1}\widetilde{\mathbf{z}}$, and $\widetilde{\mathbf{z}} = \mathbf{B}'\mathbf{V}_\epsilon^{-1}\mathbf{z}$.

Based on this posterior distribution of $\boldsymbol{\eta}$, the likelihood can be written as follows (e.g., Katzfuss and Hammerling (2017)):

$$-2\log L(\boldsymbol{\theta}) = -\log|\boldsymbol{\Lambda}| + \log|\widetilde{\boldsymbol{\Lambda}}| + \log|\mathbf{V}_\epsilon| + \mathbf{z}'\mathbf{V}_\epsilon^{-1}\mathbf{z} - \widetilde{\mathbf{z}}'\widetilde{\boldsymbol{\Lambda}}^{-1}\widetilde{\mathbf{z}}. \tag{3.5}$$

Using this expression, the likelihood can be evaluated quickly for any given value of the parameter vector $\boldsymbol{\theta}$. This allows us to conduct likelihood-based inference (e.g., maximum likelihood or Metropolis–Hastings) on the parameters in C_0 and $\mathbf{V}_\epsilon$ by computing the quantities in (3.2)–(3.5) for each parameter value.

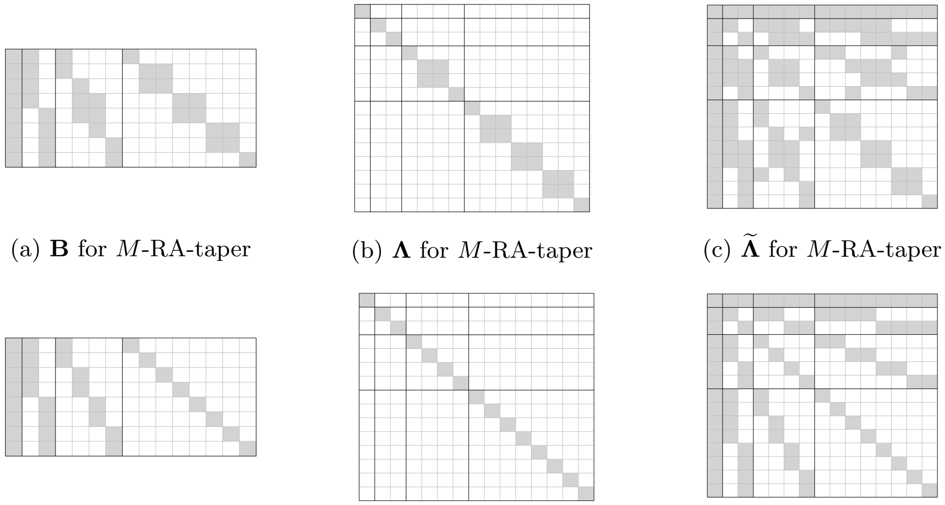
To obtain spatial predictions for fixed parameters $\boldsymbol{\theta}$, note that $\mathbf{y}_M(\mathcal{S}^P) = \mathbf{B}^P\boldsymbol{\eta}$, where $\mathbf{B}^P := \mathbf{b}(\mathcal{S}^P)$. Defining $\mathbf{W}^k_{\mathcal{S}^P,l} := v_k(\mathcal{S}^P, \mathcal{Q}_l)$, $\mathbf{B}^P = [\mathbf{B}^P_0, \ldots, \mathbf{B}^P_M]$ can be obtained based on the quantities from Section 3.1.1 by calculating $\mathbf{W}^0_{\mathcal{S}^P,m} = v_0(\mathcal{S}^P, \mathcal{Q}_m)$ and

$$\mathbf{W}^{k+1}_{\mathcal{S}^P,m} = \left(\mathbf{W}^k_{\mathcal{S}^P,m} - \mathbf{W}^k_{\mathcal{S}^P,k}\boldsymbol{\Lambda}_k^{-1}\mathbf{W}^k_{m,k}{}'\right) \circ T_{k+1}(\mathcal{S}^P, \mathcal{Q}_m),\ k = 0, \ldots, m-1,$$

and setting $\mathbf{B}^P_m = \mathbf{W}^m_{\mathcal{S}^P,m}$, for $m = 0, \ldots, M$. The posterior predictive distribution is given by

$$\mathbf{y}_M(\mathcal{S}^P) \,|\, \mathbf{z} \sim \mathcal{N}_{n_P}(\mathbf{B}^P\widetilde{\boldsymbol{\nu}}, \mathbf{B}^P\widetilde{\boldsymbol{\Lambda}}^{-1}\mathbf{B}^{P\prime}). \tag{3.6}$$

