Special Issue Paper

Environmetrics

Received: 7 December 2010, Revised: 1 September 2011, Accepted: 6 September 2011, Published online in Wiley Online Library: 24 October 2011

(wileyonlinelibrary.com) DOI: 10.1002/env.1137

In order to make spatial statistics computationally feasible, we need to forget about the covariance function†

Daniel Simpson[a], Finn Lindgren[a] and Håvard Rue[a]*

Gaussian random fields (GRFs) are the most common way of modeling structured spatial random effects in spatial statistics. Unfortunately, their high computational cost renders the direct use of GRFs impractical for large problems and approximations are commonly used. In this paper, we compare two approximations to GRFs with Matérn covariance functions: the kernel convolution approximation and the Gaussian Markov random field representation of an associated stochastic partial differential equation. We show that the second approach is a natural way to tackle the problem and is better than methods based on approximating the kernel convolution. Furthermore, we show that kernel methods, as described in the literature, do not work when the random field is not smooth. Copyright © 2011 John Wiley & Sons, Ltd.

Keywords: SPDE; GMRF; kernel methods; covariance functions

1. INTRODUCTION

Problems that have a non-negligible spatial component are ubiquitous in environmental statistics. In contrast to traditional statistical modeling, practical problems in spatial statistics are, by and large, *computational* in nature. Many applications feature large sets of data collected at irregular locations, which necessitates the development of fast and efficient methods for computing both the point estimates and the associated uncertainties of the parameters in the model.

The need to balance modeling desires with computational limits has created a large number of new and exciting methods in applied statistics. The basic aim of these methods is to construct models for which exact or approximate inference can be performed quickly. One such class of models, known as fixed-rank kriging models (Cressie and Johannesson, 2008), restricts the class of random fields to low-dimensional combinations of deterministic functions (cf. predictive process models (Banerjee *et al.*, 2008)). In this paper we restrict our attention to two other methods. The first method, which we will refer to as the *kernel method* and was originally proposed by Higdon (1998), is based on a moving average representation of the random field. The second method we will consider was proposed very recently by Lindgren *et al.* (2011) and is based on the representation of the random field as the solution to a stochastic partial differential equation (SPDE). While this appears to be a complicated way to specify a random field, it was shown by Lindgren *et al.* (2011) to not only be natural but extremely flexible. Because these two methods can be used to approximate the same random fields, it is natural to compare them and ask which is better.

In this paper, we compare the kernel approach with the SPDE approach in terms of computational efficiency and accuracy for approximating the class of Matérn random fields, which are often used in applications. What we found was surprising—kernel methods can perform very poorly for these fields, especially when the field is rough. This poor performance stems from the behavior of the kernel function at the origin, which is not taken into account when discretizing the white noise integral at the core of the kernel method. On the other hand, the method of Lindgren *et al.* (2011) behaves in a more stable manner.

This result has important practical consequences: in realistic inference problems, a practical method is required to work for *a large range* of parameter values. In Sections 3 and 4, we tested a "reasonable" (i.e., not unreasonable) set of parameter values and found them not to work. This is enough to say that naïve kernel methods (e.g., the methods found in Higdon, 1998) are unreliable for inference with Matérn fields.

In the remainder of the paper, we will first review the basic setting that we are looking at. We will then outline the link between the SPDE (2) and the convolution representation (3) and discuss the practical problems that are encountered when using convolution representations as a basis for a computational method. In Section 4, we will review the methods of Lindgren *et al.* (2011) for constructing a Gaussian Markov random field (GMRF) directly from (2). In the following section, we will compare the accuracy and performance of the SPDE and

* *Correspondence to: H. Rue, Department of Mathematical Sciences Norwegian University of Science and Technology, N-7491 Trondheim, Norway. E-mail: hrue@math.ntnu.no*

a *Department of Mathematical Sciences, Norwegian University of Science and Technology, N-7491 Trondheim, Norway*

† *This article is published in Environmetrics as a special issue on Spatio-Temporal Stochastic Modelling (METMAV), edited by Wenceslao González-Manteiga and Rosa M. Crujeiras, University of Santiago de Compostela, Spain.*

convolution field methods. Finally, we will discuss the various extensions of the SPDE method and give reasons apart from computational efficiency that SPDEs are *good models* in spatial statistics.

2. SETTING

Throughout this paper, we will consider the following generic scenario. Assume that we have observed some data $\{y_i\}_{i=1}^N$ at some spatial locations $\{s_i\}_{i=1}^N$ and that we have a hierarchical model for this observation process (Diggle and Ribeiro, 2006):

$$\begin{aligned} y_i|x_i,\boldsymbol{\theta} &\overset{\text{i.i.d.}}{\sim} \pi(y_i|x_i,\boldsymbol{\theta}) \\ \mathbf{x} &\sim N(\boldsymbol{\mu}(\boldsymbol{\theta}),\boldsymbol{\Sigma}(\boldsymbol{\theta})) \\ \boldsymbol{\theta} &\sim \pi(\boldsymbol{\theta}) \end{aligned}$$

where $\boldsymbol{\theta}$ is a vector of model parameters, $\boldsymbol{\mu}(\boldsymbol{\theta})$ is a model for the mean of the underlying and unobserved spatial process $\mathbf{x}$, $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is the covariance matrix of $\mathbf{x}$, and $\pi(\boldsymbol{\theta})$ is a prior on the vector of parameters. In most applications, we are interested in the posterior distribution of parameters $(\boldsymbol{\theta},\mathbf{x})$ given the data $\mathbf{y}$. An application of Bayes' formula shows that the posterior is given by

$$\pi(\boldsymbol{\theta},\mathbf{x}|\mathbf{y}) \propto \pi(\boldsymbol{\theta})\pi(\mathbf{x}|\boldsymbol{\theta}) \prod_{j=1}^{N} \pi(y_j|x_j,\boldsymbol{\theta})$$

In most situations, the posterior will not be of a standard form, and it is necessary to investigate it numerically.

