Covariance tapering for interpolation of large spatial datasets

Furrer, Reinhard; Genton, Marc; Nychka, Douglas

Covariance Tapering for Interpolation of Large Spatial Datasets

Reinhard Furrer, Marc G. Genton and Douglas Nychka

November 2005

Interpolation of a spatially correlated random process is used in many scientific areas. The best unbiased linear predictor, often called a kriging predictor in geostatistical science, requires the solution of a (possibly large) linear system based on the covariance matrix of the observations. In this article, we show that tapering the correct covariance matrix with an appropriate compactly supported positive definite function reduces the computational burden significantly and still leads to an asymptotically optimal mean squared error. The effect of tapering is to create a sparse approximate linear system that can then be solved using sparse matrix algorithms. Monte Carlo simulations support the theoretical results. An application to a large climatological precipitation dataset is presented as a concrete and practical illustration.

Keywords: asymptotic optimality, compactly supported covariance, kriging, large linear systems, sparse matrix.

1 Introduction

Many applications of statistics across a wide range of disciplines depend on estimating the spatial extent of a physical process based on irregularly spaced observations. In many cases the most interesting spatial problems are large, and their analysis often overwhelms traditional implementations of spatial statistics.

In this work we propose an approximation to the standard linear spatial predictor that can be justified by asymptotic theory and is both accurate and computationally efficient. Our basic idea is to taper the spatial covariance function to zero beyond a certain range using a positive definite but compactly supported function. This results in sparse systems of linear equations that can be solved efficiently. Indeed, we have found that approximate taper based methods make it possible to analyze and fit large spatial data sets in a high level and interactive computing environment. Moreover, we show that tapering can result in a linear predictor that is nearly the same as the exact solution. The effect of tapering can be analyzed using the infill asymptotic theory for a misspecified covariance and we find it interesting that in our case the "misspecification" is deliberate and has computational benefits. In addition, we believe that many large spatial datasets fit the assumptions made by infill asymptotic analysis.

What do we consider a large spatial dataset? Given our geophysical perspective, we will use the US climate data record to motivate the need for more efficient statistical methods. Although a complete analysis of these data is beyond the scope of this paper, we believe that this application is useful as a realistic testbed and for comparison of different computational approaches. Briefly, the core data record is comprised of average temperature and total precipitation observations for each month from 1894 through the present (more than 1,200 months) recorded at a network of weather stations that has a peak size of more than 5,900 irregular spaced locations. Some scientific applications require that the station data be interpolated to a fine grid and we assume a prediction grid with a spacing of approximately 4 km. With

Reinhard Furrer is Assistant Professor at the Colorado School of Mines, Golden, CO 80401–1887, rfurrer@mines.edu. Marc G. Genton is Associate Professor at Texas A&M University, College Station, TX 77843–3143, genton@stat.tamu.edu. Douglas Nychka is Senior Scientist at the Geophysical Statistics Project, National Center for Atmospheric Research, Boulder, CO 80307–3000, nychka@ucar.edu.

these constraints the goal is then to predict a spatial field on a grid of size approximately 1,000×1,000 based on data from several thousand irregularly spaced locations. Moreover, this operation must be efficient as it will be repeated for many months and possibly under different covariance models.

The size of this spatial problem for climate studies is not unusual and, in fact, geophysical datasets several orders of magnitude larger can be expected based on satellite observing systems. Because of the size of these problems it is well known that a naive implementation of spatial process prediction, such as kriging, is not feasible. In addition, more complex approaches such as Bayesian hierarchical space-time models often have a kriging-like step to sample one of the full conditional distributions in a Gibbs sampling scheme. Thus, these more flexible methods are also limited in their application to large spatial problems unless the spatial prediction step can be made more efficient.

1.1 Spatial Prediction

Assume that a random spatial field $Z(\boldsymbol{x})$ is a process with covariance function $K(\boldsymbol{x}, \boldsymbol{x}^*)$ for $\boldsymbol{x}, \boldsymbol{x}^* \in \mathcal{D} \subset \mathbb{R}^d$, and is observed at the n locations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. For the illustration in Section 4, Z corresponds to monthly total precipitation for a particular month and $\mathcal{D}$ is the conterminous US. A common problem is to predict $Z(\boldsymbol{x}^*)$ given the n observations for an arbitrary $\boldsymbol{x}^* \in \mathcal{D}$. In geostatistics the standard approach, termed *kriging*, is based on the principle of minimum mean squared error (*e.g.* Cressie, 1990, 1993) and as motivation we start with the simplest spatial model. Assume that $Z(\boldsymbol{x})$ has mean zero and is observed without any measurement error. Then the best linear unbiased prediction (BLUP) at an (unobserved) location $\boldsymbol{x}^*$ is

$$\hat{Z}(\boldsymbol{x}^*) = \boldsymbol{c}^{*\mathsf{T}} \mathbf{C}^{-1} \mathbf{Z}, \tag{1}$$

where $\mathbf{Z} = \big(Z(\boldsymbol{x}_1), \ldots, Z(\boldsymbol{x}_n)\big)^{\mathsf{T}}$, $\mathbf{C}_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ and $\boldsymbol{c}_i^* = K(\boldsymbol{x}_i, \boldsymbol{x}^*)$. The BLUP (1) can also be written as $\hat{Z}(\boldsymbol{x}^*) = \boldsymbol{c}^{*\mathsf{T}} \boldsymbol{u}$ with $\mathbf{C}\boldsymbol{u} = \mathbf{Z}$. More specifically, if we assume that Z is a Gaussian process then $\hat{Z}(\boldsymbol{x}^*)$ as given by (1) is simply the conditional expectation of $Z(\boldsymbol{x}^*)$ given the observations. If the BLUP is calculated under an assumed and probably different covariance function $\tilde{K}$, the mean-squared prediction error has the form

$$\mathrm{MSE}(\boldsymbol{x}^*, \tilde{K}) = K(\boldsymbol{x}^*, \boldsymbol{x}^*) - 2\tilde{\boldsymbol{c}}^{*\mathsf{T}} \tilde{\mathbf{C}}^{-1} \boldsymbol{c}^* + \tilde{\boldsymbol{c}}^{*\mathsf{T}} \tilde{\mathbf{C}}^{-1} \mathbf{C} \tilde{\mathbf{C}}^{-1} \tilde{\boldsymbol{c}}^*, \tag{2}$$

where the tilde terms are based on $\tilde{K}$. Here it is important to note that the covariance $\tilde{K}$ in the second argument of the MSE corresponds to an assumed covariance structure and may not necessarily be the actual covariance of the process. This distinction is important if one wants to study the performance of the kriging predictor if the covariance is misspecified, or at least deviates from the actual covariance of the process. However, if K is indeed the true covariance the $\mathrm{MSE}(\boldsymbol{x}^*, K)$ in (2) simplifies to

$$\varrho(\boldsymbol{x}^*, K) = K(\boldsymbol{x}^*, \boldsymbol{x}^*) - \boldsymbol{c}^{*\mathsf{T}} \mathbf{C}^{-1} \boldsymbol{c}^*, \tag{3}$$

the well known expression for the variance of the kriging prediction. Finally, we note that $\varrho(\boldsymbol{x}^*, \tilde{K})$ is a naive prediction variance computed assuming that $\tilde{K}$ is the true covariance function.

The computation of $\boldsymbol{u} = \mathbf{C}^{-1}\mathbf{Z}$ in (1) involves the solution of a linear system that is the size of the number of observations. The operation count for exactly solving a linear system is of order n^3 and the associated storage is of order n^2. Moreover, we wish to evaluate the prediction at many grid points and so practical applications involve finding $\boldsymbol{c}^{*\mathsf{T}} \boldsymbol{u}$ for many vectors $\boldsymbol{c}^*$. These two linear algebra steps effectively limit a straightforward calculation of the spatial prediction to small problems. Note that for our climate test case $n = 5{,}906$ and $\boldsymbol{c}^{*\mathsf{T}} \boldsymbol{u}$ must be evaluated at a number of location of order 10^6. The direct computation of the variance of the kriging prediction (3) is even more demanding as this involves

either solving a different linear system at each $\boldsymbol{x}^*$ or directly inverting the matrix $\mathbf{C}$ and performing the multiplications explicitly.

An extensive literature is concerned with the relationship between the covariance function and the linear predictor; some examples include Diamond and Armstrong (1984), Yakowitz and Szidarovszky (1985), Warnes (1986), and Stein and Handcock (1989). In a series of papers Stein (1988, 1990b, 1997, 1999b) gives a thorough theoretical discussion of the effect of misspecifying the covariance function. In his approach, "misspecified" refers to a covariance similar — in some sense — to the true underlying covariance. Although much of that work is motivated by a covariance that is in error, one might adapt these results to consider the effect of deliberately modifying the "true" covariance through a taper. We note that from a theoretical perspective Stein has also suggested that tapering could be effective (Stein, 1999a, page 53) for reducing the computational burden.

1.2 Tapering and Nearest Neighbors

The goal of our work is to give an accurate approximation to (1) and (3) but also to propose a method that scales reasonably to large spatial problems. The basic idea is simple: we deliberately introduce zeros into the matrix $\mathbf{C}$ in (1) to make it sparse. The linear system (1) with a sparse covariance matrix can be solved efficiently. How the zeros are introduced is crucial. In particular one must maintain positive definiteness of any sparse modification of the covariance matrix. Let K_θ be a covariance function that is identically zero outside a particular range described by θ. Now consider a tapered covariance that is the direct (or Schur) product of K_θ and K:

$$K_{\mathsf{tap}}(\boldsymbol{x}, \boldsymbol{x}^*) = K(\boldsymbol{x}, \boldsymbol{x}^*)K_\theta(\boldsymbol{x}, \boldsymbol{x}^*).$$

An approximate predictor is obtained by replacing the covariance matrices in (1) based on K by those defined by K_{tap}. The intuition behind this choice is both that the product K_{tap} preserves some of the shape of K and that it is identically zero outside of a fixed range. Of equal importance, K_{tap} is a valid covariance, since the Schur product of two positive definite matrices is again positive definite (Horn and Johnson, 1994, Theorem 5.2.1).

