Spatial statistics, addition to Part I. Parameter estimation and kriging for Gaussian random fields

Jo Eidsvik

Department of Mathematical Sciences, NTNU, Norway. (joeid@math.ntnu.no)

February 3, 2011

1 Introduction

The advent of remote sensing technology (satellites, radar, seismic, automatic registration sensors, etc.), along with global positioning systems and global information systems, have provided us with massive geocoded datasets. As a consequence we see increased interest in models and methods for spatial statistics. Statisticians attempt to analyze the spatial data and use it for making better decisions.

In this note we describe the frequentist version of spatial prediction and parameter estimation. The note is thus an addition to the book by Le and Zidek (2006), chapter 7-8, which gives an alternative derivation of the kriging formula (Chapter 7) and discusses Bayesian kriging (Chapter 8). Other textbooks of interest are Banerjee et al (2004), Schabenberger and Gotway (2004), and Diggle and Ribeiro (2007), which give an overview of geostatistics and hierarchical modeling, while Cressie (1993), Stein (1999) and Gaetan and Guyon (2009) are somewhat more theoretical.

In Section 2 we present the model assumption, Section 3 gives the maximum likelihood method for parameter estimation in spatial Gaussian models, Section 4 shows one way of deriving the optimal spatial predictor. Section 5 gives a synthetic example, one example with precipitation data at monitoring sites, and one from the mining industry.

2 Model assumptions

In this note we study the Gaussian random field model, which is the most common model for continuous response spatial data. We assume that the process is defined at all locations $\boldsymbol{s} \in \mathcal{D}$, where $\mathcal{D}$ denotes a continuous spatial domain in $2D$ or $3D$. The model for the response $Y(\boldsymbol{s})$ at an arbitrary site $\boldsymbol{s} \in \mathcal{D}$ is

$$Y(\boldsymbol{s}) = \boldsymbol{x}^t(\boldsymbol{s})\boldsymbol{\beta} + w(\boldsymbol{s}) + \epsilon(\boldsymbol{s}), \tag{1}$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_k)^t$ is the vector of k regression parameters and $\boldsymbol{x}^t(\boldsymbol{s}) = (x_1(\boldsymbol{s}), \dots, x_k(\boldsymbol{s}))$ is the vector of k covariates at site $\boldsymbol{s}$. The residual part is split in two parts: $w(\boldsymbol{s})$ and $\epsilon(\boldsymbol{s})$. The spatially structured residual $w(\boldsymbol{s})$ provides dependence in space, capturing the effect of unobserved covariates with spatial pattern. The non-structured spatially independent residual $\epsilon(\boldsymbol{s})$ can be interpreted as a measurement error, and is sometimes called the nugget effect.

For the independent, white noise, process we assume $\epsilon(\boldsymbol{s}) \sim N(0, \tau^2)$, for all $\boldsymbol{s}$. I.e. this is assumed to be *independent and identically didstributed (iid)* measurement noise. The covariance structure of the spatial residual $w(\boldsymbol{s})$ is typically characterized by a few parameters describing the scale and correlation range. We define the covariance by $\text{Cov}(w(\boldsymbol{s}), w(\boldsymbol{s}')) = \Sigma(\boldsymbol{s}, \boldsymbol{s}')$, and assume that $w(\boldsymbol{s})$ is a stationary proces, i.e. it only depends on the distance vector $|\boldsymbol{s} - \boldsymbol{s}'|$, and not on the locations $\boldsymbol{s}$ and $\boldsymbol{s}'$. We further assume isotropy, i.e. the covariance only depends on the absolute distance between locations, and does not depend on the direction between the locations. This entails that $\Sigma(\boldsymbol{s}, \boldsymbol{s}') = \Sigma(h)$, where $h = ||\boldsymbol{s} - \boldsymbol{s}'||$ is the absolute distance. Common examples of spatial covariance functions are the exponential $\Sigma(h) = \sigma^2 \exp(-\phi h)$, and the Matern (with smoothness $3/2$) given by $\Sigma(h) = \sigma^2(1 + \phi h)\exp(-\phi h)$, $\phi > 0$. Here, σ^2 is the variance of the spatial process at any site, while ϕ determines the decay of the covariance function. If ϕ is large, the covariance goes quickly to 0, while it decays more slowly for small ϕ. For the exponential, one sometimes parametrizes the decay by the effective spatial range $3/\phi$, since $\exp(-3) \approx 0.05$, indicating that the correlation is only 0.05 at spatial distance $h = 3/\phi$.

