J. R. Statist. Soc. B (2008)
70, *Part* 1, *pp.* 209–226

Fixed rank kriging for very large spatial data sets

Noel Cressie

The Ohio State University, Columbus, USA

and Gardar Johannesson

Lawrence Livermore National Laboratory, Livermore, USA

[Received May 2006. Final revision July 2007]

Summary. Spatial statistics for very large spatial data sets is challenging. The size of the data set, n, causes problems in computing optimal spatial predictors such as kriging, since its computational cost is of order n^3. In addition, a large data set is often defined on a large spatial domain, so the spatial process of interest typically exhibits non-stationary behaviour over that domain. A flexible family of non-stationary covariance functions is defined by using a set of basis functions that is fixed in number, which leads to a spatial prediction method that we call fixed rank kriging. Specifically, fixed rank kriging is kriging within this class of non-stationary covariance functions. It relies on computational simplifications when n is very large, for obtaining the spatial best linear unbiased predictor and its mean-squared prediction error for a hidden spatial process. A method based on minimizing a weighted Frobenius norm yields best estimators of the covariance function parameters, which are then substituted into the fixed rank kriging equations. The new methodology is applied to a very large data set of total column ozone data, observed over the entire globe, where n is of the order of hundreds of thousands.

Keywords: Best linear unbiased predictor; Covariance function; Frobenius norm; Geostatistics; Mean-squared prediction error; Non-stationarity; Remote sensing; Spatial prediction; Total column ozone

1. Introduction

Kriging, or spatial best linear unbiased prediction (BLUP), has become very popular in the earth and environmental sciences, where it is sometimes known as optimum interpolation. Matheron (1962) coined the term 'kriging' in honour of D. G. Krige, a South African mining engineer (Cressie, 1990). With its internal quantification of spatial variability through the covariance function (or variogram), kriging methodology can produce maps of optimal predictions and associated prediction standard errors from incomplete and noisy spatial data (e.g. Cressie (1993), chapter 3). Sometimes a spatial datum is expensive to obtain (e.g. drilling wells for oil reserve estimation), in which case the sample size n is typically small and kriging can be performed straightforwardly. Recently, with the ubiquity of remote sensing platforms on satellites, database paradigms have moved from small to massive, often of the order of gigabytes per day. Solving the kriging equations directly involves inversion of an $n \times n$ variance–covariance matrix Σ, where n data may require $O(n^3)$ computations to obtain Σ^{-1}. Under these circumstances, straightforward kriging of massive data sets is not possible. Our goal in this paper is to develop methodology that reduces the computational cost of kriging to $O(n)$.

Address for correspondence: Noel Cressie, Department of Statistics, Ohio State University, 1958 Neil Avenue, Columbus, OH 43210-1247, USA.
E-mail: ncressie@stat.osu.edu

1369–7412/08/70209

Even a spatial data set of the order of several thousand can result in computational slowdowns. *Ad hoc* methods of subsetting the data were formalized by the moving window approach of Haas (1995), although it appears that the local covariance functions that are fitted within the window yield incompatible covariances at larger spatial lags. The variance–covariance matrix Σ is typically sparse when the covariance function has a finite range, and hence Σ^{-1} can be obtained by using sparse matrix techniques. Rue and Tjelmeland (2002) *approximated* Σ^{-1} to be sparse, approximating it to be the precision matrix of a Gaussian Markov random field wrapped on a torus.

When data sets are large (of the order of tens of thousands) to very large (of the order of hundreds of thousands), straightforward kriging can break down and *ad hoc* local kriging neighbourhoods are typically used (e.g. Cressie (1993), pages 131–134). One avenue of recent research has been to *approximate* the kriging equations (Nychka *et al.*, 1996, 2002; Nychka, 2000; Billings *et al.*, 2002 a, b; Furrer *et al.*, 2006; Quiñonero-Candela and Rasmussen, 2005). Suggestions include giving an equivalent representation in terms of orthogonal bases and truncating the bases, doing covariance tapering, using approximate iterative methods such as conjugate gradient, implementing sparse approximations using inducing variables or replacing the data locations with a smaller set of space filling locations. Kammann and Wand (2003) took up this last idea when fitting a class of spatial models that they called geoadditive models.

Another approach has been to choose classes of covariance functions for which kriging can be done *exactly*, even though the spatial data sets are large (e.g. Huang *et al.* (2002), Johannesson and Cressie (2004a) and Johannesson *et al.* (2007)). In these papers, a multiresolution spatial (and spatiotemporal) process was constructed so that (simple) kriging can be computed iteratively and extremely rapidly, with computational complexity linear in the size of the data. In the spatial case, Johannesson and Cressie (2004a) achieved speed-ups of the order of 10^8 over directly solving the kriging equations. They could compute optimal spatial predictors and their associated mean-squared prediction errors over the entire globe in about 3 min for $n \simeq 160000$. One advantage of having a spatial model that allows exact computations is that there is no concern about how close approximate kriging predictors and approximate mean-squared prediction errors are to the corresponding theoretical values. For exact methods, two important questions are, then, how flexible are the spatial covariance functions that are used for kriging and how are they fitted?

For the multiresolution models that were referred to above, the implied spatial covariances are non-stationary and 'blocky'. In this paper, we use a different approach to achieve orders-of-magnitude speed-ups for optimal spatial prediction, using covariance functions that are very flexible and can be chosen to be smooth or not, as determined by the type of spatial dependence that is exhibited by the spatial data (in contrast with the approach of Tzeng *et al.* (2005)). We shall show that there is a very rich class of spatial covariances from which kriging of large spatial data sets can be carried out exactly, with a computational cost of $O(n)$.

In what is to follow, we consider a class of $n \times n$ variance–covariance matrices Σ such that Σ^{-1} can be obtained by inverting $r \times r$ matrices, where r is fixed; in the application to the total column ozone (TCO) data that is given in Section 4, n was 173405 and r was chosen to be 396. From the derivations that are given in Section 2.3, the number of computations per prediction location in the kriging equations is $O(nr^2)$, which increases only linearly with sample size.

Furthermore, suppose that the data set is the result of remote sensing from a satellite that achieves global coverage. Then any spatial dependences in the data will almost certainly be heterogeneous across the globe. What is new in the methodology that is presented in this paper is that we address both problems (data set size and spatial heterogeneity) directly. The result is a spatial BLUP procedure that we call fixed rank kriging (FRK), which relies on inverting $r \times r$ matrices for r fixed and independent of n.

For completeness, we mention another approach to spatial prediction, which is based on smoothing splines. In contrast with kriging, smoothing splines do not rely on a spatial stochastic process whose covariance function must be modelled, fitted and used for computing the optimal predictor. However, there are knots and a smoothing parameter to be determined and, once again, the size of the spatial data set causes computational difficulties. Hastie (1996) and Johannesson and Cressie (2004b) developed low rank spline smoothers for massive data sets.

To carry out FRK, we must specify the form of the (non-stationary) covariance function; the class that we propose is sufficiently flexible to allow multiple scales of spatial variation to be modelled and yields an $n \times n$ variance–covariance matrix $\mathbf{\Sigma}$ whose inverse can be computed straightforwardly. The spatial BLUP that minimizes the mean-squared prediction error involves $\mathbf{\Sigma}^{-1}$ in various matrix computations; we show that FRK has computational cost that is linear in n. The spatial covariance function is fitted to empirical covariances by minimizing a weighted Frobenius norm.

Section 2 presents the kriging methodology and gives the equations that define FRK. In Section 3, the class of non-stationary covariance functions that are used in FRK is investigated, including how to find the one that best fits the data. Section 4 applies the methodology to TCO data, where $n = 173405$; kriging by directly inverting the $n \times n$ theoretical variance–covariance matrix of the data is not possible. Section 5 contains discussion and conclusions, which is followed by technical Appendix A.

2. Kriging: optimal linear spatial prediction

In this section, we present the notation for kriging, and we equate it with BLUP in a spatial setting. When the spatial data sets are large, exact computation of kriging is generally not possible. In the latter part of this section, we show how choice of a particular class of non-stationary spatial covariances allows rapid computation of the kriging predictor (i.e. spatial BLUP) and the kriging standard error (i.e. root-mean-squared prediction error).

2.1. The kriging equations

Let $\{Y(\mathbf{s}) : \mathbf{s} \in D \subset \mathbb{R}^d\}$ be a real-valued spatial process. We are interested in making inference on the Y-process on the basis of data that have measurement error incorporated; consider the process $Z(\cdot)$ of actual and potential observations,

$$Z(\mathbf{s}) \equiv Y(\mathbf{s}) + \varepsilon(\mathbf{s}), \qquad \mathbf{s} \in D, \tag{2.1}$$

where $\{\varepsilon(\mathbf{s}) : \mathbf{s} \in D\}$ is a spatial white noise process with mean 0, $\text{var}\{\varepsilon(\mathbf{s})\} = \sigma^2\, v(\mathbf{s}) \in (0, \infty)$, $\mathbf{s} \in D$, for $\sigma^2 > 0$ and $v(\cdot)$ known. In fact, the process $Z(\cdot)$ is known only at a finite number of spatial locations $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$; define the vector of available data to be

$$\mathbf{Z} \equiv (Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n))'. \tag{2.2}$$

The hidden process $Y(\cdot)$ is assumed to have a linear mean structure,

$$Y(\mathbf{s}) = \mathbf{t}(\mathbf{s})'\boldsymbol{\alpha} + \nu(\mathbf{s}), \qquad \mathbf{s} \in D, \tag{2.3}$$

where $\mathbf{t}(\cdot) \equiv (t_1(\cdot), \ldots, t_p(\cdot))'$ represents a vector process of known covariates; the coefficients $\boldsymbol{\alpha} \equiv (\alpha_1, \ldots, \alpha_p)'$ are unknown, and the process $\nu(\cdot)$ has zero mean, $0 < \text{var}\{\nu(\mathbf{s})\} < \infty$, for all $\mathbf{s} \in D$, and a generally non-stationary spatial covariance function,

$$\operatorname{cov}\{\nu(\mathbf{u}), \nu(\mathbf{v})\} \equiv C(\mathbf{u}, \mathbf{v}), \qquad \mathbf{u}, \mathbf{v} \in D, \tag{2.4}$$

which for the moment is left unspecified.