The standard method for specifying the covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is to define it as

$$\Sigma_{ij} = c(s_i,s_j)$$

where $c(s_i,s_j)$ is a covariance function, which forces the covariance matrix to be positive semi-definite. In general, finding useful covariance functions is difficult and typically it is assumed that the covariance between two points only depends on the distance between the points (or in the nomenclature of spatial statistics that $\mathbf{x}$ is a stationary, isotropic random field). The class of Gaussian random fields (GRFs) that we are focusing on in this paper is the Matérn random fields, which are stationary, isotropic random fields with covariance functions given by[‡]

$$c_\nu(s_i,s_j) = \frac{\sigma^2}{\Gamma(\nu+d/2)(4\pi)^{d/2}\kappa^{2\nu}2^{\nu-1}} \left(\kappa\|s_i-s_j\|\right)^\nu K_\nu\left(\kappa\|s_i-s_j\|\right) \tag{1}$$

where $\kappa,\sigma^2,\nu>0$, and $K_\nu(\cdot)$ is the modified Bessel function of the second kind. It can be shown that the parameter ν defines the smoothness of the random field, whereas κ is a scale parameter and σ^2 is a variance parameter. Matérn random fields are popular models in many areas of applied statistics where the ability to specify the smoothness of the field is useful. Furthermore, the Matérn family includes two of the most popular covariance models—the exponential covariance function, which occurs when $\nu=1/2$, and the Gaussian covariance function, which is the limit case as $\nu\to\infty$.

The major problem with using non-compactly supported covariance functions (such as the Matérn covariance functions) is that the resulting covariance matrices are completely dense and, therefore, *inference with (1) has a computational complexity of* $\mathcal{O}(N^3)$*, where* N *is the number of data points.* This makes direct use of the Matérn covariance function manifestly unsuitable for practical inference on moderate to large data sets. This is, of course, not a new problem in spatial statistics, and a number of approximations have been developed to overcome this problem such as covariance tapering (Furrer *et al.*, 2006) and predictive process modeling (Banerjee *et al.*, 2008).

Since the 1980s, there has been a great deal of work done on GMRFs, especially over graphs. In practical situations, this Markov property essentially forces the precision matrix $\mathbf{Q}=\Sigma^{-1}$ to be *sparse*. The sparsity of the precision matrix allows for computation to be performed in $\mathcal{O}(n^{3/2})$ operations, where n is the number of vertices in the graph, if the problem is two dimensional (Rue and Held, 2005). This reduction in computational complexity allows for the solution of quite large spatial models (for instance, a two-dimensional second-order conditional autoregressive model with ~400,000 nodes). Furthermore, the Markov property allows us to use fast approximate inference methods, such as INLA (Rue *et al.*, 2009), to compute with these models.

The Markov random fields that we are considering in this paper are the stationary solutions to SPDEs of the form

$$(\kappa^2-\Delta)^{\alpha/2}x(s) = \sigma W(s), \qquad s\in\mathbb{R}^2 \tag{2}$$

where $\Delta=(\partial^2/\partial s_1^2)+(\partial^2/\partial s_2^2)$ is the Laplacian on $\mathbb{R}^2$; $W(s)$ is white noise (which in two dimensions can be thought of as the "derivative" of the Brownian sheet); α is a positive integer related to the smoothness parameter in (1) by $\nu=\alpha-d/2$, where d is the dimension of the space; and κ and σ are the scale and variance parameters from the Matérn covariance function, respectively. In order to assure that $x(s)$ is an ordinary random field, we assume that $\alpha>d/2$, which forces the associated smoothness parameter ν to be positive. We note here that model (2) can be easily extended in a number of ways to give non-stationary GRFs on the sphere and other more general spatial domains (Lindgren *et al.*, 2011).

[‡]The scaling that we are using for the Matérn covariance function is slightly non-standard. It has been chosen to reflect the solution to the SPDE defined in (2).

The major aim of this paper is to demonstrate that for this class of models, methods based on the SPDE formulation are significantly faster for practical computation than methods based on the covariance function. For comparison, we will consider a *mathematically equivalent* method based on a convolution (or moving average) representation of the random field

$$x(s) = \int_{\mathbb{R}^2} k(s, s')\, \mathrm{d}W(s') \tag{3}$$

where $k(s, s') = c_\eta(s, s')$ is the Matérn covariance function with smoothness parameter $\eta = (\alpha - d)/2$. In Section 3, we will show that this convolution representation is equivalent to the solution of the SPDE (2). Method based on approximations to (3) are very popular in spatial statistics, as the kernel function $k(s, s')$ can be quite general and can be used to generate non-stationary, anisotropic random fields. In this paper, we restrict ourselves to the case where $k(s, s')$ is homogeneous and the associated random field is stationary and isotropic.

3. STOCHASTIC PARTIAL DIFFERENTIAL EQUATIONS AND CONVOLUTION FIELDS

In order to demonstrate the link between the convolution representation of a Matérn field and the SPDE (2), we need to look at precisely what we mean by a *solution* to an SPDE. This section is, necessarily, more technical than the other parts of this paper. The message is that the solution of (2) can be written as (6).

This representation forms the basis for the computational method described in Section 3.1. There are many different concepts of a solution, but in this paper, we will be mainly interested in the solution defined by Walsh (1986). In order to define a solution, we need a class of test functions that can be used to "pick out" enough features of the equation to completely define it. The standard choice is to take the test functions to be the class of smooth functions that go to 0 at infinity. We then define a solution to be any random field $x(s)$ that satisfies

$$\int_{\mathbb{R}^2} x(s)(\kappa^2 - \Delta)^{\alpha/2}\phi(s)\, \mathrm{d}s = \sigma \int_{\mathbb{R}^2} \phi(s)\, \mathrm{d}W(s) \tag{4}$$

for *every* smooth test function $\phi(s)$ (Walsh, 1986). The integral on the right-hand side of (4) is, by definition, a Gaussian random variable with mean zero and variance $\int_{\mathbb{R}^2} |\phi(s)|^2\, \mathrm{d}s$. The link between (4) and convolution fields can be seen by a careful choice of the test function $\phi(s)$: take $\phi(s)$ to be the solution to

$$(\kappa^2 - \Delta)^{\alpha/2}\phi(s) = \psi(s)$$

for some smooth test function $\psi(s)$. It follows from the basic properties of $(\kappa^2 - \Delta)^{\alpha/2}$ that $\phi(s)$ is a smooth function. If we use this $\phi(s)$ in (4), we get

$$\int_{\mathbb{R}^2} x(s)\psi(s)\, \mathrm{d}s = \sigma \int_{\mathbb{R}^2} (\kappa^2 - \Delta)^{-\alpha/2}\psi(s)\, \mathrm{d}W(s) \tag{5}$$