Limiting the covariance function to a local neighborhood is of course not a new idea. Indeed, a very effective use of covariance tapering is well known in the data assimilation literature for numerical weather prediction (Gaspari and Cohn, 1999). Atmospheric scientists use tapering (also known as localization) in ensemble Kalman filtering (*e.g.* Houtekamer and Mitchell, 2001; Hamill *et al.*, 2001). In this application a sample covariance matrix is tapered using a compactly supported correlation function. Besides introducing computational efficiency due to the localization, tapering also has important benefits in controlling the variance of the sample covariance matrices. Although our method borrows the tapering idea from filtering applications we do not rely on the variance reduction property that is necessary for ensemble filters to be stable.

A possible objection to tapering is that it may not be effective for a spatial covariance with long range correlations. In this case the tapered covariance and the original covariance would be very different. However, one can argue qualitatively that tapering should still be useful. Although $Z(\boldsymbol{x}^*)$ may be highly correlated with distant observations it can be nearly independent of distant observations *conditional* on its neighbors. In general we expect the weights in (1) to be close to zero for observation locations that are "far" from $\boldsymbol{x}^*$. This heuristic principle is surprisingly difficult to prove but is well accepted in geostatistics based on much empirical evidence (Chilès and Delfiner, 1999, Section 3.6). Stein (2002) discusses this so-called screening effect in the case of regular lattices.

The localization of the weights in the prediction equation motivates kriging using only a neighborhood of locations. One simply calculates the spatial prediction based on a small and manageable number of

observations that are close to $\boldsymbol{x}^*$. This approach is quite useful when predicting at a limited number of locations (*e.g.* Johns *et al.*, 2003), but has several drawbacks as pointed out in Cressie (1993). However, the method involves some subjective choices and it is not clear for which single well defined statistical problem it is the exact solution. Gribov and Krivoruchko (2004) modify the kriging system such that the moving neighborhood produces continuous prediction and prediction standard error surfaces. They derive simple kriging equations based on specifically tapered covariance functions and smoothed data where the taper weights depend on the prediction and the data location. The computational cost is similar to the classical neighborhood kriging and thus the method is equally efficient.

We also acknowledge a parallel development in nearest neighbor and local estimates from nonparametric regression (*e.g.* Cleveland *et al.*, 1992). Here, the form of the estimators is justified by asymptotic optimality and usually depends on measurement error being a significant fraction of the variance in the data. For our purposes we are more concerned with the low noise situation where the fitted surface tends to interpolate or nearly interpolate the observations. However, in all of these cases the difficulty of neighborhood methods is that the neighborhood changes for each point for prediction. Although the computation is reduced for an individual point, prediction of the field without artifacts from changing neighborhoods is problematic.

Another approach to efficiently solve a linear system consists of replacing the direct inversion techniques with iterative methods, using preconditioning to lower the iteration count and computing the matrix-vector product with a fast multipole or fast moment method, *e.g.* Billings *et al.* (2002a,b). The authors claim that this method is essentially of order n, conditional on the availability of an efficient preconditioner. However, this approach is not used in this article.

We will show that the tapering and sparse matrix approach from this work has a similar operation count to nearest neighbor predictors without its disadvantages. In addition it is easy to implement and its efficiency leverages the utility of standard numerical packages for sparse linear algebra.

1.3 Outline

The paper is organized to answer the questions:

Question A. What is the increase in squared prediction error by using the taper approach?

Question B. What are the associated computational gains?

The next section answers Question A by adapting the asymptotic theory of Stein (1990b, 1997, 1999b) in order to understand the large sample properties of the proposed predictor. This is paired with some exact calculations in Section 3.2 to investigate its efficiency for finite samples. Question B can be answered by comparing standard and sparse techniques and Section 3.3 illustrates the gain in storage and computational cost when tapering is used. To emphasize the practical benefits of tapering we report timing results for the climate example in Section 4. To limit the scope of this paper we will only consider stationary processes and, in fact, restrict most of our study to the Matérn family of covariance functions. In addition we do not address the more practical spatial processes that admit some fixed effects (also known as spatial drift). The last section discusses the natural extension of tapering algorithms to nonstationary covariances and to spatial models with fixed effects.

2 Properties of Tapering

The goal of this section is to show that under specific conditions the asymptotic mean squared error of the predictions using the tapered covariance will converge to the minimal error. Following the theory of

Stein we phrase these results in terms of a misspecified covariance. Of course, the misspecification here is deliberate and involves tapering.

2.1 Matérn Covariance Functions

An important restriction throughout this analysis is that the processes and tapering functions are second order stationary and isotropic. Moreover, we will focus on the Matérn family of covariance functions. Assume that the process Z is isotropic, stationary and has an underlying Matérn covariance function defined by $K_{\alpha,\nu}(\boldsymbol{x},\boldsymbol{x}^*) = C_{\alpha,\nu}(h)$, $h = \|\boldsymbol{x}-\boldsymbol{x}^*\|$ with

$$C_{\alpha,\nu}(h) = \frac{\phi}{2^{\nu-1}\Gamma(\nu)}(\alpha h)^{\nu}\mathcal{K}_{\nu}(\alpha h), \quad \alpha>0,\ \phi>0,\ \nu>0. \tag{4}$$

Here Γ is the Gamma function and $\mathcal{K}_\nu$ is the modified Bessel function of the second kind of order ν (Abramowitz and Stegun, 1970). The process Z is m times mean square differentiable iff $\nu > m$. The parameters α and ϕ are related to the effective range and the sill, respectively. The Matérn family is a prototype for a family of covariances with different orders of smoothness and for a real argument ρ has a simple spectral density

$$\frac{\Gamma(\nu+d/2)\alpha^{2\nu}}{\pi^{d/2}\Gamma(\nu)}\cdot\frac{\phi}{(\alpha^2+\rho^2)^{\nu+d/2}}. \tag{5}$$

Without loss of generality, we assume $\phi = 1$. It is convenient to let $f_{\alpha,\nu}(\rho)$ denote the Matérn spectral density in (5) with this restriction. For certain ν the Matérn covariance function (4) has appealing forms. For example, if $\nu = 0.5$, $C_{\alpha,\nu}$ is an exponential covariance, if $\nu = n+0.5$ with n an integer, $C_{\alpha,\nu}$ is the product of an exponential covariance and a polynomial of order n. In the limit, as $\nu\to\infty$ and with appropriate scaling of α (depending on ν) $C_{\alpha,\nu}$ converges to the Gaussian covariance.

2.2 Asymptotic Equivalence of Kriging Predictors

In the following we briefly review a key result of Stein (1993) that is suitable to prove our main result in Section 2.3.

Our results are asymptotic in the context of a fixed domain size and the number of observations increasing within the domain, commonly known as infill asymptotics. (A referee pointed out, that this assumption can be weakened to assuming $\boldsymbol{x}^*$ is a limit point of the sequence below.)

Infill Condition. *Let $\boldsymbol{x}^*\in\mathcal{D}$ and $\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots$ be a dense sequence in $\mathcal{D}$ and distinct from $\boldsymbol{x}^*$.*

Asymptotic equivalence of the mean squared prediction error for two covariance functions is easiest to describe based on tail behavior of the corresponding spectral densities.

Tail Condition. *Two spectral densities f_0 and f_1 satisfy the tail condition iff*

$$\lim_{\rho\to\infty}\frac{f_1(\rho)}{f_0(\rho)} = \gamma, \quad 0<\gamma<\infty. \tag{6}$$

Based on the tail condition we have the following general result for misspecification, which can be seen as a Corollary of Theorems 1 and 2 of Stein (1993):

Theorem 2.1. *Let C_0 and C_1 be isotropic Matérn covariance functions with corresponding spectral densities f_0 and f_1. Furthermore assume that Z is an isotropic, mean zero second order stationary process with covariance function C_0 and that the Infill Condition holds. If f_0 and f_1 satisfy the Tail Condition then*

$$\lim_{n\to\infty}\frac{\mathrm{MSE}(\boldsymbol{x}^*,C_1)}{\mathrm{MSE}(\boldsymbol{x}^*,C_0)} = 1, \qquad \lim_{n\to\infty}\frac{\varrho(\boldsymbol{x}^*,C_1)}{\mathrm{MSE}(\boldsymbol{x}^*,C_0)} = \gamma.$$

The first limit indicates that the misspecified predictor using C_1 has the same convergence rate as the optimal predictor using C_0. The second limit indicates that the naive formula (3) for the prediction kriging variance also has the correct convergence rate. Finally, if $\gamma = 1$ then we have asymptotic equivalence for the MSE and the variance using the wrong covariance function. If $\gamma \neq 1$, we can divide the taper by γ to obtain asymptotic equivalence.

The theorem cited above does not identify the rate of convergence of the optimal predictor. However, these are well known for equispaced multidimensional grids (Stein, 1999a). In addition, we believe one can apply some classical interpolation theory (*e.g.* Madych and Potter, 1985) to bound the convergence rate for the kriging predictor for irregular sets of points, but this is a subject of future research.

2.3 The Taper Theorem

In order to apply Theorem 2.1 it is necessary to verify the Tail Condition for the tapered Matérn covariance function. Recall that the convolution or multiplication of two functions is equivalent to, respectively, multiplication or convolution of their Fourier transforms. Let $f_{\alpha,\nu}$ and f_θ denote the spectral densities for the Matérn covariance and taper functions. Then the spectral density f_{tap} of the tapered covariance function C_{tap} is given by

$$f_{\mathsf{tap}}(\|\boldsymbol{u}\|) = \int_{\mathbb{R}^d} f_{\alpha,\nu}(\|\boldsymbol{u}-\boldsymbol{v}\|) f_\theta(\|\boldsymbol{v}\|)\, \mathsf{d}\boldsymbol{v}. \tag{7}$$

It is reasonable to expect the two spectral densities f_{tap} and $f_{\alpha,\nu}$ to satisfy the Tail Condition when f_θ has lighter tails than $f_{\alpha,\nu}$, *i.e.* $f_\theta(\rho)/f_{\alpha,\nu}(\rho)$ converges to zero for $\rho \to \infty$. Consider the following intuitive reasoning. Suppose the spectra $f_{\alpha,\nu}$ and f_θ are the densities of independent random variables, say X and Y, respectively. Being a convolution, f_{tap} is the density of $X+Y$. The Tail Condition then implies that the variables $X+Y$ and X have the same moment properties, which is true given the initial tail assumptions on $f_{\alpha,\nu}$ and f_θ.