If we define $\boldsymbol{\varepsilon}$, $\mathbf{Y}$ and $\boldsymbol{\nu}$ in an analogous manner to $\mathbf{Z}$, then expressions (2.1)–(2.4) imply a general linear mixed model,

$$\mathbf{Z} = \mathbf{T}\boldsymbol{\alpha} + \boldsymbol{\delta}, \qquad \boldsymbol{\delta} = \boldsymbol{\nu} + \boldsymbol{\varepsilon}, \tag{2.5}$$

where $\mathbf{T}$ is an $n \times p$ matrix of covariates $(\mathbf{t}(\mathbf{s}_1), \ldots, \mathbf{t}(\mathbf{s}_n))'$. Observe from model (2.5) that the error term $\boldsymbol{\delta}$ is made up of two independent, zero-mean components, resulting in $E(\boldsymbol{\delta}) = \mathbf{0}$ and $\operatorname{var}(\boldsymbol{\delta}) = \boldsymbol{\Sigma} \equiv (\sigma_{ij})$, where

$$\sigma_{ij} = \begin{cases} C(\mathbf{s}_j, \mathbf{s}_j) + \sigma^2 v(\mathbf{s}_j), & i = j, \\ C(\mathbf{s}_i, \mathbf{s}_j), & i \neq j. \end{cases}$$

On writing $\mathbf{C} \equiv (C(\mathbf{s}_i, \mathbf{s}_j))$ and $\mathbf{V} \equiv \operatorname{diag}\{v(\mathbf{s}_1), \ldots, v(\mathbf{s}_n)\}$, it is easily seen that

$$\boldsymbol{\Sigma} = \mathbf{C} + \sigma^2 \mathbf{V}. \tag{2.6}$$

No assumptions of stationarity or isotropy of the covariance function have been made; nor will there be.

Interest is in inference on the Y-process, not the noisy Z-process. For point prediction, we wish to predict the Y-process at a location $\mathbf{s}_0$, $\mathbf{s}_0 \in D$, regardless of whether $\mathbf{s}_0$ is or is not an observation location. Cressie (1993), section 3.4.5, gave a formula for the kriging predictor of $Y(\mathbf{s}_0)$ in terms of the covariance function:

$$\hat{Y}(\mathbf{s}_0) = \mathbf{t}(\mathbf{s}_0)'\hat{\boldsymbol{\alpha}} + \mathbf{k}(\mathbf{s}_0)'(\mathbf{Z} - \mathbf{T}\hat{\boldsymbol{\alpha}}), \tag{2.7}$$

where

$$\hat{\boldsymbol{\alpha}} = (\mathbf{T}'\boldsymbol{\Sigma}^{-1}\mathbf{T})^{-1}\mathbf{T}'\boldsymbol{\Sigma}^{-1}\mathbf{Z}, \tag{2.8}$$

$$\mathbf{k}(\mathbf{s}_0)' = \mathbf{c}(\mathbf{s}_0)'\boldsymbol{\Sigma}^{-1}, \tag{2.9}$$

and $\mathbf{c}(\mathbf{s}_0) \equiv (C(\mathbf{s}_0, \mathbf{s}_1), \ldots, C(\mathbf{s}_0, \mathbf{s}_n))'$. The equivalence of equation (2.7) to kriging may not be immediately apparent, since the traditional derivation of kriging is in terms of the variogram and with no measurement error (i.e. where the ε-process in expression (2.1) is identically 0); see Journel and Huijbregts (1978), chapter V. The kriging standard error is the root-mean-squared prediction error of $\hat{Y}(\mathbf{s}_0)$, $[E\{Y(\mathbf{s}_0) - \hat{Y}(\mathbf{s}_0)\}^2]^{1/2}$, which is given by

$$\sigma_k(\mathbf{s}_0) = \{C(\mathbf{s}_0, \mathbf{s}_0) - \mathbf{k}(\mathbf{s}_0)'\boldsymbol{\Sigma}\,\mathbf{k}(\mathbf{s}_0) + (\mathbf{t}(\mathbf{s}_0) - \mathbf{T}'\,\mathbf{k}(\mathbf{s}_0))'(\mathbf{T}'\boldsymbol{\Sigma}^{-1}\mathbf{T})^{-1}(\mathbf{t}(\mathbf{s}_0) - \mathbf{T}'\,\mathbf{k}(\mathbf{s}_0))\}^{1/2}. \tag{2.10}$$

As the prediction location $\mathbf{s}_0$ in equations (2.7) and (2.10) varies over D, a kriging prediction map and a kriging standard error map respectively are generated. (In practice, prediction locations are finite in number and typically taken as nodes of a fine resolution grid superimposed on D.)

Inspection of the kriging equations (2.7) and (2.10) shows that $\boldsymbol{\Sigma}^{-1}$ is an essential component and the most obvious place where a computational bottleneck could occur. The inverse of a generic $n \times n$ symmetric positive definite matrix has a computational cost of $O(n^3)$. When n is tens of thousands and above, equations (2.7) and (2.10) will not generally be computable in any reasonable amount of time. In the next subsection, we show how choice of a rich class of covariance functions yields orders-of-magnitude speed-ups for optimal spatial prediction (i.e. kriging).

2.2. *Spatial covariance function*

In general, the covariance function $C(\mathbf{u},\mathbf{v})$ that is defined by expression (2.4) must be positive definite on $\mathbb{R}^d \times \mathbb{R}^d$. Often $C(\mathbf{u},\mathbf{v})$ is modelled as being stationary, in which case it must be a non-negative-definite function of $\mathbf{u}-\mathbf{v}$. In this paper, we take a different approach and instead try to capture the scales of spatial dependence through a set of r (not necessarily orthogonal) basis functions,

$$\mathbf{S}(\mathbf{u}) \equiv (S_1(\mathbf{u}),\ldots,S_r(\mathbf{u}))', \qquad \mathbf{u} \in \mathbb{R}^d, \tag{2.11}$$

where r is fixed. Examples of basis functions are given in Section 3.1. For any $r \times r$ *positive definite* matrix $\mathbf{K}$, we model $\operatorname{cov}\{Y(\mathbf{u}),Y(\mathbf{v})\}$ according to

$$C(\mathbf{u},\mathbf{v}) = \mathbf{S}(\mathbf{u})'\mathbf{K}\,\mathbf{S}(\mathbf{v}), \qquad \mathbf{u},\mathbf{v} \in \mathbb{R}^d, \tag{2.12}$$

which can be shown to be a non-negative-definite function (Section 3.1) and hence is a valid covariance function (Cressie and Johannesson, 2006). It is possible to add $\tau^2\, I(\mathbf{u}=\mathbf{v})$ to expression (2.12), although we do not do so in this paper; see Section 5.

It is easy to see that expression (2.12) is a consequence of writing $\nu(\mathbf{s}) = \mathbf{S}(\mathbf{s})'\boldsymbol{\eta}$, $\mathbf{s} \in D$, where $\boldsymbol{\eta}$ is an r-dimensional vector with $\operatorname{var}(\boldsymbol{\eta}) = \mathbf{K}$. We call the model for $\nu(\cdot)$ a *spatial random-effects* model. Hence, from expression(2.3), $Y(\mathbf{s}) = \mathbf{t}(\mathbf{s})'\boldsymbol{\beta} + \mathbf{S}(\mathbf{s})'\boldsymbol{\eta}$, $\mathbf{s} \in D$, which is a mixed effects linear model that we call a *spatial mixed effects* model.

2.3. *Fixed rank kriging*

From expression (2.12), we can write the $n \times n$ theoretical variance–covariance matrix of $\mathbf{Y}$ (or $\boldsymbol{\nu}$) as $\mathbf{C} = \mathbf{SKS}'$, and hence

$$\boldsymbol{\Sigma} = \mathbf{SKS}' + \sigma^2\mathbf{V}, \tag{2.13}$$

where the unknown parameters are $\mathbf{K}$, a positive definite $r \times r$ matrix, and $\sigma^2 > 0$. Both $\mathbf{S}$, the $n \times r$ matrix whose (i,l) element is $S_l(\mathbf{s}_i)$, and $\mathbf{V}$, a diagonal matrix with entries given by the measurement error variances, are assumed known. Further,

$$\operatorname{cov}\{Y(\mathbf{s}_0),\mathbf{Z}\} = \mathbf{c}(\mathbf{s}_0)' = \mathbf{S}(\mathbf{s}_0)'\mathbf{KS}', \tag{2.14}$$

i.e., on the basis of the model (2.1), (2.3) and (2.12), we can find expressions for all the components that are needed in the kriging equations (2.7) and (2.10).

There remains the problem of n being very large to massive but, as we shall now show, the choice of covariance function (2.12) allows alternative ways of computing the kriging equations involving inversion of only $r \times r$ matrices. Recall from equation (2.13) that $\boldsymbol{\Sigma} = \mathbf{SKS}' + \sigma^2\mathbf{V}$, where $\mathbf{V}$ is diagonal. Then

$$\boldsymbol{\Sigma}^{-1} = \sigma^{-1}\mathbf{V}^{-1/2}\{\mathbf{I} + (\sigma^{-1}\mathbf{V}^{-1/2}\mathbf{S})\mathbf{K}(\sigma^{-1}\mathbf{V}^{-1/2}\mathbf{S})'\}^{-1}\sigma^{-1}\mathbf{V}^{-1/2}. \tag{2.15}$$

Now, it is easy to see that, for any $n \times r$ matrix $\mathbf{P}$,

$$\mathbf{I} + \mathbf{PKP}' = \mathbf{I} + (\mathbf{I} + \mathbf{PKP}')\mathbf{PK}(\mathbf{I} + \mathbf{P}'\mathbf{PK})^{-1}\mathbf{P}'.$$

Multiplying by $(\mathbf{I} + \mathbf{PKP}')^{-1}$ yields

$$(\mathbf{I} + \mathbf{PKP}')^{-1} = \mathbf{I} - \mathbf{P}(\mathbf{K}^{-1} + \mathbf{P}'\mathbf{P})^{-1}\mathbf{P}',$$

which is a result that is covered by the Sherman–Morrison–Woodbury formulae (see Henderson and Searle (1981)). This is then used in equation (2.15) to give the computational simplification

$$\Sigma^{-1} = (\sigma^2 \mathbf{V})^{-1} - (\sigma^2 \mathbf{V})^{-1} \mathbf{S}\{\mathbf{K}^{-1} + \mathbf{S}'(\sigma^2 \mathbf{V})^{-1}\mathbf{S}\}^{-1}\mathbf{S}'(\sigma^2 \mathbf{V})^{-1}. \qquad (2.16)$$

Note that the formula (2.16) for Σ^{-1} involves inverting the *fixed rank* $r \times r$ positive definite matrices and the $n \times n$ *diagonal* matrix $\mathbf{V}$. Finally then, the kriging predictor (spatial BLUP) (2.7) is

$$\hat{Y}(\mathbf{s}_0) = \mathbf{t}(\mathbf{s}_0)'\hat{\alpha} + \mathbf{S}(\mathbf{s}_0)'\mathbf{K}\mathbf{S}'\Sigma^{-1}(\mathbf{Z} - \mathbf{T}\hat{\alpha}), \qquad (2.17)$$

where $\hat{\alpha} = (\mathbf{T}'\Sigma^{-1}\mathbf{T})^{-1}\mathbf{T}'\Sigma^{-1}\mathbf{Z}$ and Σ^{-1} is given by equation (2.16). The kriging standard error (2.10) is

$$\sigma_k(\mathbf{s}_0) = \{\mathbf{S}(\mathbf{s}_0)'\mathbf{K}\,\mathbf{S}(\mathbf{s}_0) - \mathbf{S}(\mathbf{s}_0)'\mathbf{K}\mathbf{S}'\Sigma^{-1}\mathbf{S}\mathbf{K}\,\mathbf{S}(\mathbf{s}_0) + (\mathbf{t}(\mathbf{s}_0) - \mathbf{T}'\Sigma^{-1}\mathbf{S}\mathbf{K}\,\mathbf{S}(\mathbf{s}_0))'(\mathbf{T}'\Sigma^{-1}\mathbf{T})^{-1}(\mathbf{t}(\mathbf{s}_0) - \mathbf{T}'\Sigma^{-1}\mathbf{S}\mathbf{K}\,\mathbf{S}(\mathbf{s}_0))\}^{1/2}, \qquad (2.18)$$

where Σ^{-1} is again given by equation (2.16). FRK is the name that we gave to the methodology that leads to equations (2.16)–(2.18) (Cressie and Johannesson, 2006). As the prediction location $\mathbf{s}_0$ in equations (2.17) and (2.18) varies over D, a kriging prediction map and a kriging standard error map respectively are generated.

