Mathematical Statistics
Stockholm University

Fast kriging of large data sets with Gaussian Markov random field models

Linda Werner and Ola Hössjer

Research Report 2004:12

ISSN 1650-0377

Postal address:

Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:

http://www.math.su.se/matstat

Mathematical Statistics
Stockholm University
Research Report **2004:12**,
http://www.math.su.se/matstat

Fast kriging of large data sets with Gaussian Markov random field models

Linda Werner* and Ola Hössjer†

December 2004

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. For data sets with many observations, computation time and memory requirement could be an obstacle for kriging. In this article we approximate a Gaussian field with a Gaussian Markov random field on a lattice to speed up calculations and decrease memory requirements. By using a bilinear interpolation at non-lattice locations the method is well suited for non-lattice data.

KEY WORDS: spatial interpolation, GMRF, non-lattice data

*Mathematical Stastistics, Centre for Mathematical Sciences, Lund University, Box 118, 221 00 LUND, Sweden. E-mail: linda@maths.lth.se. Financial support from the Swedish Foundation for Strategic Research (SSF), project "Spatial Statistics and Image Analysis in Environment and Medicine, A3 02:125"

†Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: ola@math.su.se. Financial support from the Swedish Research Council, contract nr. 626-2002-6286.

1 Introduction

Kriging is a methodology for spatial prediction for which the interest and applications have been growing during the recent decades. The development originated in mining industry 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 often had few observations, the applications of today can have thousands or tens of thousands of data, a fact that must be addressed by the evaluation methods.

The kriging methodology is part of the least squares linear regression framework. It was named by Matheron (1963), working with mining engineering in France, but Gandin (1963) working in meteorology in Russia simultaneously put optimal linear prediction in a spatial framework. In addition to linear kriging methods, such as simple, ordinary or universal kriging where a Gaussian model is assumed for data, the methodology has been enlarged to account for non-Gaussian data where the best predictor (in LS-sense) is not linear. Moyeed and Papritz (2002) compare both linear and non-linear kriging methods with respect to precision and success in modeling prediction uncertainty for a data set of metal concentrations. The non-linear methods were log-normal, disjunctive and indicator kriging, based on different transformations of data, as well as model-based kriging where arbitrary functionals of a latent process are predicted with a Bayesian approach implemented via Markov Chain Monte Carlo (MCMC) techniques (Diggle, Tawn, and Moyeed, 1998). Regarding precision, Moyeed and Papritz (2002) found that linear kriging was as good as any of the non-linear methods, but for skewed data some of the non-linear methods performed better in estimating prediction uncertainty. This article focuses on computationally effective models for linear kriging.

A 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 for each point of prediction only the closest observations are used. When the objective is to give an account of local features, local neighbourhood kriging with a small neighbourhood has besides large computational benefits also the advantage that no covariance values for large distances are used (they are hard to estimate due to few sample points). It also allows one to account for local departures from stationarity. However, when the objective is to describe long-range features local data are not appropriate for the analysis. In Barry and Pace (1997) this problem is addressed by using models for which the covariance matrix is sparse, e.g. spherical correlation, together with sparse matrix techniques. Furrer, Genton, and Nychka (2004) has a similar, but more general approach where 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 they by means of a space filling algorithm introduce the 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 which has rendered large interest in recent years is Markov field approximations of random fields, applied especially 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 (GMRF). Even for small Markov neighbourhoods 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 Gaussian 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. By means of this approximation we can exploit the GMRF's natural specification via its precision (or inverse covariance) matrix to make kriging computationally effective even for large data sets. A bi-linear interpolation facilitates to combine the strength of lattice GMRF with the generality of a continuous spatial index, which is common in geostatistical data sets.

2 Brief introduction to kriging

Depending on the mean value specification 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). For clarity of the discussion we here focus on the benefits of a Markov approximation for **ordinary kriging**, although the principles apply also for universal kriging, see Section 5.

Let therefore

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

be a stochastic field with mean value μ, zero-mean intermediate-scale variation $\gamma(\cdot)$ and white noise error $\varepsilon(\cdot)$ with variance σ^2_ε. Assume that $\gamma(\cdot)$ is a Gaussian process, and that $\varepsilon(\cdot)$ is Gaussian.

Given observations $(Z(\mathbf{t}_1), \ldots, Z(\mathbf{t}_n))^T = \mathbf{Z}$ a common task is to predict $Y(\mathbf{t}) = \mu + \gamma(\mathbf{t})$, often at a grid of locations, and to calculate the prediction error variance in each location.

In *ordinary kriging* linear predictors are used,

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

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

Let

$$\boldsymbol{\Sigma} = ((\mathbf{Cov}(Z(\mathbf{t}_i), \quad Z(\mathbf{t}_j)))_{i,j} = \mathbf{Cov}(\mathbf{Z})$$

denote the covariance matrix of the observations, and let

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

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

The optimal predictor is found either by minimizing the mean squared prediction error or, since a Gaussian model is assumed, by calculating the conditional expectation $\mathbf{E}(Y(\mathbf{t}_0)|\mathbf{Z})$. The resulting weights (Cressie, 1993, p. 123) are

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

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

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

$$\begin{aligned} \sigma_r^2(\mathbf{t}_0) &= \sigma_\gamma^2(\mathbf{t}_0) + \frac{1 - \underline{\mathbf{1}}_n^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\omega}}{\underline{\mathbf{1}}_n^T \boldsymbol{\Sigma}^{-1}\underline{\mathbf{1}}_n} - \left(\boldsymbol{\omega} + \underline{\mathbf{1}}_n\left(\frac{1 - \underline{\mathbf{1}}_n^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\omega}}{\underline{\mathbf{1}}_n^T \boldsymbol{\Sigma}^{-1}\underline{\mathbf{1}}_n}\right)\right)^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\omega} = \\ &= \sigma_\gamma^2(\mathbf{t}_0) - \boldsymbol{\omega}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\omega} + \frac{(1 - \underline{\mathbf{1}}_n^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\omega})^2}{\underline{\mathbf{1}}_n^T \boldsymbol{\Sigma}^{-1}\underline{\mathbf{1}}_n} \end{aligned} \tag{4}$$

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

When the prediction error variance is calculated an $1 - \alpha$ prediction interval is straightforwardly obtained as

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

where z_α is the $(1 - \alpha)$-quantile of the standard normal distribution. Alternatively the prediction error variance conditional on the observations $\mathbf{z}$ could be used for prediction intervals. In the formula for the conditional prediction error variance the last term of (4) vanishes, see Sjöstedt-de Luna and Young (2003).