We observe the spatial process $Y(\boldsymbol{s})$ and the associated covariates $\boldsymbol{x}^t(\boldsymbol{s})$ at n locations $\boldsymbol{s}_1 \in \mathcal{D}$, ..., $\boldsymbol{s}_n \in \mathcal{D}$. The set of locations is called the spatial design. A regular design attempts to place the observation sites $\boldsymbol{s}_1$, ..., $\boldsymbol{s}_n$ on a regular pattern across the domain $\mathcal{D}$. More commonly, an irregular design emerges by locations that are placed in what appears to be a more random or clustered pattern. This is for instance the case with monitoring sites for precipitation, wind or air pollution, which are typically located near roads or cities. Denote the collection of data by length n vector $\boldsymbol{Y} = (Y(\boldsymbol{s}_1), \dots, Y(\boldsymbol{s}_n))^t$, and the covariates by size $n \times k$ matrix $\boldsymbol{X}$, where row i is $\boldsymbol{x}^t(\boldsymbol{s}_i)$. Under the model assumptions the distribution of the data is

$$\boldsymbol{Y} \sim N(\boldsymbol{X\beta}, \boldsymbol{C}), \quad \boldsymbol{C} = \boldsymbol{C}(\boldsymbol{\theta}) = \boldsymbol{\Sigma} + \tau^2 \boldsymbol{I}_n, \tag{2}$$

where $\boldsymbol{C}$ is a $n \times n$ covariance matrix, and $\boldsymbol{\theta} = (\sigma^2, \phi, \tau^2)$ denotes the set of covariance parameters.

3 Parameter estimation

We assume that the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are fixed, but unknown. The most common way of estimation is then the maximum likelihood method. The resulting estimators have some optimal asymptotic properties that have shown helpful. Alternatively, one could use a Bayesian approach, which entails assigning prior distributions on $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, see Chapter 8 in Le and Zidek (2006).

The Gaussian distribution in equation (2) defines the log likelihood function of the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ which becomes

$$l(\boldsymbol{Y}; \boldsymbol{\theta}, \boldsymbol{\beta}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{C}| - \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X\beta})^t \boldsymbol{C}^{-1}(\boldsymbol{Y} - \boldsymbol{X\beta}). \tag{3}$$

Note that since $(\boldsymbol{X\beta})^t = \boldsymbol{\beta}^t \boldsymbol{X}^t$, multiplying the quadratic form gives

$$l(\boldsymbol{Y}; \boldsymbol{\theta}, \boldsymbol{\beta}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{C}| - \frac{1}{2}\boldsymbol{Y}^t \boldsymbol{C}^{-1}\boldsymbol{Y} + \boldsymbol{\beta}^t \boldsymbol{X}^t \boldsymbol{C}^{-1}\boldsymbol{Y} - \frac{1}{2}\boldsymbol{\beta}^t \boldsymbol{X}^t \boldsymbol{C}^{-1}\boldsymbol{X\beta}. \tag{4}$$

The maximum likelihood estimates of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ are obtained by maximizing the log likelihood in equation (3) or (4). I.e.

$$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}) = \text{argmax}_{\boldsymbol{\beta},\boldsymbol{\theta}} l(\boldsymbol{Y}; \boldsymbol{\theta}, \boldsymbol{\beta}). \tag{5}$$

It is helpful to start by showing the optimization with respect to $\boldsymbol{\beta}$, treating $\boldsymbol{\theta}$ as fixed. And, similarly, the optimization with respect to $\boldsymbol{\theta}$, treating $\boldsymbol{\beta}$ as fixed. A final optimization scheme iterates between the two steps.

