Available online at www.sciencedirect.com

ScienceDirect

Computational Statistics & Data Analysis 52 (2008) 2331 – 2349

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Fast kriging of large data sets with Gaussian Markov random fields

Linda Hartman[a,*,1], Ola Hössjer[b,2]

[a] Centre for Mathematical Sciences, Lund University, Box 118, SE-221 00 Lund, Sweden
[b] Department of Mathematics, Stockholm University, SE-106 91 Stockholm, Sweden

Received 12 July 2006; received in revised form 21 September 2007; accepted 21 September 2007
Available online 12 October 2007

Abstract

Spatial data sets are analysed in many scientific disciplines. Kriging, i.e. minimum mean squared error linear prediction, is probably the most widely used method of spatial prediction. Computation time and memory requirement can be an obstacle for kriging for data sets with many observations. Calculations are accelerated and memory requirements decreased by using a Gaussian Markov random field on a lattice as an approximation of a Gaussian field. The algorithms are well suited also for nonlattice data when exploiting a bilinear interpolation at nonlattice locations.

Keywords: Spatial interpolation; Markov random field; Nonlattice data; Bilinear interpolation

1. Introduction

Kriging is a methodology for spatial prediction which the interest in and applications for have been increasing during recent decades. The development originated in mining, where spatial prediction was used to estimate changes in ore grade within a mine. Nowadays the methods are used in a variety of areas in geology and other scientific disciplines. Applications include designing environmental monitoring networks, road and tunnel planning, management of soil resources in agriculture and forestry, etc. While the mining applications of kriging often had few observations, today's applications can have thousands or tens of thousands of data, a fact that must be addressed by the evaluation methods.

Kriging methodology is part of the least squares linear regression framework. In addition to linear kriging methods such as simple, ordinary or universal kriging, the methodology has been enlarged to account for models where the best predictor (in LS-sense) is not linear. Moyeed and Papritz (2002) compare both linear and nonlinear kriging methods with respect to precision and success in modelling prediction uncertainty for a data set of metal concentrations. Regarding precision, they found that linear kriging was as good as any of the nonlinear methods. For skewed data, however, some of the nonlinear methods performed better in estimating prediction uncertainty. This article focuses on computationally effective models for linear kriging.

One drawback of kriging is the computational cost for large data sets. A common way (Goovaerts, 1997; Isaaks and Srivastava, 1989) to deal with this is to use *local neighbourhood kriging* where only the closest observations are

[*] Corresponding author. Tel.: +46 46 222 47 79; fax: +46 46 222 46 23.
E-mail address: linda@maths.lth.se (L. Hartman).

[1] Supported by the Swedish Foundation for Strategic Research (SSF).

[2] Supported by the Swedish Research Council.

0167-9473/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2007.09.018

used (for each prediction). Although computationally attractive when predicting at a limited number of locations, the methods require a local neighbourhood for each location where a prediction is made, and predicting at a fine grid is still computationally demanding. Another drawback is discontinuity in prediction and standard error surfaces.

As an alternative to local models, Barry and Pace (1997) used models for which the covariance matrix is sparse (e.g. spherical correlation) together with sparse matrix techniques. Furrer et al., 2006 have a more general approach in which covariance matrices are tapered to obtain zero covariance for large distances. Kammann and Wand (2003) and Nychka and Saltzman (1998) use *reduced knot or low-rank kriging*, where by means of a space-filling algorithm, they introduce knots and reduce computation. A linear mixed model based on (hypothetical and unknown) observations at the knots is fitted to all data points.

Another field of spatial statistics that received much attention in recent years is Markov field approximations of random fields (especially applied in the context of simulation, e.g. MCMC analyses). In Rue and Tjelmeland (2002), Gaussian random fields are approximated with two-dimensional Gaussian Markov random fields (GMRFs). Even for small Markov neighbourhoods (such as 5 × 5) the fit to the non-Markov covariance functions is remarkably good.

In this article we suggest a method for kriging large data sets where global features are of interest. Instead of approximating the kriging predictor of a given field (as in local neighbourhood kriging) the approach in this paper is to develop an exact kriging predictor for a GMRF that approximates the given field. We can then exploit the GMRF's natural specification via its precision (inverse covariance) matrix to make kriging computationally effective even for large data sets. We further suggest using a bilinear interpolation that combines the strength of lattice GMRF with the generality of a continuous spatial index (which is common in geostatistical data sets). Compared to nearest neighbour interpolation, bilinear interpolation yields better predictions and coverage with (almost) no extra computational cost.

2. Linear kriging

Linear kriging is often divided into *simple kriging* (known mean), *ordinary kriging* (unknown but constant mean) and *universal kriging* (the mean is an unknown linear combination of known functions), depending on the mean value specification. For clarity of the following discussion we describe the benefits of a Markov approximation for *ordinary kriging*, although the principles apply also for universal kriging, see Section 7.

Let therefore

$$Z(\boldsymbol{t}) = \mu + \gamma(\boldsymbol{t}) + \varepsilon(\boldsymbol{t}), \quad \boldsymbol{t} = (t_1, t_2) \in D \subset \mathbb{R}^2 \tag{1}$$

be a stochastic field in the area D with mean value μ, zero-mean intermediate-scale variation $\gamma(\cdot)$ and white noise error $\varepsilon(\cdot)$ with variance σ_ε^2. Given observations $(Z(\boldsymbol{t}_1), \ldots, Z(\boldsymbol{t}_n))^{\mathrm{T}} = \boldsymbol{Z}$, a common task is to predict $Y(\boldsymbol{t}) = \mu + \gamma(\boldsymbol{t})$, often at a grid of locations, and to calculate the prediction error variance at each such location.

In *ordinary kriging* linear predictors are used, i.e.

$$\hat{Y}(\boldsymbol{t}_0) = \sum_{i=1}^{n} a_i Z(\boldsymbol{t}_i) = \boldsymbol{a}^{\mathrm{T}}\boldsymbol{Z}, \tag{2}$$

where $\boldsymbol{a} = (a_1, \ldots, a_n)^{\mathrm{T}}$ are weights such that $\sum_{i=1}^{n} a_i = 1$, in order for the predictor to be unbiased.

Let $\boldsymbol{\Sigma} = (\mathrm{Cov}(Z(\boldsymbol{t}_i), Z(\boldsymbol{t}_j)))_{i,j} = \mathrm{Cov}(\boldsymbol{Z})$ denote the covariance matrix of the observations, and let the vector

$$\boldsymbol{\omega} = (\mathrm{Cov}(Z(\boldsymbol{t}_1), Y(\boldsymbol{t}_0)), \ldots, \mathrm{Cov}(Z(\boldsymbol{t}_n), Y(\boldsymbol{t}_0)))^{\mathrm{T}} = \mathrm{Cov}(\boldsymbol{Z}, Y(\boldsymbol{t}_0))$$

contain the covariances between observations and the value at an arbitrary location $\boldsymbol{t}_0$, where we want to predict the process.

The best linear predictor is found by minimizing the mean squared prediction error. The resulting weights (Cressie, 1993, p. 123) are

$$\boldsymbol{a} = \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\omega} + \mathbf{1}_n\left(\frac{\mathbf{1} - \mathbf{1}_n^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\omega}}{\mathbf{1}_n^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{1}_n}\right)\right), \tag{3}$$

where $\mathbf{1}_n = (\mathbf{1}, \ldots, \mathbf{1})^{\mathrm{T}}$ is a column vector of size n.

The unconditional kriging variance, i.e. the variance of the prediction error $r(\boldsymbol{t}) = Y(\boldsymbol{t}) - \hat{Y}(\boldsymbol{t})$, is (Cressie, 1993, pp. 123, 128)

$$\sigma_r^2(\boldsymbol{t}_0) = \sigma_\gamma^2(\boldsymbol{t}_0) + \frac{1 - \mathbf{1}_n^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\omega}}{\mathbf{1}_n^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{1}_n} - \left(\boldsymbol{\omega} + \mathbf{1}_n\left(\frac{\mathbf{1} - \mathbf{1}_n^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\omega}}{\mathbf{1}_n^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{1}_n}\right)\right)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\omega}$$

$$= \sigma_\gamma^2(\boldsymbol{t}_0) - \boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\omega} + \frac{(1 - \mathbf{1}_n^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\omega})^2}{\mathbf{1}_n^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{1}_n}, \tag{4}$$

where $\sigma_\gamma^2(\boldsymbol{t}_0)$ is the variance of the intermediate-scale variation at $\boldsymbol{t}_0$.

Using the quantiles of the predictors, prediction intervals can be obtained. E.g. if $Z(\boldsymbol{t})$ is Gaussian, so is the predictor, and thus the $1 - \alpha$ prediction interval is obtained as

$$I_Y(\boldsymbol{t}_0) = [\hat{Y}(\boldsymbol{t}_0) \pm z_{\alpha/2}\sigma_r(\boldsymbol{t}_0)], \tag{5}$$

where z_α is the $(1 - \alpha)$-quantile of the standard normal distribution. Alternatively the conditional prediction error variance, calculated without the last term of (4), could be used for prediction intervals, see Sjöstedt-de Luna and Young (2003).

Any covariance function (i.e. any nonnegative definite function) is valid for kriging. To be able to estimate model parameters, however, most applications use a model where the covariance function is stationary (though possibly anisotropic) and of standard form, such as exponential, spherical, Gaussian, pure nugget or a linear combination of such models, see Goovaerts (1997, p. 88).

If a parametric model, $\mathrm{Cov}(Z(\boldsymbol{t}_1), Z(\boldsymbol{t}_2)) = C_Z(\boldsymbol{t}_1, \boldsymbol{t}_2; \theta)$, for the covariance function is given, where θ is a vector of unknown parameters, a common procedure is as follows: Estimate θ and then predict $Y(\boldsymbol{t}_0)$ with (2) and (5) replacing θ by a plug-in estimate $\hat{\theta}$ in $\boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\omega}$. When plug-in of estimated covariance parameters is used, the prediction error variance (4) does not reflect the uncertainty of the estimate and the plug-in prediction interval will therefore not have the prescribed coverage. Sjöstedt-de Luna and Young (2003) handle this with bootstrap calibration.

3. Kriging and GMRF

After parameter estimation, ordinary kriging requires solving a linear system (3) of size $n \times n$, an operation for which operation count and memory requirement increase by order n^3 and n^2, respectively. This implies that kriging large data sets on a global basis is very slow, or even impossible due to memory limitations.

However, for models with a sparse precision matrix $Q = \Sigma^{-1}$, the algorithms can be greatly improved. The idea of this paper is to obtain those computational benefits by *approximating* the given stochastic field with a field with a sparse precision matrix. As the kriging weights are determined by the covariance matrix (3), the key is to find a field with covariance matrix similar to the original one, while maintaining a sparse precision matrix. Turning our attention to Gaussian fields (for which the best linear unbiased predictor is optimal) this could be obtained by approximating with a GMRF, described below.