Hence, the main computational effort required for inference lies in the Cholesky decomposition of $\widetilde{\boldsymbol{\Lambda}}$, the posterior precision matrix of the basis-function weights in (3.4). Because $\boldsymbol{\Lambda}$ and $\mathbf{B}$ are both sparse, $\widetilde{\boldsymbol{\Lambda}}$ is a sparse matrix that can be decomposed quickly. Specifically, $\widetilde{\boldsymbol{\Lambda}}$ has the block structure $\widetilde{\boldsymbol{\Lambda}} = (\widetilde{\boldsymbol{\Lambda}}_{m,l})_{m,l=0,\ldots,M}$, where $\widetilde{\boldsymbol{\Lambda}}_{m,l} = \boldsymbol{\Lambda}_m \mathbb{1}_{\{m=l\}} + \mathbf{B}_m'\mathbf{V}_\epsilon^{-1}\mathbf{B}_l$ is an $r_m \times r_l$ matrix with (i,j)th element zero if $\nexists \mathbf{s} \in \mathcal{D}$ such that $T_m(\mathbf{q}_{m,i}, \mathbf{s}) \neq 0$ and $T_l(\mathbf{q}_{l,j}, \mathbf{s}) \neq 0$. Figure 2 shows the sparsity structures of $\mathbf{B}$, $\boldsymbol{\Lambda}$, and $\widetilde{\boldsymbol{\Lambda}}$ corresponding to the toy



(a) $\mathbf{B}$ for M-RA-taper

(b) $\boldsymbol{\Lambda}$ for M-RA-taper

(c) $\widetilde{\boldsymbol{\Lambda}}$ for M-RA-taper

(d) $\mathbf{B}$ for M-RA-block

(e) $\boldsymbol{\Lambda}$ for M-RA-block

(f) $\widetilde{\boldsymbol{\Lambda}}$ for M-RA-block

Figure 2. Illustration of the sparsity in the matrices $\mathbf{B}$, $\boldsymbol{\Lambda}$, and $\widetilde{\boldsymbol{\Lambda}}$ for the toy example in Figure 1. Resolutions are separated by solid black lines. Top row: M-RA-taper. Bottom row: M-RA-block.

example in Figure 1.

3.1.3. Inference in the absence of measurement error

If there is no measurement error (i.e., $\mathbf{V}_\epsilon = \mathbf{0}$), we have

$$\mathbf{z} = \mathbf{y} \sim \mathcal{N}_n(\mathbf{0}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma} = \mathbf{B}\boldsymbol{\Lambda}^{-1}\mathbf{B}'$. To ensure that $\mathbf{B}$ (and hence $\boldsymbol{\Sigma}$) has full rank, we assume for this case that $\mathcal{S} = \mathcal{Q}$ (and, thus, $n = r$) and (in light of Proposition 2) that the knots are unique. The likelihood can then be calculated as $-2\log L(\boldsymbol{\theta}) = -\log|\boldsymbol{\Sigma}| - \mathbf{y}'\boldsymbol{\Sigma}^{-1}\mathbf{y}$, where $\log|\boldsymbol{\Sigma}| = \log|\mathbf{B}\boldsymbol{\Lambda}^{-1}\mathbf{B}'| = \log|\mathbf{B}|^2 - \log|\boldsymbol{\Lambda}|$, and $\mathbf{y}'\boldsymbol{\Sigma}^{-1}\mathbf{y} = \widetilde{\mathbf{y}}'\boldsymbol{\Lambda}\widetilde{\mathbf{y}}$, with $\widetilde{\mathbf{y}} = \mathbf{B}^{-1}\mathbf{y}$.

Note that this form of M-RA inference can also be used when the M-RA is applied directly to data whose covariance, say $C_0^*(\mathbf{s}_i, \mathbf{s}_j) = C_0(\mathbf{s}_i, \mathbf{s}_j) + \tau^2 I(\mathbf{s}_i = \mathbf{s}_j)$, includes noise with variance τ^2.

3.2. Inference details for the M-RA-block

For the M-RA-block from Section 2.5.1, $\mathbf{B}$, $\boldsymbol{\Lambda}$, and $\widetilde{\boldsymbol{\Lambda}}$ are block-sparse matrices, with each block roughly of size $r_0 \times r_0$ and corresponding to (the knots at) a pair of regions.

As noted in Section 3.1.1, we have $(\mathbf{\Lambda}_m)_{i,j} = 0$ if $T_m(\mathbf{q}_{m,i}, \mathbf{q}_{m,j}) = 0$; thus, $\mathbf{\Lambda}_m$ is a block-diagonal matrix with diagonal blocks $\{v_m(\mathcal{Q}^{j_1,\ldots,j_m}, \mathcal{Q}^{j_1,\ldots,j_m}) : j_1, \ldots, j_m = 1, \ldots, J\}$, where $\mathcal{Q}^{j_1,\ldots,j_m} = \{\mathbf{q}_{m,i} : \mathbf{q}_{m,i} \in \mathcal{Q}_m \cap \mathcal{D}_{j_1,\ldots,j_m}\}$ is the set of roughly r_0 knots at resolution m that lie in $\mathcal{D}_{j_1,\ldots,j_m}$. It is well known that the inverse $\mathbf{\Lambda}_k^{-1}$ of a block-diagonal matrix $\mathbf{\Lambda}_k$ has the same block-diagonal structure as $\mathbf{\Lambda}_k$. Therefore, the prior calculations in Section 3.1.1 involving $\mathbf{\Lambda}_k^{-1}$ can be carried out at low computational cost.