Any covariance function, i.e. any non-negative definite function, is valid for kriging, but most applications use a model where the covariance function is stationary 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, $\mathbf{Cov}(Z(\mathbf{t}_1),\ Z(\mathbf{t}_2)) = C_Z(\mathbf{t}_1,\ \mathbf{t}_2;\ \theta)$, for the covariance function is given, with θ an unknown parameter, a common procedure is as follows: Estimate θ and then predict $Y(\mathbf{t}_0)$ with (3)– (5) replacing θ by a plug-inestimate $\hat{\theta}$ in $\mathbf{\Sigma}^{-1}$ and $\boldsymbol{\omega}$. See Section 4.2 for more details on estimating θ. When plug-in of estimated covariance parameters are 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.

After parameter estimation, ordinary kriging requires solving a linear system (3) of size $n \times n$, an operation for which both operation count and memory requirement increase by order n^2. This implies that kriging large data sets on a global basis is very slow, or even impossible due to memory limitations. In the next section we suggest an approach where this problem is handled with a GMRF approximation.

3 Kriging and GMRF

In ordinary kriging the model is Gaussian with a (possibly) full covariance matrix. A special case of Gaussian models are *Gaussian Markov Random Fields (GMRF)*.

Let Γ be a $N_1 \times N_2$ regular lattice in $\mathbb{R}^2$ with coordinates $\tilde{\mathbf{t}}_1, \ldots, \tilde{\mathbf{t}}_N$, $N = N_1 N_2$, and horizontal and vertical box widths δ_1 and δ_2. The Gaussian field $\boldsymbol{\gamma}$ defined on Γ, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_N)^T = (\gamma(\tilde{\mathbf{t}}_1), \ldots, \gamma(\tilde{\mathbf{t}}_N))^T$, is a GMRF if it satisfies the Markov property

$$\gamma_i | \boldsymbol{\gamma}_{(-i)} = \gamma_i | \boldsymbol{\gamma}_{N[i]} \in \mathrm{N}\left(\sum_{j=1}^{N} w_{ij} \gamma_j, \tau^2 \right) \text{for } i = 1, \ldots, N \tag{6}$$

where $N[i]$ is the *neighbourhood* of lattice point i, often chosen as the closest vicinity of point i. With

$$\mathbf{W} = (w_{ij})_{i,j=1}^{N} \text{ where } w_{ij} = \begin{cases} \alpha_{j-i}, & j \in N[i] \\ 0, & \text{otherwise} \end{cases}$$

the joint distribution of $\boldsymbol{\gamma}$ is

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

provided that the parameters α_{j-i} are such that $(\mathbf{I}_N - \mathbf{W})^{-1}$ exists and is symmetric and non-negative definite. Note that τ^2 is the conditional, not the unconditional, variance of γ_i.

The Markov property induced by the neighbourhood system $N[i], i \in \Gamma$ thus imposes a zero-pattern structure in the precision matrix
$\mathbf{Q} = (\mathbf{I}_N - \mathbf{W})\tau^{-2}$ 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.

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 the Gaussian fields most commonly used for kriging can be well approximated by GMRFs with small neighbourhoods, even for fields with long correlation length. The covariance function deviance between a Gaussian field and it's fitted GMRF was small compared to the error usually induced when estimating the covariance parameters from data. This indicates that a GMRF approximation would give competitive kriging results, to a lower computational cost.

However, since many geostatistical applications have non-lattice data, we suggest a generalization to data with continuous spatial index. We therefore extend the model to non-lattice data, while maintaining the computational benefits of the GMRF formulation. Influences for approximating non-lattice data with a grid process come from Wikle, Berliner, and Cressie (1998) where nearest-neighbour mapping was used in an MCMC application. We will use a more sophisticated version with a bilinear interpolation, as suggested in Werner (2003).

Let for an arbitrary $\mathbf{t} \in D$

$$\gamma(\mathbf{t}) = \sum_{j=1}^{N} k(\mathbf{t}, \tilde{\mathbf{t}}_j)\ \boldsymbol{\gamma}(\tilde{\mathbf{t}}_j) \tag{8}$$

where $k(\mathbf{t}, \cdot)$ are weights averaging the value of $\boldsymbol{\gamma}$ of the four nearest lattice points to an arbitrary point $\mathbf{t}$. Therefore let $k(\mathbf{t}, \tilde{\mathbf{t}}_j) = 0$ if j is not corner of the lattice square that surrounds $\mathbf{t}$, and calculate the remaining 4 weights from the areas $B_1, \ldots, B_4$ in Figure 1:

$$\begin{cases} k(\mathbf{t}, \tilde{\mathbf{t}}_{(k,l)}) & = & B_1/B \\ k(\mathbf{t}, \tilde{\mathbf{t}}_{(k+1,l)}) & = & B_2/B \\ k(\mathbf{t}, \tilde{\mathbf{t}}_{(k,l+1)}) & = & B_3/B \\ k(\mathbf{t}, \tilde{\mathbf{t}}_{(k+1,l+1)}) & = & B_4/B \end{cases} \tag{9}$$

where $B = \delta_1\delta_2 = B_1 + B_2 + B_3 + B_4$.

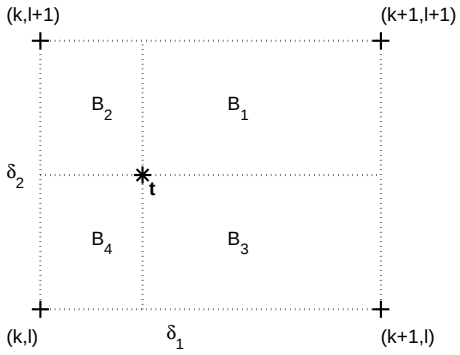

Figure 1: Areas determining the weights when averaging over the four nearest lattice points.

The above model for $\gamma(\mathbf{t})$ is a continuous Gaussian process that equals to the GMRF $\boldsymbol{\gamma}$ at grid locations. It 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 Prediction

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

Let $\mathbf{K} = (k_{ij})$ be the $n \times N$ matrix defined as $k_{ij} = k(\mathbf{t}_i, \tilde{\mathbf{t}}_j)$. Then from (7) and (8) we have

$$\boldsymbol{\Sigma} = \mathbf{K}\mathbf{Q}^{-1}\mathbf{K}^T + \sigma_\varepsilon^2 \mathbf{I}_n \tag{10}$$

where $\mathbf{I}_n$ is the identity matrix of dimension n and $\mathbf{Q} = (\mathbf{I}_N - \mathbf{W})\tau^{-2}$ is the precision matrix for the lattice GMRF.

