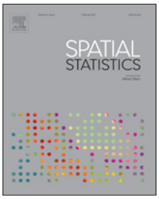
Spatial Statistics 12 (2015) 31–49

Contents lists available at ScienceDirect

Spatial Statistics

journal homepage: www.elsevier.com/locate/spasta



Equivalent kriging

William Kleiber[a,*], Douglas W. Nychka[b]

[a] *Department of Applied Mathematics, University of Colorado, Boulder, CO, United States*

[b] *Geophysical Statistics Project, Institute for Mathematics Applied to Geosciences, National Center for Atmospheric Research, United States*

ARTICLE INFO

Article history:
Received 31 July 2014
Accepted 22 January 2015
Available online 9 February 2015

Keywords:
Equivalent kernel
Kriging
Massive dataset

ABSTRACT

Most modern spatially indexed datasets are very large, with sizes commonly ranging from tens of thousands to millions of locations. Spatial analysis often focuses on spatial smoothing using the geostatistical technique known as kriging. Kriging requires covariance matrix computations whose complexity scales with the cube of the number of spatial locations, making analysis infeasible or impossible with large datasets. We introduce an approach to kriging in the presence of large datasets called *equivalent kriging*, which relies on approximating the kriging weight function using an equivalent kernel, requiring presence of a nontrivial nugget effect. Resulting kriging calculations are extremely fast and feasible in the presence of massive spatial datasets. We derive closed form kriging approximations for multiresolution classes of spatial processes, as well as under any stationary model, including popular choices such as the Matérn. The theoretical justification for equivalent kriging also leads to a correction term for irregularly spaced observations that also reduces edge effects near the domain boundary. For large sample sizes, equivalent kriging is shown to outperform covariance tapering in an example. Equivalent kriging is additionally illustrated on multiple simulated datasets, and a monthly average precipitation dataset whose size prohibits traditional geostatistical approaches.

[*] Corresponding author.
E-mail address: william.kleiber@colorado.edu (W. Kleiber).

http://dx.doi.org/10.1016/j.spasta.2015.01.004
2211-6753/© 2015 Elsevier B.V. All rights reserved.

1. Introduction

In the era of big data, spatially indexed datasets are especially prone to size-induced limitations. Indeed, modern atmospheric, hydrologic, ecological and environmental datasets are increasingly large and complex, often involving data sited at between thousands and millions of locations. A major goal when faced with such complex and noisy data is estimating the underlying physical process whose observations are subject to noise. In geostatistics, the main tool used for surface estimation is kriging, which is the linear predictor that minimizes predictive squared loss, assuming known model parameters.

The main obstacle for kriging on large datasets is solving a linear system of equations involving the spatial covariance matrix. This covariance matrix is usually dense and unstructured, and has size that scales as the square of the number of spatial locations. Over the past decade there have been a number of proposed approaches to kriging on large datasets. Many of the most popular techniques rely on a low rank representation for the spatial covariance matrix. For instance, fixed rank kriging achieves low rank by representing spatial covariances via a small set of basis functions in the observation domain (Cressie and Johannesson, 2008). Similarly, predictive processes use a conditional expectation representation at a small set of knots in the observation domain that leads to a low rank type setup (Banerjee et al., 2008). An alternative approach is covariance tapering, using a compactly supported function to impose sparsity in the covariance matrix (Furrer et al., 2006; Kaufman et al., 2008). One of the criticisms of low rank ideas is that they tend to capture low frequency behavior quite well, but are unable to model well high frequency behavior (Finley et al., 2009). To overcome this problem, an idea that retains computational feasibility is to use a low rank representation of spatial covariance, and superimpose a high frequency term that is generated by a compactly supported covariance; Sang and Huang (2012) named this approach a full scale approximation, see also Stein (2008). Finally, a simple alternative is to window the data and perform kriging locally; Stein (2014) found this approach to be favorable to low rank methods in approximating likelihoods.

A more recent idea involves approximating a Gaussian random field by a Gaussian Markov random field (Lindgren et al., 2011). This approach is computationally extremely fast for very large datasets, but is designed for processes with Matérn covariances, and can only approximate the restrictive subclass whose smoothnesses are integer plus one half values. A somewhat related approach is a multiresolution representation of the underlying stochastic field, a specific class of which has been developed very recently by Nychka et al. (in press), which they term LatticeKrig. A computationally expensive step common to many models is evaluating the likelihood (or Bayesian posterior) to determine variance and covariance parameters; some approaches to likelihood approximations have been proposed, involving an approximate gridding of the observations and using techniques for regular lattices (Fuentes, 2007). Sun et al. (2012) give an overview of some of the aforementioned techniques and others.

We propose a novel approach to kriging over large datasets called equivalent kriging. Equivalent kriging relies on approximating the kriging weights using an equivalent kernel via ideas that have previously been confined to the spline literature (Silverman, 1984). This approximation's primary limitation is that it is only valid with a nonzero nugget effect, akin to spline smoothing. The equivalent kernel is available in closed form for multiresolution processes, and has a representation as a Hankel transform for kriging with any isotropic covariance function. Specifically, we can approximate kriging under a Matérn covariance with an arbitrary smoothness, improving upon many of the previously proposed techniques. We explore both gridded and irregularly spaced data situations. Estimation can proceed by cross-validation or generalized cross-validation, as the smoothing matrix is quickly computable using the equivalent kernel approximation. We follow the technical discussion with data examples, empirically illustrating the computational advantages of equivalent kriging over traditional kriging.

As a suggestion of the timing improvements of equivalent kriging over traditional kriging, Fig. 1 illustrates a simple example. The goal is to smooth a set of noisy observations on an $n \times n$ grid by kriging or equivalent kriging using an exponential covariance model with a nugget effect. The grid is on $[0, 2\pi]^2$ with the exponential scale set to unity. Timing comparisons are shown in seconds for between approximately $n^2 = 700$ and 10 000 total locations. For even moderately large datasets, equivalent

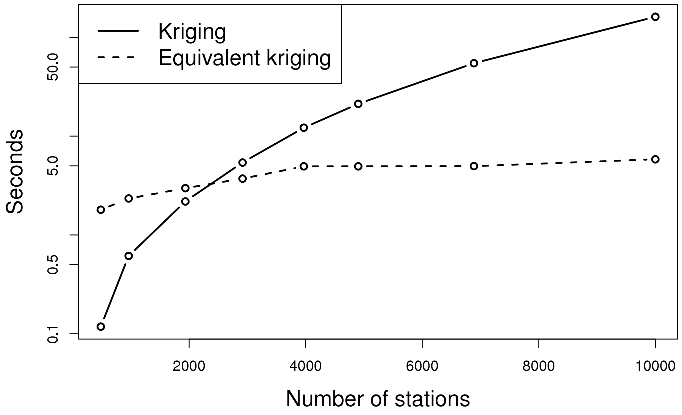


Fig. 1. Timing comparison in seconds of smoothing an $n \times n$ planar grid of noisy data using an exponential covariance using both kriging and equivalent kriging where n^2 is the number of stations; note the y-axis is indexed on the log scale. Computations are performed in R on a MacBook Pro laptop with a 2.9 GHz Intel Core i7 processor and 8 GB of 1600 MHz DDR3 RAM.

kriging improves the timing cost of traditional kriging by multiple orders of magnitude. In particular, Fig. 1 suggests that surface estimation will be possible for massive datasets where traditional kriging calculations can no longer be made. Here and throughout the manuscript, all timing comparisons are performed in a standard implementation of R (Ihaka and Gentleman, 1996) and a MacBook Pro laptop with a 2.9 GHz Intel Core i7 processor and 8 GB of 1600 MHz DDR3 RAM.

1.1. Kriging

Consider a model for an observed spatial process $Y(\mathbf{s})$ indexed by spatial location $\mathbf{s} \in \mathbb{R}^d$,

$$Y(\mathbf{s}) = \mu(\mathbf{s}) + Z(\mathbf{s}) + \varepsilon(\mathbf{s}). \tag{1}$$

The mean function $\mu(\mathbf{s})$ is typically regarded as a regression on some covariates, $\mu(\mathbf{s}) = \boldsymbol{\beta}'\mathbf{X}(\mathbf{s})$. The observations are subject to variation from this mean function by a spatially correlated stochastic term $Z(\mathbf{s})$ and a (usually) white noise process representing microscale variability and/or measurement error, $\varepsilon(\mathbf{s})$ with variance τ^2. It is common to suppose $Z(\mathbf{s})$ is a mean zero stochastic process; throughout the manuscript we denote by $k(\mathbf{s}_1, \mathbf{s}_2) = \text{Cov}(Z(\mathbf{s}_1), Z(\mathbf{s}_2))$ the covariance function for Z. For notational simplicity, we present methodology in the case $\mu(\mathbf{s}) = 0$.

Arguably the most common venture in a spatial analysis is to smooth a set of observations $\mathbf{Y} = (Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n))'$ to a location $\mathbf{s}_0$ (which may be one of $\mathbf{s}_i$, or not), or to a grid of points. In the context of the observational model (1), the most common smoother is the kriging predictor,

$$\hat{Z}(\mathbf{s}_0) = \Sigma_0' \Sigma^{-1} \mathbf{Y}, \tag{2}$$

where $\Sigma_0 = (k(\mathbf{s}_0, \mathbf{s}_1), \ldots, k(\mathbf{s}_0, \mathbf{s}_n))'$ and the (i, j)th entry of Σ is $k(\mathbf{s}_i, \mathbf{s}_j) + \tau^2 \mathbb{1}_{[i=j]}$. This predictor coincides with the conditional expectation of a multivariate normal, and is traditionally derived in the geostatistical literature as the linear predictor that minimizes the expected squared error (Cressie, 1993; Chilès and Delfiner, 1999). The main issue with the kriging predictor (2) is evaluating $\Sigma^{-1}\mathbf{Y}$, because operationally Σ is a dense and unstructured matrix of large dimension.