For the posterior covariance matrix, we have from Section 3.1.2 that $(\widetilde{\mathbf{\Lambda}}_{m,l})_{i,j} = 0$ if $\nexists \mathbf{s} \in \mathcal{D}$ such that $T_m(\mathbf{q}_{m,i}, \mathbf{s}) \neq 0$ and $T_l(\mathbf{q}_{l,j}, \mathbf{s}) \neq 0$. Therefore, the block in $\widetilde{\mathbf{\Lambda}}$ corresponding to regions $\mathcal{D}_{i_1,\ldots,i_m}$ and $\mathcal{D}_{j_1,\ldots,j_m}$ is zero if the regions do not overlap (i.e., if $\mathcal{D}_{i_1,\ldots,i_m} \cap \mathcal{D}_{j_1,\ldots,j_m} = \emptyset$). The Cholesky factor of a (appropriately reordered) matrix with this particular block-sparse structure has zero fill-in, and can thus be calculated very rapidly.

Katzfuss (2017) describes an algorithm for inference for a special case of an M-RA-block that can be extended to the more general M-RA-block considered here. This algorithm is well suited to parallel and distributed computations for massive data sets, and it leads to efficient storage of the full posterior predictive distribution in (3.6). The time and memory complexity are shown to be $\mathcal{O}(nM^2r_0^2)$ and $\mathcal{O}(nMr_0)$, respectively.

3.3. Inference details for the M-RA-taper

The M-RA-taper from Section 2.5.2 results in sparse matrices, but care must be taken to ensure computational feasibility. A crucial observation for the computational results below is that, for any location $\mathbf{s} \in \mathcal{D}$ and any resolution m, only $\mathcal{O}(r_0)$ knots from $\mathcal{Q}_m$ are within a distance of d_m from $\mathbf{s}$ (i.e., all sets of the form $\{\mathbf{q}_{m,i} \in \mathcal{Q}_m : \|\mathbf{s} - \mathbf{q}_{m,i}\| \leq d_m\}$ contain only $\mathcal{O}(r_0)$ elements). This is because we assume that the $r_m = r_0 J^m$ knots at resolution m are roughly equally spread over the domain $\mathcal{D}$, and that $d_m = d_0/J^{m/d}$.

First, consider the calculation of the prior matrices described in Section 3.1.1. The matrices $\mathbf{\Lambda}$ and $\mathbf{B}$ have $\mathcal{O}(nr_0)$ and $\mathcal{O}(nMr_0)$ nonzero elements, respectively, because $(\mathbf{\Lambda}_m)_{i,j} = 0$ if $T_m(\mathbf{q}_{m,i}, \mathbf{q}_{m,j}) = 0$, and $(\mathbf{B}_m)_{i,j} = 0$ if $T_m(\mathbf{s}_i, \mathbf{q}_{m,j}) = 0$. Before performing the inference procedures, it is helpful to pre-calculate $\mathcal{I}_{m,l} := \{(i,j) : T_l(\mathbf{q}_{m,i}, \mathbf{q}_{l,j}) \neq 0\}$, the set of nonzero indices of the matrix $\mathbf{W}_{m,l}^l$, for $l = 0, \ldots, m$ and $m = 0, \ldots, M$. This can typically be done in $\mathcal{O}(n \log n)$ time (e.g., Vaidya (1989)). In the inference procedure, we then need only calculate the $\mathcal{I}_{m,l}$-elements of the matrices $\mathbf{W}_{m,l}^k$ in (3.2). Here, the main difficulty is that although $\mathbf{\Lambda}_k$ is sparse, its inverse $\mathbf{\Lambda}_k^{-1}$ is not. However, we need only compute certain elements of $\mathbf{\Lambda}_k^{-1}$.

Proposition 7. *For $l = 0, \ldots, m$ and $m = 0, \ldots, M$, the matrix $\mathbf{W}_{m,l}^{l}$ can be obtained by computing*

$$\mathbf{W}_{m,l}^{k+1} = \left(\mathbf{W}_{m,l}^{k} - \mathbf{W}_{m,k}^{k}\mathbf{S}_k\mathbf{W}_{l,k}^{k}{}'\right) \circ T_{k+1}(\mathcal{Q}_m, \mathcal{Q}_l),\ k = 0, \ldots, l-1, \tag{3.7}$$

where $\mathbf{S}_k = \mathbf{\Lambda}_k^{-1} \circ \mathbf{G}_k$ and $(\mathbf{G}_k)_{i,j} = \mathbb{1}_{\{\|\mathbf{q}_{m,i}-\mathbf{q}_{m,j}\|<(2+2/J)d_m\}}$. Thus, the (i,j) element of $\mathbf{\Lambda}_m^{-1}$ is not required in order to calculate the prior matrices in (3.2) *if $\|\mathbf{q}_{m,i} - \mathbf{q}_{m,j}\| \geq (2+2/J)\, d_m$.*

The total time complexity for computing all prior matrices in (3.2) *is $\mathcal{O}$ $(nM^2r_0^3)$, ignoring the cost of computing the $\mathbf{S}_k$ from the $\mathbf{\Lambda}_k$.*