Inspection of equations (2.16)–(2.18) reveals that, for a fixed number of regressors p and a fixed rank r of $\mathbf{K}$ in the covariance model that is defined by expression (2.12), the computational burden of FRK is only linear in n. To see this (and without loss of generality assume here that $\sigma^2\mathbf{V} = \mathbf{I}$), computations that are associated with equations (2.17) and (2.18) involve computations of $\mathbf{S}'\Sigma^{-1}\mathbf{S}$, $\mathbf{S}'\Sigma^{-1}\mathbf{a}$ and $\Sigma^{-1}\mathbf{a}$, for given vectors $\mathbf{a}$ of length n. To carry out these computations, $\mathbf{A} \equiv \mathbf{S}'\mathbf{S}$ and $\mathbf{B} \equiv (\mathbf{K}^{-1} + \mathbf{S}'\mathbf{S})^{-1}$ are computed initially, at a maximum computational cost of $O(nr^2)$ (we assume that $n > r$). Then, from equation (2.16), $\mathbf{S}'\Sigma^{-1}\mathbf{S} = \mathbf{A} - \mathbf{ABA}$, and $\mathbf{ABA}$ requires $O(r^3)$ computations. The quantities $\mathbf{S}'\Sigma^{-1}\mathbf{a}$ and $\Sigma^{-1}\mathbf{a}$ have a computational cost that never exceeds $O(nr^2)$. Finally, equation (2.17) has $O(r)$ computations, and equation (2.18) has $O(r^2)$ computations (assuming that $p \ll r$) for a fixed $\mathbf{s}_0$. Hence, the overall computational cost is $O(nr^2)$. As confirmed by the timings that are given in Section 4, the FRK methodology makes it feasible to construct maps of kriging predictors and kriging standard errors that are based on very large spatial data sets.

The relationships between kriging methodology and smoothing methodology are quite well established by now (e.g. Cressie (1993), section 5.9, and Nychka (2000)). Indeed, FRK, which depends on the covariance model (2.12), was motivated by a fixed rank smoothing technique that evolved from regularization and ridge regression (Johannesson and Cressie, 2004b). The novelty of our methodology is the combination of a *fixed rank* positive definite matrix $\mathbf{K}$ (the parameter to be estimated) and basis functions $\{S_l(\cdot)\}$ (to be specified) that yield a very flexible spatial covariance function (2.12), followed by computationally efficient (linear in the number of data) kriging predictors and kriging standard errors for large spatial data sets.

In the next section, we consider *inter alia* estimation of $\mathbf{K}$ and the measurement error variance σ^2 in detail.

3. The class of covariance functions

Recall, from expression (2.12), the class of covariance functions that we are considering in this paper:

$$C(\mathbf{u}, \mathbf{v}) = \mathbf{S}(\mathbf{u})'\mathbf{K}\,\mathbf{S}(\mathbf{v}), \qquad \mathbf{u}, \mathbf{v} \in \mathbb{R}^d,$$

where $\mathbf{K}$ is an $r \times r$ positive definite matrix and $\mathbf{S}(\cdot)$ is an $r \times 1$ vector made up of basis functions $S_1(\cdot), \ldots, S_r(\cdot)$, where r is fixed. This is similar to the form that was given by Stroud *et al.* (2001), although they used it to motivate a spatiotemporal model and did not have in mind inverting $\boldsymbol{\Sigma}$ for kriging. In the subsections that follow, we give some properties of this class of covariance functions. We also show how, in a classical geostatistical sense, the data are used twice (e.g. Cressie (1989)). Not only are they present (linearly) in the kriging *predictor* (2.7); they are also used (non-linearly) to obtain an *estimator* of the spatial dependence parameters $\mathbf{K}$ and σ^2.

3.1. Some basic properties

Most importantly, the function $C(\mathbf{u},\mathbf{v})$ is non-negative definite, the proof of which is straightforward: for any locations $\{\mathbf{s}_i : i=1,\ldots,m\}$ in $\mathbb{R}^d$, any real $\{b_i : i=1,\ldots,m\}$, and any integer m, then (using obvious notation),

$$\sum_{i=1}^{m}\sum_{j=1}^{m} b_i b_j\, C(\mathbf{s}_i,\mathbf{s}_j)=\mathbf{b}'_m(\mathbf{S}_m\mathbf{K}\mathbf{S}'_m)\mathbf{b}_m=(\mathbf{S}'_m\mathbf{b}_m)'\mathbf{K}(\mathbf{S}'_m\mathbf{b}_m)\geqslant 0,$$

since $\mathbf{K}$ is positive definite.

A related model to expression (2.12), but different from $C(\cdot,\cdot)$ given above, is a consequence of the Karhunen–Loéve expansion (e.g. Adler (1981), section 3.3). Define the covariance function

$$C_1(\mathbf{u},\mathbf{v})\equiv\sum_{l=1}^{\infty}\lambda_l\,\phi_l(\mathbf{u})\,\phi_l(\mathbf{v}), \tag{3.1}$$

where $\{\lambda_l\}$ and $\{\phi_l(\cdot)\}$ are non-negative eigenvalues and orthonormal eigenfunctions respectively, which are obtained from the integral equation

$$\int C_1(\mathbf{u},\mathbf{v})\,\phi(\mathbf{v})\,\mathrm{d}\mathbf{v}=\lambda\,\phi(\mathbf{u}).$$

On truncating at the kth term of function (3.1), we obtain a different covariance function,

$$C_2(\mathbf{u},\mathbf{v})=\sum_{l=1}^{k}\lambda_l\,\phi_l(\mathbf{u})\,\phi_l(\mathbf{v})\equiv\boldsymbol{\phi}(\mathbf{u})'\boldsymbol{\Lambda}\,\boldsymbol{\phi}(\mathbf{v}),$$

where $\boldsymbol{\Lambda}$ is a $k\times k$ diagonal matrix of non-negative entries. Without loss of generality, assume that the truncation keeps only terms with positive eigenvalues; then clearly the truncated Karhunen–Loéve expansion is a special case of expression (2.12). Conversely, if we write $\mathbf{K}$ in its spectral form, $\mathbf{K}=\mathbf{P}\boldsymbol{\Lambda}\mathbf{P}'$, we see that $C(\mathbf{u},\mathbf{v})=(\mathbf{P}'\,\mathbf{S}(\mathbf{u}))'\boldsymbol{\Lambda}(\mathbf{P}'\,\mathbf{S}(\mathbf{v}))$, which looks like a truncated Karhunen–Loéve expansion but with non-orthogonal functions $\{\phi_l(\cdot)\}$. To sum up, model (2.12) involves a fixed rank variance–covariance matrix $\mathbf{K}$ (in general, not the identity matrix $\mathbf{I}$) to be estimated and a finite set of basis functions (in general, not orthogonal) to be chosen.

3.2. Basis functions

Because we make no requirement of orthogonality of basis functions, the choice of $S_1(\cdot),\ldots,S_r(\cdot)$ is unrestricted and may include *inter alia* the smoothing spline basis functions (e.g. Wahba (1990)), the wavelet basis functions (e.g. Vidakovic (1999)) and the radial basis functions (e.g. Hastie *et al.* (2001), pages 186–187). Whereas $\mathbf{K}$ is estimated from the data, $\mathbf{S}(\cdot)\equiv(S_1(\cdot),\ldots,S_n(\cdot))'$ is not. Nychka (2000) and Nychka *et al.* (2002) have brought together various choices of basis functions, where they assumed either $r=n$ (i.e. r is equal to the sample size and hence not fixed or if r is fixed then $\mathbf{K}$ is diagonal). The ability of the class (2.12) to approximate other covariance functions that are used in geostatistics, such as an isotropic exponential model, is convincingly demonstrated in Nychka *et al.* (2002). In fact, in Section 4 we shall be doing kriging on the globe, where we choose the basis functions to be multiresolution local bisquare functions.

Our main recommendation regarding the choice of basis functions is that they be multiresolutional. This enables the covariance function model (2.12) to capture multiple scales of variation. In Section 2.2, it is seen that model (2.12) can equivalently be thought of as a spatial random-effects model, $\mathbf{S}(\cdot)'\boldsymbol{\eta}$, where the random effects $\boldsymbol{\eta}$ have dependence structure given by $\mathbf{K}$. Hence, multiresolutional components of $\mathbf{S}(\cdot)$ allow many spatial scales of variation to be captured. Indeed, a large spatial scale that is missed by the mean function $\mathbf{t}(\cdot)'\boldsymbol{\alpha}$ in expression (2.3) can potentially be recovered by some of the spatial random-effects components of $\mathbf{S}(\cdot)'\boldsymbol{\eta}$.

Obvious classes of multiresolutional functions are different types of wavelets; equally, the class of local bisquare functions that are used in Section 4 is multiresolutional (but not orthogonal). In fact, in an analysis of satellite data on aerosols, Shi and Cressie (2007) choose the (non-orthogonal) W-wavelets as components of both the vector of mean functions $\mathbf{t}(\cdot)$ and the covariance basis functions $\mathbf{S}(\cdot)$. In this case, the problem reduces to one of model choice regarding which wavelets are used in $\mathbf{t}(\cdot)$ and which are used in $\mathbf{S}(\cdot)$, a solution to which is given by Shi and Cressie (2007). The more difficult problem of choosing which *class* of basis functions to use, from among several, is currently under investigation. Should we wish to compare two FRK maps to detect unusual differences, we recommend that the components of $\mathbf{t}(\cdot)$ and $\mathbf{S}(\cdot)$ be the same for the two maps.