Let Γ be a lattice in $\mathbb{R}^2$ with coordinates $\tilde{\boldsymbol{t}}_1, \ldots, \tilde{\boldsymbol{t}}_N$. The field $\tilde{\boldsymbol{\gamma}}$ defined on Γ is a GMRF if it is Gaussian and satisfies the Markov property

$$\tilde{\gamma}_i | \tilde{\gamma}_{(-i)} \overset{\mathrm{d}}{=} \tilde{\gamma}_i | \tilde{\gamma}_{N[i]} \sim \mathrm{N}\left(\sum_{j=1}^{N} w_{ij}\tilde{\gamma}_j, \tau^2\right) \quad \text{for } i = 1, \ldots, N, \tag{6}$$

where $\tilde{\gamma} = (\tilde{\gamma}_1, \ldots, \tilde{\gamma}_N)^{\mathrm{T}}$, and $N[i]$ is the *neighbourhood* of lattice point i. With $\mathbf{W} = (w_{ij})_{i,j=1}^{N}$ where $w_{ij} = 0$, $j \notin N[i]$, the joint distribution of $\tilde{\boldsymbol{\gamma}}$ is

$$\tilde{\boldsymbol{\gamma}} \sim \mathrm{N}(0, (\mathbf{I}_N - \mathbf{W})^{-1}\tau^2), \tag{7}$$

provided that $(\mathbf{I}_N - \mathbf{W})^{-1}$ exists and is symmetric and nonnegative definite. Note that τ^2 is the conditional, not the unconditional, variance of $\tilde{\gamma}_i$.

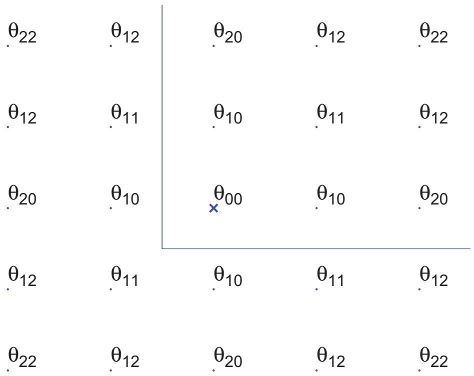

Fig. 1. Parameters for a stationary isotropic GMRF with 5 × 5 neighbourhood. Since the GMRF is stationary, each parameter determines the value of the precision matrix between a grid point and the corresponding neighbour. The parameters are defined as $\theta_{kl} = Q_{(00,kl)}$, where the double digit index refers to the grid point with the corresponding coordinates. As the precision matrix $\mathbf{Q} = (\mathbf{I}_N - \mathbf{W})\tau^{-2}$, the parameters $\boldsymbol{\theta}$ specifies the weight matrix $\mathbf{W}$.

The Markov property induced by the neighbourhood system $N[i]$, $i \in \Gamma$ thus imposes a zero-pattern structure in the precision matrix $\boldsymbol{Q} = (\mathbf{I}_N - \mathbf{W})\tau^{-2}$ viz. that $Q_{ij} = 0$ unless i and j are neighbours. As the inverse covariance matrix is essential in kriging, the GMRF model on a lattice is a model with considerable computational benefits for large data sets, provided that neighbourhoods are small in comparison to grid size.

Although the Markov model is seldom the most natural model for data it can be used as an approximation of the model at hand. Rue and Tjelmeland (2002) show that even for fields with long correlation length, the Gaussian fields most commonly used for kriging can be well approximated (on regular grids) using GMRFs with small neighbourhoods (such as a rectangle of size 5 × 5 or 7 × 7 grid points). Compared to the error usually induced when estimating the covariance parameters from data, the covariance function deviance between a Gaussian field and its fitted GMRF was small even for neighbourhoods as small as 5 × 5. This indicates that a GMRF approximation on a regular grid would give competitive kriging results at a lower computational cost.

Following Rue and Tjelmeland (2002), let Γ be a $N_1 \times N_2$ square lattice in $\mathbb{R}^2$ with coordinates $\tilde{\boldsymbol{t}}_1, \ldots, \tilde{\boldsymbol{t}}_N$, $N = N_1 N_2$, and cell width δ. Let $\tilde{\boldsymbol{\gamma}} = (\tilde{\gamma}_j)_{j=1}^N = (\tilde{\gamma}(\tilde{\boldsymbol{t}}_j))_{j=1}^N$ be a GMRF that is fitted to the Gaussian field $\gamma(\cdot)$ at lattice locations. To avoid boundary effects, the lattice can be extended on a torus. The precision matrix $\boldsymbol{Q}$ of an isotropic GMRF with a $(2m+1) \times (2m+1)$-neighbourhood then requires only $m(m+1)$ parameters. Fig. 1 displays the six parameters needed with a 5 × 5 neighbourhood. For a stationary $\gamma(\cdot)$, we will approximate $\gamma(\cdot)$ by $\tilde{\gamma}(\cdot)$ at lattice locations by requiring

$$\operatorname{Var}(\gamma(\tilde{\boldsymbol{t}}_j)) = \operatorname{Var}(\tilde{\gamma}_j), \quad j = 1, \ldots, N, \tag{8}$$

and then fitting the correlation function of $\tilde{\boldsymbol{\gamma}}$ to the correlation function of $\gamma(\cdot)$ on the lattice. The conditional precision θ_{00} is thus fitted exactly to obtain the same variance as the Gaussian field and the other parameters are fitted to minimize the distance between the correlation matrix of the GMRF and the corresponding correlation function of the Gaussian field:

$$\|\rho - \rho(\theta)\|_\omega^2 = \sum_{ij} (\rho_{ij} - \tilde{\rho}(\boldsymbol{\theta})_{ij})^2 \alpha_{ij}. \tag{9}$$

Here ρ_{ij} is the correlation of the original Gaussian field between the lattice point with coordinates $\{i, j\}$ and the lattice point with coordinates $\{0, 0\}$, and $\tilde{\rho}(\boldsymbol{\theta})_{ij}$ is the counterpart for the GMRF. The positive weights α_{ij} determine which part of the correlation function is given most weight in the fitting, e.g. if $d(\cdot, \cdot)$ is the distance, the weights $\alpha_{ij} \propto 1/d((i, j), (0, 0))$ for $ij \neq 00$, give more weight at shorter distances, see Rue and Held (2005, Chapter 5.1.2) for details. If the Gaussian field is anisotropic, the fitting procedure is preceded by a linear transformation such that isotropy is obtained. Considering untransformed coordinates, this will imply that the lattice is rotated, and that the original cell widths may be different in the two directions. In the new coordinate system, the range can be expressed in cell widths. Thus in the old coordinates, the cell width is proportionally longer in the direction in which the range was longer.

However, many geostatistical applications have nonlattice data, and as suggested by Diggle and Lophaven (2006), a regular design is often not the best choice for spatial sampling when both efficient parameter estimation and efficient spatial prediction are considered. We therefore suggest a generalization where the model is extended to nonlattice locations, while the computational benefits of the GMRF formulation are maintained. To extend the lattice model to an arbitrary $\boldsymbol{t} \in D$, let the intermediate-scale variation consist of two components

$$\gamma(\boldsymbol{t}) = \gamma_1(\boldsymbol{t}) + \gamma_2(\boldsymbol{t}). \tag{10}$$

At lattice locations let $\gamma_1(\tilde{\boldsymbol{t}}_j) = \tilde{\gamma}_j$, where $\tilde{\boldsymbol{\gamma}}$ is the GMRF fitted to the Gaussian field. For an arbitrary $\boldsymbol{t} \in D$ let

$$\gamma_1(\boldsymbol{t}) = \sum_{j=1}^{N} k(\boldsymbol{t}, \tilde{\boldsymbol{t}}_j)\ \tilde{\gamma}_j, \tag{11}$$

where $k(\boldsymbol{t}, \cdot)$ are interpolation weights fulfilling $\sum_j k(\boldsymbol{t}, \tilde{\boldsymbol{t}}_j) = 1$ and $k(\tilde{\boldsymbol{t}}_i, \tilde{\boldsymbol{t}}_i) = 1$. Two interpolation schemes, nearest neighbour and bilinear interpolation, are suggested in Section 3.1 and examined in Section 4. Due to computational aspects, only lattice locations within the used Markov neighbourhood of the closest lattice location are used in the interpolation.

For nonlattice locations, $\gamma_1(\cdot)$ within grid cells will be a weighted average of $\tilde{\boldsymbol{\gamma}}$. Thus, for interpolation schemes where the weighted sum (11) consists of more than one nonzero term, the variance of $\gamma_1(\cdot)$ will be lower within grid cells than at lattice locations. This entails that the approximation error $\gamma_2(\boldsymbol{t}) = \gamma(\boldsymbol{t}) - \gamma_1(\boldsymbol{t})$ will have larger variance within grid cells.

In order to balance good approximation of $\gamma(\cdot)$ with computational efficiency, we make the following *approximate* assumptions for $\gamma_2(\boldsymbol{t})$:

$$\gamma_2(\cdot) \text{ and } \gamma_1(\cdot) \text{ are independent fields,} \tag{12}$$

$$\gamma_2(\cdot) \text{ is a (possibly heterogeneous) white noise field} \tag{13}$$

and

$$\mathrm{Var}(\gamma_2(\boldsymbol{t})) = \mathrm{Var}(\gamma(\boldsymbol{t})) - \sum_i \sum_j k_i k_j \mathrm{Cov}(\gamma(\tilde{\boldsymbol{t}}_i), \gamma(\tilde{\boldsymbol{t}}_j)). \tag{14}$$

Note that $\mathrm{Var}(\gamma_2(\boldsymbol{t})) = 0$ at lattice locations.

Following from (10)–(13), the variance of γ_2 would be

$$\mathrm{Var}(\gamma_2(\boldsymbol{t})) = \mathrm{Var}(\gamma(\boldsymbol{t})) - \sum_{ij} k_i k_j \mathrm{Cov}(\gamma_1(\tilde{\boldsymbol{t}}_i), \gamma_1(\tilde{\boldsymbol{t}}_j)), \tag{15}$$

but (14) is chosen for computational reasons. Further, since the variance of γ and γ_1 coincide on the lattice (Eq. (8)), the covariances of the two fields for adjacent grid points will also be close. Thus the difference between (14) and (15) is small in practise, see further discussion in Section 6.

For the purpose of comparison, in the simulations of Section 4 we will also consider the kriging quality when γ_2 is set to 0, i.e. when the model does not compensate for the lower variance within grid cells.

The GMRF $\tilde{\boldsymbol{\gamma}}$ is fitted to approximate the Gaussian field $\gamma(\cdot)$ at lattice locations and the lattice locations must be chosen by the modeller. The grid size is a balance between computational cost and accurate results, as is apparent from the simulations in Section 4. As the fitting criterion (8) only compares covariances at distances present in the regular lattice, the chosen cell-width δ must be small enough to achieve good approximations at short distances. A denser grid admits fitting the GMRF to the Gaussian field at shorter distances, but will also imply heavier computations. For a field with short range, short range behaviour is emphasized, and grid points must then be closer than for a field with longer range. To avoid edge effects, the lattice is wrapped on a torus, and to avoid influence of the torus connections, the lattice should be extended to cover an area larger than the area of interest (Rue and Held, 2005, Chapter 5.1.4). When the observations or wanted predictions are from an area that is far from being rectangular, computational gains will decrease as many extra grid points outside the area of interest will be needed, just as in the soil data example in Section 5.