For fixed $\boldsymbol{\theta}$ the estimate of $\boldsymbol{\beta}$ is obtained analytically by differenting equation (4). We use that the derivative of a quadratic form is $d(\boldsymbol{\beta}^t \boldsymbol{A\beta}) = 2\boldsymbol{A\beta}$, while the derivative of a linear expression is $d(\boldsymbol{\beta}^t A) = A$. This gives

$$\frac{dl}{d\boldsymbol{\beta}} = \boldsymbol{X}^t\boldsymbol{C}^{-1}\boldsymbol{Y} - \boldsymbol{X}^t\boldsymbol{C}^{-1}\boldsymbol{X}\boldsymbol{\beta} = \underline{0}, \tag{6}$$

and the maximum likelihood estimator for $\boldsymbol{\beta}$ (for fixed $\boldsymbol{\theta}$) becomes

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\boldsymbol{Y};\boldsymbol{\theta}) = \boldsymbol{S}\boldsymbol{X}^t\boldsymbol{C}^{-1}\boldsymbol{Y}, \quad \boldsymbol{S} = [\boldsymbol{X}^t\boldsymbol{C}^{-1}\boldsymbol{X}]^{-1}. \tag{7}$$

By direct calculation, using $Var(\boldsymbol{AY}) = \boldsymbol{A}Var(\boldsymbol{Y})\boldsymbol{A}^t$, the variance of this estimator is

$$Var(\hat{\boldsymbol{\beta}}) = [\boldsymbol{X}^t\boldsymbol{C}^{-1}\boldsymbol{X}]^{-1} = \boldsymbol{S}. \tag{8}$$

The estimator extends that of standard least squares

$$\tilde{\boldsymbol{\beta}} = [\boldsymbol{X}^t\boldsymbol{X}]^{-1}\boldsymbol{X}^t\boldsymbol{Y}, \tag{9}$$

since, in the spatial context, the estimate would depend on the variability and correlation in the Gaussian process, not only the covariates $\boldsymbol{X}$ and the data $\boldsymbol{Y}$.

The maximum likelihood estimate of $\boldsymbol{\theta}$ can be obtained by numerical maximization of the likelihood in equation (3), treating $\boldsymbol{\beta}$ as fixed. The Newton-Raphson algorithm is one of the most common methods for this purpose. This algorithm finds the value of $\boldsymbol{\theta}$ where the gradient of the log likelihood is $\underline{0}$. Thus, it requires the first and second derivative of the log likelihood. Denote the gradient (score) by $dl/d\boldsymbol{\theta}$, and the second derivative (Hessian) by $d^2l/d\boldsymbol{\theta}^2$. The iterative Newton-Raphson method is described in Algorithm 1. The final output $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}(j)$ is the maximum likelihood estimator of $\boldsymbol{\theta}$ (for fixed $\boldsymbol{\beta}$).

Algorithm 1 Newton-Raphson algorithm for optimization of log-likelihood with respect to $\boldsymbol{\theta}$.

```
Set intial guess: $\boldsymbol{\theta}(0)$
for $j = 1...$ until converged do
    $\boldsymbol{\theta}(j) = \boldsymbol{\theta}(j-1) - [d^2l/d\boldsymbol{\theta}^2(j-1)]^{-1}dl/d\boldsymbol{\theta}(j-1).$
end for
```

The derivatives of the likelihood can be computed analytically for most covariance models. Let $\boldsymbol{Q} = \boldsymbol{C}^{-1}$ denote the inverse of the covariance matrix (called precision matrix). Define further a component of $\boldsymbol{\theta}$ by θ_r. Some useful rules for derivatives are

$$\frac{d\log|\boldsymbol{C}|}{d\theta_r} = \text{trace}(\boldsymbol{Q}\frac{\boldsymbol{C}}{d\theta_r}),$$

$$\frac{d\boldsymbol{Y}^t\boldsymbol{C}^{-1}\boldsymbol{Y}}{d\theta_r} = -\boldsymbol{Y}^t[\boldsymbol{Q}\frac{d\boldsymbol{C}}{d\theta_r}\boldsymbol{Q}]\boldsymbol{Y},$$

where the trace is defined as the sum of the diagonal elements of a matrix. The $n \times n$ matrix $d\boldsymbol{C}/d\theta_r$ has the derivative $dC(i,j)/d\theta_r$ as matrix entry (i,j), where $C(i,j) = \Sigma(\boldsymbol{s}_i, \boldsymbol{s}_j) + \tau^2 I(\boldsymbol{s}_i = \boldsymbol{s}_j)$. By using the above rules for derivative of a determinant and a quadratic form, the first derivative (score) of the log likelihood with respect to element θ_r becomes

$$\begin{aligned} \frac{dl}{d\boldsymbol{\theta}_r} &= -\frac{1}{2}\text{trace}(\boldsymbol{Q}\frac{d\boldsymbol{C}}{d\theta_r}) \\ &+ \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^t\boldsymbol{Q}\frac{d\boldsymbol{C}}{d\theta_r}\boldsymbol{Q}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}). \end{aligned} \tag{10}$$