Throughout the manuscript we will assume n observations are available at locations corresponding to an empirical cumulative distribution function F_n on $\mathcal{D} \in \mathbb{R}^d$, and such that $F_n \to F$ as $n \to \infty$ where F has corresponding bounded density f with respect to the Lebesgue measure. Examining (2), we see the kriging predictor can be written

$$\hat{Z}(\mathbf{s}_0) = \frac{1}{n} \sum_{i=1}^{n} w_n(\mathbf{s}_0, \mathbf{s}_i) Y(\mathbf{s}_i). \tag{3}$$

We refer to w_n as the kriging weight function, which empirically acts as a local kernel smoother. The weight function w_n depends on the sample size and exact siting of observation locations. The basic goal of this manuscript is to develop an approximation such that $w_n \approx G$, with the form of G being more readily computable than w_n.

Many years ago it was established that kriging is equivalent to a variational problem (Kimeldorf and Wahba, 1971; Wahba, 1990; Chilès and Delfiner, 1999). Although this identification provided a striking theoretical connection between spline smoothing and geostatistics, it has not found substantial service in applications. Specifically, kriging is equivalent to minimizing

$$\mathcal{L}(g) = \frac{1}{n}\sum_{i=1}^{n}(Y(\mathbf{s}_i) - g(\mathbf{s}_i))^2 + \lambda\langle g, g\rangle \tag{4}$$

over the Hilbert space of functions $\mathcal{C} = \{g \mid \langle g, g\rangle < \infty\}$ and with $\langle\cdot,\cdot\rangle$ the inner product. The formal connection between this variational problem and kriging happens when $\mathcal{C}$ defines a Hilbert space of functions such that k, the covariance function of Z, is the reproducing kernel for the inner product $\langle\cdot,\cdot\rangle$ (Wahba, 1990; Furrer and Nychka, 2007). Thus, if k is the reproducing kernel for the inner product in (4), the spline solution will be identical to the estimate defined by kriging.

The identification of kriging with the variational problem of minimizing $\mathcal{L}$ yields a useful characterization of the kriging weight function. In particular, $w_n(\cdot,\cdot)$ is the reproducing kernel for the following inner product,

$$\langle h_1, h_2\rangle_w = \int h_1(\mathbf{s})h_2(\mathbf{s})\mathrm{d}F_n(\mathbf{s}) + \lambda\langle h_1, h_2\rangle. \tag{5}$$

An original proof of this result was developed by Cox (1983) by finding the minimum of $\mathcal{L}$ via its Gâteaux derivative and the fact that w_n is a reproducing kernel was not highlighted. The following proposition relies on an algebraic proof and the reproducing property of the kernel.

Proposition 1. *If $w_n(\cdot,\cdot)$ is the reproducing kernel for the inner product (5), then the minimizing solution of the variational problem (4) is $g(\cdot) = n^{-1}\sum_{i=1}^{n} w_n(\cdot, \mathbf{s}_i)Y(\mathbf{s}_i)$.*

The proof of Proposition 1 is in the Appendix. In the following section we use the fact that the kriging weight function is the reproducing kernel for (5) to develop equivalent kriging.

2. Characterizing the equivalent kernel

The main heuristic for equivalent kriging is by noticing that as $F_n \to F$, the integral in (5) converges to $\int h_1(\mathbf{s})h_2(\mathbf{s})\mathrm{d}F(\mathbf{s})$, and the resulting inner product does not depend on sample size or sample locations. Thus, we should expect the reproducing kernel for the new inner product

$$\langle h_1, h_2\rangle_\lambda = \int h_1(\mathbf{s})h_2(\mathbf{s})\mathrm{d}F(\mathbf{s}) + \lambda\langle h_1, h_2\rangle \tag{6}$$

to be close to the weight function w_n. Call G_λ the reproducing kernel of the inner product (6); we also define G_λ as the *equivalent kernel* for w_n. By standard properties of reproducing kernels, G_λ is unique. We note that, although this heuristic is asymptotic, the results in this manuscript are exact and do not rely on $n \to \infty$.

Differencing (5) and (6) and using the reproducing properties of both w_n and G_λ, it is straightforward to show that

$$w_n(\mathbf{s},\mathbf{t}) = G_\lambda(\mathbf{s},\mathbf{t}) + \big(\mathscr{R}_n w_n(\cdot,\mathbf{t})\big)(\mathbf{s}),$$

where we introduce the integral operator

$$(\mathscr{R}_n h)(\mathbf{s}) = \int G_\lambda(\mathbf{s},\mathbf{t})h(\mathbf{t})\mathrm{d}(F - F_n)(\mathbf{t}). \tag{7}$$

If we knew the exact form of the kriging weight function, the remainder term could be used as a correction to the equivalent kernel approximation for irregularly spaced observations. Unfortunately, the kriging weight function is almost never known explicitly in practice. However, a straightforward induction argument yields the following useful generalization.

Proposition 2. *If $\mathscr{R}_n$ is a bounded operator with jth power $\mathscr{R}_n^j$ and G_λ is the equivalent kernel for w_n, then*

$$w_n(\mathbf{s},\mathbf{t}) = G_\lambda(\mathbf{s},\mathbf{t}) + \sum_{j=1}^{\infty}\big(\mathscr{R}_n^j G_\lambda(\cdot,\mathbf{t})\big)(\mathbf{s}). \tag{8}$$

The convenience of this result is that all terms on the right hand side depend on the equivalent kernel, G_λ, and not the kriging weight function. We will later illustrate that these remainder terms can sometimes be derived explicitly, leading to refined estimates of the kriging weight function.

In order to complete a rigorous theoretical link between w_n and its equivalent kernel G_λ, we require some conditions on the tail behavior of G_λ. The following *exponential envelope condition* is the key.

EEC1 assumption Suppose the equivalent kernel $G_\lambda(\mathbf{s},\mathbf{t})$ satisfies the *Exponential Envelope Condition-L^1 (EEC1)* in that there are constants $\alpha, \varepsilon, K_\lambda > 0$ with $\rho = \lambda^\gamma$ where $\gamma > 0$ such that

$$|G_\lambda(\mathbf{s},\mathbf{t})| \le (K_\lambda/\rho)\exp(-(\alpha+\varepsilon)\|\mathbf{s}-\mathbf{t}\|_1/\rho)$$
$$\left|\frac{\partial}{\partial s_i}G_\lambda(\mathbf{s},\mathbf{t})\right| \le (K_\lambda/\rho^2)\exp(-(\alpha+\varepsilon)\|\mathbf{s}-\mathbf{t}\|_1/\rho)$$
$$\left|\frac{\partial^2}{\partial s_i\partial s_j}G_\lambda(\mathbf{s},\mathbf{t})\right| \le (K_\lambda/\rho^3)\exp(-(\alpha+\varepsilon)\|\mathbf{s}-\mathbf{t}\|_1/\rho)$$

where $\|\cdot\|_1$ is the L^1-norm.

Additionally set $D_n = \sup_{\mathbf{s}}|F(\mathbf{s}) - F_n(\mathbf{s})|$.

Proposition 2 and the EEC1 allow the following key approximation result.

Proposition 3. *Suppose $d = 2$ and set $\delta_n = 4D_n(K_\lambda/\rho)(1/\varepsilon + 1/\alpha)^2$ where $2(1/\varepsilon + 1/\alpha) > 1$. Then under the* EEC1 *assumption on G_λ, we have*

$$|w_n(\mathbf{s},\mathbf{t}) - G_\lambda(\mathbf{s},\mathbf{t})| \le \frac{\delta_n K_\lambda}{\rho(1-\delta_n)}\exp(-\alpha\|\mathbf{s}-\mathbf{t}\|_1/\rho).$$

The proof of Proposition 2 is given by Kleiber and Nychka (2014), while an outline of the proof of Proposition 3 is in the Appendix. Proposition 3 is particular for the case of kriging in the plane, $d = 2$, which is the most common situation in a spatial setup. The case $d = 1$ is established by Kleiber and Nychka (2014), and we conjecture the general case $d > 2$ also holds. We believe the proof is in principle straightforward using Young (1917), but whose details become messy.

Note that our EEC1 assumption is slightly different than that appearing elsewhere in the literature (Nychka, 1995; Furrer and Nychka, 2007; Kleiber and Nychka, 2014) in that ours involves the L^1 distance, while others typically involve L^2 distance; hence we denote our exponential envelope condition as EEC1, rather than EEC as in other manuscripts. For any covariance satisfying the EEC under the L^2 distance, Proposition 3 will still hold, since $\|\mathbf{s}-\mathbf{t}\|_1/\sqrt{d} < \|\mathbf{s}-\mathbf{t}\|_2$. Common covariance models imply equivalent kernels that satisfy the EEC1 assumption, including Matérn (Furrer, 2008) and multiresolution processes (Kleiber and Nychka, 2014).

3. Equivalent kriging

We now turn to the main idea of this manuscript. The *equivalent kriging* predictor $\hat{Z}_{EK}(\mathbf{s}_0)$ for $Z(\mathbf{s}_0)$ is simply the kriging predictor having replaced the exact kriging weight function by its equivalent kernel G_λ,

$$\hat{Z}(\mathbf{s}_0) = \frac{1}{n}\sum_{i=1}^{n} w_n(\mathbf{s}_0, \mathbf{s}_i)Y(\mathbf{s}_i) \approx \frac{1}{n}\sum_{i=1}^{n} G_\lambda(\mathbf{s}_0, \mathbf{s}_i)Y(\mathbf{s}_i) = \hat{Z}_{EK}(\mathbf{s}_0). \tag{9}$$

Thus, the onus of solving $\Sigma^{-1}\mathbf{Y}$ required for kriging can be relieved if we can find convenient forms for the equivalent kernel. We cover two main approximations, that for Z having any arbitrary stationary covariance, and Z being a multiresolution process. Throughout we use the spline terminology λ for the smoothing parameter; the connection to kriging is when $\lambda = \tau^2/n$.