From a computational point of view, it is beneficial to use a class of basis functions for which it is quick to evaluate $\mathbf{S}'\mathbf{V}^{-1}\mathbf{S}$ and $\mathbf{S}'\mathbf{a}$ for any $\mathbf{a}$. Although we have seen in Section 2.3 that such computations are in general $O(nr^2)$, by using the bisquare class in Section 4 or wavelet classes and sparse matrix libraries, the computational cost can be reduced in practice to $O(kr^2)$ where $k<n$.

3.3. Fitting the covariance function

The strategy that we adopt to fit the spatial covariance function is consistent with the geostatistical approach that is found in classical expositions, like that of Matheron (1963) and Journel and Huijbregts (1978). In that approach, an empirical estimator is first obtained for $\boldsymbol{\Sigma}$, which is based on the method of moments. The resulting estimator $\hat{\boldsymbol{\Sigma}}$ is noisy and may not be positive definite. However, on the basis of a parametric class $\{\boldsymbol{\Sigma}(\boldsymbol{\theta}):\boldsymbol{\theta}\in\Theta\}$, where each member of the class is positive definite, one chooses a $\hat{\boldsymbol{\theta}}\in\Theta$ such that $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$ is 'closest' to $\hat{\boldsymbol{\Sigma}}$. Finally, the resulting $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$ is substituted into the kriging equations (2.17) and (2.18).

In what is to follow, we see that the spatial dependence parameters $\boldsymbol{\theta}$ are made up of the $r\times r$ positive definite matrix $\mathbf{K}$ and a variance component $\sigma^2\in(0,\infty)$. Estimates $\hat{\mathbf{K}}$ and $\hat{\sigma}^2$ are obtained from minimizing a Frobenius norm between an empirical variance–covariance matrix and a theoretical variance–covariance matrix.

First, we define an empirical estimator of the variances and covariances, for which we need detrended data. In the absence of initial knowledge of the spatial dependence and for computational speed, we use the ordinary least squares estimator of $\boldsymbol{\alpha}$,

$$\bar{\boldsymbol{\alpha}}\equiv(\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{Z}, \tag{3.2}$$

from which we define the *detail residuals*,

$$D(\mathbf{s}_i)\equiv Z(\mathbf{s}_i)-\mathbf{t}(\mathbf{s}_i)'\bar{\boldsymbol{\alpha}}, \qquad i=1,\ldots,n. \tag{3.3}$$

As in classical geostatistics, we 'bin' the data for computation of the method-of-moments estimator of the spatial dependence. The number of bins, M, is meant to be fixed but larger than r, the number of basis functions. Hence, for estimation and fitting of covariances, once the data have been binned the computational complexity does not depend on n. Suppose that $\{\mathbf{u}_j:j=1,\ldots,M\}$, where $r\leqslant M<n$, is a set of locations, or *bin centres*, offering good coverage of D. Around the jth bin centre $\mathbf{u}_j$, define a neighbourhood $N(\mathbf{u}_j)$ and 0–1 weights,

$$w_{ji} \equiv \begin{cases} 1, & \text{if } \mathbf{s}_i \in N(\mathbf{u}_j), \\ 0, & \text{otherwise,} \end{cases} \tag{3.4}$$

$i=1,\ldots,n$, $j=1,\ldots,M$. Denote $\mathbf{w}_j \equiv (w_{j1},\ldots,w_{jn})'$ and define

$$\bar{Z}_j \equiv \mathbf{w}_j'(\mathbf{Z}-\mathbf{T}\tilde{\boldsymbol{\alpha}})/\mathbf{w}_j'\mathbf{1}_n, \tag{3.5}$$

where $\mathbf{1}_n$ is the $n \times 1$ vector of 1s, to be the average detrended data that are associated with bin centre $\mathbf{u}_j$, $j=1,\ldots,M$.

In approximation (A.4) in Appendix A, we show how $\boldsymbol{\Sigma}_M \equiv \text{var}(\bar{Z}_1,\ldots,\bar{Z}_M)$ can be approximated by $\bar{\boldsymbol{\Sigma}}_M(\mathbf{K},\sigma^2) \equiv \bar{\mathbf{S}}\mathbf{K}\bar{\mathbf{S}}' + \sigma^2\bar{\mathbf{V}}$, where $\bar{\mathbf{S}}$ and $\bar{\mathbf{V}}$ are easily computable binned versions of $\mathbf{S}$ and $\mathbf{V}$; as in classical geostatistics, the approximation is caused by small biases due to the binning. Also, in expression (A.2) in Appendix A, we give an *empirical positive definite estimate* $\hat{\boldsymbol{\Sigma}}_M$ that is based on the detail residuals (3.3). We then choose $\mathbf{K}$ positive definite and $\sigma^2 \in (0,\infty)$ such that $\bar{\boldsymbol{\Sigma}}_M(\mathbf{K},\sigma^2)$ is as 'close' to $\hat{\boldsymbol{\Sigma}}_M$ as possible.

There are various matrix norms between two matrices $\mathbf{A}$ and $\mathbf{B}$ of the same order. The one that we shall use is the Frobenius norm,

$$\|\mathbf{A}-\mathbf{B}\|^2 \equiv \text{tr}\{(\mathbf{A}-\mathbf{B})'(\mathbf{A}-\mathbf{B})\} = \sum_{j,k}(A_{jk}-B_{jk})^2, \tag{3.6}$$

which has also been used by Hastie (1996) in deriving pseudosplines, and by Donoho *et al.* (1998) in estimating covariances. We shall ultimately use a weighted Frobenius norm, in the same spirit as weighted least squares estimation of the variogram (Cressie, 1985), but for the moment we consider the unweighted version (3.6).

When $\sigma^2=0$, $\bar{\boldsymbol{\Sigma}}_M(\mathbf{K},0)=\bar{\mathbf{S}}\mathbf{K}\bar{\mathbf{S}}'$. Using the Frobenius norm, the $\mathbf{K}$ that minimizes $\|\hat{\boldsymbol{\Sigma}}_M - \bar{\boldsymbol{\Sigma}}_M(\mathbf{K},0)\|$ is given by (see Appendix A)

$$\hat{\mathbf{K}} = \mathbf{R}^{-1}\mathbf{Q}'\hat{\boldsymbol{\Sigma}}_M\mathbf{Q}(\mathbf{R}^{-1})', \tag{3.7}$$

and the corresponding fitted variance–covariance matrix is $\bar{\boldsymbol{\Sigma}}_M(\hat{\mathbf{K}},0)=\mathbf{Q}\mathbf{Q}'\hat{\boldsymbol{\Sigma}}_M\mathbf{Q}\mathbf{Q}'$, where $\bar{\mathbf{S}}=\mathbf{Q}\mathbf{R}$ is the *Q*–*R*-decomposition of $\bar{\mathbf{S}}$ (i.e. $\mathbf{Q}$ is an $M \times r$ orthonormal matrix and $\mathbf{R}$ is a non-singular $r \times r$ upper triangular matrix). The computational cost of the *Q*–*R*-decomposition is $O(r^3)$. Since $\hat{\boldsymbol{\Sigma}}_M$ is positive definite, so also is $\hat{\mathbf{K}}$.

When $\sigma^2 \in (0,\infty)$,

$$\|\hat{\boldsymbol{\Sigma}}_M - \bar{\boldsymbol{\Sigma}}_M(\mathbf{K},\sigma^2)\| = \|\hat{\boldsymbol{\Sigma}}_M - \sigma^2\bar{\mathbf{V}} - \bar{\mathbf{S}}\mathbf{K}\bar{\mathbf{S}}'\|,$$

resulting in the optimal parameter estimate (in terms of a given σ^2)

$$\hat{\mathbf{K}} = \mathbf{R}^{-1}\mathbf{Q}'(\hat{\boldsymbol{\Sigma}}_M - \sigma^2\bar{\mathbf{V}})\mathbf{Q}(\mathbf{R}^{-1})'; \tag{3.8}$$

the corresponding fitted variance–covariance matrix is $\bar{\boldsymbol{\Sigma}}_M(\hat{\mathbf{K}},\sigma^2)$ given by

$$\begin{aligned}\bar{\boldsymbol{\Sigma}}_M(\hat{\mathbf{K}},\sigma^2) &= \mathbf{Q}\mathbf{Q}'(\hat{\boldsymbol{\Sigma}}_M - \sigma^2\bar{\mathbf{V}})\mathbf{Q}\mathbf{Q}' + \sigma^2\bar{\mathbf{V}} \\ &= \mathbf{Q}\mathbf{Q}'\hat{\boldsymbol{\Sigma}}_M\mathbf{Q}\mathbf{Q}' + \sigma^2(\bar{\mathbf{V}} - \mathbf{Q}\mathbf{Q}'\bar{\mathbf{V}}\mathbf{Q}\mathbf{Q}').\end{aligned} \tag{3.9}$$

Thus, $\hat{\sigma}^2$ can be obtained by minimizing with respect to $\sigma^2 \in (0,\infty)$:

$$\|\hat{\boldsymbol{\Sigma}}_M - \bar{\boldsymbol{\Sigma}}_M(\hat{\mathbf{K}},\sigma^2)\|^2 = \sum_{j,k}\{(\hat{\boldsymbol{\Sigma}}_M - \mathbf{P}(\hat{\boldsymbol{\Sigma}}_M))_{jk} - \sigma^2(\bar{\mathbf{V}} - \mathbf{P}(\bar{\mathbf{V}}))_{jk}\}^2,$$

where $\mathbf{P}(\mathbf{A}) \equiv \mathbf{Q}\mathbf{Q}'\mathbf{A}\mathbf{Q}\mathbf{Q}'$ for any $M \times M$ matrix $\mathbf{A}$. The computational cost of this is $O(M^3)$. Note that this is just a simple linear regression with slope σ^2 and zero intercept. Hence the minimization, which is constrained so that equation (3.8) is positive definite and $\sigma^2 \in (0,\infty)$, is easily carried out. The result is

$$\hat{\mathbf{K}} = \mathbf{R}^{-1}\mathbf{Q}'(\hat{\mathbf{\Sigma}}_M - \hat{\sigma}^2\bar{\mathbf{V}})\mathbf{Q}(\mathbf{R}^{-1})'. \tag{3.10}$$

For $r < M$, the computational cost of parameter estimation is $O(M^3)$, which implies that a good choice of M is one that is independent of n and hence does not dominate the computational cost, $O(nr^2)$, of kriging; see Section 4 for comparisons of timings.