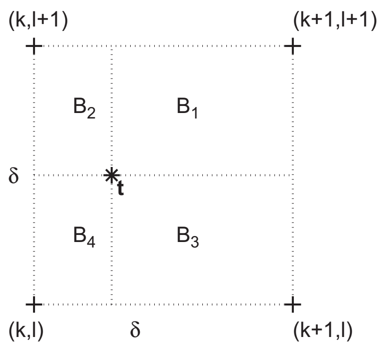

Fig. 2. Areas determining the weights when averaging over the four nearest lattice points.

3.1. Interpolation schemes

When approximating a process with continuous spatial index with a process on a grid, nonlattice locations can be associated with lattice locations through an interpolation scheme. This allows for kriging prediction for data with continuous spatial index, as in Section 2, but with considerable gain in computational time if the Markov neighbourhood is small.

3.1.1. Nearest neighbour interpolation

With this scheme, the value at each nonlattice location is set equal to the value at the closest lattice location, thus giving rise to a field which is piecewise constant. This scheme is equal to interpolation (11) with

$$k(\boldsymbol{t},\tilde{\boldsymbol{t}}_j)=\begin{cases}1 & \tilde{\boldsymbol{t}}_j \text{ is the grid location closest to location } \boldsymbol{t},\\ 0 & \text{otherwise.}\end{cases} \tag{16}$$

With this scheme $\gamma_2(\cdot)\equiv 0$ if $\gamma(\cdot)$ is stationary. Applications using this scheme are found e.g. in Wikle et al. (1998) and Rue and Held (2005, Chapter 5).

3.1.2. Bilinear interpolation

Nearest neighbour interpolation produces a discontinuous field that does not capture the behaviour of the full Gaussian field within grid cells. More sophisticated is a bilinear interpolation, used in Werner Hartman (2006). Here the value at a location inside a grid cell is equal to the mean value of the GMRF at the four surrounding grid points.

For k in Eq. (11) let $k(\boldsymbol{t},\tilde{\boldsymbol{t}}_j)=0$ if j is not a corner of the lattice square that surrounds $\boldsymbol{t}$, and calculate the remaining four weights from the areas $B_1,\ldots,B_4$ in Fig. 2: $k(\boldsymbol{t},\tilde{\boldsymbol{t}}_{(k,l)})=B_1/B$, $k(\boldsymbol{t},\tilde{\boldsymbol{t}}_{(k+1,l)})=B_2/B$, $k(\boldsymbol{t},\tilde{\boldsymbol{t}}_{(k,l+1)})=B_3/B$ and $k(\boldsymbol{t},\tilde{\boldsymbol{t}}_{(k+1,l+1)})=B_4/B$ where $B=\delta^2=B_1+B_2+B_3+B_4$. For grid locations let $k(\tilde{\boldsymbol{t}}_j,\tilde{\boldsymbol{t}}_j)=1$.

This makes $\gamma_1(\cdot)$ a continuous Gaussian process that equals the GMRF $\tilde{\boldsymbol{\gamma}}$ at grid locations. Elsewhere $\gamma_1(\cdot)$ is the weighted mean value of the four closest lattice points. The variance of $\gamma_1(\cdot)$ is tuned to $\gamma(\cdot)$ at lattice locations (Eq. (8)), and thus the variance of $\gamma_1(\cdot)$ will be lower than that of $\gamma(\cdot)$ within cells. The addition of $\gamma_2(\cdot)$ in (10) gives the stipulated variances (though not covariances) even for locations within cells.

3.2. Prediction

This section contains an algorithm for efficient use of the suggested model when kriging large data sets.

Let $\boldsymbol{K}=(k_{ij})$ be the $n\times N$ matrix defined as $k_{ij}=k(\boldsymbol{t}_i,\tilde{\boldsymbol{t}}_j)$. Then from (7)–(13) we have

$$\boldsymbol{\Sigma}=\boldsymbol{K}\boldsymbol{Q}^{-1}\boldsymbol{K}^{\mathrm{T}}+\boldsymbol{D}. \tag{17}$$

Here $\boldsymbol{Q}=(\boldsymbol{I}_N-\boldsymbol{W})\tau^{-2}$ is the precision matrix of $\tilde{\boldsymbol{\gamma}}$ and $\boldsymbol{D}$ is a (diagonal) matrix with the variance of $\gamma_2(\boldsymbol{t})+\varepsilon(\boldsymbol{t})$ at the locations where the observations are collected. Thus, from (14) it follows that

$$\boldsymbol{D}=\sigma_\gamma^2\boldsymbol{I}_n-\operatorname{diag}(\boldsymbol{K}\tilde{\boldsymbol{\Sigma}}\boldsymbol{K}^{\mathrm{T}})+\sigma_\varepsilon^2\boldsymbol{I}_n, \tag{18}$$

where $\tilde{\mathbf{\Sigma}} = (\mathrm{Cov}(\gamma(\tilde{t}_i), \gamma(\tilde{t}_j)))_{i,j=1:N}$. In order to calculate $\boldsymbol{K}\tilde{\mathbf{\Sigma}}\boldsymbol{K}^{\mathrm{T}}$ only the diagonals corresponding to adjacent lattice locations must be known in $\tilde{\mathbf{\Sigma}}$.

For an arbitrary point $\boldsymbol{t}_0$, where we want to predict Y, define the vector of k-weights for the point, $\boldsymbol{k} = (k_1, \ldots, k_N)^{\mathrm{T}}$, where $k_i = k(\boldsymbol{t}_0, \tilde{\boldsymbol{t}}_i)$, to obtain

$$\omega = \boldsymbol{K}\boldsymbol{Q}^{-1}\boldsymbol{k}. \tag{19}$$

In all equations it is assumed that the locations where predictions are wanted do not coincide with any location with an observation used in the calculations unless such a location is also a lattice location. As the index $\boldsymbol{t}$ is continuous, this restriction does not affect the usefulness of the method.

Should approximations other than (12)–(14) of $\gamma_2(\boldsymbol{t})$ be used, Eqs. (17)–(19) would change.

Insertion of expressions (17)–(19) into (3) gives the interpolation weights for our model. While direct calculation does not give any substantial computational gain, the following method of computation will considerably ease the computational burden of calculating the kriging predictions for large data sets.

(1) Initially predict $Y(\cdot)$ on all grid points. Let therefore

$$\mathbf{\Omega} = (\mathrm{Cov}(Z(\boldsymbol{t}_i),\ Y(\tilde{\boldsymbol{t}}_j)))_{ij} \text{ be a } n \times N \text{ matrix,}$$

and $\hat{\boldsymbol{Y}} = (\hat{Y}(\tilde{\boldsymbol{t}}_1), \ldots, \hat{Y}(\tilde{\boldsymbol{t}}_N))^{\mathrm{T}}$.
Using Eqs. (2) and (3) for $\boldsymbol{t}_0 = \tilde{\boldsymbol{t}}_1, \ldots, \tilde{\boldsymbol{t}}_N$ we conclude

$$\hat{\boldsymbol{Y}} = \left(\mathbf{\Omega} + \mathbf{1}_n \left(\frac{\mathbf{1}_N^{\mathrm{T}} - \mathbf{1}_n^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{\Omega}}{\mathbf{1}_n^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{1}_n}\right)\right)^{\mathrm{T}} \mathbf{\Sigma}^{-1}\boldsymbol{Z}. \tag{20}$$

(2) At an arbitrary point $\boldsymbol{t}_0$

$$\hat{Y}(\boldsymbol{t}_0) = \boldsymbol{k}^{\mathrm{T}}\hat{\boldsymbol{Y}}, \tag{21}$$

see Appendix A for details.

From (20) and (21) we get $\hat{Y}(\boldsymbol{t}_0)$ at any point $\boldsymbol{t}_0$. It is therefore enough to seek computationally effective expressions for $\mathbf{\Sigma}^{-1}\mathbf{\Omega}$ and $\mathbf{\Sigma}^{-1}\mathbf{1}_n$.

The definition of the intermediate-scale variation gives

$$\mathbf{\Omega} = \boldsymbol{K}\boldsymbol{Q}^{-1} \tag{22}$$

that together with (17) yields

$$\begin{aligned}\mathbf{\Sigma}^{-1}\mathbf{\Omega} &= \mathbf{\Sigma}^{-1}\mathbf{\Omega}(\boldsymbol{Q} + \boldsymbol{K}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{K})(\boldsymbol{Q} + \boldsymbol{K}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{K})^{-1} \\ &= \mathbf{\Sigma}^{-1}\mathbf{\Sigma}\boldsymbol{D}^{-1}\boldsymbol{K} \cdot (\boldsymbol{Q} + \boldsymbol{K}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{K})^{-1} \\ &= \boldsymbol{D}^{-1}\boldsymbol{K}(\boldsymbol{Q} + \boldsymbol{K}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{K})^{-1}.\end{aligned} \tag{23}$$

Further, since the row-sums of $\boldsymbol{K}$ equal 1,

$$\mathbf{\Sigma}^{-1}\mathbf{1}_n = \mathbf{\Sigma}^{-1}\boldsymbol{K}\mathbf{1}_N = \mathbf{\Sigma}^{-1}\mathbf{\Omega}\boldsymbol{Q}\mathbf{1}_N. \tag{24}$$

From (23) and (24) we can thus conclude that to calculate (20), it suffices to invert a *band limited* $N \times N$ matrix $\boldsymbol{Q} + \boldsymbol{K}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{K}$. When $\boldsymbol{K}$ interpolates only grid points within the Markov neighbourhood (as for nearest neighbour or bilinear interpolation), the addition of $\boldsymbol{K}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{K}$ adds no extra bandwidth to the GMRF precision matrix $\boldsymbol{Q}$, and thus no extra computational cost.

3.3. Prediction error

To calculate the prediction error variance for the Markov approach define

$$C_r(\boldsymbol{t}_1, \boldsymbol{t}_2) = \mathrm{Cov}(r(\boldsymbol{t}_1), r(\boldsymbol{t}_2))$$

to be the covariance function of the prediction errors. Let $\tilde{\omega}_i = \mathrm{Cov}(\boldsymbol{Z}, Y(\tilde{t}_i))$ be the ith column of $\boldsymbol{\Omega}$, and generalize (4) to get

$$C_r(\tilde{\boldsymbol{t}}_i, \tilde{\boldsymbol{t}}_j) = C_\gamma(\tilde{\boldsymbol{t}}_i, \tilde{\boldsymbol{t}}_j) - \tilde{\omega}_i^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \tilde{\omega}_j + \frac{(\mathbf{1}_n^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \tilde{\omega}_i - \mathbf{1})(\mathbf{1}_n^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \tilde{\omega}_j - \mathbf{1})}{\mathbf{1}_n^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbf{1}_n}.$$

As $\mathrm{Var}(\gamma_2(\tilde{\boldsymbol{t}}_j)) = 0$, it follows that $C_\gamma(\tilde{\boldsymbol{t}}_i, \tilde{\boldsymbol{t}}_j) = C_{\gamma_1}(\tilde{\boldsymbol{t}}_i, \tilde{\boldsymbol{t}}_j) = \mathrm{Cov}(\tilde{\gamma}_i, \tilde{\gamma}_j)$. Thus after some manipulations given in Appendix B, the covariance matrix $\boldsymbol{C}_{\tilde{r}} = (C_r(\tilde{\boldsymbol{t}}_i,\ \tilde{\boldsymbol{t}}_j))_{ij}$ is given by

$$\boldsymbol{C}_{\tilde{r}} = (\boldsymbol{Q} + \boldsymbol{K}^{\mathrm{T}} \boldsymbol{D}^{-1} \boldsymbol{K})^{-1} + \frac{(\mathbf{1}_n^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\Omega} - \mathbf{1}_N^{\mathrm{T}})^{\mathrm{T}} (\mathbf{1}_n^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\Omega} - \mathbf{1}_N^{\mathrm{T}})}{\mathbf{1}_n^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbf{1}_n}. \tag{25}$$

For the conditional prediction error the last term of (25) vanishes.