To calculate $\mathbf{S}_k$ from $\mathbf{\Lambda}_k$, we use a selected-inversion algorithm (Erisman and Tinney (1975); Li et al. (2008); Lin et al. (2011)) in which we regard element (i,j) as a structural zero only if $\|\mathbf{q}_{k,i} - \mathbf{q}_{k,j}\| \geq (2+2/J)d_m$. This algorithm has the same computational complexity as the Cholesky decomposition of the same matrix. For one-dimensional domains ($d = 1$), $\mathbf{\Lambda}_k$ is a banded matrix with bandwidth $\mathcal{O}(r_0)$; thus, the time complexity to compute its Cholesky decomposition (and selected inverse) is $\mathcal{O}(r_k r_0^2)$ (e.g.,Gelfand et al. (2010, p. 187)). For $d \geq 2$, the rows and columns of $\mathbf{\Lambda}$ should be ordered such that the Cholesky decomposition leads to a (near) minimal fill-in and, hence, fast computations. Functions for this reordering are readily available in most statistical or linear-algebra software. The discussion in Furrer, Genton and Nychka (2006) indicates that the resulting time complexity for the Cholesky decomposition is roughly linear in the matrix dimension for $d = 2$. Moreover, our numerical experiments showed that the selected inversions account for only a small fraction of the total time required to compute the prior matrices. Therefore, this computation time scales roughly as $\mathcal{O}(nM^2r_0^3)$.

Once the prior matrices, including $\mathbf{B}$ and $\mathbf{\Lambda}$, have been obtained, the posterior inference requires computing and decomposing the posterior precision matrix $\widetilde{\mathbf{\Lambda}} = \mathbf{\Lambda} + \mathbf{B}'\mathbf{V}_\epsilon^{-1}\mathbf{B}$ in (3.4), with (m,l)th block $\widetilde{\mathbf{\Lambda}}_{m,l} = \mathbf{\Lambda}_m \mathbb{1}_{\{m=l\}} + \mathbf{B}_m'\mathbf{V}_\epsilon^{-1}\mathbf{B}_l$. The (j,k)th element of this block is given by

$$(\widetilde{\mathbf{\Lambda}}_{m,l})_{j,k} = (\mathbf{\Lambda}_m)_{j,k}\mathbb{1}_{\{m=l\}} + \textstyle\sum_{i=1}^{n} v_m(\mathbf{s}_i, \mathbf{q}_{m,j})v_l(\mathbf{s}_i, \mathbf{q}_{l,k})(\mathbf{V}_\epsilon)_{i,i}^{-1}.$$

Because each of the n $\mathbf{s}_i$ is within distances d_m and d_l of $\mathcal{O}(r_0)$ elements of $\mathcal{Q}_m$ and $\mathcal{Q}_l$, respectively, the time complexity to compute $(\mathbf{B}'\mathbf{B})_{m,l}$ is $\mathcal{O}(nr_0^2)$; hence, computing $\widetilde{\mathbf{\Lambda}}$ requires $\mathcal{O}(nM^2r_0^2)$ time.

Proposition 8. *The number of nonzero elements in* $\widetilde{\mathbf{\Lambda}}$ *is* $\mathcal{O}(nMr_0)$.

The time complexity for obtaining the Cholesky decomposition of $\widetilde{\mathbf{\Lambda}}$ is difficult to quantify, because it depends on its sparsity structure and the chosen ordering. However, our numerical experiments showed that the contribution of the Cholesky decomposition to the overall computation time is relatively small when appropriate reordering algorithms are used.

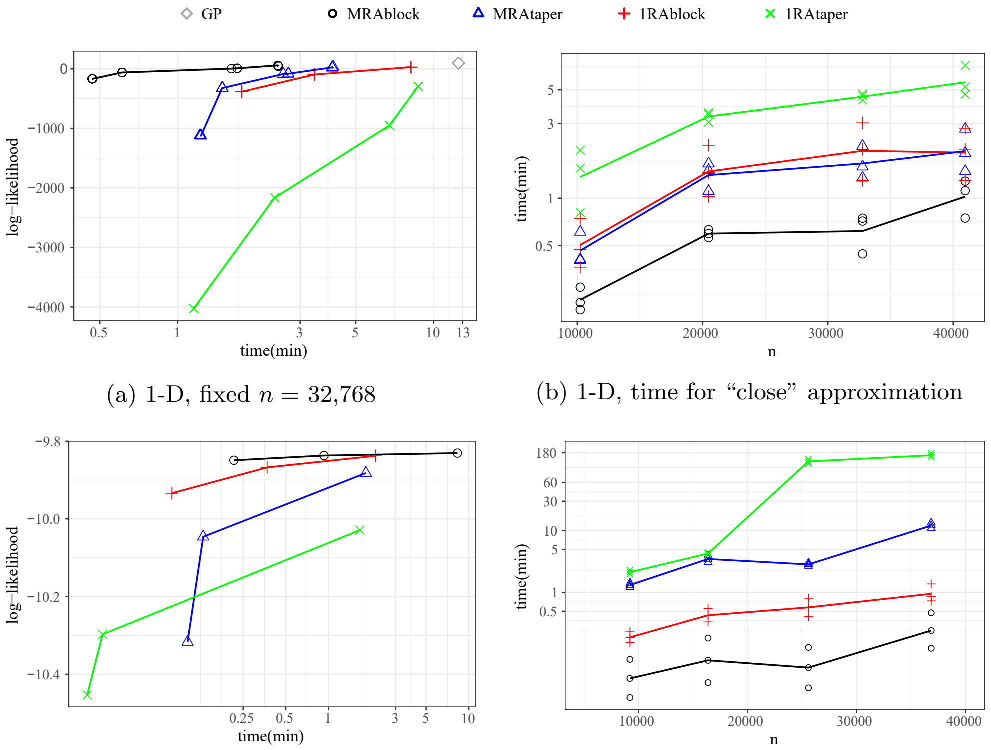
For predictions, the posterior covariance $\mathbf{B}^P\widetilde{\mathbf{\Lambda}}^{-1}\mathbf{B}^{P\prime}$ in (3.6) is dense and, hence, cannot be obtained explicitly for a large number of prediction locations. However, the posterior covariance matrix of a moderate number of linear combinations $\mathbf{Ly}(\mathcal{S}^P)$ can be obtained as $(\mathbf{LB}^P)\widetilde{\mathbf{\Lambda}}^{-1}(\mathbf{LB}^P)'$, also based on a Cholesky decomposition of $\widetilde{\mathbf{\Lambda}}$.