Finally, with spatial covariance parameter estimates $\hat{\mathbf{K}}$ and $\hat{\sigma}^2$, we can implement kriging; the estimates are substituted into equations (2.16)–(2.18). The resulting FRK involves matrix inversions of fixed rank $r \times r$ matrices and an $n \times n$ diagonal matrix, thus achieving a computational complexity that is linear in the number of data.

More weight should be given to bins that are less variable or have more data. Consider the *weighted* Frobenius norm,

$$\|\hat{\mathbf{\Sigma}}_M - \bar{\mathbf{\Sigma}}_M(\mathbf{K},\sigma^2)\|_a^2 \equiv \sum_{j,k} a_j a_k \{(\hat{\mathbf{\Sigma}}_M)_{jk} - (\bar{\mathbf{\Sigma}}_M(\mathbf{K},\sigma^2))_{jk}\}^2, \tag{3.11}$$

where $a_1, \ldots, a_M$ are known positive weights. Equivalently,

$$\|\hat{\mathbf{\Sigma}}_M - \bar{\mathbf{\Sigma}}_M(\mathbf{K},\sigma^2)\|_a^2 = \|\bar{\mathbf{A}}^{1/2}\hat{\mathbf{\Sigma}}_M\bar{\mathbf{A}}^{1/2} - \bar{\mathbf{A}}^{1/2}\bar{\mathbf{\Sigma}}_M(\mathbf{K},\sigma^2)\bar{\mathbf{A}}^{1/2}\|^2,$$

where $\bar{\mathbf{A}} \equiv \mathrm{diag}(a_1, \ldots, a_M)$, i.e. the weighted version of the Frobenius norm involves just scaling the rows and columns of $\hat{\mathbf{\Sigma}}_M$ and $\bar{\mathbf{\Sigma}}_M(\mathbf{K},\sigma^2)$ by $\{a_j : j = 1, \ldots, M\}$, and hence it is computationally no more onerous than the unweighted version. From expression (A.5) in Appendix A, we motivate statistically the choice to be

$$a_j \propto (\mathbf{w}_j'\mathbf{1}_n)^{1/2}/V_D(\mathbf{u}_j), \qquad j = 1, \ldots, M,$$

which is a data-based weight where $V_D(\mathbf{u}_j)$ is the empirical variance in the jth bin, given by expression (A.1) in Appendix A.

In summary, the problem of estimation of $\mathbf{K}$ and σ^2 is based on minimizing a weighted Frobenius norm. This is a weighted least squares criterion that is directly analogous to the approach that was given by Cressie (1985) for variogram estimation. It is moment based, not likelihood based. Under assumptions of Gaussianity, the likelihood of $\mathbf{K}$ and σ^2 depends on Σ^{-1} and $|\Sigma|$. From the Sherman–Morrison–Woodbury formula (2.16), we obtain $\mathbf{\Sigma}^{-1}$. A like formula yields $|\mathbf{\Sigma}| = |\sigma^2\mathbf{V}||\mathbf{K}||\mathbf{K}^{-1} + \mathbf{S}'(\sigma^2\mathbf{V})^{-1}\mathbf{S}|$, which involves determinants of $r \times r$ matrices, i.e. computation of the likelihood of $\mathbf{K}$ and σ^2 is feasible; however, its maximization is problematic unless $\mathbf{K}$ is further parameterized (Stein, 2008). Fuentes (2007) gives approximate likelihoods for large spatial data sets under assumptions of Gaussianity and covariance stationarity.

4. Fixed rank kriging of total column ozone satellite data

The problem of measuring TCO has been of interest to scientists for decades. Ozone depletion results in an increased transmission of ultraviolet radiation (290–400 nm wavelength) through the atmosphere. This is mostly deleterious because of damage to DNA and cellular proteins that are involved in biochemical processes, affecting growth and reproduction.

Relatively few measurements of TCO were taken in the first quarter of the 20th century. Subsequently, with the invention of the Dobson spectrophotometer, researchers gained the ability to measure efficiently and accurately TCO abundance (London, 1985). A system of ground-based stations has provided important TCO measurements for the past 40 years; however, the ground-based stations are relatively few in number and provide poor geographic coverage of the earth. The advent of polar orbiting satellites has dramatically enhanced the spatial coverage of measurements of TCO.

The Nimbus-7 polar orbiting satellite was launched on October 24th, 1978, with the total ozone mapping spectrometer instrument aboard. The instrument scanned in steps of 3° to an extreme of 51° on each side of nadir, in a direction that was perpendicular to the orbital plane (McPeters *et al.*, 1996). Each scan took roughly 8 s to complete, including 1 s for retrace (Madrid, 1978). The altitude of the satellite and scanning pattern of the total ozone mapping spectrometer instrument are such that consecutive orbits overlap, with the area of overlap depending on the latitude of the measurement. The satellite was sun synchronous, staying on the plane between the Earth and the Sun. Successive orbits moved westwards because of the rotation of the Earth, and hence the Nimbus-7 satellite covered the entire globe in a 24-h period. The instrument is a passive sensor and relies on backscattered light, which means that there are very few observations in winter near the poles.

On receiving satellite data, NASA calibrates them ('level 1') and preprocesses them to yield spatially and temporally irregular TCO measurements ('level 2'). The level 2 data are subsequently processed to yield a daily, spatially regular data product that is released widely to the scientific community ('level 3'). The level 3 data product for TCO used 1° latitude by 1.25° longitude (1°×1.25°) pixels (McPeters *et al.* (1996), page 44). Level 2 TCO data were obtained from the Ozone Processing Team of NASA–Goddard, Distributed Active Archive Center, and were stored in hierarchical data format as developed by the National Center for Supercomputing Applications at the University of Illinois.

In what is to follow, we use kriging, in particular FRK, to predict TCO data at the centre of the 1°×1.25° grid, on a daily basis, i.e. the prediction is at level 2 spatial support on the regular level 3 grid. The example is meant as an illustration of the ability of FRK to handle very large data sets; Shi and Cressie (2007) use FRK to produce a level 3 data product for aerosol optimal depth based on measurements from the multiangle imaging spectroradiometer instrument on the Terra satellite.

In this section, we use the 173405 level 2 TCO data that are available for October 1st, 1988; see Fig. 1. We implement FRK based on the kriging equations (2.16)–(2.18); some of the practical aspects of this optimal spatial prediction are now discussed.

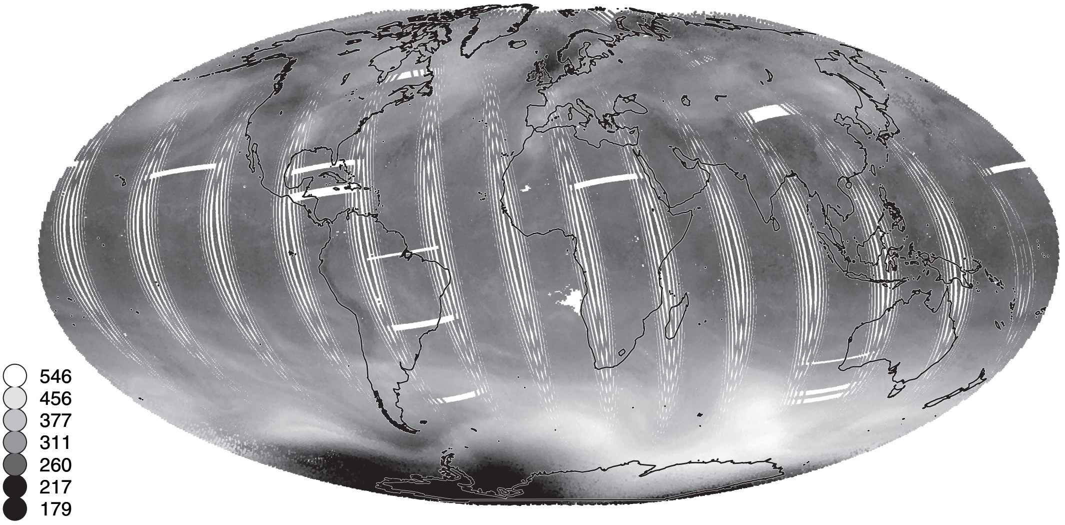

Fig. 1. Level 2 TCO data on October 1st, 1988, in Dobson units

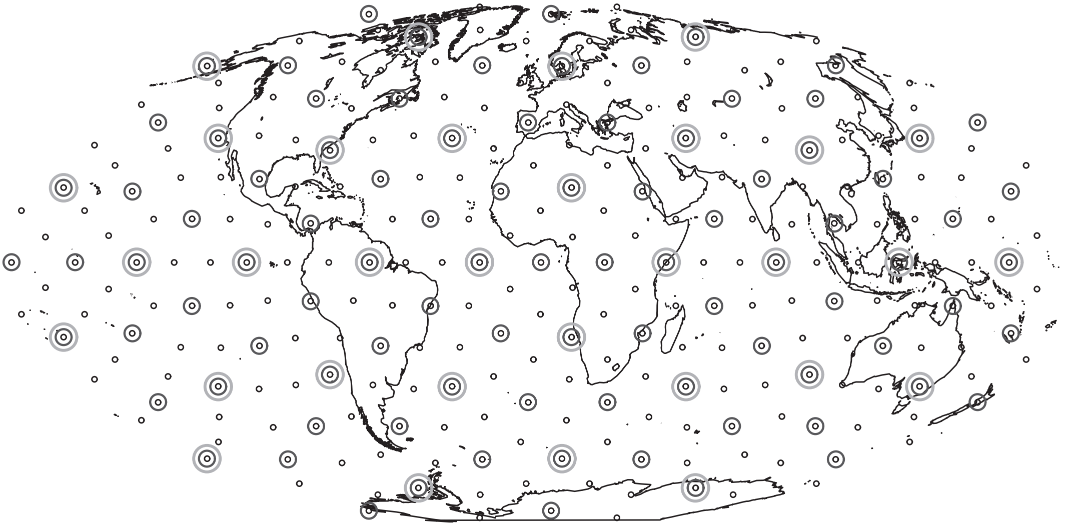

Fig. 2. Centre points of three resolutions of a discrete global grid

The basis functions that we choose in the spatial covariance model (2.12) are made up of three scales of variation. Each scale has 32, 92 and 272 functions associated with them, corresponding to the centre points of a discrete global grid (Sahr, 2001); see Fig. 2. The generic basis function in our spatial covariance model is the local bisquare function:

$$S_{j(l)}(\mathbf{u}) \equiv \begin{cases} \{1-(\|\mathbf{u}-\mathbf{v}_{j(l)}\|/r_l)^2\}^2, & \|\mathbf{u}-\mathbf{v}_{j(l)}\| \leqslant r_l, \\ 0, & \text{otherwise,} \end{cases}$$

where $\mathbf{v}_{j(l)}$ is one of the centre points of the lth resolution, $l=1,2,3$, and

$$r_l \equiv 1.5(\text{shortest great arc distance between centre points of the } l\text{th resolution}).$$

For example, for $l=1$, the shortest distance is 4165 km, and hence $r_l=6747.5$; the distances between centre points from resolution 2 and 3 are 1610 km and 1435 km respectively. Note that there are a total of $r=32+92+272=396$ basis functions.