The Tail Condition will be replaced by a condition on the behavior of the spectral density of the taper.

Taper Condition. *Let f_θ be the spectral density of the taper covariance function, C_θ with taper range θ, and for some $\epsilon \geq 0$ and $M(\theta) < \infty$*

$$0 < f_\theta(\rho) \leq \frac{M(\theta)}{(1+\rho^2)^{\nu+d/2+\epsilon}}.$$

The following proposition gives a rigorous result leveraging the simple form for the Matérn family.

Proposition 2.2. *If f_θ satisfies the Taper Condition then f_{tap} and $f_{\alpha,\nu}$ satisfy the Tail Condition.*

The proof of Proposition 2.2 consists of evaluating (7) and showing that it has the same tail behavior as $f_{\alpha,\nu}$. The technical details are given in Appendix A.

We now formulate the main result of the paper which is a direct consequence of Theorem 2.1 and Proposition 2.2.

Theorem 2.3. (Taper Theorem) *Assume that $C_{\alpha,\nu}$ is a Matérn covariance with smoothness parameter ν and the Infill and Taper Conditions hold. Then*

$$\lim_{n\to\infty} \frac{\mathrm{MSE}(\boldsymbol{x}^*, C_{\alpha,\nu}C_\theta)}{\mathrm{MSE}(\boldsymbol{x}^*, C_{\alpha,\nu})} = 1, \tag{8}$$

$$\lim_{n\to\infty} \frac{\varrho(\boldsymbol{x}^*, C_{\alpha,\nu}C_\theta)}{\mathrm{MSE}(\boldsymbol{x}^*, C_{\alpha,\nu})} = \gamma, \tag{9}$$

where $0 < \gamma < \infty$.

The analysis above has focused on spectral densities because they provide the most accessible theory. However, because tapering is done in the spatial domain it would be practical to characterize the Taper or Tail Conditions in terms of the taper covariance directly. Although a one-to-one relationship is intuitive, a rigorous formal proof, probably based on a special case of a Tauberian theorem, is not straightforward.

In order to relate the tail behavior of the spectrum and the differentiability at the origin we conjecture the central role of the concept of principal irregular term (PIT) which is a characterization of a stationary covariance function at the origin. For a stationary, isotropic covariance function, consider the series expansion of $C(h)$ about zero. An operational definition of the PIT of C is given as the first term as a function of h in this expansion about zero that is not raised to an even power (Matheron, 1971). Stein (1999a) discusses this loose definition of the PIT in more detail. In the case of a Matérn covariance function and selected polynomial tapers, we have a one-to-one relationship between the PIT and the tail behavior, but unfortunately we could not find a universal relationship (see also Appendix B).

In addition to differentiability at the origin, the taper has to be sufficiently differentiable away from zero. If this condition is not met, the spectral density of the taper may not be strictly positive, which is the case for the triangular covariance function for example.

3 Finite Sample Accuracy and Numerical Efficiency

In this section we investigate numerically the convergence of the ratios (8) and (9) for different sample sizes, shapes of covariance functions and tapers. These results are complemented by timing studies for sparse matrix implementations.

3.1 Constructing Practical Tapers

Wu (1995), Wendland (1995, 1998), Gaspari and Cohn (1999), and Gneiting (2002) give several procedures to construct compactly supported covariance functions with arbitrary degree of differentiability at the origin and at the support length. For the applications in this work we consider the spherical covariance and two of the Wendland tapers, all parameterized so that they have support in $[0, \theta)$. All three tapers are valid covariances in $\mathbb{R}^3$. The functions are plotted in Figure 1 and summarized in Table 1. Note that the spherical covariance is linear at the origin and once differentiable at θ. The Wendland tapers are of minimal degree, given smoothness and space dimension and their tail behavior is largely known (see also Gneiting, 1999a). Based on the theory from Section 2, with respect to the Matérn smoothness parameter we use the spherical covariance to taper for $\nu \leq 0.5$, Wendland$_1$ for $\nu \leq 1.5$ and Wendland$_2$ for $\nu \leq 2.5$. Appendix B gives some additional analytical results.

Other choices of positive definite taper functions could be based on closeness to the top hat function $I_{\{h \leq \theta\}}$. One approach to construct such a function is to maximize its integral over its support, within the class of correlation functions with given support. This optimization problem is known as Turán's problem, and the solution is given in Theorem 4.4 of Ehm *et al.* (2004) and the references therein. As it turns out, the optimal d-dimensional taper is Euclid's hat function (Gneiting, 1999b), which is linear at the origin. If $d = 3$ this is the spherical taper. Another approach is to minimize the curvature at the origin. This optimization problem is discussed and solved in Gneiting (2002) and Ehm *et al.* (2004).

The concept of compatibility (see Stein, 1988, 1990a or Krasnits′kiĭ, 2000) can be used to optimize the tapering performance by rescaling the range and sill parameters to α^* and $\phi^*(\alpha^*)$ leading to the taper covariance $C_{\mathsf{tap}} = C_{\alpha^*,\nu} C_\theta$. The intuition behind this rescaling is that for large ranges α, a small taper length might be less efficient than tapering a small range α^*. Numerical experiments reported in the next section indicate that adjusting the scale in concert with tapering is only slightly more efficient than tapering alone.

Table 1: Characteristics of the tapers used in the numerical experiments. ($x_+ = \max\{0, x\}$.)

Taper	$C_\theta(h)$ for $h \geq 0$	PIT	Derivative(s) at: zero	θ	Valid taper for
Spherical	$\left(1-\frac{h}{\theta}\right)_+^2\left(1+\frac{h}{2\theta}\right)$	$-\frac{3h}{2\theta}$	0	1	$\nu \leq 0.5$
Wendland$_1$	$\left(1-\frac{h}{\theta}\right)_+^4\left(1+4\frac{h}{\theta}\right)$	$\frac{20h^3}{\theta^3}$	2	3	$\nu \leq 1.5$
Wendland$_2$	$\left(1-\frac{h}{\theta}\right)_+^6\left(1+6\frac{h}{\theta}+\frac{35h^2}{3\theta^2}\right)$	$-\frac{448h^5}{3\theta^5}$	4	5	$\nu \leq 2.5$

3.2 Numerical Experiments

Throughout this section, we will focus on the stationary Matérn covariance function in $\mathbb{R}^2$. The factors in the numerical experiments related to the covariance function are smoothness ν, range α and taper length θ. The spatial domain is $\mathcal{D} = [0,1]^2$ and the data locations are sampled randomly in $\mathcal{D}$ or on a square grid. The spatial prediction is for the center location $\boldsymbol{x}^* = (0.5, 0.5)$ and the following quantities are calculated: the mean squared error (MSE) for estimates of $Z(0.5, 0.5)$ using the true covariance and using the tapered covariance, *i.e.* $\mathrm{MSE}(\boldsymbol{x}^*, C_{\alpha,\nu})$, $\mathrm{MSE}(\boldsymbol{x}^*, C_{\mathsf{tap}})$, and the naive estimate $\varrho(\boldsymbol{x}^*, C_{\mathsf{tap}})$ of the MSE. Note that the MSE can be computed exactly for a fixed configuration of observation locations and so the only random element in these experiments is due to the locations being sampled from a uniform distribution over $\mathcal{D}$.

The first experiment examines the convergence analogous to infill asymptotics. The sample size n is varied in the range $[49, 784]$. For each sample size, 100 different sets of uniformly distributed locations are generated and, additionally, a regular grid of locations is also evaluated. The covariance parameters are $\nu = 0.5, 1, 1.5$, $\theta = 0.4$ and the range α is fixed so that correlation decreases to 0.05 over the distance 0.4. Note that for the Matérn covariance, the value of α that achieves this criterion must be found numerically and will depend on the choice of ν. As a reference, Figure 1 shows the covariance and taper functions. Outside a disk of radius 0.4 centered at $(0.5, 0.5)$ the field contributes little information for the prediction and this choice also minimizes any edge or boundary effects in the numerical experiment.

Figure 2 summarizes the results. The convergence is considerably faster for ratio (8) compared to ratio (9). The variation of the MSE ratios calculated from the random locations increases with increasing smoothness. For small n the ratio calculated from the regular grid is above the mean of the ratios calculated from the random locations. This is not surprising, since for random patterns there are often more locations within the taper range. For the spherical taper with $\nu = 0.5$, the limit of equation (9) is $\gamma = 1.5$, *cf.* left lower panel of Figure 2.

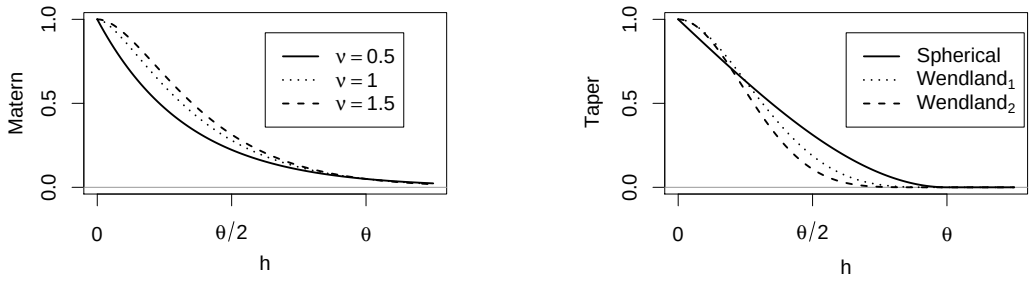

Figure 1: Matérn covariance with effective range θ (*i.e.* $\alpha = \theta/3, \theta/4, 0.21{\cdot}\theta$), sill 1 and different smoothness parameters (left). Spherical, Wendland$_1$ and Wendland$_2$ tapers with taper length θ and sill 1 (right).