3.1. Stationary covariances

For stationary covariances, $k(\mathbf{s}_1, \mathbf{s}_2) = k(\mathbf{s}_1 - \mathbf{s}_2)$, there is a convenient formula relating the Fourier transform of k to that of G_λ. Denote by $\mathcal{F}(g)$ the Fourier transform of the function g. The following lemma is the key link to the equivalent kernel.

Lemma 4. *If k is a stationary covariance whose kriging weight function has equivalent kernel G_λ,*

$$\mathcal{F}(G_\lambda)(\boldsymbol{\omega}) = \frac{1}{1 + \frac{\lambda}{\mathcal{F}(k)(\boldsymbol{\omega})}} \tag{10}$$

for $\boldsymbol{\omega} \in \mathbb{R}^d$.

See Furrer and Nychka (2007) or Sollich and Williams (2005) for a proof. Although we have found a few cases where G_λ has a closed form, this representation also suggests a way to approximate to any degree of accuracy the equivalent kernel for arbitrary stationary covariances.

The most popular covariance class is the Matérn class, due to its ability to capture smoothness structures of random fields (Stein, 1999). A Matérn correlation function k is defined by

$$k(\mathbf{s}_1, \mathbf{s}_2) = \frac{2^{1-\nu}}{\Gamma(\nu)}(a\|\mathbf{s}_1 - \mathbf{s}_2\|)^\nu \mathrm{K}_\nu(a\|\mathbf{s}_1 - \mathbf{s}_2\|).$$

Here, $a > 0$ is a length scale parameter, $\nu > 0$ is the key smoothness parameter and K_ν is a modified Bessel function of the second kind of order ν. Exponential correlations are popular in many fields, but this is a special case of the Matérn when $\nu = 1/2$. The spectral density of a Matérn correlation function is

$$\mathcal{F}(k)(\boldsymbol{\omega}) = \frac{\Gamma(\nu + d/2)a^{2\nu}}{\Gamma(\nu)\pi^{d/2}} \frac{1}{(a^2 + \|\boldsymbol{\omega}\|^2)^{\nu+d/2}},$$

where $\boldsymbol{\omega} \in \mathbb{R}^d$. Thus, according to (10), the equivalent kernel for a Matérn correlation is the inverse Fourier transform of

$$\mathcal{F}(G_\lambda)(\boldsymbol{\omega}) = \left(1 + \lambda \frac{\Gamma(\nu)\pi^{d/2}}{\Gamma(\nu + d/2)a^{2\nu}}(a^2 + \|\boldsymbol{\omega}\|^2)^{\nu+d/2}\right)^{-1}. \tag{11}$$

For the special case of $d = 1$ and $\nu = 1/2$, $Z(\mathbf{s})$ is an Ornstein–Uhlenbeck process, and $k(\mathbf{s}_1, \mathbf{s}_2) = k(\|\mathbf{s}_1 - \mathbf{s}_2\|) = k(r)$ is a function of the radius r. The equivalent kernel can be written in closed form as

$$G_\lambda(r) = \frac{a}{2\lambda} \frac{1}{\sqrt{\frac{a}{\lambda\pi} + a^2}} \exp\left(-r\sqrt{\frac{a}{\lambda\pi} + a^2}\right)$$

using 3.723.2 of Gradshteyn and Ryzhik (2000). A similar derivation gives a closed form for $d = 1$ and $\nu = 3/2$,

$$G_\lambda(r) = \sqrt{\frac{\pi}{2\pi\lambda a + 4}} \exp(-rA)\,(B\cos(rB) + A\sin(rB))$$

where $2A^2 = \sqrt{a^4 + 2a^3/(\lambda\pi)} + a^2$ and $2B^2 = \sqrt{a^4 + 2a^3/(\lambda\pi)} - a^2$. Zhang and Stein (1993) derived closed forms for the equivalent kernel of a limiting Matérn class when the process range tends to infinity, which can then be identified with a thin plate spline smoother.

Isotropic covariances in the plane

Assume $\mathbf{s} \in \mathbb{R}^2$ and for isotropic processes, $k(\mathbf{s}_1, \mathbf{s}_2) = k(\|\mathbf{s}_1 - \mathbf{s}_2\|) = k(r)$ where $r = \|\mathbf{s}_1 - \mathbf{s}_2\|$ is the distance between locations. Then evaluating the equivalent kernel reduces to finding a Hankel transform of order zero. In particular, a change of variables to cylindrical coordinates yields

$$G_\lambda(r) = \int_0^\infty \rho \mathrm{J}_0(r\rho)\mathcal{F}(G_\lambda)(\rho)\mathrm{d}\rho \tag{12}$$

where $\mathrm{J}_0(r)$ is a Bessel function of the first kind of order zero. Thus, the two-dimensional inverse Fourier transform becomes an integral over a univariate function.

For a two-dimensional process with Matérn covariance, the equivalent kernel via (12) is

$$G_\lambda(r) = \Gamma(\nu+1)a^{2\nu} \int_0^\infty \frac{\rho \mathrm{J}_0(r\rho)}{\Gamma(\nu+1)a^{2\nu} + \lambda\pi\Gamma(\nu)(a^2+\rho^2)^{\nu+1}}\mathrm{d}\rho.$$

We have not found many useful closed forms for $G_\lambda(r)$ in this situation, however, the univariate integral can be approximated to any degree of accuracy quickly using standard numerical analysis techniques. Some options for solving the Hankel transform include exploiting the fast Fourier transform (Siegman, 1977), or using quadrature (Key, 2012). Cree and Bones (1993) compare some competing numerical approaches for solving the Hankel transform. Thus, our implementation will just use an accurate numerical approximation to the exact Hankel transform.

3.2. Multiresolution covariances

Multiresolution processes form a flexible and computationally feasible class for spatial processes (Nychka et al., 2002). Recently, Nychka et al. (in press) introduced a so-called LatticeKrig framework for spatial modeling,

$$Z(\mathbf{s}) = \sum_{\ell=1}^{L}\sum_{i=1}^{m_\ell} c_{i\ell}\phi_{i\ell}(\mathbf{s}) \tag{13}$$

where $\phi_{i\ell}$ are known basis functions (Nychka et al., in press favor compactly supported Wendland functions). Thus, the process $Z(\mathbf{s})$ is broken up into L levels of resolution, with the ℓth having m_ℓ components. On each level ℓ, the stochastic coefficients $\{c_{i\ell}\}_{i=1}^{m_\ell}$ form a Gaussian Markov random field (GMRF). Order coefficients into a vector grouped by resolution level, $\mathbf{c} = (c_{11}, c_{12}, \ldots, c_{1m_1}, c_{21}, \ldots, c_{Lm_L})'$, and similarly group the basis functions using the same ordering, $\Phi = (\phi_{11}, \phi_{12}, \ldots, \phi_{1m_1}, \phi_{21}, \ldots, \phi_{Lm_L})'$. Then if Q is the precision matrix of $\mathbf{c}$, the covariance function for $Z(\mathbf{s})$ can be written $k(\mathbf{s}_1, \mathbf{s}_2) = \Phi(\mathbf{s}_1)'Q^{-1}\Phi(\mathbf{s}_2)$.

The multiresolution setup is convenient for equivalent kriging, as the equivalent kernel is available in closed form. If $P = \int \Phi(\mathbf{s})\Phi(\mathbf{s})'\mathrm{d}F(\mathbf{s})$ is the L^2 inner product matrix of basis functions, then the equivalent kernel for k is

$$G_\lambda(\mathbf{s}_1, \mathbf{s}_2) = \Phi(\mathbf{s}_1)'(P + \lambda Q)^{-1}\Phi(\mathbf{s}_2). \tag{14}$$

Choosing the basis functions as compactly supported implies a sparse inner product matrix P, and the GMRF structure on Q also yields a sparse matrix. Thus, computation of $G_\lambda(\mathbf{s}_1, \mathbf{s}_2)$ can take advantage of sparse matrix methods.

3.3. Regularly spaced locations

Equivalent kriging with stationary covariances can proceed by exploiting fast Fourier techniques when the observational network can be viewed as falling on a grid, such as most sets of physical model output or digital images in image analysis where pixels arise on a regular lattice.

The basic idea is to note that the equivalent kriging predictor (9) is a discrete convolution of $G_\lambda(\mathbf{s}_1, \mathbf{s}_2)$ and $Y(\mathbf{s})$. Thus, a discrete fast Fourier transform (FFT) can be used, given the equivalent kernel G_λ. The FFT does not require equal spacing in all axial directions, simply regular spacing for each axis. In particular, if $\hat{Z}_{EK} = n^{-1}\sum G_\lambda(\cdot, \mathbf{s}_i)Y(\mathbf{s}_i) = G_\lambda * Y$ where $*$ represents the convolution operator, then $\mathscr{F}(G_\lambda * Y) = \mathscr{F}(G_\lambda)\mathscr{F}(Y)$ where $\mathscr{F}$ represents the discrete Fourier transform. Thus, $\hat{Z}_{EK} = \mathscr{F}^{-1}(\mathscr{F}(G_\lambda)\mathscr{F}(Y))$, which can be computed quickly via FFT methods. For complete gridded data, this approach is similar to Wiener filtering, although next we consider irregularly spaced observations which cannot be achieved using traditional Wiener filtering.

3.4. Irregularly spaced locations

Many traditional observational datasets involve irregularly spaced locations, such as temperature and precipitation stations or pollution monitoring stations. In these cases, we propose two approaches to account for the irregularity of the network that do not cede the computational advantages of equivalent kriging.

The first option is to approximately grid the observation network and rely on the fast Fourier technique proposed for regularly spaced data. By choosing a fine grid, each data point can be associated with its nearest grid point. If all grid points happen to be populated by an observation, then the FFT can be implemented on the approximately gridded data. Usually, however, there will be a number (possibly a large number) of grid points with no associated observations. In this case we adopt the multiple imputation self-consistent algorithm of Lee and Meng (2005). Initially fill all missing grid cells with temporary values (possibly zeros, or a first-pass estimate of the spatial field such as the solution from a thin plate spline). The algorithm proceeds by applying equivalent kriging on the completed data, and replacing the observation grid cells with their original values, leaving the remaining grid cells with the predicted values. The algorithm then iterates through this procedure until some stopping criterion is reached. The fast performance of this algorithm rests on the speed of equivalent kriging.