Similarly define the vector of k-weights for the point, $\mathbf{t}_0$, where we want to predict Y as $\mathbf{k} = (k_1, \ldots, k_N)^T$, where $k_i = k(\mathbf{t}_0, \tilde{\mathbf{t}}_i)$, to get

$$\boldsymbol{\omega} = \mathbf{K}\mathbf{Q}^{-1}\mathbf{k}. \tag{11}$$

Insertion of expressions (10)–(11) 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, i.e. calculate (3) for $\mathbf{t}_0 = \tilde{\mathbf{t}}_1, \ldots, \tilde{\mathbf{t}}_N$. Let therefore

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

   and $\hat{\mathbf{Y}} = (\hat{Y}(\tilde{\mathbf{t}}_1), \ldots, \hat{Y}(\tilde{\mathbf{t}}_N))^T$.

   Using Equations (2) and (3) for $\mathbf{t}_0 = \tilde{\mathbf{t}}_1, \ldots, \tilde{\mathbf{t}}_N$ we conclude

   $$\hat{\mathbf{Y}} = \left(\boldsymbol{\Omega} + \underline{\mathbf{1}}_n \left(\frac{\underline{\mathbf{1}}_N^T - \underline{\mathbf{1}}_n^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\Omega}}{\underline{\mathbf{1}}_n^T \boldsymbol{\Sigma}^{-1} \underline{\mathbf{1}}_n}\right)\right)^T \boldsymbol{\Sigma}^{-1}\mathbf{Z} \tag{12}$$

2. At an arbitrary point $\mathbf{t}_0$,

   $$\hat{Y}(\mathbf{t}_0) = \mathbf{k}^T \hat{\mathbf{Y}}, \tag{13}$$

   see Appendix A for details.

From (12) and (13) we get $\hat{Y}(\mathbf{t}_0)$ at any point $\mathbf{t}_0$. It is therefore enough to seek computationally effective expressions for $\boldsymbol{\Sigma}^{-1}\boldsymbol{\Omega}$ and $\boldsymbol{\Sigma}^{-1}\underline{\mathbf{1}}_n$.

The definition of the intermediate-scale variation gives

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

that together with (10) yields

$$\begin{aligned} \boldsymbol{\Sigma}^{-1}\boldsymbol{\Omega} &= \boldsymbol{\Sigma}^{-1}\boldsymbol{\Omega}(\mathbf{Q} + \sigma_\varepsilon^{-2}\mathbf{K}^T\mathbf{K})(\mathbf{Q} + \sigma_\varepsilon^{-2}\mathbf{K}^T\mathbf{K})^{-1} \\ &= \boldsymbol{\Sigma}^{-1}(\sigma_\varepsilon^{-2}\boldsymbol{\Sigma}\mathbf{K}) \cdot (\mathbf{Q} + \sigma_\varepsilon^{-2}\mathbf{K}^T\mathbf{K})^{-1} \\ &= \sigma_\varepsilon^{-2}\mathbf{K}(\mathbf{Q} + \sigma_\varepsilon^{-2}\mathbf{K}^T\mathbf{K})^{-1}. \end{aligned} \tag{14}$$

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

$$\boldsymbol{\Sigma}^{-1}\underline{\mathbf{1}}_n = \boldsymbol{\Sigma}^{-1}\mathbf{K}\underline{\mathbf{1}}_N = \boldsymbol{\Sigma}^{-1}\boldsymbol{\Omega}\mathbf{Q}\underline{\mathbf{1}}_N. \tag{15}$$

From (14) and (15) we can thus conclude that it suffices to invert a *band limited* $N \times N$ matrix $\mathbf{Q} + \sigma_\varepsilon^{-2}\mathbf{K}^T\mathbf{K}$. As $\mathbf{K}$ interpolates only the four nearest grid points the addition of $\mathbf{K}^T\mathbf{K}$ adds no extra bandwidth to the GMRF precision matrix $\mathbf{Q}$.

3.2 Prediction error

In this section a formula for the kriging prediction error variance is derived for the Markov model approach. To calculate the prediction error variance for the Markov approach define

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

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

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

so that after some manipulations given in appendix, $\mathbf{C}_r = (C_r(\tilde{\mathbf{t}}_i,\ \tilde{\mathbf{t}}_j))_{ij}$ is given by

$$\mathbf{C}_r = (\mathbf{Q} + \sigma_\varepsilon^{-2} \mathbf{K}^T \mathbf{K})^{-1} + \frac{(\underline{\mathbf{1}}_n^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\Omega} - \underline{\mathbf{1}}_N^T)^T (\underline{\mathbf{1}}_n^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\Omega} - \underline{\mathbf{1}}_N^T)}{\underline{\mathbf{1}}_n^T \boldsymbol{\Sigma}^{-1} \underline{\mathbf{1}}_n} \tag{16}$$

Due to (13) and $Y(\mathbf{t}) = \sum_{j=1}^N k_j Y(\tilde{t}_j)$ it follows that $r(\mathbf{t}_0) = \sum_{i=1}^N k_i r(\tilde{\mathbf{t}}_i)$ and hence

$$\sigma_r^2(\mathbf{t}_0) = \mathbf{k}^T \mathbf{C}_r \mathbf{k} = \sum_{i,j=1}^N k_i k_j C_r(\tilde{\mathbf{t}}_i,\ \tilde{\mathbf{t}}_j), \tag{17}$$

is a double sum with at most 16 terms. Thus calculating the prediction error variance requires no extra matrix inversions.

The idea behind the linear interpolation (8) and (13) is to predict a more realistic value of $Y(\cdot)$ at non-lattice locations. A deficit with the above approach is that although the variance of $Y(\cdot)$ is well tuned at grid locations, since the interpolation is a weighted average of the near-by gridpoints, the variance in between gridpoints becomes too small. Thus the prediction error variance (17) will be too small at non-lattice locations, and the prediction intervals (5) will not have the prescribed coverage. To decrease this deficiency we suggest a correction of the prediction error variance. One possibility is to calculate the prediction error variance as

$$\sigma_r^2(\mathbf{t}_0) = \sum_{i=1}^N k_i \sigma_r^2(\tilde{\mathbf{t}}_i), \tag{18}$$

i.e. as a weighted average of the prediction error variances of the grid points surrounding $\mathbf{t}_0$. This amounts to multiplication with a correction factor

$$CF = \sum_i k_i \sigma_r^2(\tilde{\mathbf{t}}_i) / \sum_{ij} k_i k_j C_r(\tilde{\mathbf{t}}_i, \tilde{\mathbf{t}}_j), \tag{19}$$

and gives a remedy for the location dependent shrinking variance. As the range of $Y(\cdot)$ and therefore of $r(\cdot)$ decreases, the CF approaches $1/\sum_i k_i^2$, whereas when the range of $r(\cdot)$ approaches infinity CF approaches 1. We believe that due to the screening effect in kriging, see Cressie (1993, p. 133), the range of the residual field will stay small, and the correction will therefore have large impact even when the range of $Y(\cdot)$ is large.