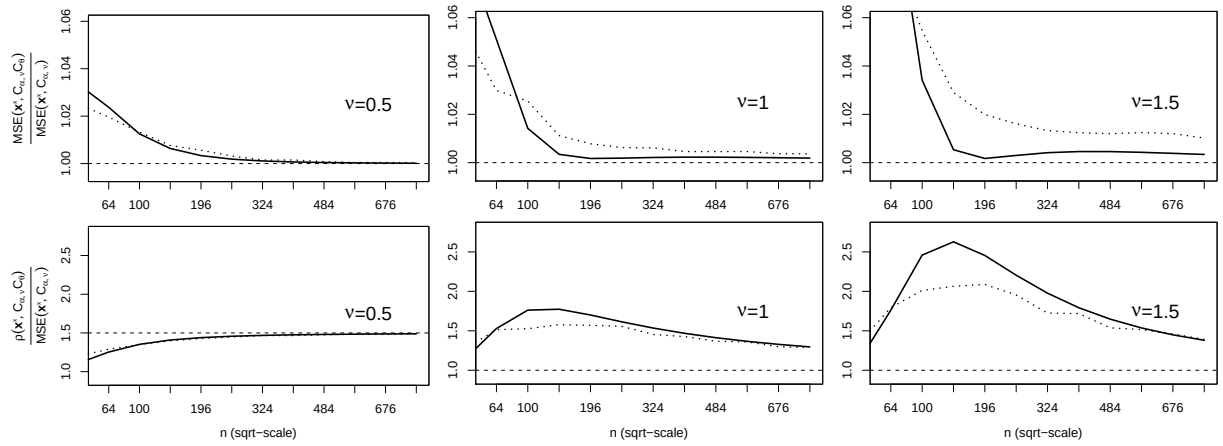

Figure 2: A comparison between taper predictor and the BLUP with respect to different true covariance functions. The ratio of the MSE of the tapered predictor to that of the BLUP is displayed in the top row and the ratio of the naive estimate of the MSE to its actual MSE in the bottom row, *cf.* ratios given by (8) and (9). The smoothness parameter is $\nu = 0.5, 1.0, 1.5$ for the left, middle and right column respectively. We use a spherical (left column) and a Wendland$_2$ tapers (middle and right columns). The solid line corresponds to the ratios calculated from a regular grid in the unit square. 100 samples of n random points in the unit square are sampled and the dotted lines show the mean of the MSE ratios.

The second numerical experiment examines the influence of the taper shape and support on accuracy. The locations are on the one hand 100 samples of uniformly distributed locations and on the other hand a 20×20 grid in the unit square and we predict at $\boldsymbol{x}^* = (0.5, 0.5)$. We calculate the ratio of $\mathrm{MSE}(\boldsymbol{x}^*, C_{\mathsf{tap}})$ and $\mathrm{MSE}(\boldsymbol{x}^*, C_{\alpha,\nu})$ for different θ, ν and different tapers. Figure 3 summarizes the results. If our goal is to be within 5% of the optimal MSE then according to Figure 3 a rule of thumb is to require 16 to 24 points within the support of the taper. A few more points should be added for very smooth fields. As a reference we also added the normalized MSE of nearest neighbor kriging with nearest neighbor distance θ to Figure 3. As expected the nearest neighbor approach performs very well, even if we include as few as 12 neighbors ($\theta = 0.1$).

We note that the ratio increases slightly for increasing smoothness; however, for $\nu = 0.5$ and $\theta > 0.15$, all three tapers perform similarly. Wendland$_1$ is slightly better than Wendland$_2$ for comparable smoothness parameters. The non-monotonic behavior using Wendland$_2$ might be explained by numerical instabilities. For the Wendland$_2$ taper, θ should be chosen slightly bigger. This may be explained by the fact that it decays much faster than the spherical taper beyond $\theta/3$.

We indicated in Section 3.1 that the original covariance could be both scaled and tapered to improve the approximation. To study this approach, consider again the same simulation setup with effective range of $C_{\alpha,\nu}(\cdot)$ of 0.4. For a fixed $\theta = 0.15$ of a Wendland$_2$ taper, we used $C_{\mathsf{tap}} = C_{\alpha^*,\nu} C_\theta$ for different values of α^*. Figure 4 shows that by reducing the range to an effective range between 0.2 and 0.3, we can gain approximately one to two percent relative accuracy. Note that the values observed at the effective range of 0.4 correspond to the values at $\theta = 0.15$ in the corresponding panels of the last column of Figure 3.

Finally, we were curious about what would happen if we simply tapered with a hard threshold, *i.e.* use the top hat taper. This is a naive approach and mimics the idea of nearest neighbors. The resulting matrices $\mathbf{C}$ in (1) are not necessarily positive definite for all θ. The top hat tapers often lead to numerical instabilities and the MSE ratios are inferior to those from positive definite tapers.

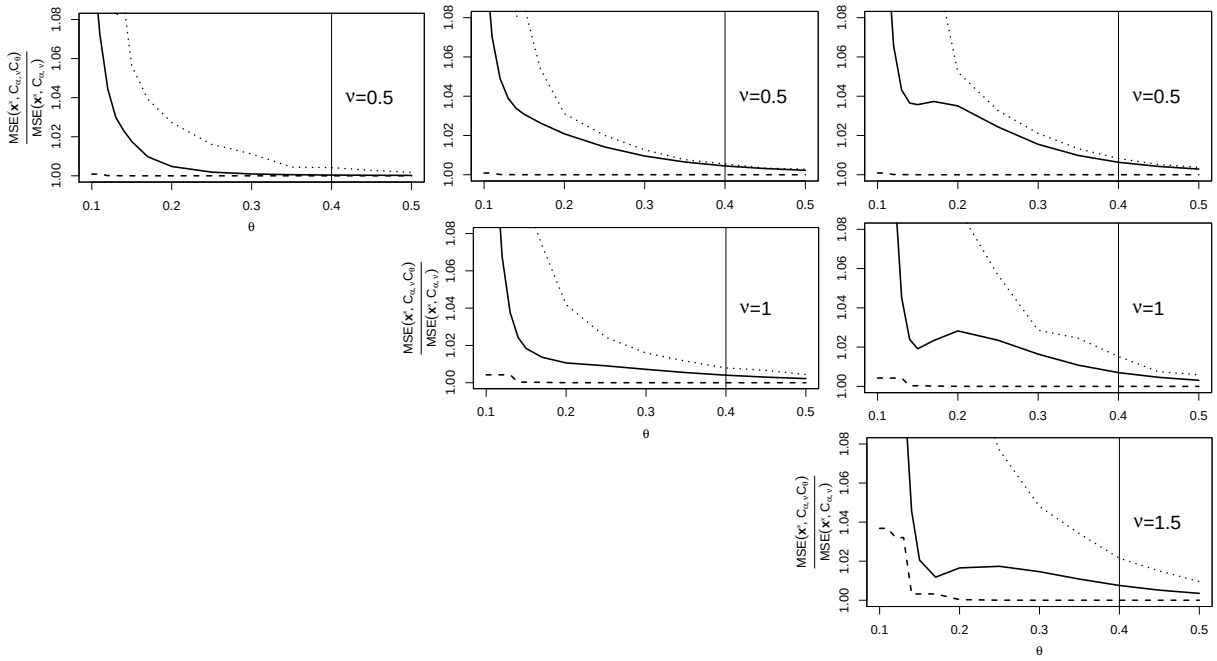

Figure 3: The ratio of the MSE using the tapered and the exact covariance functions. The columns are for spherical, Wendland$_1$, and Wendland$_2$ tapers. The rows correspond to increasing the smoothness parameter of the Matérn covariance function, fixing an effective range of 0.4. The solid line corresponds to the MSE calculated from a 20×20 regular grid in the unit square. 100 samples of $n = 400$ random points in the unit square are sampled and the dotted line shows the mean of the MSE ratios. The dashed line gives the MSE of nearest neighbor kriging with corresponding nearest neighbor distance. With $\theta = 0.1, 0.125, 0.15, 0.175, 0.2$ there are $12, 16, 24, 32, 44$ points within the taper range respectively.

3.3 Numerical Performance

For symmetric, positive definite matrices $\mathbf{C}$, the predictor in (1) is found by first performing a Cholesky factorization on $\mathbf{C} = \mathbf{A}\mathbf{A}^{\mathsf{T}}$. Then one solves the triangular systems $\mathbf{A}\boldsymbol{w} = \mathbf{Z}$ and $\mathbf{A}^{\mathsf{T}}\boldsymbol{u} = \boldsymbol{w}$ giving $\boldsymbol{u} = \mathbf{C}^{-1}\mathbf{Z}$ (also referred to as back substitution or backsolving). The final step is the calculation of the dot product $\boldsymbol{c}^{*\mathsf{T}}\boldsymbol{u}$. The common and widely used numerical software packages MATLAB and R contain a toolbox (Gilbert *et al.*, 1992) and a library SPARSEM (Koenker and Ng, 2003), respectively, with sparse matrix techniques functions to perform the Cholesky factorization.