The second option for equivalent kriging with irregularly spaced observations is to improve the approximation of the kriging weight function by a sum of remainders. In particular, Eq. (8) provides the key refinements for correcting the initial equivalent kriging solution. We see the multiresolution construction as being most useful in this case, where the remainders have explicit closed form. For instance, the first remainder can be written

$$\begin{aligned}(\mathscr{R}_n G_\lambda(\cdot, \mathbf{s}_2))(\mathbf{s}_1) &= \int G_\lambda(\mathbf{s}_1, \mathbf{t})G_\lambda(\mathbf{t}, \mathbf{s}_2)\mathrm{d}(F - F_n)(\mathbf{t}) \\ &= \Phi(\mathbf{s}_1)'(P + \lambda Q)^{-1}(P - P_n)(P + \lambda Q)^{-1}\Phi(\mathbf{s}_2) \qquad (15)\end{aligned}$$

where $P_n = \int \Phi(\mathbf{s})\Phi(\mathbf{s})'\mathrm{d}F_n(\mathbf{s})$ is just the outer product of Φ evaluated at the irregular observation locations. Formulas for additional remainder corrections follow similarly, all of which are easily computed using sparse matrix solves and the fact that remainders arise as quadratic forms.

Using the remainder formula (8) requires knowledge of the limiting observation network distribution F. In some cases, it is reasonable to assume the sampling network approaches a uniform distribution, but many observational datasets exhibit preferential sampling, placing more network locations near areas of high population, for example. In this latter case, we propose estimating the limiting network distribution F by a standard binned kernel density estimator (Wand, 1994). The kernel density estimator we consider is of the form $\hat{f}(\mathbf{s}) = \sum_{i=1}^{n} K_{\mathbf{h}}(\mathbf{s}_i - \mathbf{s})$ where K is a Gaussian density with bandwidth vector $\mathbf{h} = (h_1, \ldots, h_d)'$. Bandwidths for the density estimate can either be specified by scientifically motivated knowledge, or can be estimated. Moreover, we note that guidance in how well the F matches F_n can be gleaned from the requirement in Proposition 3 that δ_n must be less than 1 and the smaller this quantity the better the approximation.

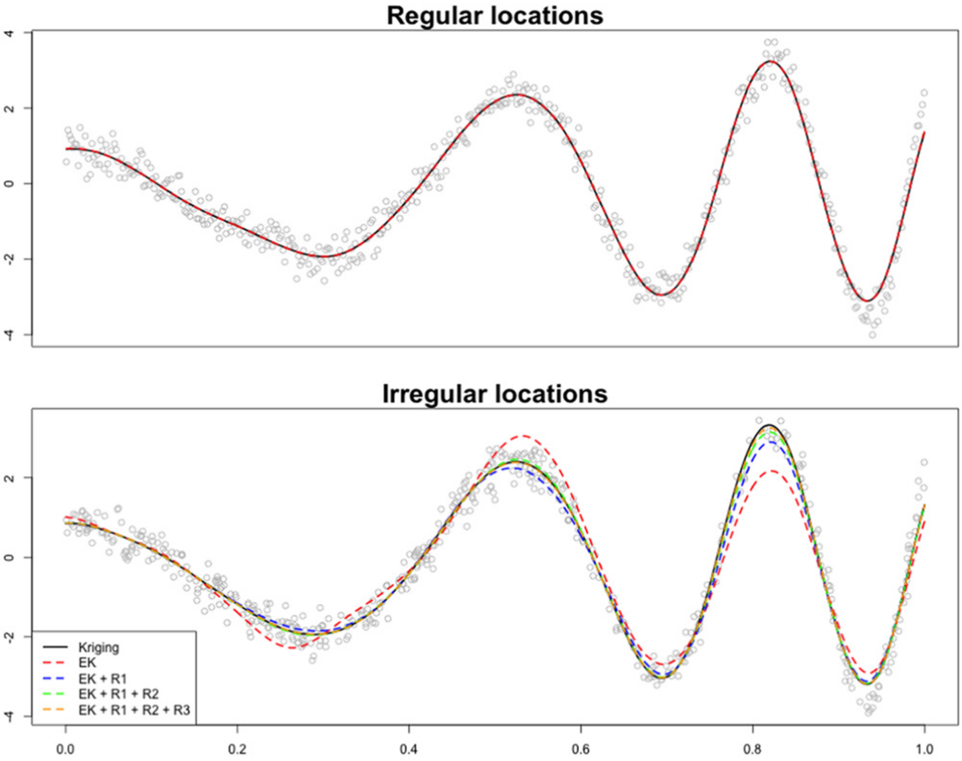


Fig. 2. Regularly and irregularly spaced observations at 500 locations. The true kriging function is shown in black, while the equivalent kriging function is a red dashed line. For irregularly spaced locations, successive refinement terms can be added to $\hat{Z}_{EK}$ that eventually converges on the kriging solution $\hat{Z}$ (see Eq. (8)). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4. Data illustrations

This section is concerned with illustrating equivalent kriging in three cases. The first is a synthetic one-dimensional example to illustrate basic properties and a correction technique for irregular networks of points. The second example is a massive dataset on a complete and incomplete grid with known truth; here we additionally make a small comparison with covariance tapering as a competing method. The third is a large irregularly spaced precipitation anomaly dataset from the National Climatic Data Center.

4.1. One-dimensional process

In this subsection we illustrate some of the basic properties of equivalent kriging, comparing an equivalent kriging solution to the true kriging surface in one dimension. Two situations are considered, the first is regularly spaced observation locations, the second irregularly spaced. For irregularly spaced locations, we show how a sequence of correction factors can be readily used to converge to the true kriging predictor. As this is an illustrative example, we use a small-to-moderately sized dataset.

Fig. 2 displays the problem setup, with $Y(\mathbf{s}) = Z(\mathbf{s}) + \varepsilon(\mathbf{s})$, $\mathbf{s} \in [0, 1]$, where $Z(\mathbf{s}) = \exp(\mathbf{s})(\cos(2\pi \mathbf{s} \exp(\mathbf{s})) - \sin(2\pi \mathbf{s} \exp(\mathbf{s})))$ and $\varepsilon(\mathbf{s})$ is Gaussian white noise process with standard deviation 0.3. The goal is to smooth the observations $Y(\mathbf{s})$ and estimate the underlying continuous function $Z(\mathbf{s})$. We consider 500 observation locations, both regularly or irregularly spaced on [0, 1]; the irregularly spaced locations are independent samples from a uniform distribution. The stochastic model

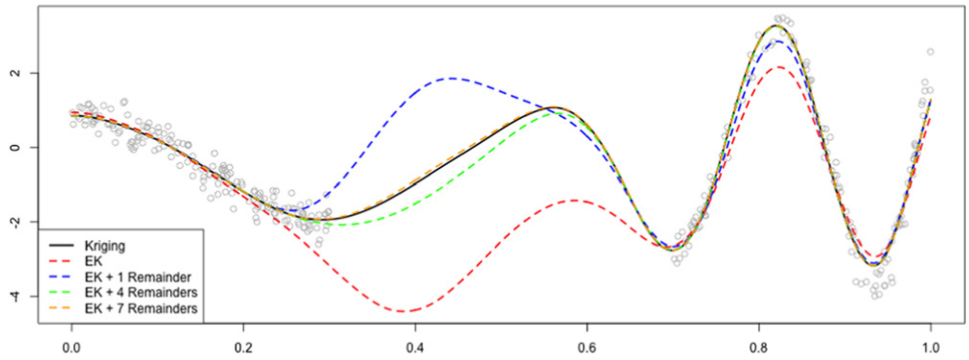


Fig. 3. Irregularly spaced observations with a missing gap in [0.3, 0.7]. The true kriging function is shown in black, while the equivalent kriging function is a red dashed line. For irregularly spaced locations, successive refinement terms can be added to $\hat{Z}_{EK}$ that eventually converges on the kriging solution $\hat{Z}$ (see Eq. (8)). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

uses a multiresolution covariance model with one level of resolution covering 20 equally spaced nodes, $Z(\mathbf{s}) = \sum_{i=1}^{20} c_i\phi_i(\mathbf{s})$, where $\phi_i(\mathbf{s}) = \phi(\mathbf{s} - \mathbf{s}_i)$ are translates of a Wendland basis function of order one with support such that $\phi(\mathbf{s}_1 - \mathbf{s}_2) = 0$ if $|\mathbf{s}_1 - \mathbf{s}_2| > 1/5$. We impose an autoregressive Markov random field structure on $\{c_i\}_{i=1}^{20}$ whose 20×20 precision matrix Q has diagonal $\text{diag}(Q) = (1, 1+\varphi^2, 1+\varphi^2, \ldots, 1+\varphi^2, 1)$ and whose upper and lower principal minor diagonals are $-\varphi$, setting $\varphi = 0.4$. Note these choices are purely for expository purposes.

Fig. 2 shows the kriging predictor (as a solid black line) along with the equivalent kriging predictor (9) (dashed red line). For equally spaced design points, the equivalent kriging solution differs from the kriging solution by at most 0.029, which occurs near the boundary, while on the interior the difference is on the order of 0.001. In the irregularly spaced design points case, $\hat{Z}_{EK}$ suffers from the irregularity of the spatial locations, and requires refinements. To correct the initial equivalent kriging solution, we use the closed form remainder for the multiresolution covariance, (15). The corrected (using up to a third order correction) equivalent kernel approximations are shown in Fig. 2, where by the third order correction the maximal error from the kriging smoother is less than 0.08, and occurs near the mode at $\mathbf{s} = 0.8$ in the surface. In practical applications with irregularly spaced data, remainders can be added to the equivalent kernel until some stopping criterion is reached.