In order to make it the rest of the way to the convolution representation, we need to know what $(\kappa^2 - \Delta)^{-\alpha/2}\psi(s)$ looks like. Fortunately, as we are working in $\mathbb{R}^2$, Fourier transform methods can be used to find an integral representation of the solution, namely

$$\sigma(\kappa^2 - \Delta)^{-\alpha/2}\psi(s) = \int_{\mathbb{R}^2} c_\eta(s, t)\psi(t)\, \mathrm{d}t$$

where $c_\eta(s, t)$ is the Matérn covariance function (1) with smoothness parameter $\eta = (\alpha - d)/2$. The link between the Matérn covariance function and fractional PDEs was noted by Whittle (1963). Plugging this integral representation into (5) and changing the order of integration, we get

$$\int_{\mathbb{R}^2} x(s)\psi(s)\, \mathrm{d}s = \int_{\mathbb{R}^2} \left(\int_{\mathbb{R}^2} c_\eta(s, t)\, \mathrm{d}W(t) \right) \psi(s)\, \mathrm{d}s$$

for every smooth test functions $\psi(s)$. This is the weak form of the equation

$$x(s) = \int_{\mathbb{R}^2} c_\eta(s, t)\, \mathrm{d}W(t) \tag{6}$$

which is our desired result, namely equation (3).

3.1. Computing with convolution representations

The typical way to use the convolution field representation (6) is to replace the integral by a sum

$$x(s) \approx \sum_{i=1}^{n} c_\eta(s, t_i)\xi_i$$

where t_i are the midpoints of the boxes B_i, $\xi_i \sim \mathrm{N}(0, |B_i|)$ are independent, and n is the number of boxes (Xia and Gelfand, 2005; Higdon, 1998). Unfortunately, this typical approach does not work for general Matérn fields—to see this, we note that the kernel function $c_\eta(t, s)$ is

singular if $\eta = (\alpha - d)/2 \leqslant 0$. This can occur as we have only assumed that $\alpha > d/2$ (which ensures the covariance itself is not singular). As α is related to the smoothness of the random field (Bolin and Lindgren, 2011), this says that the typical convolution approach is only possible for random fields that have more than $d/2$ mean-square derivatives.

This problem can be rectified by using a more appropriate discretization of (6). One possibility is to take

$$x(s) \approx \sum_{i=1}^{n} \left(\frac{1}{|B_i|} \int_{B_i} c_\eta(s,t)\,\mathrm{d}t \right) \xi_i$$

where the integral *smooths out* the singularity in the kernel function. Clearly, this is only a reasonable approach if the integral can be computed cheaply. In one dimension, where the ith box is $B_i = [t_{i,L}, t_{i,R}]$, this integral can be evaluated exactly in terms of modified Struve functions as

$$\frac{1}{t_{i,R} - t_{i,L}} \int_{t_{i,L}}^{t_{i,R}} c_\eta(s,t)\,\mathrm{d}t = \frac{\sigma^2}{2\kappa^{2\eta} D} \left(s_L I(\eta, \kappa|s_L|) - s_R I(\eta, \kappa|s_R|) \right)$$

where $s_{L/R} = s - t_{i,L/R}$, $D = t_{i,R} - t_{i,L}$, $I(\eta, t) = K_\eta(t) L_{\eta-1}(t) + L_\eta(t) K_{\eta-1}(t)$, and $L_\xi(t)$ is the modified Struve function with parameter ξ (Gradshteyn and Ryzhik, 1994, Equation 6.561.4). Figure 1 shows a kernel for $\eta = 1$ along with the corresponding smoothed version. We have been unable to analytically compute this integral for a two-dimensional field.

A simple test when assessing the quality of a method for approximating a random function is to investigate how well it can represent simple deterministic functions. Figure 2 shows the best representation in each basis to the function $x(s) \equiv 1$, when $\alpha = 2$, $\kappa = 20$, and $n = 11$. There are two things immediately apparent from this figure: the simple kernel fails to reproduce constant functions, and there are significant edge effects when using kernel methods. The first problem can be partially alleviated by taking more points—when $n = 51$, the error continues to oscillate, but its magnitude is reduced to around 1%. The second problem cannot be fixed; however, the common work-around is to enlarge the region of interest. It can be seen that the area effected is of the same order of magnitude as the range of the random field (which is around $\sqrt{8\nu}/\kappa \approx 0.17$; see Lindgren *et al.*, 2011).

Some simulated and real data examples are given in Section 5; however, Figure 2 already tells the whole story. It is clear that any basis that fails to approximate a constant function is unlikely to be able to reproduce more complicated functions, and therefore, simple kernel methods are doomed to fail. Of course, this may not happen—the behavior is not as bad when κ is small (i.e., the range is long) or α is large. In fact, the nicest thing that we can say about kernel methods is that they may not produce numerical artifacts if the range is sufficiently long or the field is very, very smooth. Although long ranges do appear in real problems, it is uncommon to use random fields smooth enough ($\nu > 3$) to make kernel methods dependable.

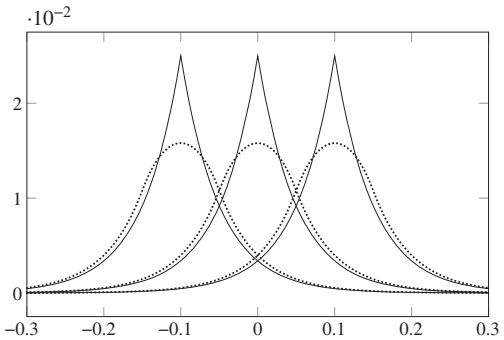


Figure 1. A comparison of the standard kernel function (solid lines) and the smoothed kernel function (dotted lines) for $\alpha = 2$ ($\eta = 1/2$) and $\kappa = 20$.