In summary, the time and memory complexity of the M-RA-taper are $\mathcal{O}(nM^2 r_0^3)$ and $\mathcal{O}(nMr_0)$, respectively, plus the cost of computing the Cholesky decompositions of $\mathbf{\Lambda}$ and $\widetilde{\mathbf{\Lambda}}$. However, these decompositions accounted for only a relatively small amount of the overall computation time in our numerical experiments. Thus, the time complexity of the M-RA-taper is roughly cubic in r_0, and it is square in r_0 for the M-RA-block. Note that the computational cost for the M-RA-taper can be reduced further if the covariance function C_0 has a small effective range relative to the size of $\mathcal{D}$, in which case, C_0 can be tapered at resolution 0 without causing a large approximation error. In contrast, for the M-RA-block, we always have $T_0(\mathbf{s}_1, \mathbf{s}_2) \equiv 1$. As explained in Katzfuss (2017), it is often appropriate to expect a good approximation for $M = \mathcal{O}(\log n)$ (and, hence, $r_0 = \mathcal{O}(1)$), which results in quasi-linear complexity as a function of n for the M-RA.

4. Simulation Study

For this section, we used data simulated from a true GP to compare the M-RA-block and M-RA-taper with full-scale approximations, FSA-block (Sang, Jun and Huang (2011)) and FSA-taper (Sang and Huang (2012)), which correspond to the 1-RA-block and 1-RA-taper, respectively. An implementation of the methods in Julia (`http://julialang.org`) version 0.4.5 was run on a 16-core machine with 64GB RAM.

The true GP was assumed to have mean zero and an exponential covariance



(a) 1-D, fixed $n = 32{,}768$

(b) 1-D, time for "close" approximation

(c) 2-D, fixed $n = 36{,}864$

(d) 2-D, time for "close" approximation

Figure 3. Summary of results from the simulation study. Top row: $\mathcal{D} = [0, 1]$. Bottom row: $\mathcal{D} = [0, 1]^2$. Left column: Log-score versus computation time for different versions of the M-RA for fixed n. Right column: Computation time required to get a "close" approximation to the truth (or best approximation) for different n; lines connect the means of the three times for each model and each n. Note that all time axes are on a log scale. Additional results can be found in the Supplementary Material.

function,

$$C_0(\mathbf{s}_1, \mathbf{s}_2) = \sigma^2 \exp\left(-\frac{\|\mathbf{s}_1 - \mathbf{s}_2\|}{\kappa}\right), \ \mathbf{s}_1, \mathbf{s}_2 \in \mathcal{D}, \tag{4.1}$$

with $\sigma^2 = 0.95$ and $\kappa = 0.05$ on a one-dimensional ($\mathcal{D} = [0, 1]$) or two-dimensional ($\mathcal{D} = [0, 1]^2$) domain. We assumed a nugget or measurement-error variance of $\tau^2 = 0.05$ (i.e., $\mathbf{V}_\epsilon = 0.05\,\mathbf{I}$). The results for Matérn covariances with different range, smoothness, and variance parameters (see Supplementary Material) showed similar patterns to those presented below.

All comparisons are based on the log-score (i.e., the log-likelihood at the true parameter values), which is a strictly proper scoring rule that is uniquely maximized in expectation by the true model (e.g., Gneiting and Katzfuss (2014)). All results were averaged over five replications.