Before moving to higher dimensions, we explore the effect of a region with zero limiting observations, that is, a nontrivial set $\{s \mid f(s) = \mathrm{d}F/\mathrm{d}s = 0\}$. We keep the same setup as the irregularly spaced data in Fig. 2, except where we remove all observations in the interval [0.3, 0.7]. For datasets exhibiting substantial gaps in the observation network, including the remainder terms of (8) is crucial. It also seems apparent from this example that a greater number of remainder terms are required to converge to a tolerable error of the kriging predictor (see Fig. 3).

4.2. *Massive completely and incompletely gridded data*

We now turn to the most common case, observation locations occurring in the plane, $d = 2$. If the locations fall on a regular grid, fast Fourier techniques can be used, even with missing data. We first entertain a 1000×1000 regular grid over $[0, 2\pi] \times [0, 2\pi]$. The observations are generated by $Y(\mathbf{s}) = Z(\mathbf{s}) + \varepsilon(\mathbf{s})$, where $Z(\mathbf{s})$ is a mean zero Gaussian process with exponential covariance function $k(\mathbf{s}_1, \mathbf{s}_2) = \exp(-\|\mathbf{s}_1 - \mathbf{s}_2\|/2)$ and $\varepsilon(\mathbf{s})$ is a Gaussian white noise process with standard deviation 0.3. Standard kriging is difficult to implement in this situation, as the covariance matrix is dense and has dimension $1{,}000{,}000 \times 1{,}000{,}000$.

If the goal is to smooth the high dimensional surface by exploiting fast Fourier methods, first note that the equivalent kriging predictor (9) is a discrete convolution of $G_\lambda(\|\mathbf{s}_1 - \mathbf{s}_2\|)$and $Y(\mathbf{s})$. As

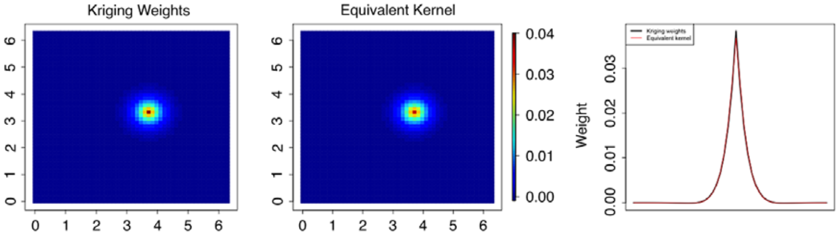


Fig. 4. Comparison of kriging weight function and equivalent kernel approximation in two dimensions on a 50 × 50 grid for an exponential covariance with range of 1/100 and nugget standard deviation 0.3. The first two plots are heatmaps of the weight functions, with the third plot a one-dimensional transect across the curve at its mode.

suggested in Section 3.1, we numerically estimate G_λ at a fine grid of 1000 equally spaced points based on a Gauss–Kronrod quadrature using the built-in `integrate` function in R, and use a cubic interpolating spline to estimate $G_\lambda(\|\mathbf{s}_1 - \mathbf{s}_2\|)$ at all interpoint distances required for the fast Fourier transform.

On this grid totaling 1,000,000 points, computing the equivalent kriging solution took approximately 19.1 s on the author's MacBook Pro laptop. Numerical integration accounts for about half of the total equivalent kriging timing, and can be reduced by judiciously choosing quadrature points or adaptive integration techniques; these ideas seem worth exploring in future research.

In this gridded situation, explicitly solving the system $\Sigma^{-1}\mathbf{Y}$ for the kriging predictor is not feasible due to the dimensionality of the problem. A sensible alternative is to calculate $\Sigma^{-1}\mathbf{Y}$ by embedding Σ in a circulant matrix, and exploiting the fast Fourier transform within the preconditioned conjugate gradient algorithm (Golub and Van Loan, 2012; Stein et al., 2013). For comparison, we implemented this idea using the circulant embedded matrix inverse as the preconditioner, which then required 407 s and 116 iterations to calculate $\Sigma^{-1}\mathbf{Y}$; we stopped the algorithm at a tolerance of unity of the 1-norm of the residual vector. Thus, even for this feasible alternative approach to calculating the exact kriging estimator, equivalent kriging exhibits substantial computational gains.

Equivalent kriging for gridded data with missing observations can proceed using similar FFT techniques. We consider the same simulation setup as before, but now remove 75% of the observations randomly, leaving a partial grid of 250,000 irregularly spaced locations. Exploiting the missing data algorithm of Lee and Meng (2005) as discussed in Section 3.4, we equivalent krige the observations to the same grid of one million locations. The missing observations increase computation time to 27.8 s; however it should be noted that in effect we are finding spatial predictions on the full 1000 × 1000 grid. We set the algorithm tolerance to stop after the solution changes by no more than one tenth, requiring four passes over the data.

We illustrate the equivalent kernel approximation on a smaller grid where the weight function is wider to facilitate a visual comparison. We consider a two dimensional example on $[0, 2\pi]^2$ for an exponential covariance with range 1/100 and noise standard deviation of 0.3. Fig. 4 shows the true kriging weights and the equivalent kernel approximation for kriging at a location (3.7, 3.3). The equivalent kernel is an accurate approximation to the true kriging weight function, capturing both the shape and approximate size of the peak.

We end this section with a small timing comparison between equivalent kriging and covariance tapering. The goal is to krige a dataset with an exponential covariance with range of 3, unity marginal variance and nugget standard deviation 0.3. The taper is a spherical taper whose width is chosen to contain approximately 20 observations within the positive support of the tapered covariance, following the suggestion of Furrer et al. (2006). Timing comparisons are for completely gridded data and incompletely gridded data where 50% of the observations are randomly held out. In this small study, equivalent kriging begins outperforming covariance tapering at between 40,000 and 50,000 observations. Hence, for large datasets, equivalent kriging appears to be preferable computationally (see Fig. 5).

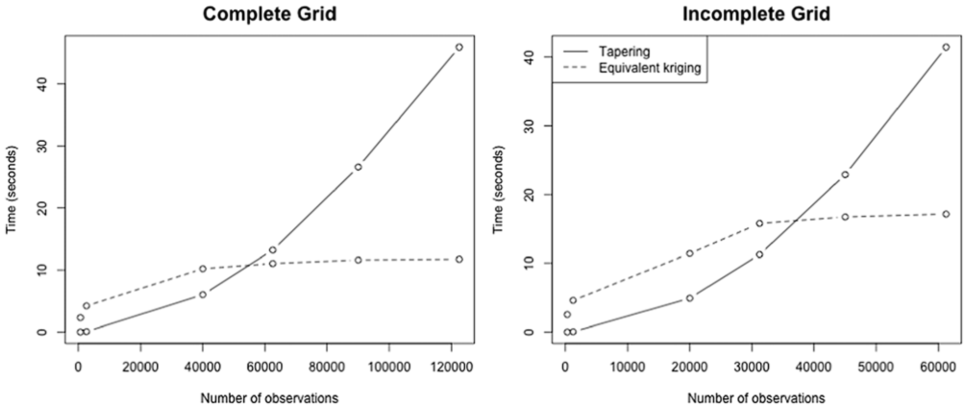


Fig. 5. Timing comparison between covariance tapering and equivalent kriging. Data are kriged under an exponential covariance with nugget. Comparisons are for completely gridded data and incompletely gridded data where 50% of the observations are randomly held out.

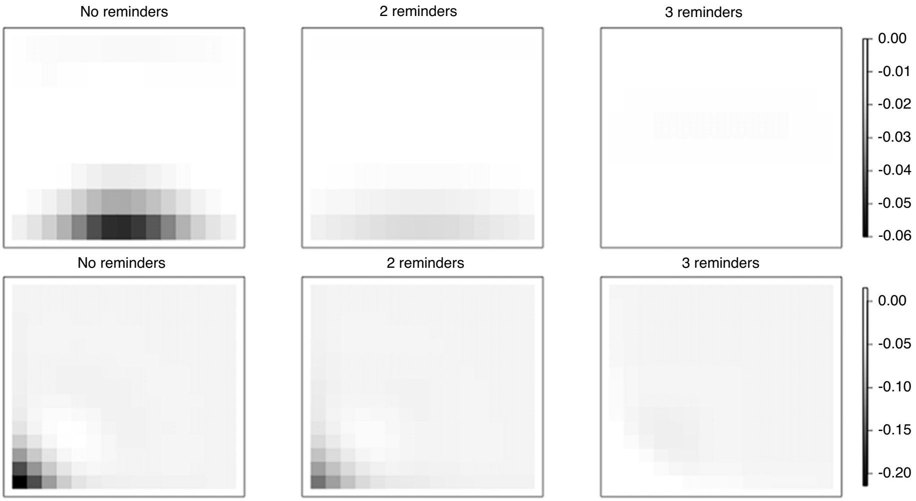


Fig. 6. Approximation error of the equivalent kernel with remainder terms in approximating the kriging weight function on the lower edge of a rectangular domain (top row) and in the corner (bottom row).

4.3. Approximations near the boundary

The equivalent kernel approximation without remainder terms is best when the observed process' spatial locations are regularly spaced; we have illustrated that for irregularly spaced data the remainder terms can improve the equivalent kernel approximation. In fact, the remainder terms also serve to remove boundary effects of the kernel approximation. As an illustration, we consider kriging under a multiresolution process with one level of 400 basis functions, a nugget standard deviation of 0.3, and 900 regularly spaced observations in $[0, 1]^2$. The stochastic behavior of the coefficients is a Gaussian Markov random field based on a spatial autoregressive model. Fig. 6 shows various approximation errors for approximating the kriging weight function on the edge of the domain and in the corner of the domain. The approximation error is substantially reduced once three remainder terms are included. Thus, the remainder terms serve the dual purpose of adjusting equivalent kriging to irregularly spaced data as well as removing boundary effects of kernel approximation.