The second derivative (Hessian) with respect to two components θ_r and θ_s is

$$\begin{aligned} \frac{d^2 l}{d\theta_r d\theta_s} &= -\frac{1}{2}\text{trace}(\boldsymbol{Q}\frac{d^2\boldsymbol{C}}{d\theta_r d\theta_s}) + \frac{1}{2}\text{trace}(\boldsymbol{Q}\frac{d\boldsymbol{C}}{d\theta_s}\boldsymbol{Q}\frac{d\boldsymbol{C}}{d\theta_r}) \\ &- \boldsymbol{Y}^t\boldsymbol{Q}\frac{d\boldsymbol{C}}{d\theta_s}\boldsymbol{Q}\frac{d\boldsymbol{C}}{d\theta_r}\boldsymbol{Q}\boldsymbol{Y} + \frac{1}{2}\boldsymbol{Y}^t\boldsymbol{Q}\frac{d^2\boldsymbol{C}}{d\theta_s d\theta_r}\boldsymbol{Q}\boldsymbol{Y}, \end{aligned} \tag{11}$$

where we use the kernel rule for each of the two expressions at the right of equation (10). Moreover, we use that $d\text{trace}(A(\boldsymbol{\theta})) = \text{trace}(dA(\boldsymbol{\theta}))$.

The second derivative can be quite unbalanced because of the randomness induced by $\boldsymbol{Y}$. A numerically more stable expression is obtained by taking the expected value of the random observations $\boldsymbol{Y}$. The expectation of a quadratic form is

$$E(\boldsymbol{Y}^t A \boldsymbol{Y}) = \text{trace}(A\text{Var}(\boldsymbol{Y})), \quad \textit{assuming } E(\boldsymbol{Y}) = \underline{0}.$$

The expectation of the Hessian is a simpler expression because several terms cancel:

$$\begin{aligned} E(\frac{d^2 l}{d\theta_r d\theta_s}) &= -\frac{1}{2}\text{trace}(\boldsymbol{Q}\frac{d^2\boldsymbol{C}}{d\theta_r d\theta_s}) + \frac{1}{2}\text{trace}(\boldsymbol{Q}\frac{d\boldsymbol{C}}{d\theta_s}\boldsymbol{Q}\frac{d\boldsymbol{C}}{d\theta_r}) \\ &- \text{trace}(\boldsymbol{Q}\frac{d\boldsymbol{C}}{d\theta_s}\boldsymbol{Q}\frac{d\boldsymbol{C}}{d\theta_r}) + \frac{1}{2}\text{trace}(\boldsymbol{Q}\frac{d^2\boldsymbol{C}}{d\theta_s d\theta_r}) \\ &= -\frac{1}{2}\text{trace}(\boldsymbol{Q}\frac{d\boldsymbol{C}}{d\theta_s}\boldsymbol{Q}\frac{d\boldsymbol{C}}{d\theta_r}). \end{aligned} \tag{12}$$

We next use the expected Hessian in the optimization algorithm. This is often referred to as the Fisher scoring algorithm.

The final algorithm for obtaining the maximum likelihood estimator of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ is in Algorithm 2. The score $dl/d\boldsymbol{\theta}$ and the expected Hessian $E(d^2l/d\boldsymbol{\theta}^2)$ are evaluated at $\boldsymbol{\beta}(j)$ and $\boldsymbol{\theta}(j-1)$.