The data were binned to carry out parameter estimation; see Section 3.3. We chose $M=812$ and $\{\mathbf{u}_1,\ldots,\mathbf{u}_{812}\}$ to be the centre points of resolution 4 of the discrete global grid that was referred to above. After computing the method-of-moments estimator $\hat{\Sigma}_M$, $\mathbf{K}$ and σ^2 were estimated assuming a constant mean, $E\{Y(\mathbf{s})\}\equiv\alpha$ (i.e. $\mathbf{t}(\mathbf{s})\equiv\mathbf{1}$), and $\mathbf{V}=\mathbf{I}$. Fig. 3 shows excellent fits of the theoretical semivariograms to the empirical semivariograms, at six locations on the globe. At a given location on the globe, the empirical semivariogram, as a function of spatial lag, was calculated from all data within a radius of 3000 km of the location. The same averages were taken of the (non-stationary) theoretical variogram values that are implied by the covariance function (2.12). The result is an averaged theoretical semivariogram that is now a function of spatial lag, and it is this that is compared with the empirical semivariogram in Fig. 3. This is a diagnostic summary that is easier to assess than empirical and fitted covariance functions at individual locations on the globe.

On substituting the estimates of $\mathbf{K}$ and σ^2 into equations (2.7) and (2.10), we obtain the FRK predictor and the FRK standard error respectively. On the regular 1°×1.25° grid, this yields Fig. 4 and Fig. 5 respectively. Fig. 4 shows a smooth map of TCO with the characteristic large TCO values around the $-55°$ latitudinal zone, from which they decrease precipitously to

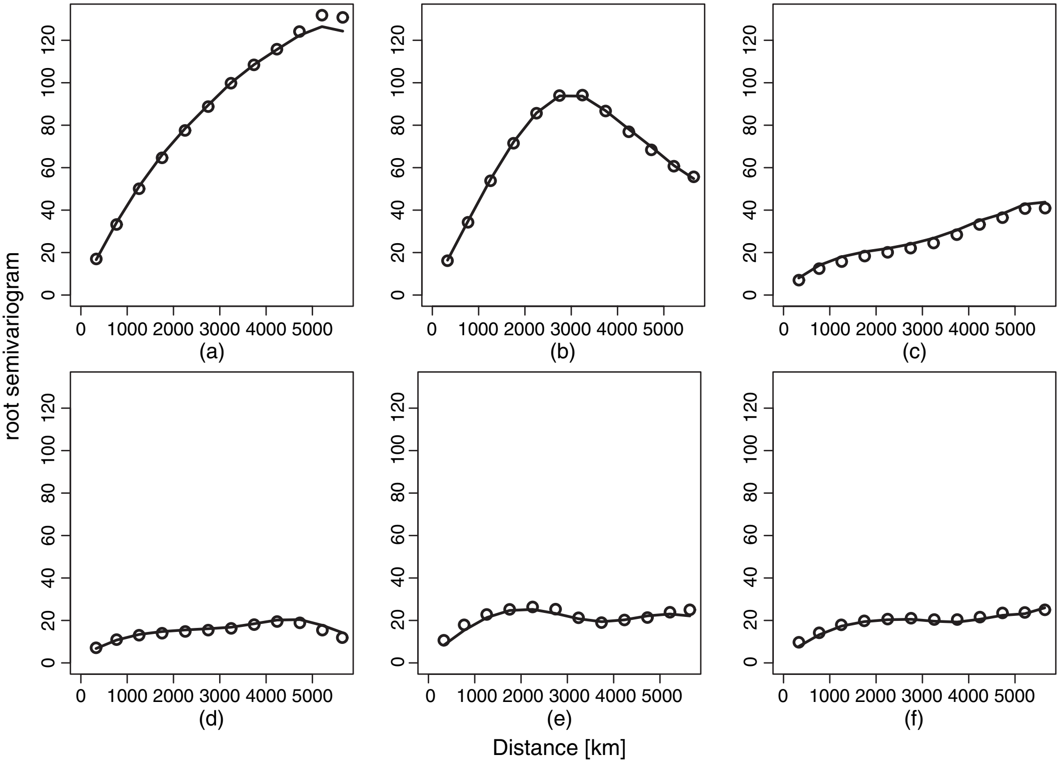

Fig. 3. Root semivariograms (i.e. semivariograms$^{1/2}$) for six different locations on the globe (location longitudes are all 0° for each plot) (○, empirical values; ———, theoretical values implied by equation (2.13)): (a) latitude −90°; (b) latitude −54°; (c) latitude −18°; (d) latitude 18°; (e) latitude 54°; (f) latitude 90°

form the ozone hole over the South Pole. In Fig. 5, the satellite swaths are evident owing to the variable sampling density that is caused by the geometry and the overlap of orbits; and the role of the basis functions in the prediction standard errors that are shown is to be expected.

We emphasize that *all* the 173405 data were used to produce Figs 4 and 5, that the covariance function that we used is *non-stationary* and that matrix inversions involved only 396 × 396 matrices. Importantly, the map in Fig. 4 is the statistically *optimal* predictor (for squared error loss) of TCO on the $1^\circ \times 1.25^\circ$ grid.

The following timings were carried out on a 1.8 GHz laptop with an Intel M processor, and they are for FRK on 173405 data at all $180 \times 228 = 51840$ prediction locations. Timings, in seconds, are given in parentheses: $\mathbf{S}$ (21 s), $\{\mathbf{S}(\mathbf{s}_0)\}$ (6 s), $\{\hat{Y}(\mathbf{s}_0)\}$ (4 s) and $\{\sigma_k(\mathbf{s}_0)\}$ (19 s). Since the computational cost is linear in n, it is clear that FRK is scalable to massive data sets of the order of gigabytes. Consider now the estimation of the parameters $\mathbf{K}$ and σ^2: $\hat{\Sigma}_M$ (35 s); $\hat{\mathbf{K}}$ and $\hat{\sigma}^2$ (56 s). Although this takes more time than the prediction part, it depends on M, the number of bins, which does not depend on n.

5. Discussion

This paper presents exact kriging (spatial BLUP) methodology when spatial data sets are very large. From the computational cost calculations, FRK is linear scalable and can handle massive data sets (of the order of gigabytes). Our results rely on using a class of non-stationary covariance

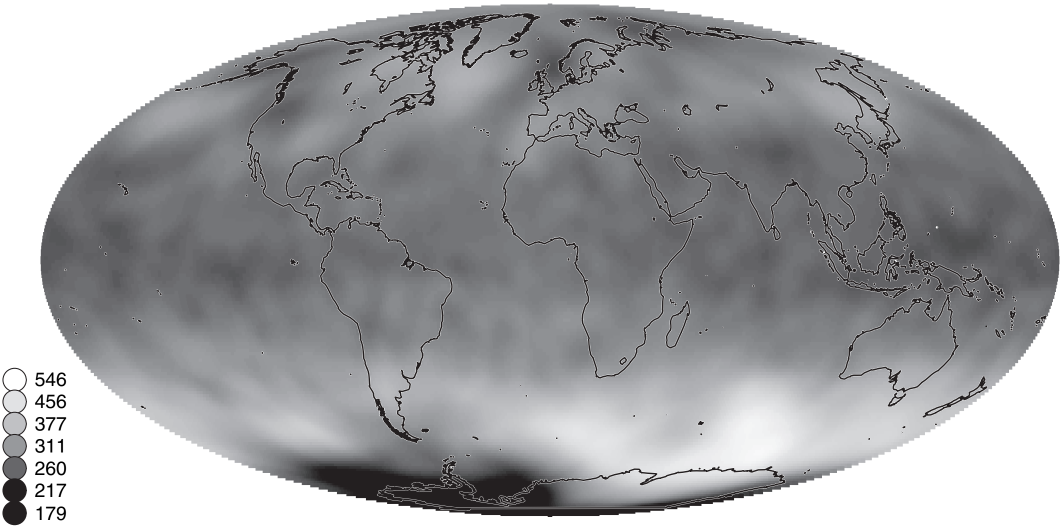

Fig. 4. FRK prediction of TCO for October 1st, 1988, in Dobson units

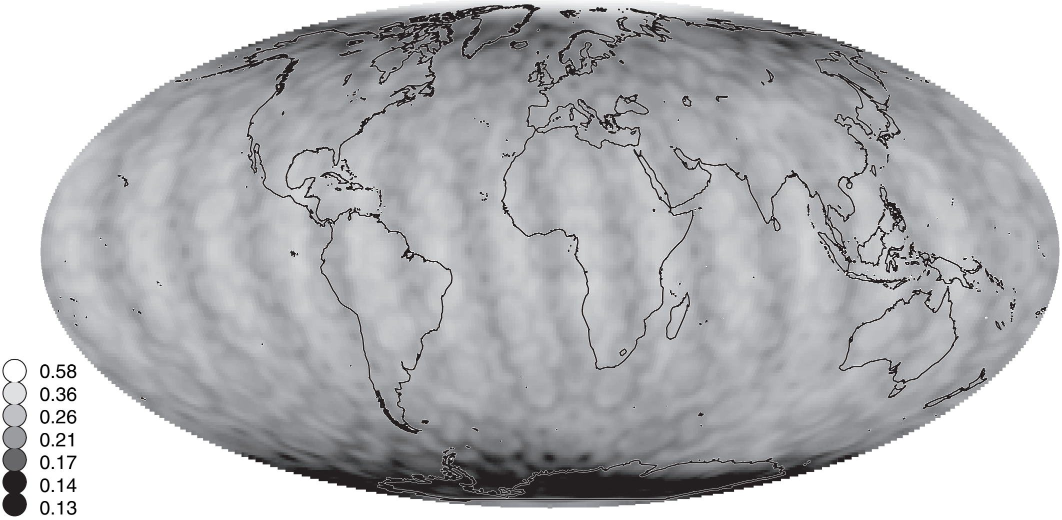

Fig. 5. FRK standard errors of the TCO predictions that are shown in Fig. 4, in Dobson units

functions that arise from a spatial random-effects model. Recall from equation (2.12) that

$$C(\mathbf{u},\mathbf{v}) \equiv \mathbf{S}(\mathbf{u})'\mathbf{K}\,\mathbf{S}(\mathbf{v}), \qquad \mathbf{u},\mathbf{v} \in \mathbb{R}^d,$$

where $\mathbf{S}(\cdot) \equiv (S_1(\cdot),\ldots,S_r(\cdot))'$ is a vector of basis functions. The $r \times r$ positive definite matrix $\mathbf{K}$ is a spatial dependence parameter, which we estimate in a classical geostatistical manner by using weighted least squares; maximum likelihood estimation of $\mathbf{K}$ is a topic of future research. A Bayesian approach, which is under development, puts a prior (e.g. a Wishart distribution) on $\mathbf{K}$. The Bayesian model specification might then be completed by assuming, for example, that $Y(\cdot)$ is a Gaussian process that is independent of the Gaussian white noise process $\varepsilon(\cdot)$, and that the prior on σ^2 is a gamma distribution.