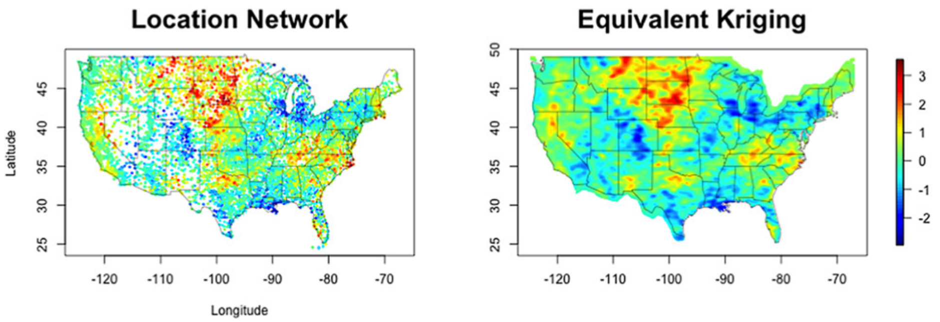


Fig. 7. Equivalent kriging solution for the 7352 stations precipitation anomaly dataset. The covariance model is a Matérn, and data are projected to a 512 × 1024 grid; parameters were estimated by generalized cross-validation. Equivalent kriging the grid of 512 × 1024 points took 37.7 s on the author's MacBook Pro laptop.

4.4. *Precipitation dataset*

The last example we consider is a set of precipitation anomalies that has previously been considered by Johns et al. (2003) and Kaufman et al. (2008). Climate change analyses often require complete gridded historical fields that can be compared to climate model output. Records of historical observations are only available at sparsely located stations, and moreover are subject to observational error as well as microscale variability. The major contribution of a statistical approach to creating a climatic data product is a formal approach to spatial smoothing with accurate assessments of the uncertainty in creating such a product. For instance, with temperature observations, the preferred covariance model for spatial modeling is a Matérn with smoothness equal to one (North et al., 2011).

The dataset we consider consists of precipitation anomalies for the year 1962 at 7352 irregularly located stations from the National Climatic Data Center. As pointed out by Kaufman et al. (2008), the year 1962 has a relatively complete record. We follow these authors in calculating anomalies as yearly totals standardized by the long-run mean and local empirical standard deviation. The data are plotted in Fig. 7 and we note that there is apparent small scale variability. Additionally, there are large regions with no historical observations, particularly in the Western United States.

We entertain two models for surface estimation based on these irregularly observed precipitation anomalies. The first involves a Matérn covariance with smoothness fixed at one, and the second is a multiresolution process. For the Matérn case, we set the marginal variance to unity, and estimate the range and the nugget-to-marginal variance ratio by generalized cross-validation. Our goal is to exploit fast Fourier techniques, and thus require a gridding of the observation locations; we approximately grid the observations to a grid with 1024 equally spaced longitudinal points and 512 equally spaced latitudinal points by associating observations with their nearest (in the sense of Euclidean distance) grid point; this yields grid spacings of approximately 0.05° in both axial directions. All other grid points are given temporary values as estimated by an initial thin plate spline; the missing gridded data algorithm we exploit converges to a true equivalent kriging surface after a sequence of iterations. Thus, operationally we are finding predictions at $512 \times 1024 \approx 5 \times 10^5$ locations, two orders of magnitude more than the original dataset. We standardize the observation domain to $[0, 2\pi]^2$, after which the range and nugget-to-marginal variance parameters were estimated as 1.18 and 0.064, respectively. Even though we dramatically increase the number of observation locations, the fast Fourier techniques result in substantial computational savings as compared to traditional kriging on the original 7352 data locations. For instance, equivalent kriging on these over half a million data locations using the missing data algorithm outlined in Section 3.4 took approximately 37.7 s on the author's MacBook Pro. Fig. 7 shows the equivalent kriging surface based on this covariance model, while 8 shows the associated standard errors based on thirty conditional simulations. The conditional simulations again rely on the approximate gridding, using circulant embedding.

The second model we consider is a multiresolution decomposition as in (13). We suppose three levels of resolution each on a regular grid with 180, 663 and 2541 basis functions respectively, thus

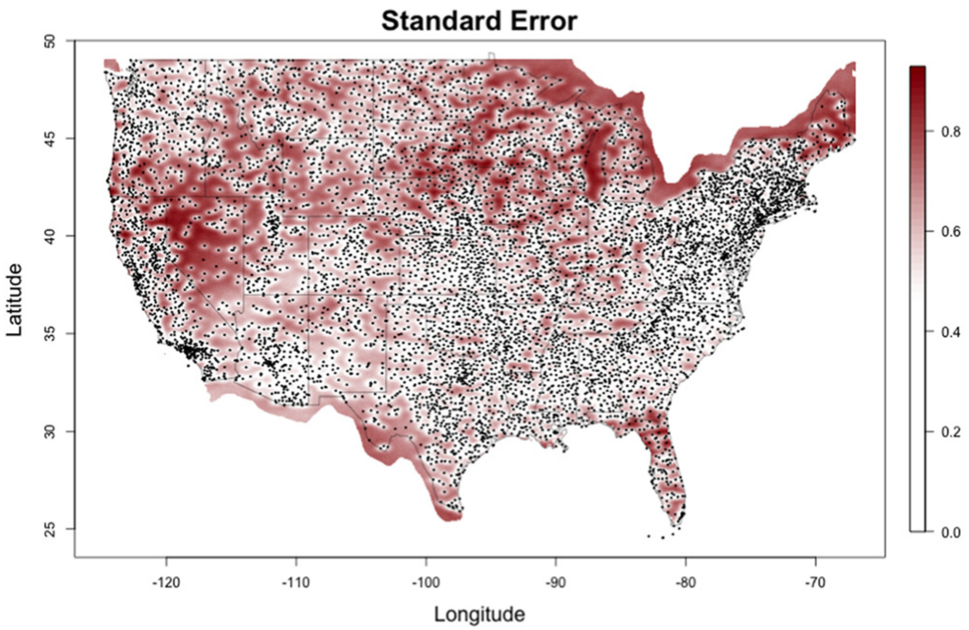


Fig. 8. Equivalent kriging standard errors for the 7352 stations precipitation anomaly dataset under a Matérn covariance model.

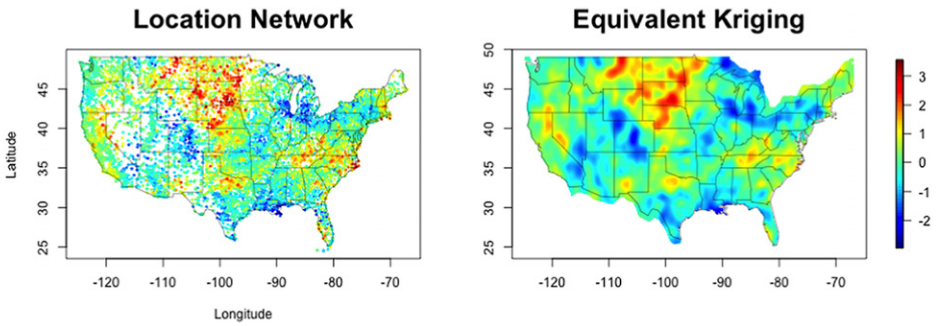


Fig. 9. Equivalent kriging solution and associated standard errors for the 7352 station precipitation anomaly dataset. The covariance model is multiresolution with three levels, parameters were estimated by cross-validation. Equivalent kriging to the grid of 512 × 1024 points took 0.33 s on the author's MacBook Pro laptop.

yielding 3384 total. We experimented with more levels of resolution, but found very similar solutions as in this setup. We follow Nychka et al. (in press) and Lindgren et al. (2011) by building the precision matrix for the ℓth level ($\ell = 1, 2, 3$), Q_ℓ as $Q_\ell = B_\ell' B_\ell / \alpha_\ell$, where B_ℓ specifies a spatial autoregressive structure with diagonal elements $4 + \kappa_\ell^2$ and off diagonals of -1, and $\alpha_\ell > 0$ is the weight assigned to the ℓth level. For this dataset, we set $\kappa_\ell^2 = 1/2$ for each level, similar to the estimate obtained on a precipitation anomaly dataset explored by Nychka et al. (in press). We estimate $\{\alpha_\ell\}_{\ell=1}^3$ and τ^2 by cross-validation, randomly choosing 20% of the data to hold out. The estimated parameters are thus $\alpha_1 = 34.2$, $\alpha_2 = 0.002$, $\alpha_3 = 11.25$ and $\tau = 0.88$. Figs. 9 and 10 show the equivalent kriging surface and associated standard errors when predicting on a 512×1024 grid. Standard errors are based on thirty conditional simulations. For the multiresolution case, equivalent kriging is extremely fast, requiring 0.33 s on the author's laptop, and estimating the standard error by conditional simulation took 16.4 s.

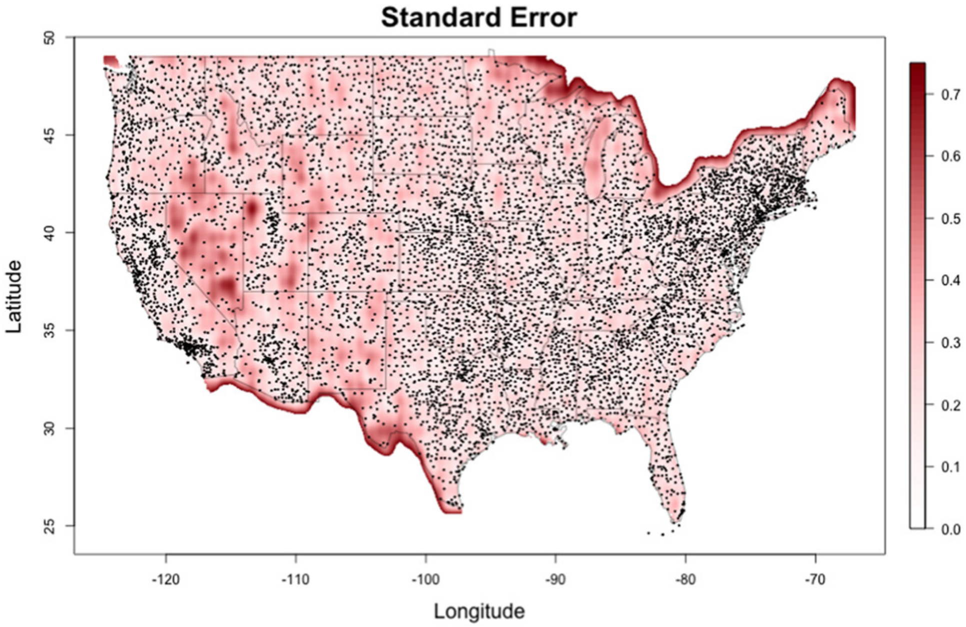


Fig. 10. Equivalent kriging standard errors for the 7352 station precipitation anomaly dataset under a multiresolution covariance model.