Algorithm 2 Fisher scoring algorithm for optimization of log-likelihood with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$.

```
Set intial guess: $\boldsymbol{\theta}(0)$
for $j = 1...$ until converged do
    $\boldsymbol{C} = \boldsymbol{C}(\boldsymbol{\theta}(j-1))$.
    $\boldsymbol{\beta}(j) = [\boldsymbol{X}^t\boldsymbol{C}^{-1}\boldsymbol{X}]^{-1}\boldsymbol{X}^t\boldsymbol{C}^{-1}\boldsymbol{Y}$
    for $r = 1, 2, 3$, compute $dl/d\theta_r$ in equation (10).
    for $r, s = 1, 2, 3$, compute $E(d^2l/\theta_r\theta_s)$ in equation (12).
    $\boldsymbol{\theta}(j) = \boldsymbol{\theta}(j-1) - [E(d^2l/d\boldsymbol{\theta}^2)]^{-1}dl/d\boldsymbol{\theta}$.
end for
```

The outputs $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}(j)$ and $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}(j)$ are the maximum likelihood estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. The Fisher-scoring method in Algorithm 2 usually reaches machine precision within a few iterations. Typically, $5-10$ iterations are enough for convergence. It is convenient to parameterize the model parameters $\boldsymbol{\theta}$ on the real line. For instance, using $\boldsymbol{\theta} = (\log\sigma^{-2}, \log\phi, \log\tau^{-2})$ tends to give a robust optimization. The initial guess $\boldsymbol{\theta}(0)$ can be specified from an empirical estimate of the variogram. The variogram is defined by

$$\gamma(h) = \frac{1}{2}\text{Var}(Y(\boldsymbol{s}) - Y(\boldsymbol{s}')), \quad ||\boldsymbol{s} - \boldsymbol{s}'|| = h. \tag{13}$$

An empirical estimate is obtained by binning the distance h into intervals $h_2 - h_1$, $h_3 - h_2$, etc., and for every interval constructing an estimate $\gamma^*(h)$ by averaging the variances of response differences for all pair of sites $\boldsymbol{s}$, $\boldsymbol{s}'$ wihtin the right binning interval distance, $B_i = \{h_i < h < h_i + \delta\}$, i.e.

$$\gamma^*(h_i) = \frac{1}{2N_i} \sum_{\boldsymbol{s},\boldsymbol{s}' \in B_i} (Y(\boldsymbol{s}) - Y(\boldsymbol{s}'))^2, \quad (\boldsymbol{s} - \boldsymbol{s}') \in B_i. \tag{14}$$

Here, N_i is the number of pairs $\boldsymbol{s}$ and $\boldsymbol{s}'$ with pair $(\boldsymbol{s}, \boldsymbol{s}') \in B_i$.

Asymptotically, as the number of data $n \to \infty$, the maximum likelihood estimator is unbiased and Gaussian distributed with variance obtained from the inverse Hessian $E[d^2l/d\boldsymbol{\theta}^2]^{-1}$ and $E[d^2l/d\boldsymbol{\beta}^2]^{-1} = \boldsymbol{S}$. The speed of convergence to this asymptotic result depends on the spatial design.

4 Kriging: Optimal spatial prediction

Kriging is presented for instance in Chapters 7-8 of Le and Zidek (2006). It is a method for spatial prediction. Under expected square loss, the optimal spatial predictor of $Y(\boldsymbol{s}_0)$, given $\boldsymbol{Y}$, is $E(Y(\boldsymbol{s}_0)|\boldsymbol{Y})$. This must hold since for any other predictor $g(\boldsymbol{Y})$, for arbitrary function g, we have

$$\begin{aligned} E(Y(\boldsymbol{s}_0) - g(\boldsymbol{Y}))^2 &= E(Y(\boldsymbol{s}_0) - E(Y(\boldsymbol{s}_0)|\boldsymbol{Y}) + E(Y(\boldsymbol{s}_0)|\boldsymbol{Y}) - g(\boldsymbol{Y}))^2 \\ &= E(Y(\boldsymbol{s}_0) - E(Y(\boldsymbol{s}_0)|\boldsymbol{Y}))^2 + E(E(Y(\boldsymbol{s}_0)|\boldsymbol{Y}) - g(\boldsymbol{Y}))^2 \\ &+ 2E(Y(\boldsymbol{s}_0) - E(Y(\boldsymbol{s}_0)|\boldsymbol{Y}))E(E(Y(\boldsymbol{s}_0)|\boldsymbol{Y}) - g(\boldsymbol{Y})) \\ &= E(Y(\boldsymbol{s}_0) - E(Y(\boldsymbol{s}_0)|\boldsymbol{Y}))^2 + E(E(Y(\boldsymbol{s}_0)|\boldsymbol{Y}) - g(\boldsymbol{Y}))^2. \end{aligned} \tag{15}$$