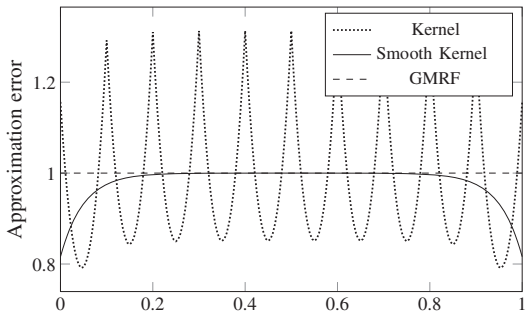


Figure 2. This figure shows the error in the approximations to $x(s) \equiv 1$ for all three sets of basis functions for $\alpha = 2$ and $\kappa = 20$. The dash-dotted line shows the simple kernel approximation. The smoothed kernel approximation (solid line) behaves much better, although it does demonstrate edge effects. The finite element basis used for the GMRF representation of the SPDE (dashed line), which is discussed in the next section, reproduces constant functions exactly.

4. GAUSSIAN MARKOV RANDOM FIELD REPRESENTATIONS OF STOCHASTIC PARTIAL DIFFERENTIAL EQUATIONS

In this section, we will outline the method of Lindgren *et al.* (2011) for constructing efficient GMRF representations of (2) for integer values of α. In the interests of clarity, we will particularly focus on the case $\alpha = 2$, which corresponds to a second-order conditional autoregression. For the general case, we refer the interested reader to Lindgren *et al.* (2011).

As we cannot represent $\mathbb{R}^2$ on a computer, the first thing that we need to do is restrict our attention to some bounded region D in $\mathbb{R}^2$, usually a rectangle. This rectangle needs to be chosen large enough to avoid any effects from the artificial boundary infecting the solution. Given such a bounded region D, the next step is to cover it in triangles in such a way that the n vertices are well distributed throughout D. It is important to note that the location of these vertices is in no way related to the location of the data points. With this in place, we will construct an approximate solution to (2) using the finite element method.

The key point in the finite element method is to replace the smooth test functions in (4) with n well-chosen piecewise linear functions. For the ith vertex in our triangularization, we define the test function $\phi_i(s)$ to be the piecewise linear function that is equal to 1 on vertex i and 0 on all other vertices. The dashed line in Figure 2 is the best representation of $x(s) \equiv 1$ in the finite element basis. This basis, which *does not depend on the parameters of the random field*, can exactly represent constant (as well as linear) functions.

Using these finite element test functions in (4) for $\alpha = 2$ leads to the linear equation

$$\left(\kappa^2 \tilde{\mathbf{C}}_n + \mathbf{G}_n\right) \tilde{\mathbf{x}}_n \stackrel{d}{=} \mathrm{N}(\mathbf{0}, \tilde{\mathbf{C}}_n)$$

where the matrices are given by

$$[\tilde{\mathbf{C}}_n]_{ij} = \int_D \phi_i(s)\phi_j(s)\,\mathrm{d}s$$
$$[\mathbf{G}_n]_{ij} = \int_D \nabla\phi_i(s) \cdot \nabla\phi_j(s)\,\mathrm{d}s$$

These integrals can be computed explicitly (see, for example, Lindgren *et al.* (2011)) and are only non-zero if vertex i is a neighbor of vertex j. Therefore, all of these matrices are *sparse* (and, in fact, symmetric positive semi-definite), which makes them amenable to fast numerical methods.

In order to complete the approximation, we note that

$$\tilde{\mathbf{x}}_n \stackrel{d}{=} \mathrm{N}(\mathbf{0}, \tilde{\mathbf{Q}}^{-1})$$

where $\tilde{\mathbf{Q}} = \left(\kappa^2 \tilde{\mathbf{C}}_n + \mathbf{G}_n\right)^{\mathrm{T}} \tilde{\mathbf{C}}_n^{-1} \left(\kappa^2 \tilde{\mathbf{C}}_n + \mathbf{G}_n\right)$. Unfortunately, as $\tilde{\mathbf{C}}_n$ is not a diagonal matrix, $\tilde{\mathbf{Q}}_n$ is not sparse; however, we follow Lindgren *et al.* (2011) and replace $\tilde{\mathbf{C}}_n$ by the diagonal matrix $\mathbf{C}_n$ that has on its diagonal the row sums of $\tilde{\mathbf{C}}_n$. The resulting GMRF representation is

$$\mathbf{x}_n \stackrel{d}{=} \mathrm{N}(\mathbf{0}, \mathbf{Q}_n^{-1}) \tag{7}$$

where $\mathbf{Q}_n = \left(\kappa^2 \mathbf{C}_n + \mathbf{G}_n\right)^{\mathrm{T}} \mathbf{C}_n^{-1} \left(\kappa^2 \mathbf{C}_n + \mathbf{G}_n\right)$.

We also note that, assuming you can even compute it in a stable and reliable manner, the (smoothed) kernel approximation is *significantly more computationally expensive* than the GMRF representation—in terms of computational complexity, $\mathcal{O}(n^3)$ versus $\mathcal{O}(n)$, where n is the number of knots/basis functions. Although the reduction in computational complexity in two dimensions is smaller—$\mathcal{O}(n^3)$ versus $\mathcal{O}(n^{3/2})$—it is more significant as n is typically much larger in two dimensions than it is in one (cf. the curse of dimensionality). Table 1 gives the asymptotic operation counts for both methods. We note that in order to make any meaningful complexity calculations, it is necessary to separate the number of data points, the number of kriging locations, and the number of basis functions, as these can vary

Table 1. Asymptotic operation counts for the calculation of the n weights for the kriging estimator $\mathbb{E}(x(s)|Y, \theta) = \sum_{i=1}^{n} w_i \phi_i(s)$, where $\phi_i(s)$ are the basis functions

Dimension	Kernel	GMRF
1	$\mathcal{O}(Nn^2 + mn + n^3)$	$\mathcal{O}(N + m + n)$
2	$\mathcal{O}(Nn^2 + mn + n^3)$	$\mathcal{O}(N + m + n^{3/2})$

Here, N is the number of data points and m is the number of kriging locations. It is assumed that $N > n$.

independently of each other. We note that when the number of data points (N) is significantly smaller than the number of basis functions (n), the kernel methods may be faster than the SPDE method, although in this regime it is likely that the asymptotic operation count will be a poor indicator of performance.