At nonlattice locations use $\tilde{\boldsymbol{r}} = (r(\tilde{\boldsymbol{t}}_1), \dots, r(\tilde{\boldsymbol{t}}_N))$ and the independence assumptions (12) and (13) to obtain

$$\begin{aligned} r(\boldsymbol{t}_0) &= Y(\boldsymbol{t}_0) - \hat{Y}(\boldsymbol{t}_0) = \mu + \gamma(\boldsymbol{t}_0) - \boldsymbol{k}^{\mathrm{T}} \hat{\boldsymbol{Y}} = \mu + \gamma_1(\boldsymbol{t}_0) + \gamma_2(\boldsymbol{t}_0) - \boldsymbol{k}^{\mathrm{T}} \hat{\boldsymbol{Y}} \\ &= \boldsymbol{k}^{\mathrm{T}} \mu \mathbf{1}_N + \boldsymbol{k}^{\mathrm{T}} \tilde{\boldsymbol{\gamma}} + \gamma_2(\boldsymbol{t}_0) - \boldsymbol{k}^{\mathrm{T}} \hat{\boldsymbol{Y}} = \gamma_2(\boldsymbol{t}_0) + \boldsymbol{k}^{\mathrm{T}} (\mu \mathbf{1}_N + \tilde{\boldsymbol{\gamma}} - \hat{\boldsymbol{Y}}) \\ &= \gamma_2(\boldsymbol{t}_0) + \boldsymbol{k}^{\mathrm{T}} \tilde{\boldsymbol{r}}. \end{aligned} \tag{26}$$

Thus

$$\sigma_r^2(\boldsymbol{t}_0) = \boldsymbol{k}^{\mathrm{T}} \boldsymbol{C}_{\tilde{r}} \boldsymbol{k} + \mathrm{Var}(\gamma_2(\boldsymbol{t}_0)). \tag{27}$$

Using an interpolation scheme where only grid points within a Markov neighbourhood are used, only terms of $C(r(\tilde{\boldsymbol{t}}_i), r(\tilde{\boldsymbol{t}}_j))$ where $\tilde{\boldsymbol{t}}_i \sim \tilde{\boldsymbol{t}}_j$ must be calculated to obtain the first term of (27). To do this we use the efficient algorithm of Rue and Martino (2007) implemented in GMRFLib, where marginal variances and covariances for neighbours are calculated with $N \log(N)^2$ operations.

3.3.1. *Alternative prediction error variances for the bilinear interpolation*

The independence assumptions (12)–(13) make $\gamma_2(\boldsymbol{t})$ behave as a nugget effect. This gives a term $\mathrm{Var}(\gamma_2(\boldsymbol{t}_0))$ in the prediction error variance (27).

In many practical problems, predictions on a grid are wanted. Since at lattice locations $\mathrm{Var}(\gamma_2(\boldsymbol{t}_0)) = 0$, the additional variance term does not affect the prediction error variance as long as the GMRF is defined on that same grid. Thus, for these problems, bilinear interpolation is only used to incorporate data (that are typically not collected on a grid). Quality of predictions and prediction error variances on the grid are both better by this method than if each data point is simply assigned to the closest grid point as in nearest neighbour interpolation, see results in Section 4. However, should predictions and prediction error variances be calculated even at nonlattice locations, a deficiency with assumptions (12)–(13) is that the prediction error variance will be overestimated. This is because $\mathrm{Var}(\gamma_2(\cdot))$ is defined directly from the variance of the underlying field Y, which has larger variance than the prediction errors $\hat{Y} - Y$. As the prediction error variance at *lattice locations* is well tuned to what it would be with the full Gaussian model, it follows that the prediction error variance will be overestimated at locations *within cells*.

However, if high-quality prediction error variances are wanted at nonlattice locations, a heuristically sound alternative could be to calculate the prediction error variance as the weighted average of the prediction error variances at the grid points surrounding $\boldsymbol{t}_0$,

$$\sigma_r^2(\boldsymbol{t}_0) = \sum_{j=1}^{N} k_j \sigma_r^2(\tilde{\boldsymbol{t}}_j) = \boldsymbol{k}^{\mathrm{T}} \boldsymbol{V}_{\tilde{r}}, \tag{28}$$

where $\boldsymbol{V}_{\tilde{r}} = \mathrm{diag}(\boldsymbol{C}_{\tilde{r}})$ is a diagonal matrix containing the prediction error variances on the grid. This alternative, although quite ad hoc, stems from prediction error variances being well tuned at grid points to those of a Gaussian model, and in the same way prediction error variance being continuous, just as with a continuous Gaussian model.

For approximations with nearest neighbour interpolation where the weight $k_j = 1$ for the closest grid point $\tilde{\boldsymbol{t}}_j$, the used approximation (14) produces $\mathrm{Var}(\gamma_2(\boldsymbol{t})) = 0$ everywhere, and thus (27) and (28) are equal.

4. Simulation study

In order to study the performance of GMRF kriging, we conducted a simulation study in two parts with calculations done in Matlab and GMRFLib, a c-library of routines for GMRFs on graphs (Rue, 2001). Due to the large memory requirements of kriging with the full model, the simulated data sets are smaller than the sample size that the GMRF approach can handle. The intention with the paper is, however, to argue for the usefulness of the Markov approach when kriging a data set that is too big to evaluate with a full model. The simulations serve as a small scale study that verifies the quality of the method.

Although all Gaussian fields cannot be fitted well by a GMRF, Rue and Tjelmeland (2002) found that the Gaussian fields most commonly used in kriging could. As an example of such fields, we have used a stationary isotropic Gaussian field with an exponential covariance function both for data generation and estimation in our study.

Following (1), we simulated several samples **z** with different values of the variables $(n, r, \sigma_\gamma^2, \sigma_\varepsilon^2)$. The simulations were done in R (R Development Core Team, 2003) with the routine `GaussRF` in package RandomFields. For each realization **z**, new positions were simulated from a uniform density. For the intermediate-scale variation $\gamma(\cdot)$ we used a stationary Gaussian field with exponential covariance function $C(|\boldsymbol{h}|) = \mathrm{Cov}(\gamma(\boldsymbol{t}), \gamma(\boldsymbol{t}+\boldsymbol{h})) = \sigma_\gamma^2 \exp(-3|\boldsymbol{h}|/r)$ where r is the practical range, i.e. $C(r) = 0.05 \cdot C(0)$. The nugget term $\varepsilon(\cdot)$ was Gaussian white noise with variance σ_ε^2. The sample was divided at random into an estimation set of size n and a validation set of size M. Thus both the estimation set and the validation set had nonlattice positions uniformly spread in the area.

As an adjustment between good fit and reasonable computational cost related to the GMRF as reported in Rue and Tjelmeland (2002), we choose a 5 × 5 Markov neighbourhood for the GMRF approximation.

For each sample **z** we performed kriging with four different models:

(1) Using the full model, i.e. a Gaussian random field with exponential covariance function and added measurement error.
(2) GMRF with bilinear interpolation and added measurement error, with $\gamma_2(\boldsymbol{t})$ as in (12)–(14).
(3) GMRF with bilinear interpolation and added measurement error, with $\gamma_2(\boldsymbol{t}) \equiv 0$, i.e. without compensation for the lower variance within grid cells.
(4) GMRF with nearest neighbour interpolation and added measurement error.

For the GMRF approximations above, we used a GMRF with 5 × 5 neighbourhood on a $\sqrt{N} \times \sqrt{N}$-grid, i.e. filling a square with side length $\Delta = \delta\sqrt{N}$. The precision parameters of the inverted correlation matrix were fitted in Rue and Tjelmeland (2002) to match an exponential correlation with range r. The precision matrix was then scaled with σ_γ^{-2} to fulfil (8). (As the correlation structure is fitted, the variances will not be exactly equal, although in practise the difference is negligible.) For the nugget, the known variance σ_ε^2 was used. In order to use pre-calculated precision parameters, the fitted parameters must not depend on the torus, at which they are calculated. In accordance with the suggestions of Rue and Held (2005, Chapter 5) we accomplished this by ensuring that the grid side was at least 6 times the range. Further, in order to avoid edge effects, the grid was extended to cover an area larger than the area where observations were taken and predictions were wanted. The frame used was 0.5 times the range at all four sides of the lattice, following Rue and Held (2005).

To measure the performance of the kriging predictors, the *prediction error sum of squares* (PRESS)

$$\mathrm{PRESS} = \sum_{i=1}^{M} (y_i - \hat{y}_i)^2 \tag{29}$$

was calculated for all approaches. Here y_i is the simulated value of the Y-field at location number i in the validation set and $\hat{y}_i$ is the prediction of Y in the same point calculated by (2)–(3) and (20)–(21) for $\mathrm{PRESS}_{\mathrm{exponential}}$ and $\mathrm{PRESS}_{\mathrm{Markov}}$, respectively.

In Section 4.1 we used the known covariance parameters when computing the kriging weights, whereas in Section 4.2 we estimated the covariance parameters from the estimation set and then used the plug-in of the estimated covariances in the kriging computations.

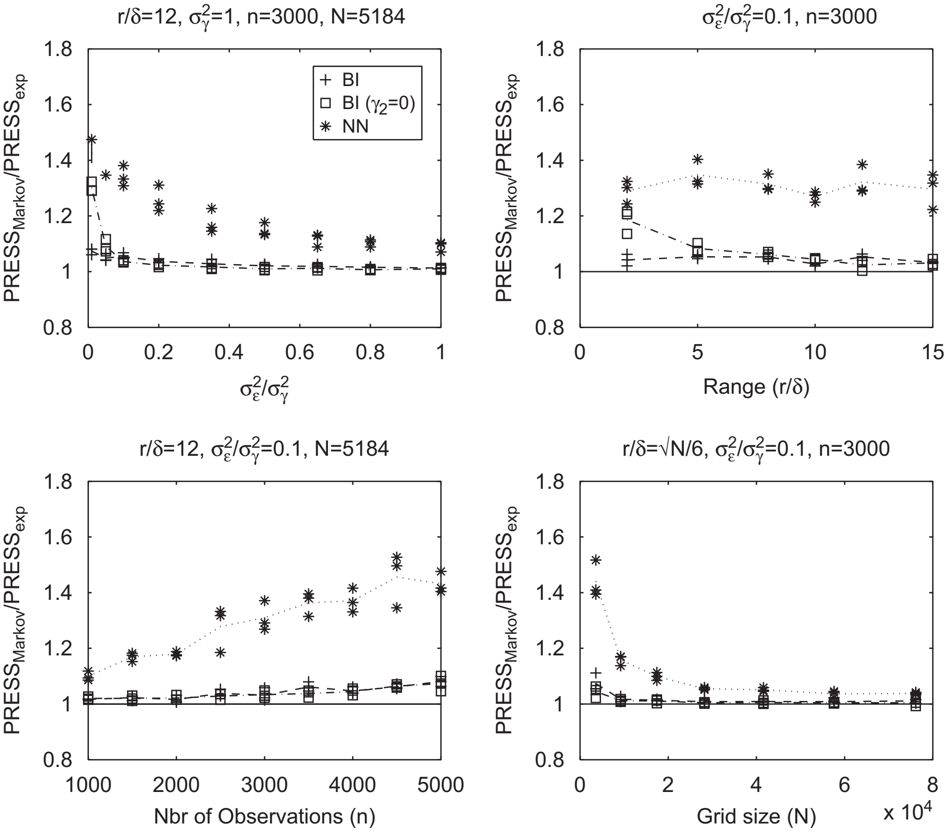

Fig. 3. Quotient of prediction error sum of squares, $\frac{\text{PRESS}_{\text{Markov}}}{\text{PRESS}_{\text{exponential}}}$, for different parameter values and number of observations, etc. The known covariance parameter values were used for the exponential Gaussian field, while fitted precision parameters were used for the GMRF. There was $M = 2000$ observations in the validation set. Each run was iterated three times. Individual results are shown for (i) GMRF with bilinear interpolation and $\gamma_2(\cdot)$ as in (12)–(14) [BI], (ii) GMRF with bilinear interpolation and $\gamma_2(\cdot) \equiv 0$ [BI ($\gamma_2 = 0$)] and (iii) GMRF with nearest neighbour interpolation [NN]. The respective means are indicated by a line. In each subplot, the parameters displayed at the x-axis were varied while the others were kept constant at values displayed above each graph.