The performance of the factorization depends on the number of non-zero elements of $\mathbf{C}$ and on how the locations are ordered. We first discuss the storage gain of sparse matrices techniques. A sparse matrix is stored as the concatenation of the vectors representing its rows. The non-zero elements are identified by two integer vectors. An $n \times m$ sparse matrix $\mathbf{S}$ with z non-zeros entries requires $8z + 4z + 4n + 1$ bytes, if we have "typical" precision with 8-byte reals and 4-byte integers. For a regular equispaced $n \times m$ grid with spacing h and taper support θ, the number of non-zero elements in the associated covariance matrix is given by

$$z = \sum_{l=0}^{n-1}(1 + I_{\{l>0\}})(n-l)\sum_{k=0}^{K_l-1}(1 + I_{\{k>0\}})(m-k), \quad K_l = \min\Big(m, \Big\lceil \big((\theta/h)^2 - l^2\big)_+^{1/2} \Big\rceil\Big), \qquad (10)$$

with $(x)_+ = \max\{0, x\}$ and $\lceil\cdot\rceil$ the biggest integer function. For irregular grids, we cannot determine directly the number of non-zero elements, but the formula can be used as a fairly good approximation

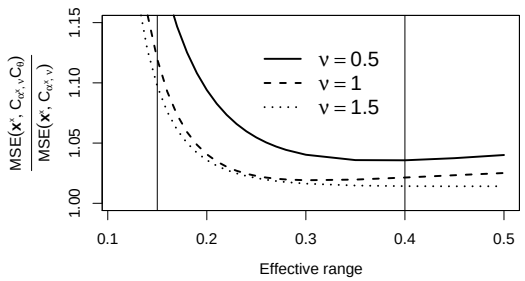

Figure 4: Mean squared errors ratios with tapered and exact covariance functions. Similar to the last column of Figure 3 but using $C_{\mathsf{tap}} = C_{\alpha^*,\nu} C_\theta$. The abscissa denotes the effective range associated with α^*. The true covariance function has an effective range of 0.4. The taper length is $\theta = 0.15$.

if the locations are uniformly distributed within a rectangle. The reduction in storage space allows us to work with much bigger problems but one must distinguish between the limitations due to the physical restrictions (RAM, available access memory) and the limitations due to the software (addressing of arrays). Currently physical limitations tend to determine the upper bound of the problem size[1].

A key assumption is that the Cholesky factor $\mathbf{A}$ of a sparse matrix will also be sparse. Surprisingly this is true, provided the matrix is properly permuted. The inverse of $\mathbf{C}$, however, is not necessarily sparse.

[1] MATLAB, for example, can handle matrices with up to $2^{28} - 1$ elements , or sparse matrices with up to roughly 2^{29} non-zero entries; `http://www.mathworks.com/support/solutions/data/1103.shtml`

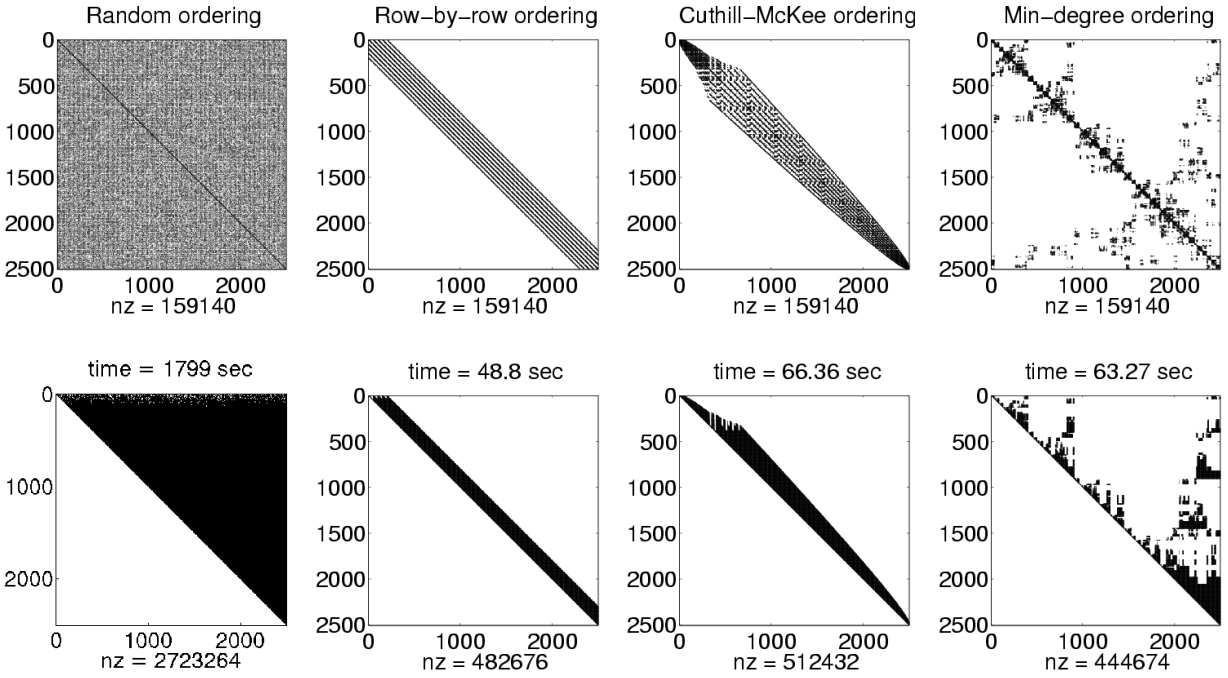

Figure 5: Influence of ordering on the performance. The top row shows the structure of a spherical covariance matrix, the bottom row its upper triangular Cholesky factor. The first column is for an arbitrary numbering, the second for a row-by-row numbering, the third column is after a reverse Cuthill–McKee reordering, the last after a minimum-degree reordering. We considered an equispaced 50×50 grid in the unit square with taper length 0.05. The indicated time is for solving 100 linear systems in MATLAB and nz states the number of non-zero elements in the matrix.

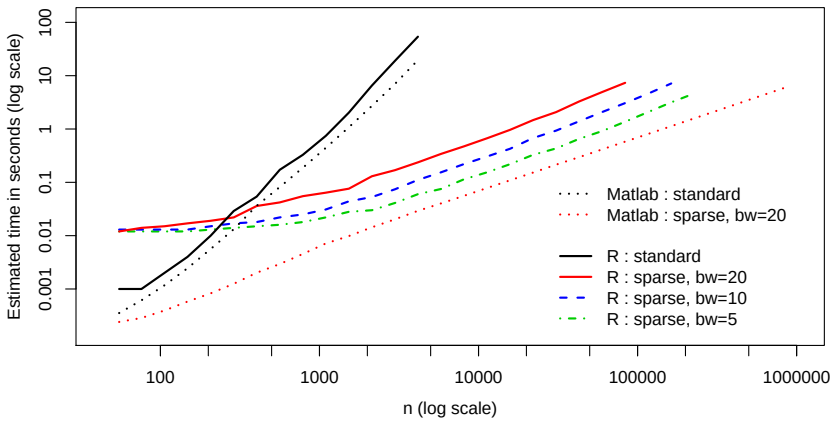

Figure 6: Comparison of the performance between R and MATLAB. A positive definite taper was applied to an equispaced one-dimensional grid of size n in $[0, 1]$. The range of the taper was such that the semi-bandwidth (bw) is 20, 10 or 5. Standard refers to Cholesky decomposition and two backsolves.

Define the semi-bandwidth s of a symmetric matrix $\mathbf{S}$ as the smallest value for which $\mathbf{S}_{i,i+s} = 0$, for all i. Then the Cholesky factor $\mathbf{A}$ has a semi-bandwidth of at most s. If the locations are not "ordered", then $\mathbf{A}$ is virtually "full". But by ordering the locations deliberately, sparsity of the Cholesky factor can be guaranteed. For ordered $n \times m$ grids with the numbering along the smaller dimension first, say n, the semi-bandwidth is

$$(n-1)L + K_L - 1, \quad L = \underset{l}{\operatorname{argmin}}\{K_l \geq 0\},$$

where K_l is given by (10). Other possible permutations are the Cuthill–McKee or minimum-degree ordering. See Figure 5 for an illustration of the effect of ordering. Although having a much larger semi-bandwidth, the minimum degree ordering[2] performs slightly better in computational cost and storage than the reverse Cuthill–McKee ordering (George and Liu, 1981). In the R library SPARSEM, there exist no explicit permutation functions and the sparse Cholesky decomposition relies on the sparse factorization and permutation algorithm by Ng and Peyton (1993).

The relative performance of the SPARSEM package of R and the SPARSE toolbox of MATLAB is evaluated by solving the linear system $\mathbf{C}\boldsymbol{u} = \mathbf{Z}$, where $\mathbf{C}$ is a tapered covariance matrix obtained form locations on a one-dimensional grid (Linux powered 2.6 GHz Xeon processor with 4 Gbytes RAM). Here, specifics of the covariance are irrelevant. The result is displayed in Figure 6. The sparse and standard approaches are approximately of the order of n and n^3 respectively. We notice that for all grid sizes MATLAB outperforms R, suggesting that the sparse methods in MATLAB are more efficient than the SPARSEM package.

4 Interpolation of a climate data set

In understanding recent climate change it is important to consider monthly temperature or precipitation fields over the past century. These (monthly) surfaces are estimated from the historical record of meteorological measurements taken at irregular station locations. The complete surfaces facilitate direct comparison with numerical climate models[3] or they can serve as inputs to ecological and vegetation models. The reader is referred to Johns *et al.* (2003) and Fuentes *et al.* (2005) for a more detailed statistical analysis of these data and discussion of the uses for the predicted fields. The methods in Johns *et al.*

[2] `http://www.mathworks.com/access/helpdesk/help/techdoc/ref/symmmd.shtml`
[3] Atmosphere-Ocean General Circulation Models

(2003) serve as a first step in creating a final data product on roughly a 4km grid for monthly total precipitation. This gridded version is an important standard observation-based resource used by the geophysical research community (see `www1.ncdc.noaa.gov/pub/data/prism100`) and is maintained and distributed by the National Climatic Data Center, a part of the National Oceanic and Atmospheric Administration (NOAA).