5. COMPARISON

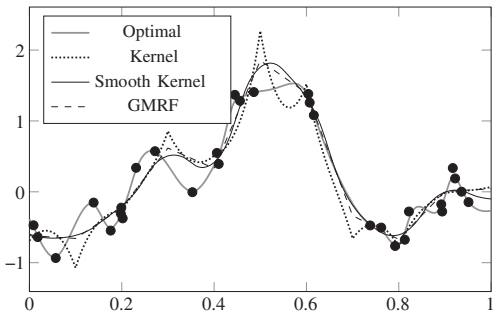
5.1. Simulated data

In this section, we compute the kriging estimate to a one-dimensional function reconstruction problem with simulated data and compare the results from the typical kernel approximation, the smoothed kernel approximation, and the GMRF representation. In order to demonstrate the differences between the methods, we have taken only 11 equally spaced knots/basis functions at locations that do not correspond to the data. We have taken $\alpha = 2$, which results in a field with one mean-square derivative. The instability of the common kernel approximation is immediately clear from Figure 3(a), especially when compared with the approximation based on the smoothed kernel. We also note that error in the GMRF representation is comparable with the error from the smoothed kernel approximation and is never worse than the common kernel estimator.

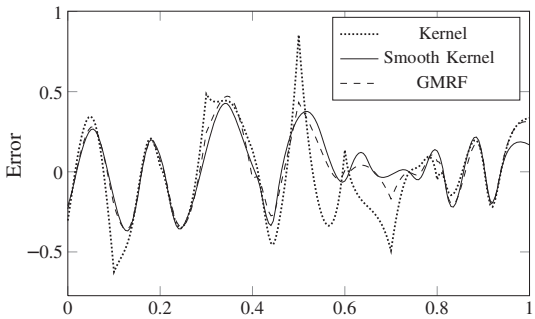
A comparison of the kernel estimate and the GMRF representation for a two-dimensional kriging problem is presented in Figure 4. In this case, we have taken our parameters to be $\kappa = 20$ and $\alpha = 3$, which results in the field having two mean-square derivatives. This is much smoother than the fields usually used in practice—the most common value being $\alpha = 3/2$, which corresponds to the exponential covariance function. It is clear from Figure 4(a) that the error in the kernel estimator is strongly related to the location of the kernels. This does not happen in the GMRF case.

5.2. Real data: precipitation anomaly in the USA

In this section, we apply the methods previously discussed to a real dataset. The precipitation anomaly for 1962 was calculated at 7352 stations across the USA, and these data were analyzed by Kaufman *et al.* (2008) and are available at http://www.image.ucar.edu/Data/Taper/. Kaufman *et al.* (2008) found that these anomaly data were normally distributed and did not deviate greatly from the assumption that the underlying field is stationary and isotropic. We will, therefore, fit a Matérn model with $\alpha = 3$ (or $\nu = 2$) using both the GMRF method and the kernel approximation. It is worth commenting on the choice of smoothing parameter. Kaufman *et al.* (2008) fitted a Matérn model with $\alpha = 3/2$ (i.e., an exponential covariance function), which *neither* of the methods considered in the paper can compute, although we note in passing that the SPDE method can construct a good approximation to it (see the authors' response to the discussion in Lindgren *et al*., 2011). Given the freedom to choose, we would take $\alpha = 2$, which gives the prior field a conceptually pleasant Markov property. Unfortunately, if



(a) Comparison of the Kriging estimates



(b) The deviation from the optimal Kriging estimates

Figure 3. The figures show the one-dimensional kriging when $\alpha = 2$ and $\kappa = 20$ with 11 knots. The GMRF representation, detailed in Section 4, is almost as accurate as the smoothed kernel approximation and is much cheaper and much more straightforward to compute. The erratic behavior of the kernel method can be clearly seen

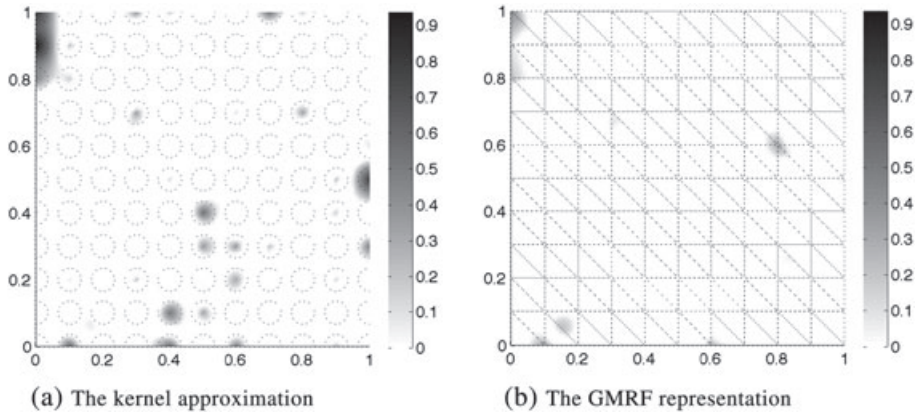


Figure 4. The figures show the deviation from the optimal kriging estimate for a two-dimensional problem with $\alpha = 3$ and $\kappa = 20$ with 11 knots in each direction. The dashed circles in (a) show the location of the integration points t_i, whereas the dashed lines in (b) show the triangles over which the finite element approximation is calculated. The dark areas in each figure show the regions of large deviation from the optimal estimate