A Bayesian analysis would also allow optimal spatial prediction in non-linear geostatistical models of the sort that were considered by Diggle *et al.* (1998). Although such models may be computationally heavy for large spatial data sets, owing to Markov chain Monte Carlo computations that are used in the analysis, there is still an opportunity to achieve computational speed-ups by using the spatial model (2.12). As evidence of this, Hrafnkelsson and Cressie (2003) compared a standard geostatistical covariance model with a model where inverse covariance matrices were modelled directly, and they reported that the latter led to a factor of more than 5 in increased computational efficiency.

Microscale variation in the (hidden) Y-process could be modelled by including another diagonal matrix in equation (2.6). When both diagonal matrices are proportional to each other, the measurement error parameter σ^2 and the microscale parameter τ^2 are not individually identifiable, although their sum $\tau^2 + \sigma^2$ is. The sum is often referred to as the 'nugget effect' in the geostatistical literature. In this paper we have assumed that $Y(\cdot)$ is smooth (i.e. $\tau^2 = 0$) and the rest of the variability is due to measurement error.

The non-stationary covariance functions that are given by expression (2.12) have remarkable change-of-support properties. Let $B \subset \mathbb{R}^d$ and define $Y(B) \equiv \int_B Y(\mathbf{s})\, d\mathbf{s}/|B|$, where $|B|$ is the d-dimensional volume of B. Then

$$\operatorname{cov}\{Y(B_1), Y(B_2)\} = \mathbf{S}(B_1)'\mathbf{K}\,\mathbf{S}(B_2), \qquad B_1, B_2 \subset \mathbb{R}^d,$$

where $\mathbf{S}(B) \equiv (S_1(B), \ldots, S_r(B))'$ and $S_l(B) \equiv \int_B S_l(\mathbf{s})\, d\mathbf{s}/|B|$, for $B \subset \mathbb{R}^d$. Thus, no matter the support of the data and the predictor, the kriging equations take the same form as equations (2.16)–(2.18). In practice, the basis functions would be integrated off line.

Finally, there is a natural generalization from the spatial random-effects model, $\nu(\mathbf{s}) = \mathbf{S}(\mathbf{s})'\boldsymbol{\eta}$, which is given in Section 3.2, to a spatiotemporal random-effects model $\nu(\mathbf{s}, t) = \mathbf{S}(\mathbf{s})'\,\boldsymbol{\eta}(t)$, where $\{\boldsymbol{\eta}(t) : t = 0, 1, 2, \ldots\}$ is an r-dimensional time series with mean 0 and $\operatorname{cov}\{\boldsymbol{\eta}(t_1), \boldsymbol{\eta}(t_2)\} \equiv \mathbf{K}(t_1, t_2)$, $t_1, t_2 = 0, 1, 2, \ldots$. Spatiotemporal kriging and spatiotemporal Kalman filtering that are based on this model, and a mixed model version involving trend, are currently under investigation.

Acknowledgements

We thank the Associate Editor and the referees for their helpful comments. This research was supported by the Office of Naval Research under grant N00014-05-1-0133. Partial support was also received from the US National Science Foundation under award DMS-0706997.

Appendix A

In Section 3.3, it was assumed that a positive definite estimate $\hat{\boldsymbol{\Sigma}}_M$ was available. We now define such an estimate. For any two bin centres $\mathbf{u}_j$ and $\mathbf{u}_k$, define the empirical covariances that are based on the detail residuals that are defined by expression (3.3):

$$\begin{aligned} C_D(\mathbf{u}_j, \mathbf{u}_k) &\equiv \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} w_{ji_1} w_{ki_2}\, D(\mathbf{s}_{i_1})\, D(\mathbf{s}_{i_2}) / (\mathbf{w}_j'\mathbf{1}_n)(\mathbf{w}_k'\mathbf{1}_n) \\ &= \bar{D}(\mathbf{u}_j)\, \bar{D}(\mathbf{u}_k), \qquad j, k = 1, \ldots, M, \end{aligned}$$

where, for $j = 1, \ldots, M$,

$$\bar{D}(\mathbf{u}_j) \equiv \sum_{i=1}^{n} w_{ji}\, D(\mathbf{s}_i) / \mathbf{w}_j'\mathbf{1}_n.$$

Further, for any bin centre $\mathbf{u}_j$, define the empirical variance,

$$V_D(\mathbf{u}_j) \equiv \sum_{i=1}^{n} w_{ji}\, D(\mathbf{s}_i)^2/\mathbf{w}_j'\mathbf{1}_n, \qquad j=1,\ldots,M. \tag{A.1}$$

Finally, define the $M \times M$ *empirical variance–covariance matrix* $\hat{\boldsymbol{\Sigma}}_M \equiv (\hat{\boldsymbol{\Sigma}}_M(\mathbf{u}_j,\mathbf{u}_k))$, where

$$\hat{\boldsymbol{\Sigma}}_M(\mathbf{u}_j,\mathbf{u}_k) \equiv \begin{cases} V_D(\mathbf{u}_k), & j=k, \\ C_D(\mathbf{u}_j,\mathbf{u}_k), & j \neq k. \end{cases} \tag{A.2}$$

Hence,

$$\hat{\boldsymbol{\Sigma}}_M = \mathbf{C}_D + \operatorname{diag}\{V_D(\mathbf{u}_1) - \bar{D}(\mathbf{u}_1)^2, \ldots, V_D(\mathbf{u}_M) - \bar{D}(\mathbf{u}_M)^2\},$$

where $\mathbf{C}_D \equiv (C_D(\mathbf{u}_j,\mathbf{u}_k))$. Further, $\hat{\boldsymbol{\Sigma}}_M$ is positive definite, because $\mathbf{C}_D$ is positive definite, and, for $j = 1,\ldots,M$,

$$V_D(\mathbf{u}_j) - \bar{D}(\mathbf{u}_j)^2 = \sum_i w_{ji}\{D(\mathbf{s}_i) - \bar{D}(\mathbf{u}_j)\}^2/\mathbf{w}_j'\mathbf{1}_n \geqslant 0.$$

For very large spatial data sets, n is of the order of hundreds of thousands and above, and M can be of the order of 1000.

Recall from expression (3.5) the definition of the average residual $\bar{Z}_j$, averaged over the jth bin, $j = 1,\ldots,M$. On the basis of the spatial covariance model (2.12), where r is of the order of hundreds, we calculate an approximation to $\operatorname{var}(\bar{Z}_1,\ldots,\bar{Z}_M)$ in terms of $\mathbf{K}$ and σ^2. For any two bin centres $\mathbf{u}_j$ and $\mathbf{u}_k$,

$$\begin{aligned} \operatorname{cov}(\bar{Z}_j,\bar{Z}_k) &\simeq \{\mathbf{w}_j'(\mathbf{S}\mathbf{K}\mathbf{S}')\mathbf{w}_k + \sigma^2(\mathbf{w}_j'\mathbf{V}\mathbf{w}_k)\}/(\mathbf{w}_j'\mathbf{1}_n)(\mathbf{w}_k'\mathbf{1}_n) \\ &\equiv \bar{\mathbf{S}}_j'\mathbf{K}\bar{\mathbf{S}}_k + \sigma^2\bar{V}_{jk}, \end{aligned} \tag{A.3}$$

where the approximation is due to replacing errors with detail residuals; this is common practice in geostatistics and has little effect when n is large, as is the case here (for example, see Cressie (1993), pages 169 and 195). Also, for $j \neq k$, $\bar{V}_{jk} \simeq 0$, which becomes an equality when bins do not overlap. As a consequence, if we define the $M \times 1$ vector $\bar{\mathbf{Z}} \equiv (\bar{Z}_1,\ldots,\bar{Z}_M)'$, then from expression (A.3)

$$\operatorname{var}(\bar{\mathbf{Z}}) \simeq \bar{\mathbf{S}}\mathbf{K}\bar{\mathbf{S}}' + \sigma^2\bar{\mathbf{V}} \equiv \boldsymbol{\Sigma}_M(\mathbf{K},\sigma^2). \tag{A.4}$$

In expression (A.4), $\bar{\mathbf{S}}$ is the $M \times r$ matrix that is defined by $\bar{\mathbf{S}} \equiv (\bar{\mathbf{S}}_1,\ldots,\bar{\mathbf{S}}_M)'$, $\bar{\mathbf{S}}_j \equiv \mathbf{S}'\mathbf{w}_j$, and $\mathbf{w}_j \equiv \{w_{ji}\}$ is a vector consisting of 0–1 weights given by expression (3.4), for $j = 1,\ldots,M$; further, $\bar{\mathbf{V}}$ is the $M \times M$ diagonal matrix that is defined by

$$\bar{\mathbf{V}} \equiv \operatorname{diag}(\bar{V}_{11},\ldots,\bar{V}_{MM}),$$

where $\bar{V}_{jj} \equiv \mathbf{w}_j'\mathbf{V}\mathbf{w}_j/(\mathbf{w}_j'\mathbf{1}_n)^2$, $j = 1,\ldots,M$.

We now give calculations that motivate the choice of weights $\{a_j\}$ in expression (3.11). Because of their form, they are determined by $\operatorname{var}\{\hat{\boldsymbol{\Sigma}}_M(\mathbf{u}_j,\mathbf{u}_j)\} = \operatorname{var}\{V_D(\mathbf{u}_j)\}$, where $V_D(\mathbf{u}_j)$ is given by expression (A.1). By invoking the same approximations as were made by Cressie (1985) in obtaining weights for estimating the variogram, we have

$$\begin{aligned} \operatorname{var}\{V_D(\mathbf{u}_j)\} &\simeq \sum_{i=1}^{n} w_{ji}^2 \operatorname{var}\{D(\mathbf{s}_i)^2\}/(\mathbf{w}_j'\mathbf{1}_n)^2 \\ &\simeq 2\sum_{i=1}^{n} w_{ji}^2 \operatorname{var}\{D(\mathbf{s}_i)\}^2/(\mathbf{w}_j'\mathbf{1}_n)^2, \end{aligned}$$

where the latter approximation is an equality if the data are Gaussian. Now $\{w_{ji}\}$ are 0–1 and in the jth bin we approximate $\operatorname{var}\{D(\mathbf{s}_i)\}$ with the natural estimate $V_D(\mathbf{u}_j)$. Then, since a_j^2 is inversely proportional to $\operatorname{var}\{V_D(\mathbf{u}_j)\}$, we have

$$\begin{aligned} a_j &\propto \operatorname{var}\{V_D(\mathbf{u}_j)\}^{-1/2} \\ &\simeq 2^{-1/2}(\mathbf{w}_j'\mathbf{1}_n)^{1/2}/V_D(\mathbf{u}_j), \end{aligned}$$

which leads to data-based weights,

$$a_j \propto (\mathbf{w}_j'\mathbf{1}_n)^{1/2}/V_D(\mathbf{u}_j), \qquad j=1,\ldots,M. \tag{A.5}$$

Finally, some technical results are needed to show how $\mathbf{K}$ and σ^2 can be estimated by minimizing the Frobenius norm.