4 Simulation study

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 the full model approach the simulated data sets are relatively small compared to 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 in our study used a stationary isotropic Gaussian field with an exponential covariance function both for data generation and estimation.

Following (1), we simulated several samples $\mathbf{z}$ with different values of the variables $(N, n, r, \sigma_\gamma^2, \sigma_\varepsilon^2)$. The simulations were done in R (R Development Core Team, 2003) with the routine `GaussRF` in package RandomFields. For the intermediate-scale variation $\gamma(\cdot)$ we used a stationary Gaussian field with exponential covariance function $C(h) = Cov(\gamma(\mathbf{t}), \gamma(\mathbf{t}+\mathbf{h})) = \sigma_\gamma^2 \exp(-3h/r)$ where $h = |\mathbf{h}|$ and 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.

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-neighbourhood for the GMRF approximation.

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 plug-in of the estimated covariances in the kriging computations.

4.1 Known covariance parameters

For each sample $\mathbf{z}$ we performed kriging twice:

1. Using the full model with known parameters, i.e. a Gaussian random field with exponential covariance function and added measurement error.
2. As described in Section 3, with an interpolated GMRF with 5×5 neighbourhood on a $\sqrt{N} \times \sqrt{N}$- grid with $\delta_1 = \delta_2 = \delta$, i.e. filling a square with sidelength $\Delta = \delta\sqrt{N}$. 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}. For the nugget the known variance σ_ε^2 was used.

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

$$\text{PRESS} = \sum_{i=1}^{M} (y_i - \hat{y}_i)^2 \tag{20}$$

was calculated for both 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 (3) and (2) ($\text{PRESS}_{exponential}$) and (12) and (13)($\text{PRESS}_{Markov}$). For the full Gaussian fields the predictions (3) were calculated in Matlab, while for the GMRF approximation (2)the calculations were done in c, with the library GMRFLib.

To compare the kriging predictors to the naive predictor $\hat{y}_i = \bar{y}_e$ (average of the estimation set) the *relative mean square error of prediction* (rMSEP)

$$\text{rMSEP} = \frac{\sum_{i=1}^{M}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{M}(y_i - \bar{y}_e)^2} = \frac{\text{PRESS}}{\sum_{i=1}^{M}(y_i - \bar{y}_e)^2} \tag{21}$$

was calculated. This was suggested in Moyeed and Papritz (2002), although they predict Z, whereas we predict Y, the field without measurement error. We believe that most often it is the underlying stochastic field that is of interest, and since we use simulated data to study the quality of our methods, we have the possibility to focus on the Y-field.

Figure 2 shows rMSEP for GMRF kriging for different parameters. For data with reasonably large signal to noise ratio and range the GMRF kriging predictor performs well, with rMSEP $\approx 20\%$.

To compare GMRF kriging to kriging with the full model, the quotient $\text{PRESS}_{Markov}/\text{PRESS}_{exponential}$ was calculated, see Figure 3. Although the

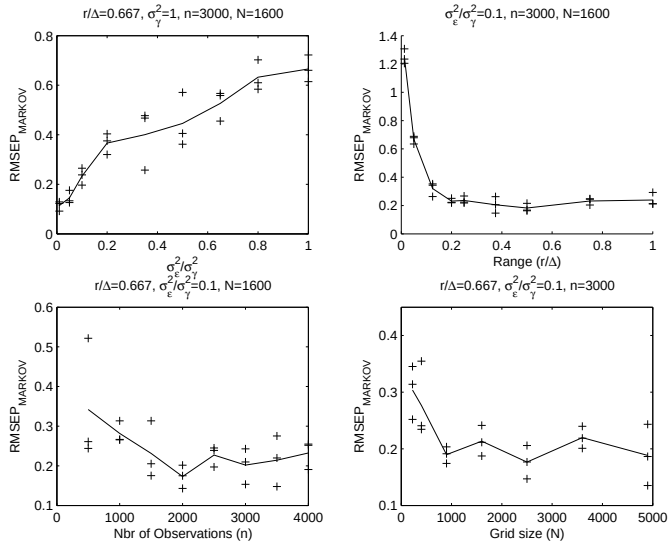

Figure 2: rMSEP $= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 / \sum_{i=1}^{n}(y_i - \bar{y}_e)^2$ for the Markov model for different parameter values and number of observations. rMSEP is in all runs calculated from the kriging result in M=2000 validation points distributed in the area according to a uniform distributions. For each set of parameters 3 simulations were created and evaluated, labeled with +, and their mean is marked by the solid line.

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. In the interesting scenarios where the range is reasonably long (otherwise global prediction will not be interesting) PRESS_{Markov} is less than 10% larger than $\text{PRESS}_{exponential}$ provided that the grid is dense enough. Examining Figure 3 acceptable results are obtained for $N \approx n/2$. However, as noted in Table 1 and discussion below, the main computational benefits does not emerge from choosing $N < n$ but from exploiting the sparse precision matrix of the GMRF.

The computations were performed on a 2.7 GHz Intel Xion with 2GB RAM. For dense linear systems Matlab's solvers are well optimized, whereas for the GMRF model we take advantage of the c-library GMRFLib (Rue, 2001) to exploit the sparse structure of the precision matrix of the GMRF, in setup and solutions of the linear systems. Therefore, all calculations for the full Gaussian field where done in Matlab, while for the Markov approach initial work, such as reading the data and calculating the interpolation matrix, was done in Matlab, the sparse systems were setup and solved in a c-routine utilizing the library GMRFLib, and plots of the results were obtained in Matlab. The computation times for kriging with the full model compared to exploiting a Markov approximation are displayed in Figure 4 for the same samples

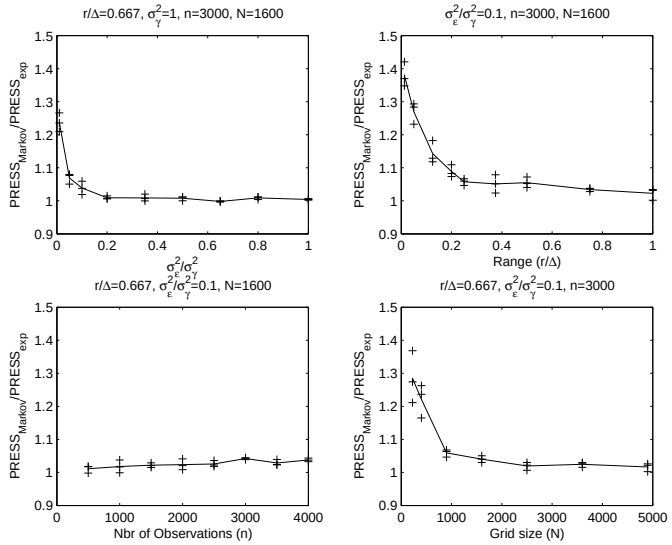

Figure 3: Quotient of prediction error sum of squares, $\frac{PRESS_{Markov}}{PRESS_{exponential}}$, for different parameter values, number of observations etc. For the exponential Gaussian field the known covariance parameter values were used while for the GMRF fitted precision parameters were used. PRESS is in all runs calculated from the kriging result in M=2000 validation points distributed in the area according to a uniform distribution. For each set of parameters 3 simulations were created and evaluated, labeled with +, and their mean is marked by the solid line.