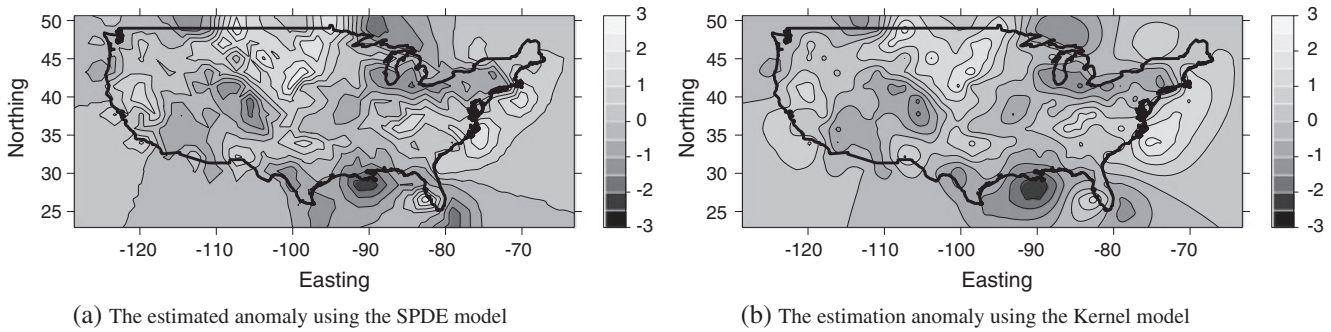


Figure 5. The figures show the kriging estimates for the 1962 precipitation anomaly computed using the SPDE approach (a) and the kernel method (b). Both surfaces contain the same features, although the SPDE reconstruction clearly demonstrates the extremely low-resolution model that has been used

we are to have even the slightest hope that the kernel methods will work, we cannot take this value of α—the kernel function is singular at $s = t$ for this value, and therefore, the Kernel method as stated will not work. This essentially forces us to use $\alpha = 3$ in order to give a fair comparison, even though this value goes against the prevailing practice in spatial statistics of using fields with only one or fewer continuous spatial derivatives. Of course, this choice can (and should) be interrogated further in practice; however, that is beyond the scope of this paper.

In order to compare the SPDE and kernel estimates, we compute them on the same grid and compare the estimates. We constructed the grid by computing a bounding box around the data points and extending it by 6% in each direction. A regular grid with 38 points in the Easting direction and 16 points in the Northing direction was then constructed. The corresponding sparse matrices for the SPDE method were of dimension 608×608, whereas the dense matrix required to compute the kernel estimate was of dimension 7352×608. The kriging surfaces, which display very similar features, are shown in Figure 5. The main differences are in the resolution of the images: because of the piecewise linear interpolation inherent in the SPDE method, the kriging estimate hews far closer to the extremely low-dimensional nature of the reconstruction than the kernel method does. The jagged nature of Figure 5(a) could easily be rectified by computing the plug-in kriging surface using a finer grid at the cost of one fairly inexpensive sparse linear solve.

The estimates in Figure 5 were produced using an almost ludicrously low resolution. This was chosen in order to be able to compute the kernel-based maximum likelihood estimates in a reasonable amount of time. Even with this low resolution, the kernel method took almost 6 min to compute the maximum likelihood estimate of the parameters in `R-2.12` on a 2010 MacBook Pro, which was 49 times longer than the SPDE method required! For larger problems, the discrepancy is even worse. This is a demonstration of the operation counts displayed in Table 1: with N data points and n points in the computational lattice, kernel methods require $\mathcal{O}(Nn^2 + n^3)$ multiplications, which compares unfavorably with the SPDE approach, which only requires $\mathcal{O}(N + n^{3/2})$ multiplications.

The tests previously performed are slightly unrealistic: it is much more useful to compare the quality of estimates that require similar amounts of time to compute. With this in mind, we computed the maximum likelihood parameter estimates and corresponding plug-in kriging surface using the SPDE method on a much finer, irregular grid. This grid had 18,203 nodes, and the parameter estimations took around 80% of the time required to compute the low resolution kernel estimate. The results can be found in Figure 6. The corresponding parameter estimates are given in Table 2. It can be seen clearly from the parameter estimates, as well as by comparing Figures 5 and 6, that the low resolution models only capture the behavior of the temperature anomalies at a very coarse level and that the parameter estimates are not stable to the choice of grid. This is a consequence of the appropriate convergence results (Lindgren *et al.*, 2011; Xia and Gelfand, 2005), and the increasing κ estimates for finer grids are also an expected effect if a smaller α value would have been more appropriate. The fundamental

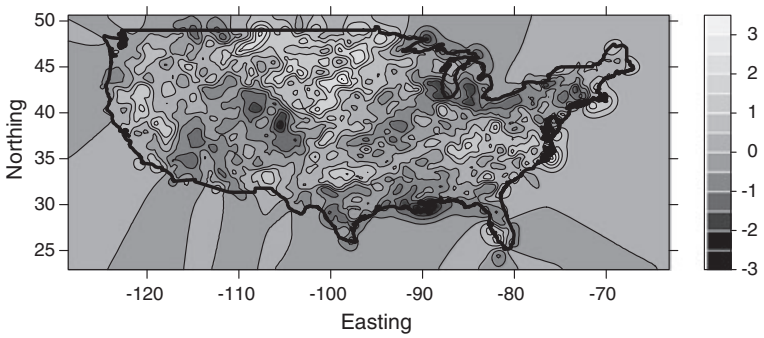


Figure 6. The kriging estimate to the 1962 precipitation anomaly computed using the SPDE approach on an irregular mesh with 18,203 nodes. This reconstruction has, appropriately, a great deal more fine-scale information than is found in the low-resolution surfaces in Figure 5

Table 2. Parameter estimates for stochastic partial differential equation (SPDE) and kernel method with various grid sizes

Method	n	κ^2	τ^2	σ^2
SPDE	348 (29×12)	0.423	0.319	0.323
Kernel	348 (29×12)	0.120	0.622	0.331
SPDE	608 (38×16)	0.718	0.121	0.289
Kernel	608 (38×16)	0.280	0.296	0.304
SPDE	4553 (unstructured)	2.977	8.7×10^{-3}	0.196
SPDE	18,203 (unstructured)	4.619	3.7×10^{-3}	0.171

Here, σ^2 is the variance of the measurement error, and τ^2 is the precision of the Gaussian field (i.e., $\mathbf{x}|\tau^2, \kappa^2 \sim N(\mathbf{0}, (\tau^2\mathbf{Q}(\kappa^2))^{-1})$).

point is that, unlike kernel methods, the SPDE approach is not designed for use at extremely low resolutions; the GMRF structure can be exploited to allow us to take several (hundred) thousand mesh points before the computation becomes unmanageable.