Here, the cross term cancels since $E(E(Y(\boldsymbol{s}_0)|\boldsymbol{Y})) = E(Y(\boldsymbol{s}_0))$. The last term in equation (15) is 0 for $g(\boldsymbol{Y}) = E(Y(\boldsymbol{s}_0)|\boldsymbol{Y})$, but positive for any other g function. The kriging predictor is the best *linear* predictor; $g(\boldsymbol{Y}) = \boldsymbol{\alpha}^t\boldsymbol{Y}$, with weights $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^t$. When we assume a Gaussian model, the conditional mean $E(Y(\boldsymbol{s}_0)|\boldsymbol{Y})$ is linear in $\boldsymbol{Y}$ and identical to the kriging predictor. For non-Gaussian data, kriging would still provide the best *linear* predictor, but there exists a non-linear predictor that has smaller expected square loss.

We want to predict the response $\boldsymbol{Y}_0 = (Y(\boldsymbol{s}_1^0), \ldots, Y(\boldsymbol{s}_m^0))^t$ at m prediction sites $\boldsymbol{s}_j^0$, $j = 1, \ldots, m$, given data $\boldsymbol{Y} = (Y(\boldsymbol{s}_1), \ldots, Y(\boldsymbol{s}_n))^t$ at n observation sites. The joint distribution of $\boldsymbol{Y}_0$ and $\boldsymbol{Y}$ is

$$\begin{bmatrix} \boldsymbol{Y}_0 \\ \boldsymbol{Y} \end{bmatrix} \sim N\left[\begin{pmatrix} \boldsymbol{X}_0 \\ \boldsymbol{X} \end{pmatrix}\boldsymbol{\beta}, \begin{pmatrix} \boldsymbol{C}_0 & \boldsymbol{C}_{0,\cdot} \\ \boldsymbol{C}_{0,\cdot}^t & \boldsymbol{C} \end{pmatrix}\right]. \tag{16}$$

Here, the $m \times k$ matrix $\boldsymbol{X}_0$ contains the covariates at the prediction sites. Moreover, the $m \times m$ matrix $\boldsymbol{C}_0$ is the covariance between the responses at all prediction sites, the $m \times n$ matrix $C_{0,\cdot}$ gives the covariances between the m response variables at prediction sites and the n observation sites. The covariances are of course dependent on the parameters $\boldsymbol{\theta}$, but note that the cross-covariance matrix $C_{0,\cdot}$ does not depend on τ^2.

When the joint distribution is Gaussian, the conditional distribution of $\boldsymbol{Y}_0$, given $\boldsymbol{Y}$ (and for fixed $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$) is also Gaussian. The length m vector of conditional means is

$$\begin{aligned} E(\boldsymbol{Y}_0|\boldsymbol{Y}) &= \boldsymbol{X}_0\boldsymbol{\beta} + \boldsymbol{C}_{0,\cdot}\boldsymbol{C}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}), \\ &= \boldsymbol{C}_{0,\cdot}\boldsymbol{C}^{-1}\boldsymbol{Y} + (\boldsymbol{X}_0 - \boldsymbol{C}_{0,\cdot}\boldsymbol{C}^{-1}\boldsymbol{X})\boldsymbol{\beta}, \end{aligned} \tag{17}$$

and the $m \times m$ conditional covariance is

$$\text{Var}(\boldsymbol{Y}_0|\boldsymbol{Y}) = \boldsymbol{C}_0 - \boldsymbol{C}_{0,\cdot}\boldsymbol{C}^{-1}\boldsymbol{C}_{0,\cdot}^t, \tag{18}$$

and the conditional variances are defined by the m diagonal elements of this matrix. For sites that are close to other data, there is much correlation in $\boldsymbol{C}_{0,\cdot}$, and the conditioning will reduce the variance $\boldsymbol{C}_0$ a lot. Sites that are further from data, will get a smaller reduction. The reduction of uncertainty also depends on the clustering of the observation sites according to $\boldsymbol{C}^{-1}$. Two data at almost the same location will not count as double information, since the two observations will be so correlated. It is remarkable that the variances in expression (18) do not depend on the data. This property holds only for the Gaussian distribution.