used in Figure 3 and 2. The times include the calculation of prediction error variances at all predictive locations. Since this involves solving a $n \times n$ system M times this is the most time consuming part of the calculations. Table 1 displays the time consumption for different steps of the analysis, for both the full Gaussian model and the GMRF approximation for two different densitys of the grid. The Markov approach is faster by a factor of over 15 and 45 for the two grid densities respectively. However, calculating only the predictions (step 4), which is what Furrer et al. (2004) reported, the Markov approach is faster by a factor of over 1000 for both grid densities.

The computation time for the Markov approach does only increase slightly with n. Since the calculations are done on the grid it is the number of gridpoints N that are crucial. With a neighbourhood of order 1, the precision matrix $\mathbf{Q}$ will (after ordering the gridpoints column- or rowwise) get bandwidth $p = \min\{k : Q_{ij} = 0, |i-j| = k\} \sim \sqrt{N}$, why solving $\mathbf{Q}^{-1}\mathbf{x}$ requires $\sim N(p^2 + 3p) \sim N^2$ operations (Rue and Held, 2005, p. 45). If the number of gridpoints is chosen proportional to the number of observational points, $N \propto n$, the increase in computations is only increasing as $N^2 \propto n^2$ as com-

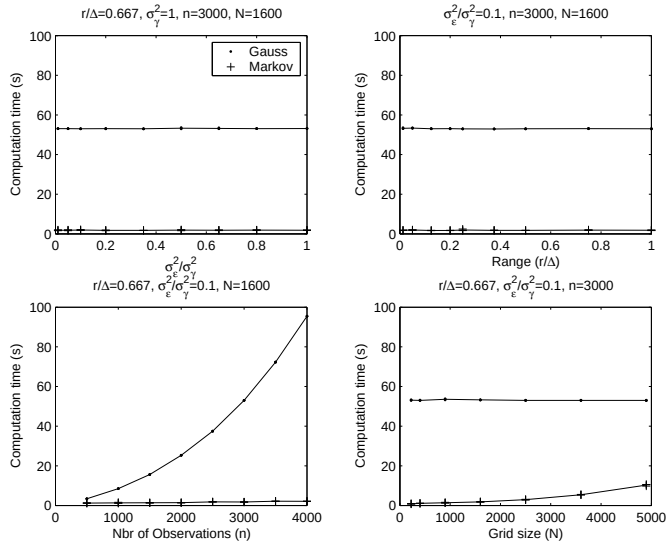

Figure 4: Computation time for kriging with Gaussian and GMRF model for different parameter values, cf. Figure 3. M=2000 observations in the validation set. For each set of parameters 3 simulations were created and evaluated, labeled with +, and their mean is marked by the solid line.

pared to n^3 for the full system. Since further the precision matrix $\mathbf{Q}$ of a lattice GMRF have many zeros even for $|i-j| < p$, the above calcuations are an upper bound of the computational demand for the GMRF model.

We also evaluated how well the Markov model performs regarding prediction error variance and coverage of prediction intervals. For 10 simulated fields with $n = 3000$ observations at random points on a square with side Δ, $N = 40^2 = 1600$ grid points covering the square, $r = 2\Delta/3$, $\sigma_\gamma^2 = 1$, and $\sigma_\varepsilon^2 = 0.1$ we compared the kriging predictors to simulated values in $M = 2000$ random locations. The observational and predictive locations were chosen dfferently in each of the ten runs. We calculated prediction error variances for both the exponential and Markov model, according to (4) and (16)- (18) respectively, and the corresponding 95%-prediction intervals (5). Similar calculations can be done with the conditional prediction error variances mentioned on Page 4. In our setting the difference between conditional and unconditional prediction error variance is less than 10^{-5} times the conditional one, and we therefore only report the results with conditional prediction error variances.

To evaluate the coverage of the prediction intervals for each of the two kriging approaches, we checked whether the prediction intervals for each predicted location contained the true value at that location or not. As argued on Page 9, the variance for the interpolated Markov model depends on where within the grid box the prediction point is located. Figure 5 contains prediction

Table 1: Computation time for kriging with full Gaussian and GMRF model. Kriging predictions and prediction error variances are calculated at $M = 2000$ locations from data with $n = 3000$ observations. (Linux, 2.7GHz Intel Xion with 2GB RAM)

	Time (s)			
	$N = 1600$		$N = 3025$	
Action	Full	GMRF	Full	GMRF
1. Reading data, variable setup	0.16	0.16	0.18	0.18
2. Creating Σ and Ω	8.12		8.00	
2'. Creating grid and interp. matrix $\mathbf{K}$		0.56		0.66
2". Creating $\mathbf{Q} + \sigma^{-2}\mathbf{K}^T\mathbf{K}$		0.07		0.20
3. Calc. kriging predictions $\hat{\mathbf{Y}}$	12.78	0.01	12.87	0.01
4. Calc. pred. err. var.	37.75	0.68	37.20	2.48
Total	58.81	1.30	58.85	3.53

error variances for both Markov and exponential covariance definition. The unwanted decrease of prediction error variance (17) for the Markov model at points closer to a grid box center is distinct. In Figure 6 the alternative calculation (18) was used for the Markov prediction error variance. This rectifies the shrinkage for locations in the interior of a grid box, but unfortunately the Markov prediction error variance is still smaller than for kriging with a full model. This will make the prediction intervals more narrow, and thus the coverage will be somewhat smaller than prescribed.

For the prediction intervals, a smoothed estimate of the coverage vs distance to nearest grid point was for each kriging approach calculatedin Figure 7 by local linear regression of the data pairs

$$(d_{ij}, 1_{\{y_{ij} \in I_{ij}\}}), \quad i = 1, \ldots, 10,\ j = 1, \ldots, 2000.$$

Here d_{ij} is the distance to the nearest grid point (in units of δ), y_{ij} is the true value and I_{ij} is the kriging prediction interval for prediction j in field i. Simulations from the same model but with different choices of r/δ gave similar results except for small r/δ, about 10, when the coverage was lower.