A key step in the creation of this data product is the ability to make spatial predictions from a large number of locations to a fine grid. As a test case we consider the monthly total precipitation record in the conterminous US for April 1948 consisting of 5,909 stations[4]. Instead of working with the raw data, we standardize the square root values. The resulting standardized values, known as anomalies, are closer to a Gaussian distribution than the raw data (Johns *et al.*, 2003). Further, there is evidence that the anomaly field is closer to being second order stationary compared to the raw scale (Fuentes *et al.*, 1998, 2005). Of course, the predicted anomaly field can always be back-transformed using predicted or interpolated climatological means and standard deviations into the raw data scale and so there is no loss of information in working with the anomaly spatial field. For the estimation of the covariance of the anomalies we used a mixture of two exponential covariances with range parameters α of 40.73 and 523.73 miles with respective sills ϕ of 0.277 and 0.722. We rescale the resulting covariance structure by a factor of 5 as explained in Section 2 and taper with a spherical covariance with a range of 50 miles. On average, each point has approximately 20 neighbor locations within 50 miles. The resulting sparse covariance matrix $\mathbf{C}$ has only 0.35% non-zero elements. The prediction surface is evaluated on a regular 0.025×0.05 latitude/longitude grid within the conterminous US, roughly at the resolution of the NOAA data product. Figure 7 shows the kriged anomaly field consisting of more than 6.6×10^5 predicted points. Table 2 summarizes the required times to construct the predicted field and the displayed figure with sparse and classical techniques. The sparse approach is faster by a factor of over 560 for step 3. While tapering reduces the time of the matrix inversion, the multiplication $\mathbf{C}^{-1}\mathbf{Z}$ still requires considerable computing time although the sparseness also reduces the amount of calculations. If the locations are on a rectangular grid, multiplication of a stationary covariance function can be done quickly by the equivalence with a convolution. Using this FFT approach we can speed up the time consuming step 4 considerably (column Sparse+FFT). The Classic+OPT approach represents the baseline to which we compare and it consists of classical techniques where costly loops are optimized (OPT) and programmed in Fortran. Applying the FFT approach to Classic+OPT, step 4 also

[4] Available at `http://www.image.ucar.edu/GSP/Data/US.monthly.met/`

Table 2: Necessary times to create the precipitation anomaly field in R with sparse and classical techniques. The result of the sparse approach is shown in Figure 7. The matrix $\tilde{\mathbf{C}}$ contains as columns the vectors $\boldsymbol{c}^*$ for the different points on the prediction grid. (Linux, 2.6 GHz Xeon processor with 4 Gbytes RAM, SparseM, Fields and Base libraries.)

Action	Time (sec)				
	Sparse	Sparse +FFT	Classic +OPT	Classic +OPT+FFT	Classic
1 Reading data, variable setup	0.54	0.54	0.54	0.54	0.54
2 Creating the matrix $\mathbf{C}$	6.35	6.35	21.59	21.59	41.34
3 Solving $\mathbf{C}\boldsymbol{x} = \mathbf{Z}$ Cholesky	0.28	0.28	169.09	169.09	169.09
3 Solving $\mathbf{C}\boldsymbol{x} = \mathbf{Z}$ Backsolve	0.03	0.03	6.13	6.13	6.13
4 Multiplying $\tilde{\mathbf{C}}^{\mathsf{T}}$ with $\mathbf{C}^{-1}\mathbf{Z}$	733.82	26.99	1830.86	26.99	4638.01
5 Creating the figure	6.19	6.19	6.19	6.19	6.19
Total	747.12	40.92	2034.40	230.53	4859.81

takes around 30 seconds due to the scalability of the problem. It would be unfair to compare our method with one based on built-in functions only (Classic). An R package, KrigSp has been posted[5] to allow readers to reproduce these results and apply the methods to other spatial data sets.

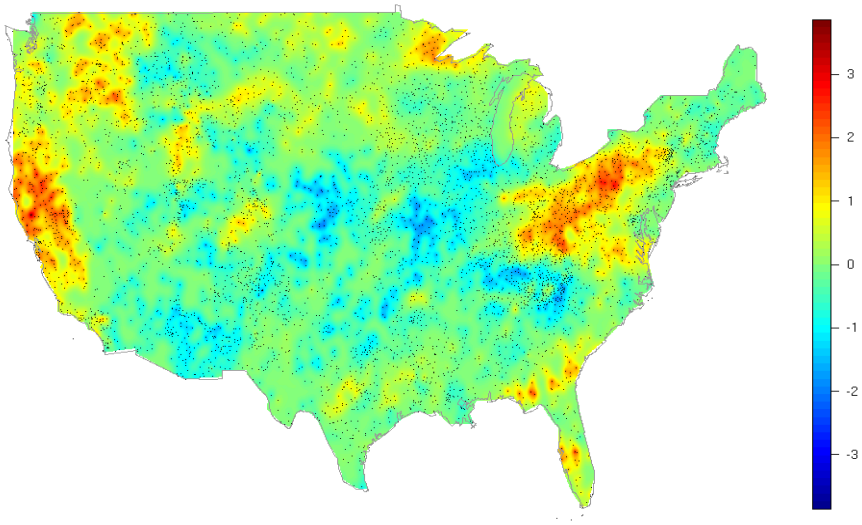

Figure 7: Kriged surface of the precipitation anomaly field of April 1948. The dots represent the 5,906 locations of the observations.

5 Discussion

In this article, we showed that truncating the covariance function to zero with appropriate polynomial tapers preserves asymptotic optimality and results in tremendous gains in computational efficiency. For sparse matrices, one can use well established algorithms to handle the sparse systems. Commonly used software packages such as R or Matlab contain libraries or toolboxes with the required functions. We showed that for large fields tapering results in a significant gain in storage and computation. In the precipitation dataset we can solve the linear system more than 500 times faster and the manageable size of the observed and predicted fields can be far bigger than with classical approaches.

Although we developed the theory for zero-mean processes with continuous covariance functions, these assumptions are not restrictive and we outline how the method can be adapted to include a fixed linear component. Consider a spatial process of the form

$$Y(\boldsymbol{x}) = \boldsymbol{m}(\boldsymbol{x})^{\mathsf{T}}\boldsymbol{\beta} + Z(\boldsymbol{x}), \tag{11}$$

where $\boldsymbol{m}$ is a known function in $\mathbb{R}^p$ and $\boldsymbol{\beta}$ is an unknown parameters in $\mathbb{R}^p$. Similar to equation (1), the BLUP of $Y(\boldsymbol{x}^*)$ is then given by

$$\hat{Y}(\boldsymbol{x}^*) = \boldsymbol{c}^{\mathsf{T}}\mathbf{C}^{-1}\left(\mathbf{Y} - \mathbf{M}\hat{\boldsymbol{\beta}}\right) + \boldsymbol{m}(\boldsymbol{x}_0)^{\mathsf{T}}\hat{\boldsymbol{\beta}}, \qquad \text{where } \hat{\boldsymbol{\beta}} = (\mathbf{M}^{\mathsf{T}}\mathbf{C}^{-1}\mathbf{M})^{-1}\mathbf{M}^{\mathsf{T}}\mathbf{C}^{-1}\mathbf{Y} \tag{12}$$

with $\mathbf{M} = \big(\boldsymbol{m}(\boldsymbol{x}_1), \ldots, \boldsymbol{m}(\boldsymbol{x}_n)\big)^{\mathsf{T}}$. The sparse approach could be used together with an iterative procedure as follows. To begin the algorithm, one estimates the mean structure, *i.e.* the vector $\boldsymbol{\beta}$ in (11), via ordinary least squares (OLS). Given the estimate $\hat{\boldsymbol{\beta}}^\star$ of the mean structure, $\mathbf{Y} - \mathbf{M}\hat{\boldsymbol{\beta}}^\star$ is kriged yielding $\mathbf{Z}^\star$. Now $\hat{\boldsymbol{\beta}}^\star$ is updated using OLS on $\mathbf{Y} - \mathbf{Z}^\star$ and we obtain a second estimate $\hat{\boldsymbol{\beta}}^\star$. These two steps are repeated

[5] `http://www.mines.edu/~rfurrer/research/programs.shtml`

until both $\hat{\boldsymbol{\beta}}^\star$ and $\mathbf{Z}^\star$ converge according to some criterion. This back-fitting procedure converges to the BLUP and we have found empirically that a few iterations usually suffice to obtain precise results. If p is not too big, the BLUP can also be obtained by solving $p+2$ linear systems as given by equation (12) using the approach described in this paper. If $\mathbf{M}$ is of full rank, the BLUP could also be written as the solution of a single linear system (Stein, 1999a, page 10). However the associated matrix is not positive definite and the system cannot be solved with standard Cholesky routines. From a theoretical perspective, Yadrenko (1983, page 138), and Stein (1990a) show that if the difference of the true mean structure and the presumed mean structure is sufficiently smooth, then Theorem 2.1 still holds. Further, Stein (1999a, Theorem 4.2), gives analogous results for processes with a nugget effect.

Our work is based on the assumption of Matérn covariances. For Theorem 2.1, this condition could be weakened, but not entirely eliminated (Stein, 1993). However, we believe that the Matérn family is sufficiently flexible to model a broad class of processes.

An entirely different approach is to approximate the covariance matrix with a compactly supported function directly, such as with truncated power function, $\phi(1-\alpha t)_+^\nu$. Numerical experiments indicate a similar performance compared with tapering. This technique can be considered as an alternative which is not based on the methodological idea of having an underlying Matérn covariance. Tapering as presented in this article also works for nonstationarity or for anisotropic processes at least with conservative taper ranges, whereas the direct approach will include more tuning.

It remains an open question how accurate the tapering approximation will be for nonstationary problems. However, our numerical results suggest that tapering is effective for different correlation ranges. A possible strategy is to choose a conservative taper that is accurate for the smallest correlation range in the domain and use this for all locations. Of course the identification of nonstationary covariances is itself difficult for large datasets but perhaps sparse techniques will also be useful in covariance estimation.

Although there are still many open questions regarding the theoretical properties of tapering and its practical application, we believe that this work is a useful step toward the analysis of large spatial problems that often have substantial scientific importance.