4.1. Known covariance parameters

To compare GMRF kriging to kriging with the full model, the quotient $\text{PRESS}_{\text{Markov}}/\text{PRESS}_{\text{exponential}}$ was calculated both for bilinear interpolation (with and without nonzero γ_2) and for nearest neighbour interpolation, see Fig. 3. Although the full Gaussian approach is clearly favoured as it uses *the correct model with known parameters*, whereas the Markov approach only uses an *approximation of the correct model*, the Markov model manages well. The better quality of the bilinear interpolation is appreciable when compared to the nearest neighbour interpolation, except in the case of very large grids. The difference when using nonzero γ_2 is almost negligible to when $\gamma_2(\cdot) \equiv 0$.

We calculated conditional prediction error variances for both exponential and Markov models following (4) and (25)–(27), respectively, and the corresponding 95%-prediction intervals (5). Fig. 4 shows that the coverage for the GMRF approximation with bilinear interpolation and nonzero $\gamma_2(\cdot)$ is appreciably better than for nearest neighbour interpolation or for bilinear interpolation with $\gamma_2(\cdot) \equiv 0$. The coverage for the full Gaussian model is displayed to show the variation between different simulations. The coverage of the GMRF approximation with bilinear interpolation is close to the prescribed value (95%) for most of the parameter values (although somewhat overestimated, as discussed in Section 3.3.1). Although the prediction quality (Fig. 3) is similar regardless of whether $\gamma_2(\cdot) = 0$ or not, the deficit of $\gamma_2(\cdot) = 0$ is apparent here. This is natural as bilinear interpolation with $\gamma_2(\cdot) = 0$ by definition produces lower variance within grid cells. $\gamma_2(\cdot)$ was introduced to compensate for this.

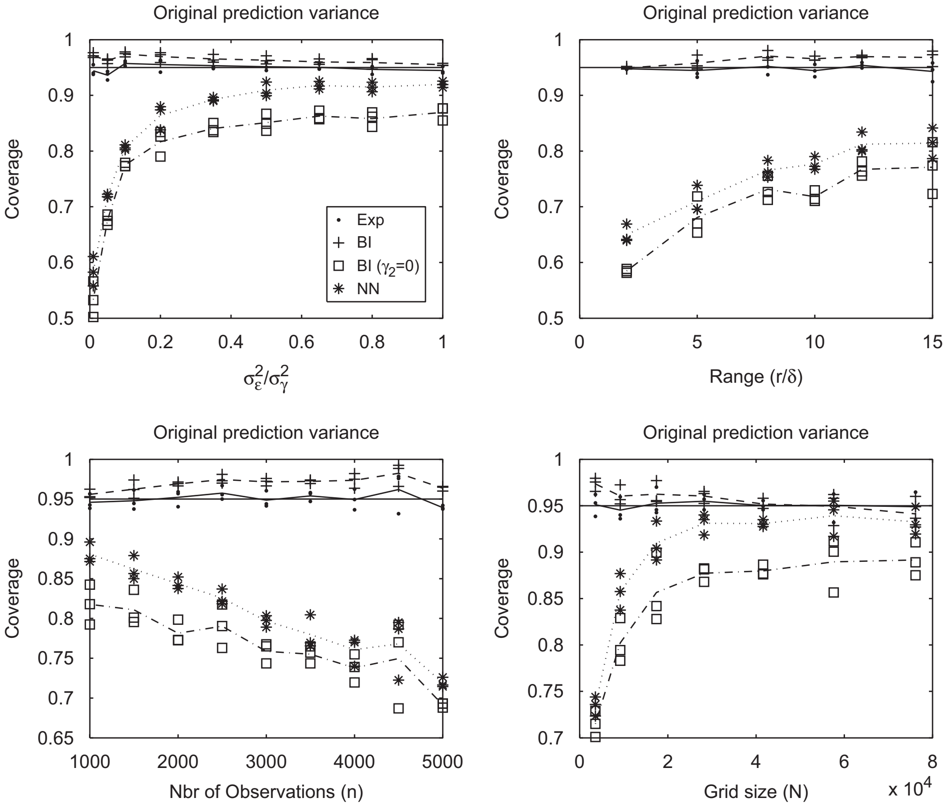

Fig. 4. Coverage of 95%-prediction intervals for kriging with Gaussian and GMRF models. For the bilinear interpolation, the original prediction error variance (27) was used. For each subplot, the same parameter values as in Fig. 3 were used. Each run was iterated three times, the individual results are shown, with their respective means indicated by a line. Legends are explained in Fig. 3, except [Exponential] which amounts to kriging with a Gaussian field with exponential covariance function.

Fig. 5 shows the coverage when the prediction error variance for the bilinear interpolation is calculated as in Section 3.3.1. Here the coverage of the GMRF approximation with bilinear interpolation and nonzero $\gamma_2(\cdot)$ is in quality almost indistinguishable from the coverage using the full Gaussian model. The model with nonzero $\gamma_2(\cdot)$ produces better prediction error variances than with $\gamma_2(\cdot) = 0$, even though the prediction error variance is here the weighted mean over the prediction error variances at the grid points (without addition of $\mathrm{Var}(\gamma_2(\cdot))$). For the more commonly used nearest neighbour interpolation, the results are not of as high a quality as for the bilinear interpolation. In fact, in our study, in order to get coverage reasonably close to 95%, such dense grids are required that the computational gain of the Markov model is lost (Fig. 8).

4.2. *Estimated covariance parameters*

When kriging real data it is rare to know the model parameters, and consequently they must be estimated from the data set. Plug-in of estimated parameters into the kriging equations is a common approach, which we use here to test the performance of the Markov model kriging. To do this, the same data sets as in the previous section are used for kriging, but here the covariance parameters are estimated from data. The performance of the two models regarding prediction intervals is calculated and assessed just as if parameters were known. It should, however, be kept in mind that the prediction error variance is then underestimated (both for the full and the Markov model), which will cause prediction intervals that are too narrow. This can be rectified by bootstrap calibration (Sjöstedt-de Luna and Young, 2003).

Regarding parameter estimation, Cressie (1993, pp. 69–73) argues that variogram estimation is preferable to covariance estimation because of its lower bias and because the variogram is defined for some processes that are not second

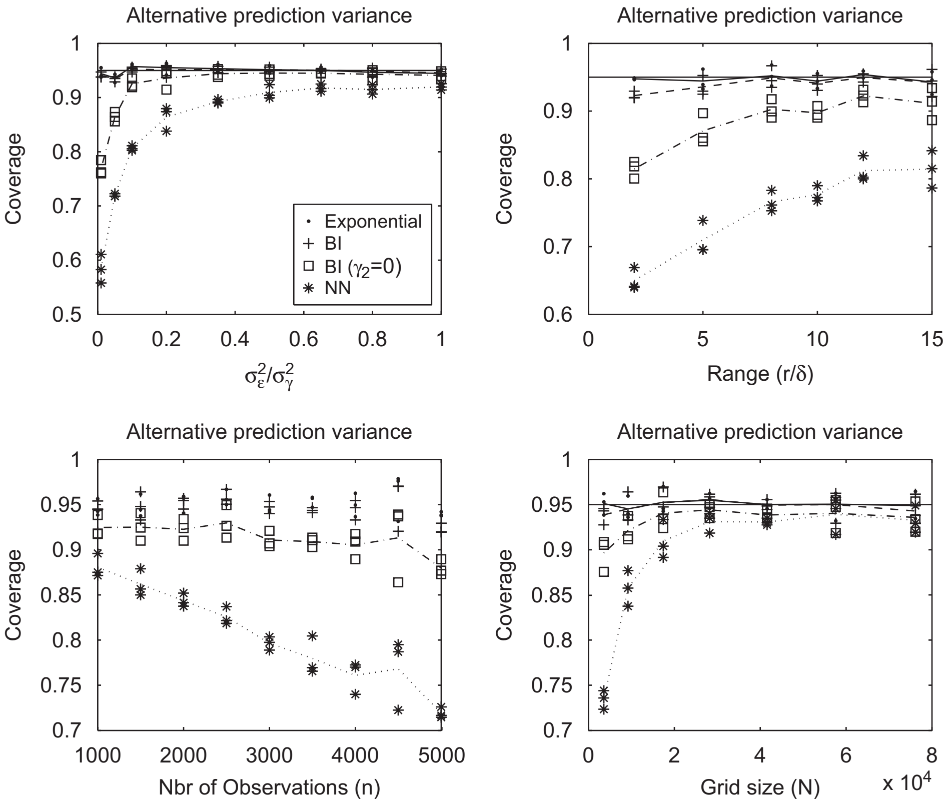

Fig. 5. Coverage of 95%-prediction intervals for kriging with Gaussian and GMRF models. For the bilinear interpolation the alternative prediction error variance (28) was used. In each subplot, the same parameter values as in Fig. 3 were used. Each run was iterated three times, the individual results are shown, with their respective means indicated by a line. Legends are explained in Fig. 3, except [Exponential] which amounts to kriging with a Gaussian field with exponential covariance function.

order stationary. When, as in the models we investigate, the covariance is defined there is a simple relationship between covariance and variogram, such that the variogram can be used for estimation while the covariance formulation is used for the kriging equations as in Eq. (3).

From the empirical variogram calculated for points in the estimation set separated less than half the maximal distance, we estimated the parameters r and σ_γ^2 for the exponential model, and the nugget variance σ_ε^2, by means of the package geoR (Ribeiro and Diggle, 2001). Although the GMRF-parameters could also be estimated directly from data, we used a simpler solution where the GMRF-parameters where chosen to fit an exponential model (as we did in Section 4.1) but now with estimated range parameter $\hat{r}$. This issue is discussed further in Section 6.