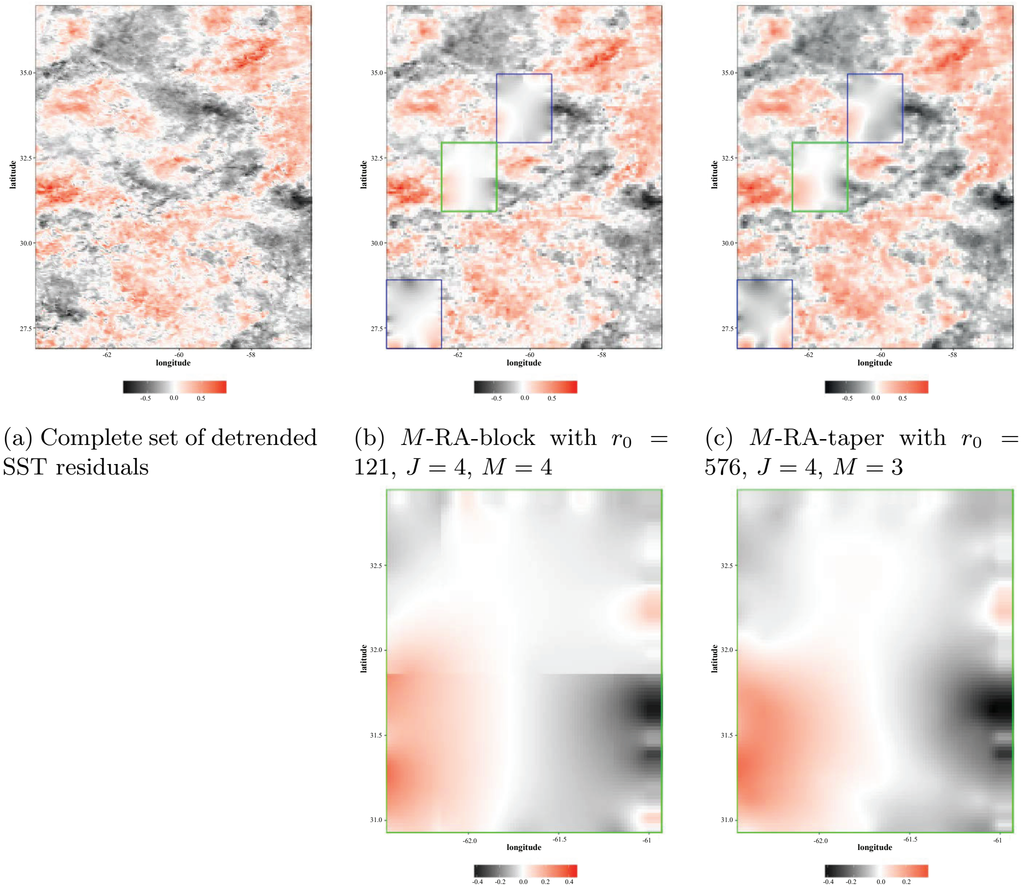
For M-RA-taper, some experimentation showed that there are general guidelines to follow in order to get a close approximation to a true GP. For a true covariance function C_0 with effective range ρ, we recommend setting the M-RA-taper range at resolution 0 to $d_0 = 2\rho$, and the distance between two adjacent knots at resolution 0 to be at most 2/3 of ρ. For example, the covariance in (4.1) has an effective range of $\rho \approx 0.15$, and so we set $d_0 = 0.3$ and the distance between adjacent knots at resolution 0 to 0.1.

First, we simulated data sets of different sizes on an equidistant grid in one dimension with $\mathcal{D} = [0, 1]$, which permitted fast simulation using the Davies–Harte algorithm, and evaluation of the exact likelihood using the Durbin–Levinson algorithm for comparison (McLeod, Yu and Krougly (2007)). For each data set, we recorded the computation times and log-scores for different versions of the M-RA (i.e., with different r_0, J, and M). We also considered the computation times required to achieve particular levels of approximation accuracy, specifically, the time required to obtain an average log-score within a difference of $0.003n$, $0.005n$, and $0.007n$ of the log-score of the true model. We then repeated the simulation study in two dimensions, $\mathcal{D} = [0, 1]^2$. Because it was infeasible to compute the true log-likelihood for large n, we use the best approximation (i.e., the largest approximated log-likelihood) as the basis on which to compare the relative performance of the various methods, with cutoff values of $0.008n$, $0.01n$, and $0.012n$.

The results are summarized in Figure 3. The computation times scaled roughly as expected. The M-RA-block was consistently better than the other methods, while M-RA-taper and 1-RA-block performed similarly. The 1-RA-taper was not competitive.

5. Application

In this section, we applied the four methods from Section 4 to a real satellite data set. We considered $n = 44{,}711$ Level-3 daytime sea-surface-temperature (SST) data from August 2016 over a region in the North Atlantic Ocean, as measured by the Moderate Resolution Imaging Spectroradiometer on board the Terra satellite. The data are freely available at `https://giovanni.gsfc.nasa.gov`. More specifically, the data (shown in Figure 4a) were taken to be the residuals of the SST data after removing longitudinal and latitudinal trends. The



(a) Complete set of detrended SST residuals

(b) M-RA-block with $r_0 = 121$, $J = 4$, $M = 4$

(c) M-RA-taper with $r_0 = 576$, $J = 4$, $M = 3$

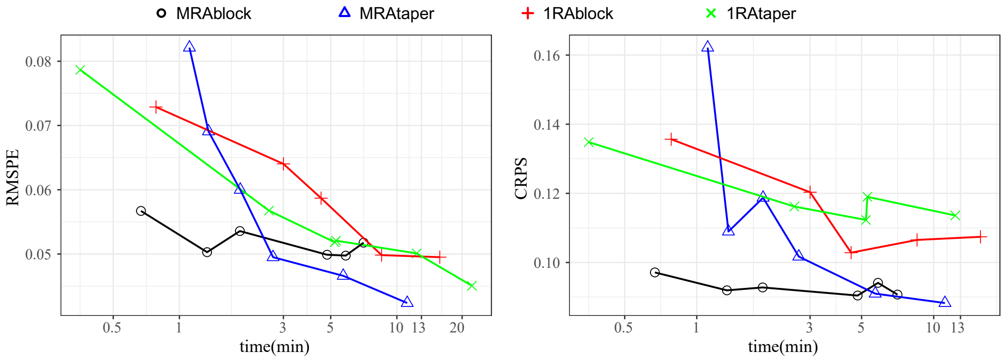
(d) Missing-area prediction for M-RA-block