5. Discussion

In this manuscript we have introduced an approximation to kriging called equivalent kriging. The equivalent kriging predictor is based on an equivalent kernel approximation to the kriging weight function. The equivalent kernel for any stationary covariance is available numerically, and in two dimensions reduces to computing a Hankel transform. On the other hand, the equivalent kernel for a multiresolution model is available in closed form. Gridded observations (even with missing data) yield fast computations via fast Fourier techniques. Equivalent kriging on irregularly spaced locations requires a series of correction terms that are available in closed form for the multiresolution approach, and can be numerically approximated for stationary covariances.

We anticipate a number of future research opportunities arising from these ideas, including optimizing numerical estimation of the equivalent kernel as well as searching for closed form solutions to particular Hankel transforms. Although we have suggested cross-validation and generalized cross-validation as feasible estimation approaches, it is desirable to explore the theoretical implications of an estimation scheme such as these, as compared to traditional geostatistical techniques such as variogram fitting and likelihood-based approaches. Additional future research goals should also include exploring equivalent kernel representations for other classes of covariances, especially nonstationary, space–time and multivariate constructions.

Acknowledgments

Kleiber's work was supported by National Science Foundation grant DMS-1406536. Nychka's work was supported by National Science Foundation grant DMS-0707069. The National Center for Atmospheric Research is managed by the University Corporation for Atmospheric Research under the sponsorship of NSF.

Appendix

We begin by proving Proposition 1.

Proof of Proposition 1. Write $f(\mathbf{s}) = g(\mathbf{s}) + \delta(\mathbf{s})$ where $g(\mathbf{s}) = \sum_{i=1}^{n} w_n(\mathbf{s}, \mathbf{s}_i)Y(\mathbf{s}_i)$, and $\delta(\mathbf{s})$ is any arbitrary function $\delta : \mathbb{R}^d \to \mathbb{R}$. Then,

$$\begin{aligned}
\mathcal{L}(f) &= \mathcal{L}(g+\delta) = \frac{1}{n}\sum_{i=1}^{n}(Y(\mathbf{s}_i) - g(\mathbf{s}_i) - \delta(\mathbf{s}_i))^2 + \lambda\langle g+\delta, g+\delta\rangle \\
&= \frac{1}{n}\sum_{i=1}^{n}(Y(\mathbf{s}_i) - g(\mathbf{s}_i))^2 - \frac{2}{n}\sum_{i=1}^{n}\delta(\mathbf{s}_i)(Y(\mathbf{s}_i) - g(\mathbf{s}_i)) + \frac{1}{n}\sum_{i=1}^{n}\delta(\mathbf{s}_i)^2 \\
&\quad + \lambda\langle g, g\rangle + 2\lambda\langle g, \delta\rangle + \lambda\langle \delta, \delta\rangle \\
&= \mathcal{L}(g) + 2\langle g, \delta\rangle_w - \frac{2}{n}\sum_{i=1}^{n}Y(\mathbf{s}_i)\delta(\mathbf{s}_i) + \frac{1}{n}\sum_{i=1}^{n}\delta(\mathbf{s}_i)^2 + \lambda\langle \delta, \delta\rangle.
\end{aligned}$$

Now note

$$\begin{aligned}
\langle g, \delta\rangle_w &= \left\langle \frac{1}{n}\sum_{i=1}^{n} w_n(\cdot, \mathbf{s}_i), \delta \right\rangle_w \\
&= \frac{1}{n}\sum_{i=1}^{n}\langle w_n(\cdot, \mathbf{s}_i)Y(\mathbf{s}_i), \delta\rangle_w \\
&= \frac{1}{n}\sum_{i=1}^{n} Y(\mathbf{s}_i)\delta(\mathbf{s}_i)
\end{aligned}$$

using the reproducing property of $w_n(\cdot, \cdot)$. Thus, $\mathcal{L}(f) = \mathcal{L}(g) + \langle \delta, \delta\rangle_w$ and for f to minimize $\mathcal{L}$, we necessarily have $\delta \equiv 0$. □

Next, we outline the proof of Proposition 3, which requires two intermediate results. Below we use the notation $\frac{\partial}{\partial x} f = f_x$.

Lemma 5. *For any bounded function h on $\mathcal{D} \subseteq \mathbb{R}^2$ whose derivatives up to order 2 are integrable, with any empirical cdf F_n such that $\sup_{\mathbf{t}} |(F - F_n)(\mathbf{t})| = D_n$, we have*

$$\begin{aligned}
&\left| \int h(\mathbf{t})\mathrm{d}(F - F_n)(\mathbf{t}) \right| \\
&\le D_n \int_0^1 \int_0^1 |h_{xy}(x, y)|\mathrm{d}x\mathrm{d}y + D_n \int_0^1 |h_x(x, 1)|\mathrm{d}x + D_n \int_0^1 |h_y(1, y)|\mathrm{d}y. \qquad (16)
\end{aligned}$$

Proof. For notational simplicity, write $D = F - F_n$. Integrating by parts twice we have

$$\begin{aligned}
&\int_0^1 \int_0^1 h(x, y)\mathrm{d}D(x, y) \\
&= \int_0^1 \left(h(x, y)D_y(x, y)\Big|_{x=0}^{1} - \int_0^1 D_y(x, y)h_x(x, y)\mathrm{d}x \right) \mathrm{d}y \\
&= \int_0^1 (h(1, y)D_y(1, y) - h(0, y)D_y(0, y))\mathrm{d}y - \int_0^1 \int_0^1 h_x(x, y)D_y(x, y)\mathrm{d}y\mathrm{d}x \\
&= h(1, y)D(1, y)\Big|_{y=0}^{1} - \int_0^1 D(1, y)h_y(1, y)\mathrm{d}y
\end{aligned}$$

$$\begin{aligned} &- h(0,y)D(0,y)\Big|_{y=0}^{1} + \int_0^1 D(0,y)h_y(0,y)\mathrm{d}y \\ &- \int_0^1 (h_x(x,1)D(x,1) - h_x(x,0)D(x,0))\,\mathrm{d}x + \int_0^1 \int_0^1 D(x,y)h_{xy}(x,y)\mathrm{d}x\mathrm{d}y. \end{aligned}$$

The first resulting term is $h(1,1)D(1,1) - h(1,0)D(1,0)$. Note $D(1,1) = (F - F_n)(1+,1+) = 0$ and $D(1,0) = 0$ since $F(1+,0-) = F_n(1+,0-) = 0$. Similar reasoning implies the third term is also zero. Additionally, $D(x,0) = D(0,y) = 0$, so we are left with

$$\int_0^1 \int_0^1 D(x,y)h_{xy}(x,y)\mathrm{d}x\mathrm{d}y - \int_0^1 h_x(x,1)D(x,1)\mathrm{d}x - \int_0^1 h_y(1,y)D(1,y)\mathrm{d}y. \tag{17}$$

Now taking absolute values and noting $|D(x,y)| \le \sup_{x,y}|F(x,y) - F_n(x,y)| = D_n$, we have (17) is bounded by

$$D_n \int_0^1 \int_0^1 |h_{xy}(x,y)|\mathrm{d}x\mathrm{d}y + D_n \int_0^1 |h_x(x,1)|\mathrm{d}x + D_n \int_0^1 |h_y(1,y)|\mathrm{d}y. \quad \square$$

Lemma 6. *Suppose $d = 2$ and G_λ satisfies the* EEC1. *Define $\delta_n = 4D_n(K_\lambda/\rho)(1/\varepsilon + 1/\alpha)^2$ where $2(1/\varepsilon + 1/\alpha) > 1$. Then for $\rho = \lambda^\gamma$ and $j \ge 0$, we have*

$$\begin{aligned} &|(\mathscr{R}_n^j G_\lambda(\cdot,\mathbf{s}))(\mathbf{t})| < \delta_n^j (K_\lambda/\rho)\exp(-\alpha\|\mathbf{s}-\mathbf{t}\|_1/\rho) \\ &\left|\frac{\partial}{\partial s_i}(\mathscr{R}_n^j G_\lambda(\cdot,\mathbf{s}))(\mathbf{t})\right| < \delta_n^j (K_\lambda/\rho^2)\exp(-\alpha\|\mathbf{s}-\mathbf{t}\|_1/\rho) \\ &\left|\frac{\partial^2}{\partial s_i \partial s_k}(\mathscr{R}_n^j G_\lambda(\cdot,\mathbf{s}))(\mathbf{t})\right| < \delta_n^j (K_\lambda/\rho^3)\exp(-\alpha\|\mathbf{s}-\mathbf{t}\|_1/\rho). \end{aligned}$$

Proof. The result will follow by induction, with the case for $j = 0$ being clear, using the EEC1 assumption on G_λ. Suppose the result is true for some $j \ge 0$. Note

$$(\mathscr{R}_n^{j+1}h)(\mathbf{s}) = \int G_\lambda(\mathbf{s},\mathbf{t})(\mathscr{R}_n^j h)(\mathbf{t})\mathrm{d}(F - F_n)(\mathbf{t}).$$