The comparison of $\text{PRESS}_{\text{Markov}}$ and $\text{PRESS}_{\text{exponential}}$ when parameters are estimated from data (see Fig. 6) was similar to the results achieved with known parameters, in Section 4.1. Naturally, a larger variability in the results occurs when parameters are fitted from data. The GMRF kriging results were good compared to kriging with the full model, and the better quality of bilinear interpolation (compared to nearest neighbour interpolation) was still appreciable.

Regarding coverage with estimated parameters (Fig. 7), the quality of the GMRF-approximation is similar to that of the full Gaussian model. The higher quality of the bilinear, when compared to nearest neighbour interpolation, is still perceivable although not as marked as for known parameter values.

4.3. Computation times

The computations were performed on a 2.7 GHz Intel Xeon with 2 GB RAM. Matlab's solvers are well optimized for dense linear systems. Therefore, all calculations for the full Gaussian field were done in Matlab. For the GMRF model

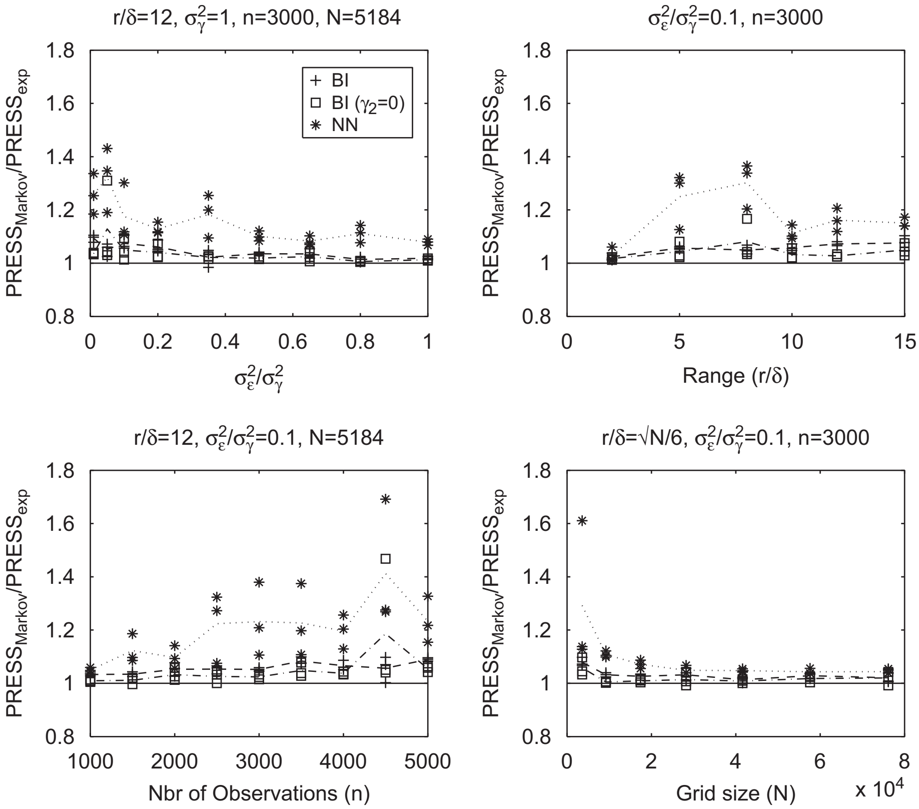

Fig. 6. Quotient of prediction error sum of squares, $\frac{\text{PRESS}_{\text{Markov}}}{\text{PRESS}_{\text{exponential}}}$, for different parameter values and number of observations, etc. For the exponential Gaussian field, estimated covariance parameter values were used while for the GMRF the precision parameters were chosen to fit an exponential with the above estimated range. PRESS is calculated from the kriging result in $M = 2000$ validation points distributed in the area according to a uniform distribution. Legends are explained in Fig. 3.

the sparse systems were setup and solved in a c-routine utilizing the library GMRFLib (Rue, 2001). Initial work (such as reading the data and calculating the interpolation matrix) and plots of results were done in Matlab. All simulations and parameter estimations were done in R, and the package R.matlab (Bengtsson, 2004) was utilized to automatize the procedures.

Computation times for kriging with the full model are in Fig. 8 compared to the times yielded by exploiting a Markov approximation (for the same samples as in Fig. 3). The times include the calculation of prediction error variances at all predictive locations. This, the most time consuming part of the calculations for the GMRF approximation, was not reported in Furrer et al. (2006). The computation demand is of order n^3 for kriging with the full model, while for the GMRF-approximation, it is of order $N^{3/2}$ for the predictions and $N \log(N)^2$ for the prediction error variances. Thus, even grids with many more nodes than the number of observations are useful. The computation time for the Markov approach does only increase slightly with n. If the number of grid points is chosen so as to be proportional to the number of observational points, i.e. $N \propto n$, the computational complexity for predictions is still only of order $n^{3/2}$ (as compared to n^3 for the full system).

5. Application to soil data

We apply the methods from the preceding sections to a data set of elemental composition of forest soil. The data consist of 9606 data points of magnesium oxide concentrations measured by the soil geochemical mapping programme of the Geological Survey of Sweden. Collection was done at irregular forest locations, with roughly 1 observation

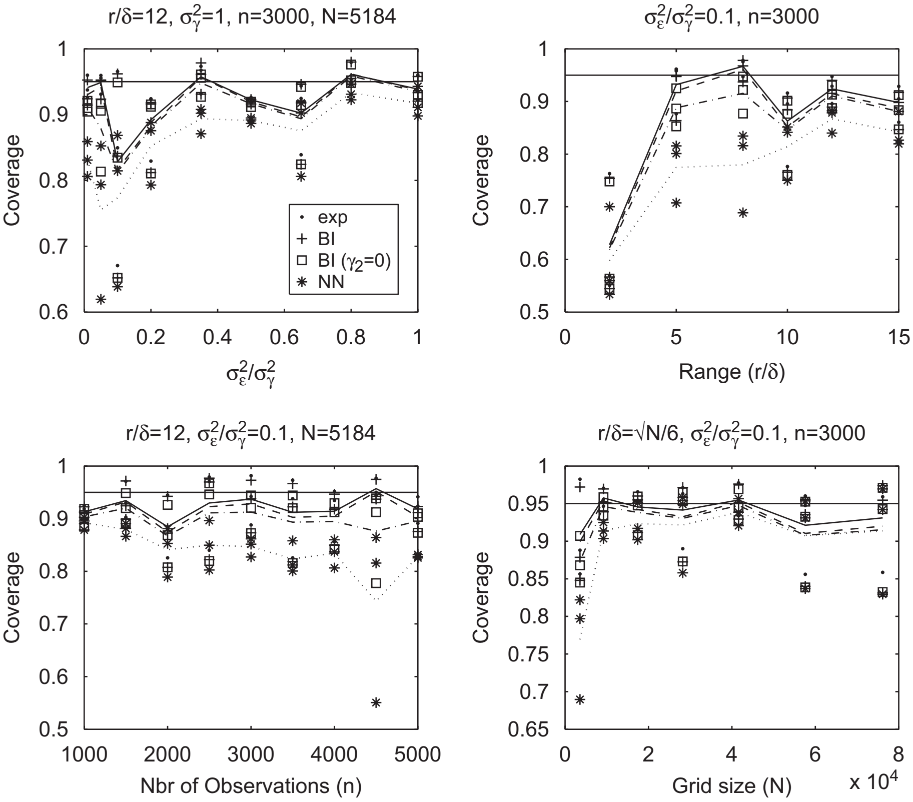

Fig. 7. Coverage of 95%-prediction intervals for kriging with Gaussian and GMRF models. For the bilinear interpolation, the alternative prediction error variance (28) was used. The same parameter values as in Fig. 3 were used. Each run was iterated three times, the individual results are shown, with their respective means indicated by a line.

per 6–7 km^2, and covered all of southern Sweden, from the southern coast up to about the latitude of Stockholm. (The coastal line is apparent in the sampling locations displayed in Fig. 10, upper left.) The total size of the area is about 300 × 400 km. The samples were taken at the C-horizon, i.e. at approximately 1 m's depth, and the elemental composition was analysed by X-ray fluorescence spectroscopy (XRF) of the fine fraction of the moraine. More information regarding the data is found in Akselsson et al. (2004).

A preliminary analysis of data showed that after taking the logarithm, the data could be fitted well with a Gaussian distribution. After subtraction of a linear trend estimated by least squares, we analysed residuals of log(MgO-content). Inspection of omni-directional empirical variograms showed no apparent anisotropy. We will illustrate our methods by performing kriging of de-trended logarithm of MgO-content, under the assumption that detrended data are stationary and isotropic. To do this, the empirical variogram calculated using every fifth data point with distances up to half the maximal distance was fitted to an exponential variogram (Fig. 9), as described in Section 4.2. The fitted parameters were range $r = 1.19 \times 10^5$ m, partial sill $\sigma_\gamma^2 = 0.138$ and nugget $\sigma_\varepsilon^2 = 0.087$.

We calculated kriging predictions on a fine grid of size 300 × 300 using GMRF-kriging with bilinear interpolation. To avoid edge effects, as discussed in Section 4.1, the grid was extended outside the measurement area.

The observations, predictions and prediction error variances are displayed in Fig. 10. The upper right subplot displays predictions for all grid locations, and thus indicates the extension of the grid outside the measurement area. The lower subplots display predictions and prediction error variances in the area of interest. The excluded gridpoints reside outside the measurement area, most of them in the sea. Calculation of the $300 \times 300 = 90,000$ predictions and prediction error variances took about 100 s, on a 2.7 GHz Intel Xeon with 2 GB RAM. If predictions and prediction error variances are wanted at nongrid locations this is done at (essentially) no extra cost by following the methods presented in Sections 3.2 and 3.3 or 3.3.1.

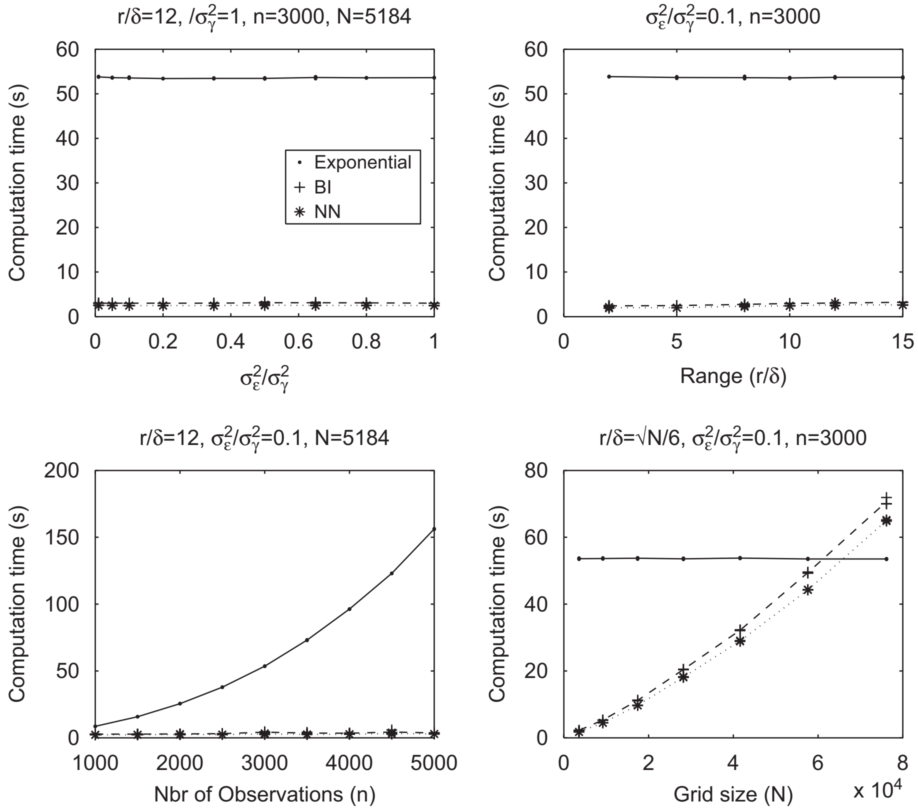

Fig. 8. Computation time for kriging with Gaussian and GMRF model for different parameter values, cf. Fig. 3. $M = 2000$ observations in the validation set. Each run was iterated three times, the individual results are shown, with their respective means indicated by a line.