With the alternative formula (18) for the prediction error variance at non-lattice locations, we obtain more reasonable results for the coverage as seen in Figure 8, although the prediction intervals still are somewhat narrow. The correction factor (19) has been close to $1/\sum_i k_i^2$ in all our simulations, i.e. the value that corresponds to a small range of the r-field. This has been the case even when r/δ for the Y-field has been 40 or larger. We believe that this is because the range of the residual field is much shorter, due to the screening

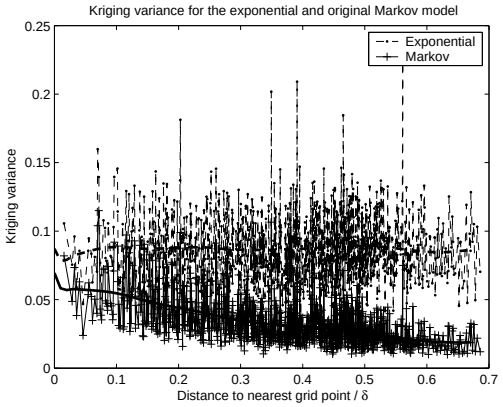

Figure 5: Prediction error variance as function of distance to nearest grid point. Every second data point is shown. The lines are smoothed from coverage data by means of local linear regression with an Epanechnikov kernel with bandwidth $h = 0.05$. Grid size $= 40 \times 40$, $\sigma_\gamma^2 = 1$, $\sigma_\varepsilon^2 = 0.1$ and $r/\delta = 20$.

effect in kriging, see (Cressie, 1993, p. 133).

4.2 Estimated covariance parameters

When kriging real data it is rare to know the model parameters, and they must consequently be estimated from the dataset. Plug-in of estimated parameters into the kriging quations is a common approach, for which we will here test the performance of the Markov model kriging. To do this, the same data sets as in the previous section were used for kriging, but here the covariance parameters were estimated from data. The performance of the two models regarding prediction intervals can be calculated and assessed just as when parameters are 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 too narrow prediction intervals. This could be rectified by boostrap calibration (Sjöstedt-de Luna and Young, 2003). We have here confined ourselves to reporting the perfomance of $PRESS$ for the Markov and exponential covariance structures. Prediction intervals simulated for the same model as in Section 4.1 but calculated by plug-in of estimated parameters show roughly the same coverage as for known parameters.

Regarding parameter estimation Cressie (1993, p. 69–73) argues that variogram estimation is preferable to covariance estimation due to lower bias and

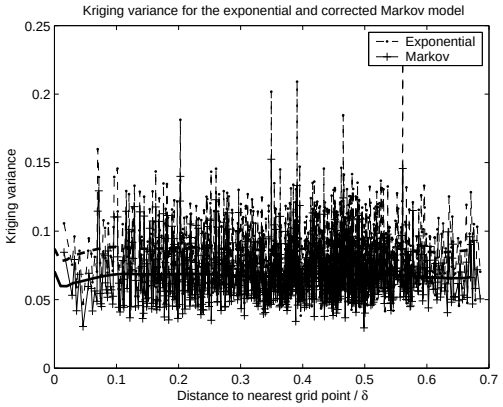

Figure 6: Prediction error variance as function of distance to nearest grid point. The Markov prediction error variances were corrected according to (18). Every second data point is shown. The lines are smoothed from coverage data by means of local linear regression with an Epanechnikov kernel with bandwidth $h = 0.05$. Grid size $= 40 \times 40$, $\sigma_\gamma^2 = 1$, $\sigma_\varepsilon^2 = 0.1$ and $r/\delta = 20$.

that the variogram is defined for some processes that are not second order stationary. When, as in the models we investigate, the covariance is defined there is a simple relationship between covariance and variogram, so that the variogram can be used for estimation while the covariance formulation is used for the kriging equations as in Equation (3). This procedure combines the low bias of variogram estimation with the computational advantage of the sparsity which is found in certain covariance matrices, see Barry and Pace (1997) for an example.

Let therefore

$$\hat{v}_{ij} = \frac{1}{2}(Z(\mathbf{t}_i) - Z(\mathbf{t}_j))^2$$

and

$$\begin{aligned} v_{ij} &= v(\mathbf{t}_i, \mathbf{t}_j; \theta) = \frac{1}{2}\operatorname{Var}(Z(\mathbf{t}_i) - Z(\mathbf{t}_j))^2 \\ &= \frac{1}{2}\mathbf{E}(Z(\mathbf{t}_i) - Z(\mathbf{t}_j))^2 \end{aligned}$$

be the empirical and theoretical *semi-variogram* respectively.

When observations are restricted to a regular grid, it is customary to estimate θ in the following way. Assuming stationarity and isotropy, all observational

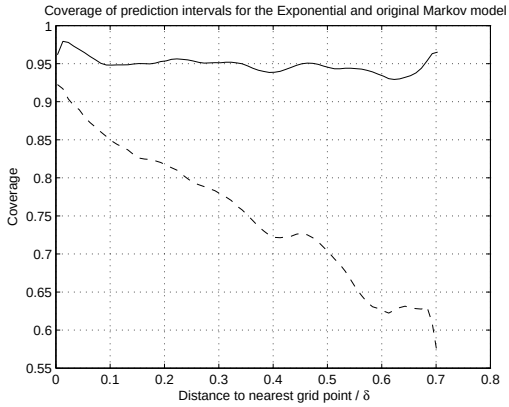

Figure 7: Coverage of the 95% kriging prediction intervals as function of distance to nearest grid point for Exponential model (Solid line) and Markov model (dotted line). Prediction variance (17)as used for the Markov model.The lines are smoothed from coverage data by means of local linear regression with an Epanechnikov kernel with bandwidth $h = 0.05$.

pairs $(\mathbf{t}_i, \mathbf{t}_j)$ with the same $|\mathbf{t}_j - \mathbf{t}_i| = h$, are clumped and used to estimate the variogram at lag h . The estimated variograms are then fitted to $v(\mathbf{t}, \mathbf{t} + \mathbf{h}; \theta) = v(h; \theta)$ using some loss function, see Lahiri, Lee, and Cressie (2002) for details.

For irregularly spaced data most distances are unique. Although θ can be estimated from the full collection $\{\hat{v}_{ij}\}$, called the variogram cloud, see Müller (1999) for details, it is common to smooth the variogram estimator by using bins:

$$2\bar{v}_k = \frac{1}{N_k} \sum_{|\mathbf{t}_i - \mathbf{t}_j| \in I_k} (Z(\mathbf{t}_i) - Z(\mathbf{t}_j))^2$$

where I_k is an interval and N_k is the number of observation pairs for which $|\mathbf{t}_i - \mathbf{t}_j| \in I_k$.