(e) Missing-area prediction for M-RA-taper

Figure 4. Top row: Complete data set of sea-surface temperature, along with posterior predictive means for M-RA-taper and M-RA-block based on removing three areal test regions and additional randomly selected values. Bottom row: Zoomed-in view of the green rectangle in the upper prediction plots. Color scales are in units of degrees Celsius.

exploratory analysis showed that an exponential covariance fit the data well, and so all methods used were approximating the covariance in (4.1). We assumed a constant noise variance τ^2 (i.e., $\mathbf{V}_\epsilon = \tau^2\mathbf{I}$).

To compare the different approximation methods, we created five different data sets by randomly splitting the complete data set of residuals into training data, areal test data, and random test data, containing 78%, 12%, and 10%, respectively, of the values in the full data set. The split of the complete data into training and test sets was designed to mimic the typical setting of Level-2 satellite



(a) RMSPE (b) CRPS

Figure 5. For the satellite SST data, comparison of scores (lower is better) for predictions of areal test data for different settings of the M-RA

data, with unobserved areas over which the satellite did not fly in a particular time period, and observed areas with some missing values (e.g., due to clouds). Specifically, the areal test locations were obtained by splitting the domain into $5 \times 5 = 25$ equal-area rectangles, and then removing three of these rectangles at random. The remaining test locations were obtained by simple random sampling of the remaining locations.

Using each of the five training sets, and for a range of settings for each of the four approximation methods, we carried out maximum-likelihood estimation of the unknown parameters σ^2, κ, and τ^2, and obtained posterior predictive distributions at the held-out test locations. We compared the pointwise (i.e., marginal) posterior distributions obtained by the methods to the held-out test data in terms of the root mean squared prediction error (RMSPE) and the continuous ranked probability score (CRPS), a strictly proper scoring rule that quantifies the fit of the entire predictive distribution to the data (e.g., Gneiting and Katzfuss (2014)). The scores for the random test data were almost zero for all methods. The scores for the areal test data are shown in Figure 5 (averaged over the five data sets). In general, the scores for M-RA-taper and M-RA-block were better than those for the full-scale approximations. M-RA-taper produced some RMSPEs that were even lower than those for M-RA-block.

Perhaps more important than the differences in the prediction scores, are the differences in the prediction plots. Figure 4 shows an example of the posterior means obtained by M-RA-taper and M-RA-block for versions of the two methods that took a similar time to run (five to seven minutes) and that resulted in similar RMSPEs in Figure 5a. Despite the good approximation accuracy and low RMSPE of M-RA-block, Figure 4d shows clearly visible artifacts due to discontinuities of the M-RA-block at the region boundaries (see Proposition 5), which do not appear for the continuous M-RA-taper in Figure 4e. Avoiding these kinds of nonphysical artifacts is often of paramount importance to domain scientists.

6. Conclusion

We have proposed and studied a general approach for obtaining multi-resolution approximations of GPs based on an orthogonal decomposition of the GP of interest into processes at multiple resolutions. We considered two specific cases of this approach: The M-RA-taper, which achieves sparsity and computational feasibility by applying increasingly compact isotropic tapering functions as the resolution increases, and the M-RA-block, which is based on a recursive block-partitioning of the spatial domain and assumes conditional independence between the spatial subregions at each resolution. We have provided algorithms for inference, along with the computational complexity of the methods. Within our framework, one could also consider other partitioning schemes or nonstationary tapering, which might be especially useful when approximating nonstationary processes.

We have shown theoretically and numerically that both M-RA versions have useful properties, and can outperform related existing approaches. The M-RA-block achieves more accurate approximations to a given covariance function for a given computation time, and its block-sparse structure allows it to approximate the likelihood for truly massive data sets on modern distributed computing systems. However, the M-RA-block process is discontinuous at the subregion boundaries, which can be undesirable in prediction maps. The M-RA-taper can be useful for real-world applications in which the true covariance function is unknown anyway, and, hence, it might be more important to have a “smooth” model that avoids the potential artifacts and discontinuities inherent to the M-RA-block, owing to its domain partitioning. The M-RA-taper’s prediction accuracy can be highly competitive, especially when the effective range of the covariance model is small relative to the domain size. Note that posterior inference involving the M-RA-taper only requires general sparse matrices, which would allow for a relatively straightforward treatment of areal-averaged measurements (e.g., satellite footprints).

Future work will consider multivariate, spatio-temporal, and nonGaussian extensions of the methodology. Also of interest is a more precise quantification of the approximation error, and a further investigation of how to choose the number of resolutions and the knots, depending on the covariance to be approximated. Although our methods are, in principle, also applicable in the context of GP regression, some additional consideration of the choice of knots and partitions in high-dimensional covariate spaces is warranted.

Supplementary Material

The online Supplementary Material contains all proofs, as well as additional settings for the simulation study in Section 4.