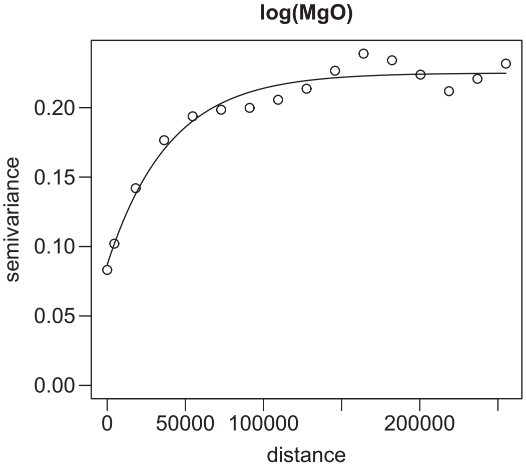

Fig. 9. Empirical variogram and fitted exponential variogram for detrended log(MgO), distances measured in metres. The fitted parameters were: range $r = 1.19 \times 10^5$ m, partial sill $\sigma_\gamma^2 = 0.138$ and nugget $\sigma_\varepsilon^2 = 0.087$.

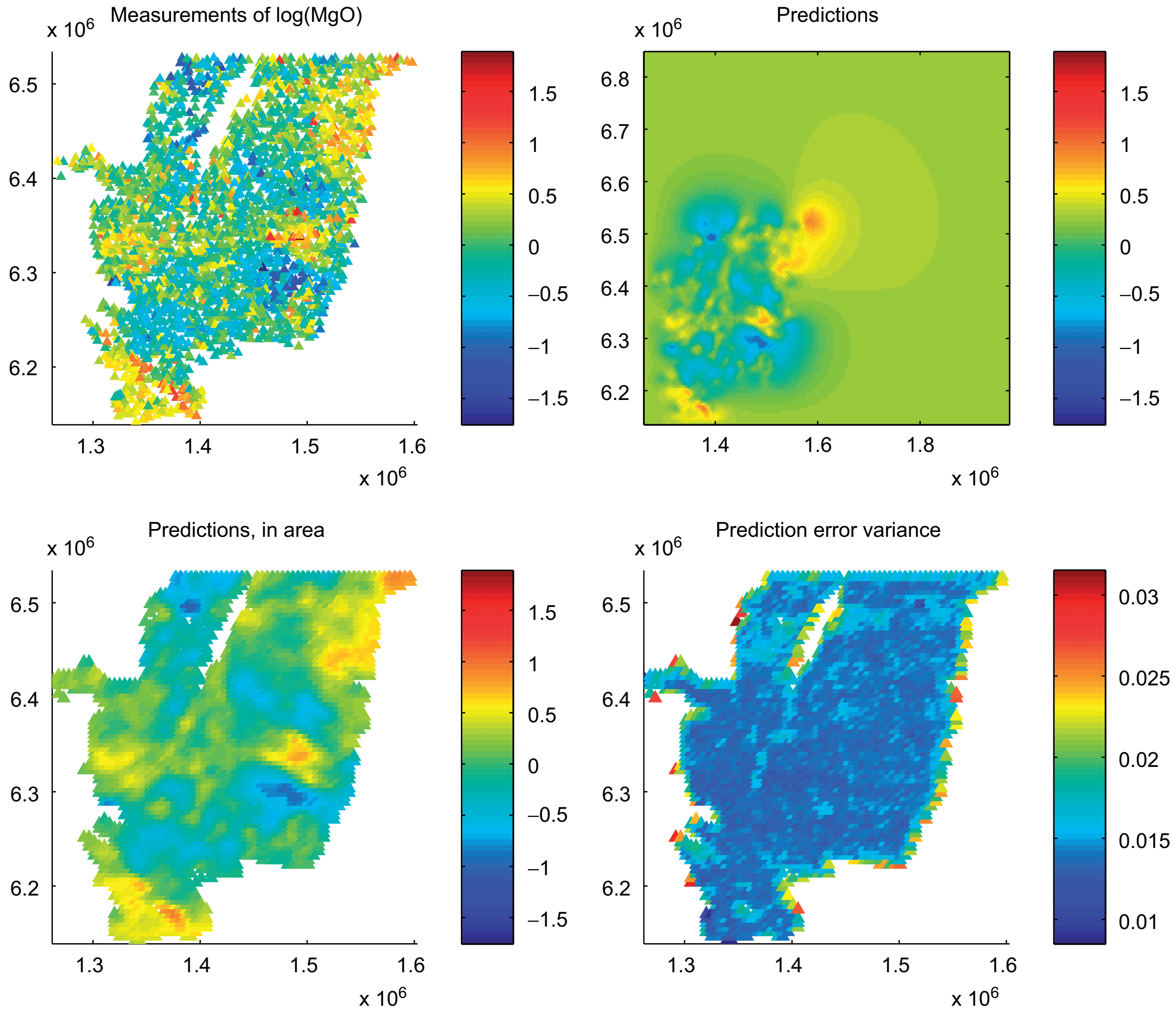

Fig. 10. Observations, i.e. detrended log(MgO), and the predictions and prediction error variance calculated on a 300 × 300 grid covering the extended area. The lower subplots display predictions and prediction error variances in southern Sweden, i.e. the extension of grid is not displayed. Coordinates are in the Swedish grid (RT90) with units 1 m. (True white color in the upper left plot indicates locations where no measurements were taken (mainly at sea and in lakes). In order to decrease image size, images were down sampled, i.e. not all observations or predictions are displayed.)

6. Discussion

In this article we have suggested using GMRF approximations of Gaussian fields as a computationally effective means of kriging. The fundamental course of the analysis is as follows: establish a reasonably dense grid, approximate the Gaussian field with a GMRF on this grid, and then obtain kriging predictions on the grid. Predictions between grid locations are then found through interpolation of the kriging predictions at grid locations. By fitting a stationary field, which is approximated by a GMRF on a regular grid, we obtain the computational benefits of a Markov field, but avoid the disadvantages (pointed out by Wall, 2004) of conditional autoregressions for irregularly spaced data. Further note, that although we use an approximate field to obtain sparse precision matrices, we do not change the formula of the kriging weights (3). Thus, the predictions will fulfil all properties that are general for kriging. These include, for example, kriging to the mean and screening of distant observations when there are closer observations in the same direction. Moreover, since we assume nonlattice locations, and since the basis of the analysis is a Gaussian field on continuous coordinates, the estimation of the nugget and the covariance parameters can utilize the important information extracted from closely positioned observations. However, in the actual fitting of the GMRF to the Gaussian field, the respective correlations are compared only at distances present in the lattice. For distances shorter than one gridcell, the interpolation determines the correlation. The approximate assumptions on $\gamma_2(\boldsymbol{t})$ are chosen as a balance between good approximation of the Gaussian field and computational efficiency. That $\gamma_2(\boldsymbol{t})$ is modelled as white noise gives computational efficiency, but, as argued in Section 3.3.1, the accompanying addition of $\mathrm{Var}(\gamma_2(\boldsymbol{t}_0))$ produces overestimated prediction error variance at locations within cells. In many practical problems, predictions are wanted on a grid, but data are collected at arbitrary locations. The problem with overestimation of the prediction error variance is then eliminated if one chooses the grid on which predictions are wanted, as the lattice for the GMRF. Bilinear interpolation is then employed to better utilize the data collected at nonlattice locations. This yields predictions and prediction error variances of higher quality than with nearest neighbour interpolation.

The assumptions also include formulation of $\mathrm{Var}(\gamma_2(\boldsymbol{t}))$, where two possibilities were mentioned in Section 3. To calculate the GMRF covariances in (15) the precision matrix, $\boldsymbol{Q}$, must be inverted, while for the covariances of the Gaussian field in (14) no such (computationally demanding) matrix inversion is needed. Further. $\mathrm{Var}(\gamma_2(\boldsymbol{t}))$ as in (15) runs the risk of being negative. Another possibility is to define $\mathrm{Var}(\gamma_2)$ directly from the GMRF, i.e.

$$\mathrm{Var}(\gamma_2(\boldsymbol{t})) = \sum_j k_j \mathrm{Var}(\gamma_1(\tilde{\boldsymbol{t}}_j)) - \sum_{ij} k_i k_j \,\mathrm{Cov}(\gamma_1(\tilde{\boldsymbol{t}}_i), \gamma_1(\tilde{\boldsymbol{t}}_j)), \tag{30}$$

where $k_j = k(\boldsymbol{t}, \boldsymbol{t}_j)$. Also here $\boldsymbol{Q}$ must be inverted to calculate $\mathrm{Var}(\gamma_2(\boldsymbol{t}))$.

Using nearest neighbour interpolation both (14) and (30) give $\mathrm{Var}(\gamma_2(\boldsymbol{t})) = 0$ and thus $\gamma_2(\boldsymbol{t}) \equiv 0$ everywhere.

In a small unpublished kriging simulation study with $\mathrm{Var}(\gamma_2(\boldsymbol{t}))$ as in (14), (15) and (30), respectively, the difference in PRESS and coverage was found to be marginal. Regarding the GMRF not as a first hand model, but rather as a computationally effective approximation of a Gaussian field, the used definition (14) is natural, and fast and easy to calculate.

When unknown parameters were used, the GMRF-parameters were fitted in a two-step procedure: (1) Fit the parameters of a Gaussian field with nugget and exponential covariance function. (2) Fit the parameters of the GMRF to match the covariance function in (1). In the second step Rue and Tjelmeland (2002) have precalculated GMRF parameters fitted to the most common covariance functions, ensuring that the GMRF has a valid covariance matrix. As these fitted parameters are available in GMRFLib, the implementation is easy. Further, the definition (14) of $\gamma_2(\boldsymbol{t})$ requires that for bilinear interpolation the parameters of the full Gaussian model are estimated even for GMRF kriging. For (15), only the variance of $\gamma(\boldsymbol{t})$ is needed, while for (30) only GMRF parameters are used. Although the nugget and exponential covariance function are always fitted using all distances present among the observations (thus also close pairs), it is important that the grid is dense enough so that even correlations on small distances are used when fitting the GMRF to the Gaussian field. However, as the results are surprisingly good even for quite small N (see Fig. 3 bottom right), the bilinear interpolation seems to capture the correlation structure well enough to produce accurate results. The drawback of the two-step fitting procedure is that the GMRF could theoretically get a better fit to data if it was fitted directly, and thus the Markov model is somewhat handicapped in the simulation study. Still we believe that the difference is minimal.