6. DISCUSSION

The main thesis of this paper is that *methods based on GMRF approximations to SPDEs are superior to methods based on the corresponding kernel methods*. Of course, such blanket statements need to be unpacked carefully and this section is devoted to doing just that.

The restriction to the SPDE models. To begin with, this statement strongly restricts the class of random field models to models based on SPDEs and, in particular, to the Matérn class of random fields. It has been argued by others, in particular Stein (1999), that the Matérn class is the only class of covariance functions that you need for practical spatial statistics. We have further restricted the smoothness parameter in the Matérn field to be of the form $\nu = \alpha - d/2$ for an *integer* α. We need to critically assess whether this is a practical restriction—that is, we must, for a given application, ask whether we truly need values of ν outside of this class. This is not a simple question to answer: for very large problems, the computational advantages of GMRFs make a very convincing argument for their use over kernel methods and, therefore, for restricting the smoothness parameter. This question is further complicated by the strong relationship between the scale and smoothness parameters (Zhang, 2004). In the end, the restriction on ν is not, in our opinion, a handicap, but rather, it reflects the very real fact that there are no universally appropriate methods for hard computational problems.

What do we mean by a superior method? The discussion in the previous paragraph implies a definition of a "superior" method. When saying that the GMRF representation of the SPDE model is superior to the kernel approximation, we are really saying that it is more computationally efficient. This is a reasonable definition under the condition that the two methods compute pretty much the same thing (we will discuss this in the next paragraph)—we have, after all, predicated the entire discussion on the idea that "practical problems in spatial statistics are, by and large, computational in nature." By this criterion, the GMRF method is superior based simply on operation counts. This, along with the mathematical equivalence of the two methods, highlights an important tenet of computational statistics: *equivalent mathematical statements do not lead to equally useful computational algorithms*. Of course, real life is more complicated than operation counts, and it is necessary to consider the availability of efficient software for these SPDE methods. This is discussed in the next paragraphs.

Is the GMRF representation an *approximation* to the SPDE? There is an important issue that we skirted in Section 4: *does this representation converge to the solution of the SPDE?* The simple answer is yes, but a more detailed answer is in order. If we are to use the GMRF approximation to a Matérn field, we would like to be able to control the approximation error. It can be shown using some finite element theory that the error in the approximation can be bounded above by some constant times h^2, where h is the diameter of the largest circle that can be inscribed in a triangle in the mesh (Lindgren *et al.*, 2011). This tells us that, as long as the vertices of the triangles are distributed in a reasonable way over the domain, the approximation will be good. A similar requirement is needed in order to make the kernel method converge (Xia and Gelfand, 2005). Ideally, whenever using a kernel method or a GMRF approximation to an SPDE as a prior model, the sensitivity of the marginal posteriors to the computational mesh should be tested. Realistically, however, we expect that the posterior is insensitive to small perturbations in the prior and simply treat the methods as exact. This assumption is necessary as it is common for the size of the prior field to be as large as is computationally feasible, which makes refining the mesh any further impossible.

There is a second way of thinking about the correctness of the GMRF representation that bypasses the inconvenient and technical discussion about convergence. The GMRF representation of the SPDE is (up to the diagonal approximation of $\tilde{\mathbf{C}}$) the *best* approximation to $x(s)$ over the space of piecewise linear functions defined over our triangularization of D. From this point of view, we can simply define our triangles and feel comfortable in the knowledge that the we have the best possible GMRF representation of the full random field over these triangles.

The approximation properties of the Kernel methods are troubling. Figure 2 shows that the simple kernel approximation can behave very poorly. This implies that great care must be taken when using the kernel methods, especially with respect to the selection of integration points. Furthermore, the quality of these approximations depend strongly on the smoothness and scale parameters which, again, suggests caution is in order. Finite element basis functions do not share this problem, and their approximation properties have been very well studied (Brenner and Scott, 2007).

Will these results hold for other classes of kernel functions? The main finding of this paper was that kernel methods can produce horrendous numerical artifacts when approximating Matérn fields. It is important, therefore, to ask if this is a general property or something unique to the Matérn class. The instability arises from the "sharpness" of the kernel at zero, which is not taken into consideration when discretizing the integral (3). If we recall that the kernel function is the convolution square root of the covariance function, it is easy to see that the Fourier transform of the kernel function will decay as the frequency approaches infinity at half the rate of the power spectrum of the random field. Therefore, it follows from an Abelian theorem that the kernel function will be "half as differentiable" at zero as the covariance function of the random field (Stein, 1999, Theorem 2.3) . Therefore, it follows that this instability may be present if the kernel method is being used to approximate a random field with low mean–square differentiability.

Should we also abandon kernel methods? In light of their instability, it is tempting to suggest that *in order to make computational statistics reliable, we need to forget about kernel methods*. Of course, this is not the case! Kernel methods have been used successfully in a number of practical situations, and in this paper, we have shown when instability can be expected. We do, however, firmly believe that better numerics are necessary in order to make kernel methods robust—it is unsatisfactory to deploy a kernel method and simply *hope* that the parameters do not venture into the region of instability. Although it is usually not possible to construct smoothed kernel approximations, as we did in Section 3.1, there are numerical integration schemes designed for problems with certain types of singularities at $s = 0$, and these methods can be deployed to stabilize kernel methods. That is, *in order to make computational statistics reliable, we need to stop using inappropriate numerical methods*.

The SPDE *means* something. There is a very strong conceptual advantage that SPDE methods have over kernel methods that we have not mentioned—the differential operator on the left-hand side of (2) has a strong physical interpretation. The first advantage of this is pedagogical: it allows for a new point of departure when explaining these models to people with a strong physics, engineering, or applied mathematics background. The second advantage is that the aforementioned groups of people use partial differential equations to model all manner of phenomena and this knowledge can be incorporated into building more complex models of spatial dependence. This idea can be used to easily incorporate (spatially dependent) drift terms or anisotropy into the model. Although this can also be done using convolution kernels, the SPDE methodology mirrors the way in which these processes are usually modeled in physics (Ockendon *et al.*, 2003) and computational biology (Murray, 2003).