Define

$$\hat{\theta} = \arg\min_{\theta} \sum_k b_k (\bar{v}_k - v(h_k, \theta))^2 \tag{22}$$

where $v(h_k, \theta)$ is the theoretical variogram calculated at h_k, the midpoint of I_k, and the numbers b_k are nonnegative weights. The choice of b_k is a compromise between simplicity and statistical efficiency of the estimate $\hat{\theta}$, see Cressie (1993) for examples. At large lags $\bar{v}_k$ tends to be an unreliable estimate of v_k due to small N_k, so following the guidelines of Journel and

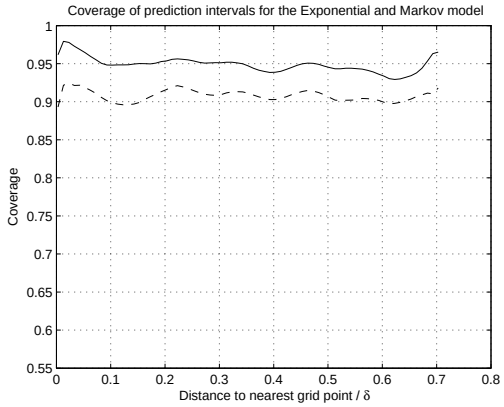

Figure 8: Coverage of the 95% kriging prediction intervals as function of distance to nearest grid point for Exponential model (Solid line) and corrected Markov model (dotted line). Prediction variance (18) was used for the Markov model.The lines are smoothed from coverage data by means of local linear regression with an Epanechnikov kernel with bandwidth $h = 0.05$.

Huighbregts (1978) we choose

$$b_k = 1_{\{h_k \leq c\}} \text{ where } c = \frac{1}{2} \max_{i,j}(|\mathbf{t}_i - \mathbf{t}_j|). \tag{23}$$

Following the methods above, we estimated the parameters for the exponential model r, σ_γ^2 and σ_ε^2 from the estimation set by means of the package geoR (Ribeiro Jr. 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 $\hat{r}$.

The comparison of $PRESS_{Markov}$ and $PRESS_{exponential}$ when parameters are estimated from data, see Figure 9, were similar to the results achieved with known parameters, in Section 4.1. Naturally a larger variability in the results occur when parameters are fitted from data. In most simulations the GMRF kriging results were good, but for occasional data sets PRESS_{Markov} was almost 50% larger than $\text{PRESS}_{exponential}$. The results for the Markov-model would probably be more competitive if the precision parameters were fitted directly from data instead of fitted to an exponential covariance function, since this favours the latter model. However, as our intention is to show that a Markov approximation give competitive results with little extra work, we want to use a simple method.

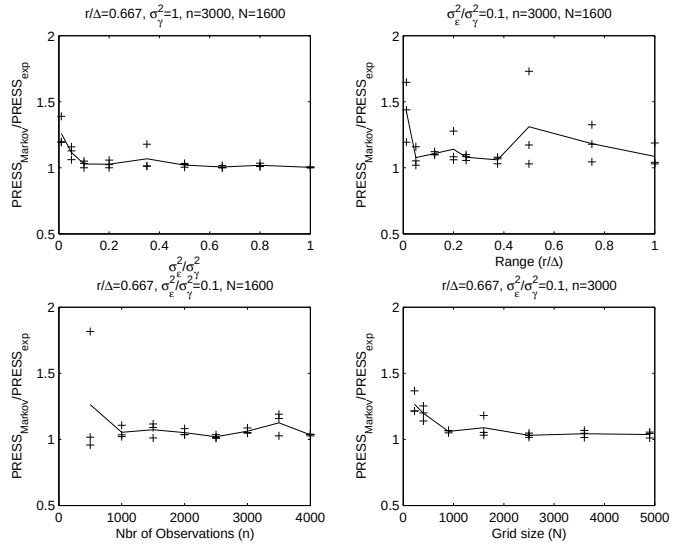

Figure 9: Quotient of prediction error sum of squares, $\frac{PRESS_{Markov}}{PRESS_{exponential}}$, for different parameter values, number of observations etc. For the exponential Gaussian field estimated covariance parameter values were used while for the GMRF the precision parameters were chosen to fit an exponential with the above estimated range. PRESS is in all runs calculated from the kriging result in M=2000 validation points distributed in the area according to a uniform distribution.

5 Conclusions

We argue that for kriging applications with large data sets where global features are of interest, a GMRF model approximation can be a practical way to reduce computational burden and memory requirements. The simulations where data from a Gaussian process with exponential covariance function are kriged with a GMRF model show that the quality of the predictions are 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 when only kriging predictions are needed. But even when the prediction error variances are needed the reductions are considerable.

Although we in this article used ordinary kriging to examplify the use of GMRFs for kriging, the method is not restricted to the setting where the mean is constant. In universal kriging the mean is an unknown linear combination of known functions, i.e. $Z(t) = \mathbf{X}\boldsymbol{\beta} + \gamma(\mathbf{t}) + \varepsilon(t)$ where $\mathbf{X} = (f_j(\mathbf{t}_i))_{i,j}$ is an $n \times p$-matrix of known quantities. In this setting a global model can more

often be fitted, since the mean value specification can take care of large scale variability in the area and related departure from stationarity.

If the covariance structure is known, the predictor at point $\mathbf{t}_0$ ($Z(\mathbf{t}_0) = \boldsymbol{f}_0^T\boldsymbol{\beta} + \gamma(\mathbf{t}_0) + \varepsilon(\mathbf{t}_0)$) has kriging weights (see e.g. Cressie, 1993, p. 154)

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

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

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

Thus to get fast calculations with the Markov approach we must now find a good expression for $\boldsymbol{\Sigma}^{-1}\boldsymbol{\Omega}$ and $\boldsymbol{\Sigma}^{-1}\mathbf{X}$ when predicting simultanously on the grid on which the GMRF is defined. However $\boldsymbol{\Sigma}^{-1}\boldsymbol{\Omega}$ is calculated as in (14) and in analogy with (15) the expressions simplify if $\mathbf{KX} = \mathbf{X}_N$ to

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

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

The interpolated Markov model is thus suitable for fast universal kriging (at least) as long as $\mathbf{KX} = \mathbf{X}_N$. This is true for $\mathbf{K}$ in (8) and (9) not only for $\mathbf{X} = \underline{\mathbf{1}}$ (ordinary kriging) but also when a linear trend in two dimensions is assumed ($\mathbf{X}_{i\cdot} = (1\ t_{i1}\ t_{i2})$ where $\mathbf{t}_i = (t_{i1}\ t_{i2})$) or when a bilinear term is added ($\mathbf{X}_{i\cdot} = (1\ t_{i1}\ t_{i2}\ t_{i1}t_{i2})$). For some other definitions of $\mathbf{X}$ it will be possible to ensure $\mathbf{KX} = \mathbf{X}_N$ by modifying the choice of interpolation matrix $\mathbf{K}$.