Although we used ordinary kriging in this article to exemplify the use of GMRFs for kriging, the method is not restricted to a setting in which the mean is constant. In universal kriging, the mean is an unknown linear combination of known functions, i.e. $Z(t) = \boldsymbol{X\beta} + \gamma(\boldsymbol{t}) + \varepsilon(t)$ where $\boldsymbol{X} = (f_j(\boldsymbol{t}_i))_{i,j}$ is an $n \times p$-matrix of known quantities. Since the mean value specification can take care of large scale variability in the area and the related departure from stationarity, a global model can more often be fitted than if using a model with constant mean.

The predictor at point $\boldsymbol{t}_0$ ($Z(\boldsymbol{t}_0) = \boldsymbol{f}_0^{\mathrm{T}}\boldsymbol{\beta} + \gamma(\boldsymbol{t}_0) + \varepsilon(\boldsymbol{t}_0)$) has kriging weights (see e.g. Cressie, 1993, p. 154)

$$\boldsymbol{a} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\omega} + \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{f}_0 - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\omega})),$$

where $\boldsymbol{f}_0 = (f_1(\boldsymbol{t}_0), \ldots, f_p(\boldsymbol{t}_0))^{\mathrm{T}}$. The unconditional prediction error variance is then

$$\sigma_r^2(\boldsymbol{t}_0) = \sigma_\gamma^2(\boldsymbol{t}_0) - \boldsymbol{\omega}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\omega} + (\boldsymbol{f}_0 - \boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\omega})'(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{f}_0 - \boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\omega}).$$

Thus, to accomplish fast calculations with the Markov approach we must now find a good expression for $\boldsymbol{\Sigma}^{-1}\boldsymbol{\Omega}$ and $\boldsymbol{\Sigma}^{-1}\boldsymbol{X}$ when predicting simultaneously on the grid on which the GMRF is defined. However, $\boldsymbol{\Sigma}^{-1}\boldsymbol{\Omega}$ is still calculated as in (23). In analogy with (24) $\boldsymbol{\Sigma}^{-1}\boldsymbol{X}$ simplify if $\boldsymbol{K}\boldsymbol{X}_N = \boldsymbol{X}$ to

$$\boldsymbol{\Sigma}^{-1}\boldsymbol{X} = \boldsymbol{\Sigma}^{-1}\boldsymbol{K}\boldsymbol{X}_N = \boldsymbol{\Sigma}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}\boldsymbol{X}_N,$$

where $\boldsymbol{X}_N$ is an $N \times p$-matrix $(f_j(\tilde{\boldsymbol{t}}_i))_{ij}$.

The interpolated Markov model is thus suitable for fast universal kriging (at least) as long as $\boldsymbol{K}\boldsymbol{X}_N = \boldsymbol{X}$. This is true for $\boldsymbol{K}$ defined in Sections 3.1.1 and 3.1.2, not only for $\boldsymbol{X} = \mathbf{1}$ (ordinary kriging), but also when a linear trend in two dimensions is assumed such that $\boldsymbol{X}_{i\cdot} = (1\ t_{i1}\ t_{i2})$ where $\boldsymbol{t}_i = (t_{i1}\ t_{i2})$, or when a bilinear term is added such that $\boldsymbol{X}_{i\cdot} = (1\ t_{i1}\ t_{i2}\ t_{i1}t_{i2})$. For some other definitions of $\boldsymbol{X}$ it will be possible to ensure $\boldsymbol{K}\boldsymbol{X}_N = \boldsymbol{X}$ by modifying the choice of interpolation matrix $\boldsymbol{K}$.

7. Conclusions

We argue that for kriging applications with large data sets where global features are of interest, a GMRF model approximation is useful in reducing computational burden and memory requirements. The resulting predictor is an exact kriging predictor, though it be calculated from an approximated covariance function. The resulting predictions will thus fulfil all general kriging properties, such as kriging to the mean. The simulations in which data from a Gaussian process with exponential covariance function are kriged with a GMRF model show that the quality of the predictions is comparable to the results of kriging with the full model. The intended field of application is, however, for data sets that are too large to handle with the full model. The largest reduction in computational demand is found for applications in which only kriging predictions are needed. But even when prediction error variances are calculated, the reduction in computational demand is considerable.

We also argue that when a GMRF is used to approximate a Gaussian field with continuous spatial index, a bilinear interpolation within grid cells produces predictions and prediction error variances with appreciably better quality. This is accomplished at the same computational cost, as the more commonly used nearest neighbour interpolation.

Acknowledgements

The authors would like to thank Håvard Rue for helpful discussions and for sharing the precision parameters for GMRFs fitted to exponential covariance functions. Two anonymous referees helped improve the final version of the manuscript; their comments are most appreciated.

Linda Werner Hartman's work has been funded in part by the Swedish Foundation for Strategic Research (SSF) project "Spatial Statistics and Image Analysis in Environment and Medicine, A3 02:125". Ola Hössjer's work has been supported by the Swedish Research Council, contract 626-2002-6286.

Appendix A. Motivation of Eq. (21)

Let $\mathbf{a}$ be the vector of weights (Eq. (3)) for prediction at location $\mathbf{t}$ and $\tilde{\mathbf{a}}_i$ the corresponding vector at (grid) location $\tilde{\mathbf{t}}_i$ obtained by $\boldsymbol{\omega} \leftarrow \tilde{\boldsymbol{\omega}}_i = \mathrm{Cov}(\boldsymbol{Z}, Y(\tilde{\boldsymbol{t}}_i))$. It suffices to show that

$$\mathbf{a} = \sum_{i=1}^{N} k_i \tilde{\mathbf{a}}_i. \tag{A.1}$$

This follows immediately from (3), (12), and that $\sum_1^N k_i = 1$ and $\sum_1^N k_i \tilde{\boldsymbol{\omega}}_i = \boldsymbol{\omega}$ (as long as locations where predictions are wanted do not coincide with an observation used in the calculation).

Appendix B. Motivation of Eq. (25)

At lattice locations $\mathrm{Var}(\gamma_2(\cdot)) = 0$. Thus, the covariance matrix of the prediction errors on the lattice, $\boldsymbol{C}_{\tilde{r}} = (\mathrm{Cov}(r(\tilde{\boldsymbol{t}}_i),\ r(\tilde{\boldsymbol{t}}_j))_{ij}$, from Section 3.3 is

$$\boldsymbol{C}_{\tilde{r}} = \boldsymbol{Q}^{-1} - \boldsymbol{\Omega}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Omega} + \frac{(\mathbf{1}_n^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Omega} - \mathbf{1}_N^{\mathrm{T}})^{\mathrm{T}}(\mathbf{1}_n^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Omega} - \mathbf{1}_N^{\mathrm{T}})}{\mathbf{1}_n^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{1}_n}.$$

Insertion of expressions (22) and (23) gives us Eq. (25):

$$\begin{aligned}\boldsymbol{Q}^{-1}-\boldsymbol{\Omega}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Omega} &= (\boldsymbol{Q}^{-1}-\boldsymbol{Q}^{-1}\boldsymbol{K}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{K}(\boldsymbol{Q}+\boldsymbol{K}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{K})^{-1})\\ &= \boldsymbol{Q}^{-1}(\boldsymbol{I}_N-\boldsymbol{K}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{K}(\boldsymbol{Q}+\boldsymbol{K}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{K})^{-1})\\ &= \boldsymbol{Q}^{-1}((\boldsymbol{Q}+\boldsymbol{K}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{K})-\boldsymbol{K}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{K})(\boldsymbol{Q}+\boldsymbol{K}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{K})^{-1}\\ &= \boldsymbol{Q}^{-1}\boldsymbol{Q}(\boldsymbol{Q}+\boldsymbol{K}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{K})^{-1}\\ &= (\boldsymbol{Q}+\boldsymbol{K}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{K})^{-1}.\end{aligned}$$

Appendix C. Supplementary data

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csda.2007.09.018.

References

Akselsson, C., Holmqvist, J., Alveteg, M., Kurz, D., Sverdrup, H., 2004. Scaling and mapping regional calculations of soil chemical weathering rates in Sweden. Water, Air, and Soil Pollution: Focus 4 (2–3), 671–681.

Barry, R.P., Pace, R.K., 1997. Kriging with large data sets using sparse matrix techniques. Comm. Statist. Simulation Comput. 26, 619–629.

Bengtsson, H., 2004. R.matlab—local and remote Matlab connectivity in R. Department of Mathematical Statistics, Centre for Mathematical Sciences, Lund University, Sweden, 2004, published online: ⟨http://www.maths.lth.se/help/R/R.matlab/⟩.

Cressie, N.A.C., 1993. Statistics for Spatial Data. Wiley, New York.

Diggle, P., Lophaven, S., 2006. Bayesian geostatistical design. Scand. J. Statist. 33 (1), 53–64.

Furrer, R., Genton, M.G., Nychka, D., 2006. Covariance tapering for interpolation of large spatial datasets. J. Comput. Graph. Statist. 15 (3), 502.

Goovaerts, P., 1997. Geostatistics for Natural Resources Evaluation. Oxford University Press, New York.

Isaaks, E., Srivastava, R., 1989. An Introduction to Applied Geostatistics. Oxford University Press, New York.

Kammann, E.E., Wand, M.P., 2003. Geoadditive models. J. Roy. Statist. Soc. Ser. C 52 (1), 1–18.

Moyeed, R.A., Papritz, A., 2002. An empirical comparison of kriging methods for nonlinear spatial point prediction. Math. Geol. 34 (4), 365–386.

Nychka, D., Saltzman, N., 1998. Design of air quality monitoring networks. In: Lecture Notes in Statistics, vol. 132. Springer, Berlin, pp. 51–76.

R Development Core Team, 2003. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Ribeiro Jr., P., Diggle, P., 2001. geoR: a package for geostatistical analysis. R-NEWS 1 (2), 15–18.

Rue, H., 2001. Fast sampling of Gaussian Markov random fields. J. Roy. Statist. Soc. Ser. B 63 (2), 325–338.

Rue, H., Held, L., 2005. Gaussian Markov Random Fields: Theory and Applications. Chapman & Hall, London.

Rue, H., Martino, S., 2007. Approximate Bayesian inference for hierarchical Gaussian Markov random field models. J. Statist. Plann. Inference 137 (10), 3177–3192.

Rue, H., Tjelmeland, H., 2002. Fitting Gaussian Markov random fields to Gaussian fields. Scand. J. Statist. 29 (1), 31–49.

Sjöstedt-de Luna, S., Young, A., 2003. The bootstrap and kriging prediction intervals. Scand. J. Statist. 30 (1), 175–192.

Wall, M., 2004. A close look at the spatial structure implied by the CAR and SAR models. J. Statist. Plann. Inference 121 (2), 311–324.

Werner Hartman, L., 2006. Bayesian modelling of spatial data using Markov random fields, with application to elemental composition of forest soil. Math. Geol. 38 (2), 113–133.

Wikle, C., Berliner, L.M., Cressie, N., 1998. Hierarchical Bayesian space–time models. Environ. Ecol. Statist. 5 (2), 117–154.