The SPDE is independent of geometry. The great theoretical advantage of defining a GRF via an SPDE is that the form of the SPDEs that we are considering does not depend on the underlying geometry of the physical space. This is a distinct advantage when compared with GRFs defined through covariance functions or kernel methods, in which the physical geometry is of vital importance. Furthermore, we note that the finite element method depends only on the *local* behavior of the operator and the field and can therefore be applied on a complicated domain or even on a manifold (Lindgren *et al.*, 2011). In particular, Bolin and Lindgren (2011) use the GMRF representation to define a random field on a sphere in the context of environmental modeling. We note that it is possible to define a kernel representation for a Matérn GRF on a sphere using similar arguments to those in Section 3; however, the resulting kernel does not have a closed form (it follows from Fourier analysis that it will be an infinite series of spherical harmonics). It is also worth noting that all of the extensions that have been previously mentioned can also be constructed using the same methods on a manifold; that is, the GMRF representation can be used to construct non-stationary, anisotropic GRFs over general manifolds (Lindgren *et al.*, 2011).

Is there a software? The availability of a freely available software is a vital part of the development of new methods in computational statistics. As such, we have implemented the GMRF representation described in this paper as part of the R-package for INLA available from http://r-inla.org. Furthermore, the code for all of the examples in this paper, which were computed using MATLAB, is available in the supplementary material.

Do we really need to forget about the covariance function? Throughout this paper, we have barely mentioned the covariance function. This is because modern methods in spatial statistics are cannily designed to avoid using it. Of course, this does not mean we do not have one—whenever a GRF is used, there is a covariance function lying around. In fact, covariance functions could have been used to show the equivalence of kernel methods and the SPDE methods of Lindgren *et al.* (2011), rather than reducing the SPDE solution to the kernel convolution, as we did in Section 3.

Of the other main methods for spatial statistics on large data sets, covariance tapering obviously does not abandon the covariance function. Although the asymptotic properties of tapered maximum likelihood estimates are consistent with the "true" model (Furrer *et al.*, 2006), Bolin and Lindgren (2009) found that the construction of the taper and the prediction step can be extremely expensive relative to other methods. The predictive process methods of Banerjee *et al.* (2008) are constructed using a covariance function, which is quickly abandoned and never heard from again. Conversely, the fixed-rank kriging method of Cressie and Johannesson (2008) constructs a covariance function for which the kriging equations can be solved efficiently. We argue that the covariance function constructed by Cressie and Johannesson (2008) is merely a byproduct of modeling the *covariance matrix* to be of a fixed rank. Therefore, we do not see the covariance function as an integral part of the fixed rank kriging methodology—it is simply hanging around.

To summarize, with the partial exception of covariance tapering (which actually does throw away the original covariance function) and predictive processes (which only uses it in passing), no computationally efficient method for spatial statistics actually uses the covariance function directly. Of course, this is not to suggest that covariance functions are not important in the theoretical analysis of these methods—they are, after all, at the heart of Gaussian process theory. We simply do not believe that they are in any way, shape, or form useful in computational statistics.

REFERENCES

Banerjee S, Gelfand AE, Finley AO, Sang H. 2008. Gaussian predictive process models for large spatial datasets. *Journal of the Royal Statistical Society, Series B* **70**(4): 825–848.

Bolin D, Lindgren F. 2009. Spatial wavelet Markov models are more efficient than covariance tapering and process convolution, Lund University, Lund, Sweden. Preprints in Mathematical Sciences 13:2009.

Bolin D, Lindgren F. 2011. Spatial models generated by nested stochastic partial differential equations, with an application to global ozone mapping. *Annals of Applied Statistics* **5**(1): 523–550.

Brenner SC, Scott R. 2007. *The Mathematical Theory of Finite Element Methods*, 3rd ed. Springer: New York.

Cressie NAC, Johannesson G. 2008. Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society, Series B* **70**(1): 209–226.

Diggle PJ, Ribeiro PJ. 2006. *Model-based Geostatistics*, Springer Series in Statistics. Springer: New York.

Furrer R, Genton MG, Nychka D. 2006. Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics* **15**(3): 502–523.

Gradshteyn IS, Ryzhik IM. 1994. *Table of Integrals, Series, and Products*, 5th ed. Academic Press: New York.

Higdon D. 1998. A process-convolution approach to modelling temperatures in the North Atlantic Ocean. *Environmental and Ecological Statistics* **5**(2): 173–190.

Kaufman C, Schervish M, Nychka D. 2008. Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association* **103**(484): 1545–1555.

Lindgren F, Rue H, Lindström J. 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach (with discussion). *Journal of the Royal Statistical Society. Series B. Statistical Methodology* **73**(4): 423–498.

Murray J. 2003. *Mathematical Biology: Spatial Models and Biomedical Applications*. Springer Verlag: New York.

Ockendon J, Howison S, Lacey A, Movchan A. 2003. *Applied Partial Differential Equations*. Oxford University Press: USA.

Rue H, Held L. 2005. *Gaussian Markov Random Fields: Theory and Applications*, Monographs on Statistics and Applied Probability 104. Chapman & Hall: London.

Rue H, Martino S, Chopin N. 2009. Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society, Series B* **71**(2): 319–392.

Stein ML. 1999. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer-Verlag: New York.

Walsh J. 1986. An introduction to stochastic partial differential equations. *École d'Été de Probabilités de Saint Flour XIV-1984* **1180**: 265–439.

Whittle P. 1963. Stochastic processes in several dimensions. *Bulletin of the Institute of International Statistics* **40**: 974–994.

Xia G, Gelfand A. 2005. Stationary process approximation for the analysis of large spatial datasets. *ISDS Discussion Paper 2005-24*, Duke University, Durham, NC.

Zhang H. 2004. Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association* **99**(465): 250–261.