Acknowledgments

The authors would like to thank T. Gneiting for a discussion and insight about optimal taper functions, and A. Olenko for providing information about Tauberian theorems for correlation functions. We also acknowledge the suggestions from the associate editor and three anonymous referees that refined and improved the manuscript.

The research of Furrer and Nychka was supported in part by the Geophysical Statistics Project at the National Center for Atmospheric Research under the NSF grants DMS-9815344 and DMS-0355474. The work of Genton was partially supported by NSF grant DMS-0504896

Appendix A: Proofs

Proof. (Theorem 2.1) The spectral density of the Matérn covariance satisfies Condition (2.1) of Stein (1993) and with the Tail Condition, *i.e.* (6), Theorems 1 and 2 of Stein (1993) hold. □

Proof. (Proposition 2.2) Without loss of generality, we suppose that $\alpha = 1$ and so $f_{1,\nu}(\|\boldsymbol{\omega}\|) = M_1/(1+\|\boldsymbol{\omega}\|^2)^{\nu+d/2}$, $\nu > 0$, see also (5). We need to prove that the limit

$$\lim_{\|\boldsymbol{\omega}\|\to\infty} \frac{\int_{\mathbb{R}^d} f_{1,\nu}(\|\boldsymbol{x}\|) f_\theta(\|\boldsymbol{x}-\boldsymbol{\omega}\|)\,\mathrm{d}\boldsymbol{x}}{f_{1,\nu}(\|\boldsymbol{\omega}\|)} \tag{13}$$

exists and is not zero. As the spectral densities are radially symmetric, we choose an arbitrary direction for $\boldsymbol{\omega}$ and we set $\|\boldsymbol{x}\| = r\|\boldsymbol{u}\|$ and $\|\boldsymbol{\omega}\| = \rho\|\boldsymbol{v}\|$, with $\|\boldsymbol{u}\| = \|\boldsymbol{v}\| = 1$. The convolution reduces to

$$\int_{\mathbb{R}^d} f_{1,\nu}(\|\boldsymbol{x}\|) f_\theta(\|\boldsymbol{x} - \boldsymbol{\omega}\|)\,\mathsf{d}\boldsymbol{x} = \int_{\partial B_d} \int_0^\infty f_{1,\nu}(r) f_\theta(\|r\boldsymbol{u} - \rho\boldsymbol{v}\|) r^{d-1}\,\mathsf{d}r\,\mathsf{d}U(\boldsymbol{u}),$$

where ∂B_d is the surface of the unit sphere in $\mathbb{R}^d$ and U is the uniform probability measure on ∂B_d. We integrate over the three annuli A, B, C described by the radii $[0, \rho - \Delta], [\rho - \Delta, \rho + \Delta], [\rho + \Delta, \infty)$ (as illustrated in Figure 8 for $d = 2$). We will bound each part under the ansatz of choosing a sufficiently large ρ and $\Delta = \mathcal{O}(\rho^\delta)$, for some $(2\nu + d)/(2\nu + d + 2\epsilon) < \delta < 1$. The basic idea is that we can bound the inner integral independently of $\boldsymbol{u}$ for the respective intervals. Then the outer integrals are simply the surface of the hypersphere, *i.e.* $2\pi^{d/2}/\Gamma(n/2)$, times the inner bound. Within the ball A, the Taper Condition implies that $f_\theta(\|r\boldsymbol{u} - \rho\boldsymbol{v}\|)$ is bounded by $M/(1 + \Delta^2)^{\nu+d/2+\epsilon}$. Hence,

$$\int_0^{\rho-\Delta} f_{1,\nu}(r) f_\theta(\|r\boldsymbol{u} - \rho\boldsymbol{v}\|) r^{d-1}\,\mathsf{d}r \le \frac{M}{(1 + \Delta^2)^{\nu+d/2+\epsilon}} \int_0^{\rho-\Delta} f_{1,\nu}(r) r^{d-1}\,\mathsf{d}r.$$

As $f_{1,\nu}$ is a density in $\mathbb{R}^d$ the last integral is finite. Since $f_{1,\nu}$ is monotonically decreasing in ρ, we have for the second part

$$\int_{\rho-\Delta}^{\rho+\Delta} f_{1,\nu}(r) f_\theta(\|r\boldsymbol{u} - \rho\boldsymbol{v}\|) r^{d-1}\,\mathsf{d}r \le f_{1,\nu}(\rho - \Delta) \int_{\rho-\Delta}^{\rho+\Delta} f_\theta(\|r\boldsymbol{u} - \rho\boldsymbol{v}\|) r^{d-1}\,\mathsf{d}r. \tag{14}$$

Again, as f_θ is a density in $\mathbb{R}^d$ the last integral is finite and is positive for all $\Delta > 0$.

For the last term, we have

$$\int_{\rho+\Delta}^\infty f_{1,\nu}(r) f_\theta(\|r\boldsymbol{u} - \rho\boldsymbol{v}\|) r^{d-1}\,\mathsf{d}r \le f_{1,\nu}(\rho) \int_{\rho+\Delta}^\infty f_\theta(\|r\boldsymbol{u} - \rho\boldsymbol{v}\|) r^{d-1}\,\mathsf{d}r.$$

As ρ tends to infinity, the integral tends to zero.

Now, as $\rho, \Delta \to \infty$ with $\Delta/\rho \to 0$, the fraction (13) is bounded by

$$\lim_{\rho\to\infty} \frac{M(1 + \rho^2)^{\nu+d/2}}{M_1(1 + \Delta^2)^{\nu+d/2+\epsilon}} \int_{\partial B_d} \int_0^{\rho-\Delta} f_{1,\nu}(r) r^{d-1}\,\mathsf{d}r\,\mathsf{d}U(\boldsymbol{u}) + \frac{(1 + \rho^2)^{\nu+d/2}}{(1 + (\rho - \Delta)^2)^{\nu+d/2}} \times$$
$$\int_{\partial B_d} \int_{\rho-\Delta}^{\rho+\Delta} f_\theta(\|r\boldsymbol{u} - \rho\boldsymbol{v}\|) r^{d-1}\,\mathsf{d}r\,\mathsf{d}U(\boldsymbol{u}) + \int_{\partial B_d} \int_{\rho+\Delta}^\infty f_\theta(\|r\boldsymbol{u} - \rho\boldsymbol{v}\|) r^{d-1}\,\mathsf{d}r\,\mathsf{d}U(\boldsymbol{u}) = 1. \tag{15}$$

To show that the limit is strictly positive, consider annulus B

$$\int_{\mathbb{R}^d} f_{1,\nu}(\|\boldsymbol{x}\|) f_\theta\|\boldsymbol{x} - \boldsymbol{\omega}\|)\,\mathsf{d}\boldsymbol{x} \ge f_{1,\nu}(\rho + \Delta) \int_{\partial B_d} \int_{\rho-\Delta}^{\rho+\Delta} f_\theta(\|r\boldsymbol{u} - \rho\boldsymbol{v}\|) r^{d-1}\,\mathsf{d}r\,\mathsf{d}U(\boldsymbol{u}),$$

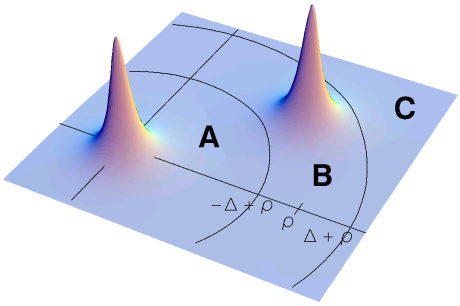

Figure 8: Separation of the convolution integral into three annuli, illustration for two dimensions.

then for all $\rho > \rho_0$, the integral is positive and has a lower bound. Further, the fraction (13) has limit

$$\lim_{\rho\to\infty} \frac{(1+\rho^2)^{\nu+d/2}}{(1+(\rho+\Delta)^2)^{\nu+d/2}} \int_{\partial B_d} \int_{\rho-\Delta}^{\rho+\Delta} f_\theta(\|r\boldsymbol{u} - \rho\boldsymbol{v}\|) r^{d-1} \,\mathsf{d}r \,\mathsf{d}U(\boldsymbol{u}) = 1. \tag{16}$$

Now, suppose that $\epsilon = 0$ and choose $\Delta = \rho/2$. The approach is similar as above and the fraction (13) is bounded by

$$\lim_{\rho\to\infty} \frac{M(1+\rho^2)^{\nu+d/2}}{M_1(1+\rho^2/4)^{\nu+d/2}} \int_{\partial B_d} \int_0^{\rho/2} f_{1,\nu}(r) r^{d-1} \,\mathsf{d}r \,\mathsf{d}U(\boldsymbol{u}) + \frac{(1+\rho^2)^{\nu+d/2}}{(1+\rho^2/4)^{\nu+d/2}} \times$$
$$\int_{\partial B_d} \int_{\rho/2}^{3\rho/2} f_\theta(\|r\boldsymbol{u} - \rho\boldsymbol{v}\|) r^{d-1} \,\mathsf{d}r \,\mathsf{d}U(\boldsymbol{u}) + \int_{\partial B_d} \int_{3\rho/2}^{\infty} f_\theta(\|r\boldsymbol{u} - \rho\boldsymbol{v}\|) r^{d-1} \,\mathsf{d}r \,\mathsf{d}U(\boldsymbol{u}) \leq \Big(\frac{M}{M_1} + 1\Big) 2^{2\nu+d}.$$

In a similar way, we can show that the limit is strictly positive. □

Proof. (Theorem 2.3) The proof of the theorem is a direct consequence of the Theorem 2.1 and Proposition 2.2. □

Appendix B: Spectral Densities of Taper Functions

Let C be an isotropic covariance function in $\mathbb{R}^d$. The corresponding spectral density can be obtained by

$$f(\rho) = (2\pi)^{-d/2} \int_0^\infty (\rho r)^{-(d-2)/2} \mathcal{J}_{(d-2)/2}(\rho r) r^{d-1} C(r) \,\mathsf{d}r,$$

where $\mathcal{J}_c$ is the Bessel function of the first kind of order c. For $d = 1$ and $d = 3$, $\mathcal{J}_{(d-2)/2}$ can be written as a function of r, a cosine and a sine function respectively. For polynomial tapers, it is thus straightforward to obtain the spectral densities for $d = 1$ and $d = 3$. For example, in one dimension, the covariance functions have the following tail behavior

$$\text{Spherical: } \lim_{\rho\to\infty} \rho^2 f_\theta(\rho) = \frac{3}{2\pi\theta}, \quad \text{Wendland}_1\text{: } \lim_{\rho\to\infty} \rho^4 f_\theta(\rho) = \frac{120}{\pi\theta^3}, \quad \text{Wendland}_2\text{: } \lim_{\rho\to\infty} \rho^6 f_\theta(\rho) = \frac{17920}{\pi\theta^5}.$$

If we write their PIT as Bh^μ (*cf.* Table 1), their tail behavior is given by $|B\mu!/\pi|$.