Acknowledgement

The authors would like to thank Håvard Rue for helpful discussions and for sharing the precision parameters for GMRFs fitted to exponential covariance functions.
Linda Werner's work have been partially founded by the Swedish Foundation for Strategic Research (SSF), project "Spatial Statistics and Image Analysis in Environment and Medicine, A3 02:125" and Ola Hössjer's work was supported by the Swedish Research Council, contract 626-2002-6286.

Appendix A: Motivation of Equation (13)

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

$$\mathbf{a} = \sum_{i=1}^{N} k_i \tilde{\mathbf{a}}_i \tag{24}$$

But this follows immediately from (3) and the facts that $\sum_1^N k_i = 1$ and $\sum_1^N k_i \tilde{\boldsymbol{\omega}}_i = \boldsymbol{\omega}$.

Appendix B: Motivation of Equation (16)

The covariance matrix of the prediction errors on the lattice, $\mathbf{C}_r = (C_r(\tilde{\mathbf{t}}_i,\ \tilde{\mathbf{t}}_j))_{ij}$, from Section 3.2 is

$$\mathbf{C}_r = \mathbf{Q}^{-1} - \boldsymbol{\Omega}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Omega} + \frac{(\underline{\mathbf{1}}_n^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Omega} - \underline{\mathbf{1}}_N^T)^T(\underline{\mathbf{1}}_n^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Omega} - \underline{\mathbf{1}}_N^T)}{\underline{\mathbf{1}}_n^T\boldsymbol{\Sigma}^{-1}\underline{\mathbf{1}}_n}.$$

Insertion of expression (10) and (11) give us Equation (16):

$$\begin{aligned}
\mathbf{Q}^{-1} - \boldsymbol{\Omega}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Omega} &= (\mathbf{Q}^{-1} - \mathbf{Q}^{-1}\mathbf{K}^T\sigma_\varepsilon^{-2}\mathbf{K}(\mathbf{Q} + \sigma_\varepsilon^{-2}\mathbf{K}^T\mathbf{K})^{-1}) = \\
&= \mathbf{Q}^{-1}(\mathbf{I}_N - \sigma_\varepsilon^{-2}\mathbf{K}^T\mathbf{K}(\mathbf{Q} + \sigma_\varepsilon^{-2}\mathbf{K}^T\mathbf{K})^{-1}) = \\
&= \mathbf{Q}^{-1}((\mathbf{Q} + \sigma_\varepsilon^{-2}\mathbf{K}^T\mathbf{K}) - \sigma_\varepsilon^{-2}\mathbf{K}^T\mathbf{K})(\mathbf{Q} + \sigma_\varepsilon^{-2}\mathbf{K}^T\mathbf{K})^{-1} = \\
&= \mathbf{Q}^{-1}\mathbf{Q}(\mathbf{Q} + \sigma_\varepsilon^{-2}\mathbf{K}^T\mathbf{K})^{-1} = (\mathbf{Q} + \sigma_\varepsilon^{-2}\mathbf{K}^T\mathbf{K})^{-1}
\end{aligned}$$

References

Barry, R. P. and Pace, R. K. (1997), "Kriging with Large Data Sets Using Sparse Matrix Techniques," *Communications in Statistics, Part B – Simulation and Computation*, 26, 619–629.

Cressie, N. A. C. (1993), *Statistics for spatial data*, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, New York: John Wiley & Sons Inc.

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998), "Model-based geostatistics," *Journal of the Royal Statistical Society, Series C, Applied Statistics*, 47, 299–350.

Furrer, R., Genton, M. G., and Nychka, D. (2004), "Covariance Tapering for Interpolation of Large Spatial Datasets," Submitted to Journal of Computational and Graphical Statistics.

Gandin, L. (1963), *Objective Analysis of Meteorological Fields*, Leningrad: Gidrometeorologicheskoe Izdatel'stvo (GIMIZ), translated by Israel Program for Scientific Translations, Jerusalem, 1965.

Goovaerts, P. (1997), *Geostatistics for Natural Resources Evaluation*, Oxford University Press.

Isaaks, E. and Srivastava, R. (1989), *An Introduction to Applied Geostatistics*, New York: Oxford University Press.

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

Kammann, E. E. and Wand, M. P. (2003), "Geoadditive models," *Journal of the Royal Statistical Society, Series C, Applied Statistics*, 52, 1–1.

Lahiri, S. N., Lee, Y., and Cressie, N. (2002), "On asymptotic distribution and asymptotic efficiency of least squares estimators of spatial variogram parameters," *Journal of Statistical Planning and Inference*, 103, 65–85, c. R. Rao 80th birthday felicitation volume, Part I.

Matheron, G. (1963), "Traité de Géostatistique Appliquée, Tome. 2: Le Krigeage," Editions Bureau de Recherches Geologiques et Minieres Paris.

Moyeed, R. A. and Papritz, A. (2002), "An empirical comparison of kriging methods for nonlinear spatial point prediction," *Mathematical Geology*, 34, 365–386.

Müller, W. G. (1999), "Least-squares fitting from the variogram cloud," *Statistics & Probability Letters*, 43, 93–98.

Nychka, D. and Saltzman, N. (1998), "Design of air quality monitoring networks," in *Lect. Notes Statist*, Springer Verlag, vol. 132, pp. 51–76.

R Development Core Team (2003), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, iSBN 3-900051-00-3.

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

Rue, H. (2001), "Fast sampling of Gaussian Markov random fields," *Journal of the Royal Statistical Society, Series B, Statistical Methodology*, 63, 325–338.

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

Rue, H. and Tjelmeland, H. (2002), "Fitting Gaussian Markov random fields to Gaussian fields," *Scandinavian Journal of Statistics*, 29, 31–49.

Sjöstedt-de Luna, S. and Young, A. (2003), "The Bootstrap and Kriging Prediction Intervals," *Scandinavian Journal of Statistics*, 30, 175–192.

Werner, L. (2003), "Bayesian modelling of spatial data using Markov random fields, with application to elemental composition of forest soil," Accepted for publication in Mathematical geology.

Wikle, C., Berliner, L. M., and Cressie, N. (1998), "Hierarchical Bayesian space-time models," *Environmental and Ecological Statistics*, 5, 117 – 154.