Proposition 1. Let $\mathbf{C}$ be a given $n_1 \times n_2$ matrix whose entries are the covariances between an n_1-dimensional and an n_2-dimensional random vector. Let $\mathbf{S}_1$ and $\mathbf{S}_2$ be any given $n_1 \times r$ and $n_2 \times r$ matrices of rank $r \leqslant \min(n_1, n_2)$. Define $\mathbf{C}^*(\mathbf{K}) \equiv \mathbf{S}_1\mathbf{K}\mathbf{S}_2'$, where $\mathbf{K}$ is an $r \times r$ positive definite matrix. Consider the matrix norm, $\|\mathbf{A}\| \equiv \operatorname{tr}(\mathbf{A}'\mathbf{A})^{1/2}$. Then minimizing $\|\mathbf{C} - \mathbf{C}^*(\mathbf{K})\|$ with respect to $\mathbf{K}$ yields

$$\hat{\mathbf{K}} = \mathbf{R}_1^{-1}\mathbf{Q}_1'\mathbf{C}\mathbf{Q}_2(\mathbf{R}_2^{-1})',$$

where $\mathbf{S}_1 = \mathbf{Q}_1\mathbf{R}_1$ and $\mathbf{S}_2 = \mathbf{Q}_2\mathbf{R}_2$ are the *Q–R*-decompositions of $\mathbf{S}_1$ and $\mathbf{S}_2$ respectively. The minimized norm is

$$\|\mathbf{C} - \mathbf{C}^*(\hat{\mathbf{K}})\| = \operatorname{tr}(\mathbf{C}'\mathbf{C}) - \operatorname{tr}\{\mathbf{C}'\mathbf{C}^*(\hat{\mathbf{K}})\},$$

and the $\mathbf{C}^*(\cdot)$ closest to $\mathbf{C}$ is

$$\mathbf{C}^*(\hat{\mathbf{K}}) = \mathbf{Q}_1\mathbf{Q}_1'\mathbf{C}\mathbf{Q}_2\mathbf{Q}_2'.$$

Proof. Write $\mathbf{S}_i = \mathbf{Q}_i\mathbf{R}_i$, where $\mathbf{Q}_i$ is an $n_i \times r$ orthonormal matrix ($\mathbf{Q}_i'\mathbf{Q}_i = \mathbf{I}$) and $\mathbf{R}_i$ is a non-singular upper triangular $r \times r$ matrix, $i = 1, 2$. Define $\mathbf{K}^* \equiv \mathbf{R}_1\mathbf{K}\mathbf{R}_2'$. Then,

$$\|\mathbf{C} - \mathbf{C}^*(\mathbf{K})\|^2 = \operatorname{tr}(\mathbf{C}'\mathbf{C}) + \operatorname{tr}\{(\mathbf{K}^*)'\mathbf{K}^*\} - 2\operatorname{tr}\{(\mathbf{Q}_1'\mathbf{C}\mathbf{Q}_2)'\mathbf{K}^*\}.$$

Taking the derivative with respect to $\mathbf{K}^*$ gives

$$\frac{\partial}{\partial \mathbf{K}^*}\|\mathbf{C} - \mathbf{C}^*(\mathbf{K})\| = 2\mathbf{K}^* - 2(\mathbf{Q}_1'\mathbf{C}\mathbf{Q}_2),$$

and setting it equal to the zero matrix gives

$$\mathbf{K}^* = \mathbf{Q}_1'\mathbf{C}\mathbf{Q}_2.$$

From this, the results of the proposition follow.

Corollary 1. The statement containing equation (3.7) is true.

Proof. Put $\mathbf{C} = \hat{\mathbf{\Sigma}}_M$, $\mathbf{C}^*(\mathbf{K}) = \bar{\mathbf{\Sigma}}_M(\mathbf{K}, 0)$, $\mathbf{S}_1 = \mathbf{S}_2 = \bar{\mathbf{S}}$, and $n_1 = n_2 = M$, in proposition 1. Then the statement containing equation (3.7) follows.

References

Adler, R. J. (1981) *The Geometry of Random Fields*. Chichester: Wiley.

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

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

Cressie, N. (1985) Fitting variogram models by weighted least squares. *J. Int. Ass. Math. Geol.*, **17**, 563–586.

Cressie, N. (1989) Geostatistics. *Am. Statistn*, **43**, 197–202.

Cressie, N. (1990) The origins of kriging. *Math. Geol.*, **22**, 239–252.

Cressie, N. (1993) *Statistics for Spatial Data*, revised edn. New York: Wiley.

Cressie, N. and Johannesson, G. (2006) Spatial prediction of massive datasets. In *Proc. Australian Academy of Science Elizabeth and Frederick White Conf.*, pp. 1–11. Canberra: Australian Academy of Science.

Diggle, P. J., Tawn, J. A. and Moyeed, R. A. (1998) Model-based geostatistics. *Appl. Statist.*, **47**, 299–326.

Donoho, D. L., Mallet, S. and von Sachs, R. (1998) Estimating covariances of locally stationary processes: rates of convergence of best basis methods. *Technical Report 517*. Stanford University, Stanford.

Fuentes, M. (2007) Approximate likelihoods for large irregularly spaced spatial data. *J. Am. Statist. Ass.*, **102**, 321–331.

Furrer, R., Genton, M. G. and Nychka, D. (2006) Covariance tapering for interpolation of large spatial datasets. *J. Computnl Graph. Statist.*, **15**, 502–523.

Haas, T. C. (1995) Local prediction of a spatio-temporal process with an application to wet sulfate deposition. *J. Am. Statist. Ass.*, **90**, 1189–1199.

Hastie, T. (1996) Pseudosplines. *J. R. Statist. Soc.* B, **58**, 379–396.

Hastie, T., Tibshirani, R. and Friedman, J. (2001) *Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.

Henderson, H. V. and Searle, S. R. (1981) On deriving the inverse of a sum of matrices. *SIAM Rev.*, **23**, 53–60.

Hrafnkelsson, B. and Cressie, N. (2003) Hierarchical modeling of count data with application to nuclear fall-out. *Environ. Ecol. Statist.*, **10**, 179–200.

Huang, H.-C., Cressie, N. and Gabrosek, J. (2002) Fast, resolution-consistent spatial prediction of global processes from satellite data. *J. Computnl Graph. Statist.*, **11**, 63–88.

Johannesson, G. and Cressie, N. (2004a) Variance-covariance modeling and estimation for multi-resolution spatial models. In *geoENV IV—Geostatistics for Environmental Applications* (eds X. Sanchez-Vila, J. Carrera and J. Gomez-Hernandez), pp. 319–330. Dordrecht: Kluwer.

Johannesson, G. and Cressie, N. (2004b) Finding large-scale spatial trends in massive, global, environmental datasets. *Environmetrics*, **15**, 1–44.

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

Journel, A. G. and Huijbregts, C. (1978) *Mining Geostatistics*. London: Academic Press.

Kammann, E. E. and Wand, M. P. (2003) Geoadditive models. *Appl. Statist.*, **52**, 1–18.

London, J. (1985) The observed distribution of atmospheric ozone and its variations. In *Ozone in the Free Atmosphere* (eds R. C. Whitten and S. S. Prasad), pp. 11–80. New York: Van Nostrand Reinhold.

Madrid, C. R. (1978) *The Nimbus-7 User's Guide*. Greenbelt: NASA.

Matheron, G. (1962) *Traite de Geostatistique Appliqueé*, vol. I. Paris: Technip.

Matheron, G. (1963) Principles of geostatistics. *Econ. Geol.*, **58**, 1246–1266.

McPeters, R. D., Bhartia, P. K., Krueger, A. J., Herman, J. R., Schlesinger, B. M., Wellemeyer, C. G., Seftor, C. J., Jaross, G., Taylor, S. L., Swissler, T., Torres, O., Labow, G., Byerly, W. and Cebula, R. P. (1996) *The Nimbus-7 Total Ozone Mapping Spectrometer (TOMS) Data Products User's Guide*. Greenbelt: NASA.

Nychka, D. (2000) Spatial-process estimates as smoothers. In *Smoothing and Regression: Approaches, Computation, and Application* (ed. M. G. Schimek), pp. 393–424. New York: Wiley.

Nychka, D., Bailey, B., Ellner, S., Haaland, P. and O'Connell, M. (1996) *FUNFITS: Data Analysis and Statistical Tools for Estimating Functions*. Raleigh: North Carolina State University.

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

Quiñonero-Candela, J. and Rasmussen, C. E. (2005) A unifying view of sparse approximate Gaussian process regression. *J. Mach. Learn. Res.*, **6**, 1939–1959.

Rue, H. and Tjelmeland, H. (2002) Fitting Gaussian Markov random fields to Gaussian fields. *Scand. J. Statist.*, **29**, 31–49.

Sahr, K. (2001) *DGGRID Version 3.1b: User Documentation for Discrete Global Grid Generation Software*. Ashland: Southern Oregon University. (Available from `http://www.sou.edu/cs/sahr/dgg/`.)

Shi, T. and Cressie, N. (2007) Global statistical analysis of MISR aerosol data: a massive data product from NASA's Terra satellite. *Environmetrics*, **18**, 665–680.

Stein, M. (2008) A modeling approach for large spatial datasets. *J. Kor. Statist. Soc.*, **37**, in the press.

Stroud, J. R., Müller, P. and Sansó, B. (2001) Dynamic models for spatiotemporal data. *J. R. Statist. Soc.* B, **63**, 673–689.

Tzeng, S., Huang, H.-C. and Cressie, N. (2005) A fast, optimal spatial-prediction method for massive datasets. *J. Am. Statist. Ass.*, **100**, 1343–1357.

Vidakovic, B. (1999) *Statistical Modeling by Wavelets*. New York: Wiley.

Wahba, G. (1990) *Spline Models for Observational Data*. Philadelphia: Society for Industrial and Applied Mathematics.