For $d = 3$, the tail behavior is decreased by ρ^2 but the limit does not exist. However, it can be shown that there exists constants M_2 and M_3 such for a sufficiently large ρ the spectral densities have a lower and upper bound of $M_2\rho^{-4-2i}$ and $M_3\rho^{-4-2i}$ respectively with $i = 0$ for the spherical taper and $i = 1, 2$, for the Wendland$_i$ taper (see also Theorem 3.6 in Wendland, 1998).

For $d = 2$, we use the following reasoning. The considered covariances are positive definite in $\mathbb{R}^3$ and have therefore a one-dimensional spectral density $f_{d=1}(\rho)$ that is non-increasing for $\rho > 0$ (Yaglom, 1987, page 361). This implies that $(\rho^2 - r^2)^{-1/2} \,\mathsf{d}f_{d=1}(\rho)/\,\mathsf{d}\rho \leq 0$, for all $r < \rho$ and the two-dimensional spectral density given by

$$f_{d=2}(r) = -\frac{1}{\pi} \int_r^\infty \frac{\mathsf{d}f_{d=1}(\rho)}{\mathsf{d}\rho} \frac{1}{\sqrt{\rho^2 - r^2}} \,\mathsf{d}\rho$$

is therefore strictly positive. Further, notice that

$$f_{d=2}(r) \leq \frac{1}{\pi} \int_r^\infty \Big|\frac{\mathsf{d}f_{d=1}(\rho)}{\mathsf{d}\rho}\Big| \frac{1}{\sqrt{\rho^2 - r^2}} \,\mathsf{d}\rho \leq M_3 \int_r^\infty \frac{1}{\rho^{3+2i}} \frac{1}{\sqrt{\rho^2 - r^2}} \,\mathsf{d}\rho < \frac{M_4}{r^{3+2i}},$$

with $i = 0$ for the spherical taper and $i = 1, 2$, for the Wendland$_i$ taper.

References

Abramowitz, M. and Stegun, I. A., editors (1970). *Handbook of Mathematical Functions*. Dover, New York. 5

Billings, S. D., Beatson, R. K., and Newsam, G. N. (2002a). Interpolation of geophysical data using continuous global surfaces. *Geophysics*, **67**, 1810–1822. 4

Billings, S. D., Beatson, R. K., and Newsam, G. N. (2002b). Smooth fitting of geophysical data using continuous global surfaces. *Geophysics*, **67**, 1823–1834. 4

Chilès, J.-P. and Delfiner, P. (1999). *Geostatistics: Modeling Spatial Uncertainty*. John Wiley & Sons Inc., New York. 3

Cleveland, W. S., Grosse, E., and Shyu, W. (1992). Local regression models. In Chambers, J. and Hastie, T., editors, *Statistical Models in S*, 309–376. Wadsworth and Brooks, Pacific Grove. 4

Cressie, N. A. C. (1990). The origins of kriging. *Mathematical Geology*, **22**, 239–252. 2

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. John Wiley & Sons Inc., New York, revised reprint. 2, 4

Diamond, P. and Armstrong, M. (1984). Robustness of variograms and conditioning of kriging matrices. *Journal of the International Association for Mathematical Geology*, **16**, 809–822. 3

Ehm, W., Gneiting, T., and Richards, D. (2004). Convolution roots of radial positive definite functions with compact support. *Transactions of the American Mathematical Society*, **356**, 4655–4685. 7

Fuentes, M., Kelly, R., Kittel, T., and Nychka, D. (1998). Spatial prediction of climate fields for ecological models. Technical report, Geophysical Statistics Project, National Center for Atmospheric Research, Boulder, CO. 13

Fuentes, M., Kittel, T. G. F., and Nychka, D. (2005). Sensitivity of ecological models to spatial-temporal estimation of their climate drivers: Statistical ensembles for forcing. To appear in *Ecological Applications*. 12, 13

Gaspari, G. and Cohn, S. E. (1999). Construction of correlation functions in two and three dimensions. *Quarterly Journal of the Royal Meteorological Society*, **125**, 723–757. 3, 7

George, A. and Liu, J. W. H. (1981). *Computer solution of large sparse positive definite systems*. Prentice-Hall Inc., Englewood Cliffs, N.J. 12

Gilbert, J. R., Moler, C., and Schreiber, R. (1992). Sparse matrices in MATLAB: design and implementation. *SIAM Journal on Matrix Analysis and Applications*, **13**, 333–356. 10

Gneiting, T. (1999a). Correlation functions for atmospheric data analysis. *Quarterly Journal of the Royal Meteorological Society*, **125**, 2449–2464. 7

Gneiting, T. (1999b). Radial positive definite functions generated by Euclid's hat. *Journal of Multivariate Analysis*, **69**, 88–119. 7

Gneiting, T. (2002). Compactly supported correlation functions. *Journal of Multivariate Analysis*, **83**, 493–508. 7

Gribov, A. and Krivoruchko, K. (2004). Geostatistical mapping with continuous moving neighborhood. *Mathematical Geology*, **36**, 267–281. 4

Hamill, T. M., Whitaker, J. S., and Snyder, C. (2001). Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Monthly Weather Review*, **129**, 2776–2790. 3

Horn, R. A. and Johnson, C. R. (1994). *Topics in Matrix Analysis*. Cambridge University Press, Cambridge. 3

Houtekamer, P. L. and Mitchell, H. L. (2001). A sequential ensemble Kalman filter for atmospheric data assimilation. *Monthly Weather Review*, **129**, 123–137. 3

Johns, C., Nychka, D., Kittel, T., and Daly, C. (2003). Infilling sparse records of spatial fields. *Journal of the American Statistical Association*, **98**, 796–806. 4, 12, 13

Koenker, R. and Ng, P. (2003). SparseM: A sparse matrix package for R. *Journal of Statistical Software*, **8**, 1–9. 10

Krasnits′kiĭ, S. M. (2000). On a spectral condition for the equivalence of Gaussian measures corresponding to homogeneous random fields. *Theory of Probability and Mathematical Statistics*, **60**, 95–104. 7

Madych, W. R. and Potter, E. H. (1985). An estimate for multivariate interpolation. *Journal of Approximation Theory*, **43**, 132–139. 6

Matheron, G. (1971). The theory of regionalized variables and its applications. *Cahiers du Centre de Morphologie Mathématiques*, **No. 5**, Fontainebleau, France. 7

Ng, E. G. and Peyton, B. W. (1993). Block sparse Cholesky algorithms on advanced uniprocessor computers. *SIAM Journal on Scientific Computing*, **14**, 1034–1056. 12

Stein, M. L. (1988). Asymptotically efficient prediction of a random field with a misspecified covariance function. *The Annals of Statistics*, **16**, 55–63. 3, 7

Stein, M. L. (1990a). Bounds on the efficiency of linear predictions using an incorrect covariance function. *The Annals of Statistics*, **18**, 1116–1138. 7, 15

Stein, M. L. (1990b). Uniform asymptotic optimality of linear predictions of a random field using an incorrect second-order structure. *The Annals of Statistics*, **18**, 850–872. 3, 4

Stein, M. L. (1993). A simple condition for asymptotic optimality of linear predictions of random fields. *Statistics & Probability Letters*, **17**, 399–404. 5, 15

Stein, M. L. (1997). Efficiency of linear predictors for periodic processes using an incorrect covariance function. *Journal of Statistical Planning and Inference*, **58**, 321–331. 3, 4

Stein, M. L. (1999a). *Interpolation of Spatial Data*. Springer-Verlag, New York. 3, 6, 7, 15

Stein, M. L. (1999b). Predicting random fields with increasing dense observations. *The Annals of Applied Probability*, **9**, 242–273. 3, 4

Stein, M. L. (2002). The screening effect in kriging. *The Annals of Statistics*, **30**, 298–323. 3

Stein, M. L. and Handcock, M. S. (1989). Some asymptotic properties of kriging when the covariance function is misspecified. *Mathematical Geology*, **21**, 171–190. 3

Warnes, J. J. (1986). A sensitivity analysis for universal kriging. *Mathematical Geology*, **18**, 653–676. 3

Wendland, H. (1995). Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Advances in Computational Mathematics*, **4**, 389–396. 7

Wendland, H. (1998). Error estimates for interpolation by compactly supported radial basis functions of minimal degree. *Journal of Approximation Theory*, **93**, 258–272. 7, 17

Wu, Z. M. (1995). Compactly supported positive definite radial functions. *Advances in Computational Mathematics*, **4**, 283–292. 7

Yadrenko, M. Ĭ. (1983). *Spectral theory of random fields*. Translation Series in Mathematics and Engineering. Optimization Software Inc. Publications Division, New York. 15

Yaglom, A. M. (1987). *Correlation theory of stationary and related random functions. Vol. I.* Springer Series in Statistics. Springer-Verlag, New York. 17

Yakowitz, S. J. and Szidarovszky, F. (1985). A comparison of Kriging with nonparametric regression methods. *Journal of Multivariate Analysis*, **16**, 21–53. 3