Acknowledgments

The authors were partially supported by National Science Foundation (NSF) Grant DMS–1521676. Katzfuss' research was also partially supported by NSF CAREER Grant DMS–1654083. We would like to thank Tim Davis, Shiyuan He, and several anonymous reviewers for their helpful comments. The data used in the application were acquired by NASA's Earth Science Division, and archived and distributed by the Goddard Earth Sciences (GES) Data and Information Services Center (DISC).

References

Ambikasaran, S., Foreman-Mackey, D., Greengard, L., Hogg, D. W. and O'Neil, M. (2016). Fast direct methods for Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **38**, 252–265.

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

Chui, C. (1992). *An Introduction to Wavelets.* Academic Press, San Diego.

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.

Erisman, A. M. and Tinney, W. F. (1975). On computing certain elements of the inverse of a sparse matrix. *Communications of the ACM* **18**, 177–179.

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

Gelfand, A., Diggle, P., Guttorp, P. and Fuentes, M., editors (2010). *Handbook of Spatial Statistics.* CRC Press.

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

Gneiting, T. and Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application* **1**, 125–151.

Heaton, M. J., Datta, A., Finley, A. O., Furrer, R., Guinness, J., Guhaniyogi, R., Gerber, F., Gramacy, R. B., Hammerling, D., Katzfuss, M., Lindgren, F., Nychka, D. W., Sun, F. and Zammit-Mangion, A. (2019). A case study competition among methods for analyzing large spatial data. *Journal of Agricultural, Biological, and Environmental Statistics* **24**, 398–425.

Higdon, D. (1998). A process-convolution approach to modelling temperatures in the North Atlantic Ocean. *Environmental and Ecological Statistics* **5**, 173–190.

Johannesson, G., Cressie, N. and Huang, H.-C. (2007). Dynamic multi-resolution spatial models. *Environmental and Ecological Statistics* **14**, 5–25.

Jurek, M. and Katzfuss, M. (2018). Multi-resolution filters for massive spatio-temporal data. *arXiv:1810.04200.*

Kanter, M. (1997). Unimodal spectral windows. *Statistics & Probability Letters* **34**, 403–411.

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 Cressie, N. (2009). Maximum likelihood estimation of covariance parameters in the spatial-random-effects model. In *Proceedings of the Joint Statistical Meetings*, 3378–3390. Alexandria, VA. American Statistical Association.

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

Katzfuss, M. and Cressie, N. (2012). Bayesian hierarchical spatio-temporal smoothing for very large datasets. *Environmetrics* **23**, 94–107.

Katzfuss, M. and Guinness, J. (2017). A general framework for Vecchia approximations of Gaussian processes. *arXiv:1708.06302.*

Katzfuss, M. and Hammerling, D. (2017). Parallel inference for massive distributed spatial data using low-rank models. *Statistics and Computing* **27**, 363–375.

Kaufman, C. G., Schervish, M. 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.

Li, S., Ahmed, S., Klimeck, G. and Darve, E. (2008). Computing entries of the inverse of a sparse matrix using the FIND algorithm. *Journal of Computational Physics* **227**, 9408–9427.

Lin, L., Yang, C., Meza, J., Lu, J., Ying, L. and Weinan, E. (2011). SelInv - An algorithm for selected inversion of a sparse symmetric matrix. *ACM Transactions on Mathematical Software* **37**, 1–19.

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.

Mardia, K., Goodall, C., Redfern, E. and Alonso, F. (1998). The kriged Kalman filter.

Test **7**, 217–282.

McLeod, A. I., Yu, H. and Krougly, Z. (2007). Algorithms for linear time series analysis: With R package. *Journal of Statistical Software* **23**, 1–26.

Nychka, D. W., Bandyopadhyay, S., Hammerling, D., Lindgren, F. and Sain, S. R. (2015). A multi-resolution Gaussian process model for the analysis of large spatial data sets. *Journal of Computational and Graphical Statistics* **24**, 579–599.

Nychka, D. W., Wikle, C. K. and Royle, J. A. (2002). Multiresolution models for nonstationary spatial covariance functions. *Statistical Modelling* **2**, 315–331.

Quiñonero-Candela, J. and Rasmussen, C. (2005). A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research* **6**, 1939–1959.

Sang, H. and Huang, J. Z. (2012). A full scale approximation of covariance functions. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **74**, 111–132.

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

Snelson, E. and Ghahramani, Z. (2007). Local and global sparse Gaussian process approximations. In *Artificial Intelligence and Statistics 11 (AISTATS)*, 524–531.

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

Stein, M. L. (2011). 2010 Rietz lecture: When does the screening effect hold? *Annals of Statistics* **39**, 2795–2819.

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

Vaidya, P. M. (1989). An O(n log n) algorithm for the all-nearest-neighbors problem. *Discrete & Computational Geometry* **4**, 101–115.

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

Wikle, C. K. and Cressie, N. (1999). A dimension-reduced approach to space-time Kalman filtering. *Biometrika* **86**, 815–829.

3143 TAMU, College Station, TX 77843, USA.

E-mail: katzfuss@tamu.edu

3143 TAMU, College Station, TX 77843, USA.

E-mail: wgong@stat.tamu.edu

(Received July 2018; accepted February 2019)