We apply Lemma 5 with $\mathbf{t} = (x,y)'$ and $h(x,y) = G_\lambda(\mathbf{s}_1,\mathbf{t})(\mathscr{R}_n G_\lambda(\cdot,\mathbf{s}_2))(\mathbf{t})$. We outline the approach for the integral involving $h_x(x,y)$, with $h_y(x,y)$ following analogously. In particular,

$$\begin{aligned} h_x(x,1) &= G_{\lambda x}(\mathbf{s}_1,\mathbf{t})(\mathscr{R}_n^j G_\lambda(\cdot,\mathbf{s}_2))(\mathbf{t}) + G_\lambda(\mathbf{s}_1,\mathbf{t})(\mathscr{R}_n^j G_\lambda(\cdot,\mathbf{s}_2))_x(\mathbf{t}) \\ &\le 2\delta_n^j (K_\lambda^2/\rho^3)\exp(-(\alpha+\varepsilon)\|\mathbf{s}_1-\mathbf{t}\|_1/\rho)\exp(-\alpha\|\mathbf{s}_2-\mathbf{t}\|_1/\rho) \end{aligned}$$

by the induction step and the EEC1 assumption. Now plug this expression into (16) and apply Lemma 4.2 of Nychka (1995) along with the triangle inequality to the second entry yielding the bound

$$\begin{aligned} D_n \int_0^1 |h_x(x,1)|\mathrm{d}x &< 2D_n\delta_n^j (K_\lambda^2/\rho^2)(1/\varepsilon + 1/\alpha)\exp(-\alpha\|\mathbf{s}_1-\mathbf{s}_2\|_1/\rho) \\ &< \delta_n^{j+1}(K_\lambda/\rho)\exp(-\alpha\|\mathbf{s}_1-\mathbf{s}_2\|_1/\rho) \end{aligned}$$

using $\delta_n > 2D_n(K_\lambda/\rho)(1/\varepsilon + 1/\alpha)$ since $2(1/\varepsilon + 1/\alpha) > 1$. Now,

$$\begin{aligned} h_{xy}(x,y) &= G_{\lambda xy}(\mathbf{s}_1,\mathbf{t})(\mathscr{R}_n^j G_\lambda(\cdot,\mathbf{s}_2))(\mathbf{t}) + G_{\lambda x}(\mathbf{s}_1,\mathbf{t})(\mathscr{R}_n^j G_\lambda(\cdot,\mathbf{s}_2))_y(\mathbf{t}) \\ &\quad + G_{\lambda y}(\mathbf{s}_1,\mathbf{t})(\mathscr{R}_n^j G_\lambda(\cdot,\mathbf{s}_2))_x(\mathbf{t}) + G_\lambda(\mathbf{s}_1,\mathbf{t})(\mathscr{R}_n^j G_\lambda(\cdot,\mathbf{s}_2))_{xy}(\mathbf{t}) \\ &< 4\delta_n^j (K_\lambda^2/\rho^4)\exp(-(\alpha+\varepsilon)\|\mathbf{s}_1-\mathbf{t}\|_1/\rho)\exp(-\alpha\|\mathbf{s}_2-\mathbf{t}\|_1/\rho). \end{aligned}$$

Using Lemma 4.2 of Nychka (1995) twice again yields the bound

$$\begin{aligned} D_n \int_0^1 \int_0^1 |h_{xy}(x,y)| \mathrm{d}x\mathrm{d}y &< 4D_n \delta_n^j (K_\lambda^2/\rho^2)(1/\varepsilon + 1/\alpha)^2 \exp(-\alpha \|\mathbf{s}_1 - \mathbf{s}_2\|_1/\rho) \\ &= \delta_n^{j+1}(K_\lambda/\rho)\exp(-\alpha\|\mathbf{s}_1 - \mathbf{s}_2\|_1/\rho) \end{aligned}$$

using $\delta_n = 4D_n(K_\lambda/\rho)(1/\varepsilon + 1/\alpha)^2$. The same proof technique yields analogous results for the first and second partials of $\mathscr{R}_n^j G_\lambda$. □

Proof of Proposition 3. The result follows since

$$\begin{aligned} |w_n(\mathbf{s},\mathbf{t}) - G_\lambda(\mathbf{s},\mathbf{t})| &\le \sum_{j=1}^{\infty} |(\mathscr{R}_n^j G_\lambda(\cdot,\mathbf{t}))(\mathbf{s})| \\ &\le (K_\lambda/\rho)\sum_{j=1}^{\infty} \delta_n^j \exp(-\alpha\|\mathbf{s}-\mathbf{t}\|_1/\rho) \\ &= \frac{\delta_n K_\lambda}{\rho(1-\delta_n)} \exp(-\alpha\|\mathbf{s}-\mathbf{t}\|_1/\rho) \end{aligned}$$

using Lemma 6. □

References

Banerjee, S., Gelfand, A.E., Finley, A.O., Sang, H., 2008. Gaussian predictive process models for large spatial data sets. J. R. Stat. Soc. Ser. B 70, 825–848.

Chilès, J.P., Delfiner, P., 1999. Geostatistics: Modeling Spatial Uncertainty. Wiley, New York.

Cox, D.D., 1983. Asymptotics for M-type smoothing splines. Ann. Statist. 11, 530–551.

Cree, M.J., Bones, P.J., 1993. Algorithms to numerically evaluate the Hankel transform. Comput. Math. Appl. 26, 1–12.

Cressie, N.A.C., 1993. Statistics for Spatial Data, revised ed. Wiley.

Cressie, N., Johannesson, G., 2008. Fixed rank kriging for very large spatial data sets. J. R. Stat. Soc. Ser. B 70, 209–226.

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

Fuentes, M., 2007. Approximate likelihood for large irregularly spaced spatial data. J. Amer. Statist. Assoc. 102, 321–331.

Furrer, E.M., 2008. Asymptotic behavior of a continuous approximation to the kriging weighting function, NCAR Technical Note NCAR/TN-476+STR.

Furrer, R., Genton, M.G., Nychka, D., 2006. Covariance tapering for interpolation of large datasets. J. Comput. Graph. Statist. 15, 502–523.

Furrer, E.M., Nychka, D.W., 2007. A framework to understand the asymptotic properties of kriging and splines. J. Korean Stat. Soc. 36, 57–76.

Golub, G.H., Van Loan, C.F., 2012. Matrix Computations, fourth ed. Johns Hopkins University Press.

Gradshteyn, I.S., Ryzhik, I.M., 2000. Table of Integrals, Series, and Products, sixth ed. Academic Press, London.

Ihaka, R., Gentleman, R., 1996. R: a language for data analysis and graphics. J. Comput. Graph. Statist. 5, 299–314.

Johns, C., Nychka, D., Kittel, T., Daly, C., 2003. Infilling sparse records of spatial fields. J. Amer. Statist. Assoc. 98, 796–806.

Kaufman, C.G., Schervish, M.J., Nychka, D.W., 2008. Covariance tapering for likelihood-based estimation in large spatial data sets. J. Amer. Statist. Assoc. 103, 1545–1555.

Key, K., 2012. Is the fast Hankel transform faster than quadrature? Geophysics 77, F21–F30.

Kimeldorf, G., Wahba, G., 1971. Some results on Tchebycheffian spline functions. J. Math. Anal. Appl. 33, 82–95.

Kleiber, W., Nychka, D., 2014. Local asymptotics for kriging, Unpublished Manuscript.

Lee, T.C.M., Meng, X.L., 2005. A self-consistent wavelet method for denoising images with missing pixels. In: Proceedings of the 30th IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, pp. 41–44.

Lindgren, F., Rue, H., Lindström, J., 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. J. R. Stat. Soc. Ser. B 73, 423–498.

North, G.R., Wang, J., Genton, M.G., 2011. Correlation models for temperature fields. J. Clim. 24, 5850–5862.

Nychka, D., 1995. Splines as local smoothers. Ann. Statist. 23, 1175–1197.

Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F., Sain, S., 2014. A multi-resolution Gaussian process model for the analysis of large spatial data sets. J. Comput. Graph. Statist. (in press).

Nychka, D., Wikle, C., Royle, J.A., 2002. Multiresolution models for nonstationary spatial covariance functions. Stat. Model. 2, 315–331.

Sang, H., Huang, J.Z., 2012. A full-scale approximation of covariance functions for large spatial data sets. J. R. Stat. Soc. Ser. B 74, 111–132.

Siegman, A.E., 1977. Quasi fast Hankel transform. Opt. Lett. 1, 13–15.

Silverman, B.W., 1984. Spline smoothing: the equivalent variable kernel method. Ann. Statist. 12, 898–916.

Sollich, P., Williams, C.K.I., 2005. Using the equivalent kernel to understand Gaussian process regression. In: Saul, L.K., Weiss, Y., Bottou, L. (Eds.), Advances in Neural Information Processing Systems. MIT Press, Cambridge, MA, pp. 1313–1320.

Stein, M.L., 1999. Interpolation of Spatial Data: Some Theory for Kriging. Springer-Verlag, New York.
Stein, M.L., 2008. A modeling approach for large spatial datasets. J. Korean Stat. Soc. 37, 3–10.
Stein, M.L., 2014. Limitations on low rank approximations for covariance matrices of spatial data. Spat. Stat. 8, 1–19.
Stein, M.L., Chen, J., Anitescu, M., 2013. Stochastic approximation of score functions for Gaussian processes. Ann. Appl. Stat. 7, 1162–1191.
Sun, Y., Li, B., Genton, M.G., 2012. Geostatistics for large datasets. In: Porcu, E., Montero, J.M., Schlather, M. (Eds.), Space–Time Processes and Challenges Related to Environmental Problems. Springer, pp. 55–77.
Wahba, G., 1990. Spline Models for Observational Data. Society for Industrial and Applied Mathematics, Philadelphia, PA.
Wand, M.P., 1994. Fast computation of multivariate kernel estimators. J. Comput. Graph. Statist. 3, 433–445.
Young, W.H., 1917. On multiple integration by parts and the second theorem of the mean. Proc. Lond. Math. Soc. 16, 273–293.
Zhang, B., Stein, M., 1993. Kernel approximations for universal kriging predictors. J. Multivariate Anal. 44, 286–313.