The formula for the conditional mean and variance can be written in many mathematically equivalent ways. For instance, one can get equivalent formulas by using the inverse of the covariance matrix, called the precision matrix. We do this exercise next. Denote the joint covariance matrix by

$$\tilde{\boldsymbol{C}} = \begin{bmatrix} \boldsymbol{C}_0 & \boldsymbol{C}_{0,\cdot} \\ \boldsymbol{C}_{0,\cdot}^t & \boldsymbol{C} \end{bmatrix}. \tag{19}$$

Denote further the joint precision matrix by $\tilde{\boldsymbol{Q}} = \tilde{\boldsymbol{C}}^{-1}$. Since $\tilde{\boldsymbol{Q}}\tilde{\boldsymbol{C}} = \boldsymbol{I}_{m+n}$, the block structure of the precision matrix equals

$$\tilde{\boldsymbol{Q}} = \begin{bmatrix} \boldsymbol{Q}_0 & \boldsymbol{Q}_{0,\cdot} \\ \boldsymbol{Q}_{0,\cdot}^t & \boldsymbol{Q} \end{bmatrix}, \quad \begin{matrix} \boldsymbol{Q}_0 = [\boldsymbol{C}_0 - \boldsymbol{C}_{0,\cdot}\boldsymbol{C}^{-1}\boldsymbol{C}_{0,\cdot}^t]^{-1} & \boldsymbol{Q}_{0,\cdot} = -\boldsymbol{Q}_0\boldsymbol{C}_{0,\cdot}\boldsymbol{C}^{-1} \\ \boldsymbol{Q} = [\boldsymbol{C} - \boldsymbol{C}_{0,\cdot}^t\boldsymbol{C}_0^{-1}\boldsymbol{C}_{0,\cdot}]^{-1} & \boldsymbol{Q}_{0,\cdot}^t = -\boldsymbol{Q}\boldsymbol{C}_{0,\cdot}^t\boldsymbol{C}_0^{-1}. \end{matrix} \tag{20}$$

We can next simplify the quadratic form of the conditional of $\boldsymbol{Y}_0$, given $\boldsymbol{Y}$. First, the joint density of $\boldsymbol{Y}$ and $\boldsymbol{Y}_0$ can be written

$$\begin{aligned} \pi(\boldsymbol{Y}, \boldsymbol{Y}_0) &= \exp\{-\frac{1}{2}\log|\tilde{\boldsymbol{C}}| - \frac{1}{2}(\boldsymbol{Y}_0 - \boldsymbol{X}_0\boldsymbol{\beta})^t\boldsymbol{Q}_0(\boldsymbol{Y}_0 - \boldsymbol{X}_0\boldsymbol{\beta}) \\ &- \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^t\boldsymbol{Q}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) - (\boldsymbol{Y}_0 - \boldsymbol{X}_0\boldsymbol{\beta})^t\boldsymbol{Q}_{0,\cdot}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})\}. \end{aligned} \tag{21}$$

Next, the conditional of $\boldsymbol{Y}_0$ given $\boldsymbol{Y}$ is proportional to this joint

$$\begin{aligned} \pi(\boldsymbol{Y}_0|\boldsymbol{Y}) &= \frac{\pi(\boldsymbol{Y}, \boldsymbol{Y}_0)}{\pi(\boldsymbol{Y})} \propto \pi(\boldsymbol{Y}, \boldsymbol{Y}_0) \\ &\propto \exp\{-\frac{1}{2}(\boldsymbol{Y}_0 - \boldsymbol{X}_0\boldsymbol{\beta})^t\boldsymbol{Q}_0(\boldsymbol{Y}_0 - \boldsymbol{X}_0\boldsymbol{\beta}) - (\boldsymbol{Y}_0 - \boldsymbol{X}_0\boldsymbol{\beta})^t\boldsymbol{Q}_{0,\cdot}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})\} \\ &\propto \exp\{-\frac{1}{2}\boldsymbol{Y}_0^t\boldsymbol{Q}_0\boldsymbol{Y}_0^t + \boldsymbol{Y}_0^t[\boldsymbol{Q}_0\boldsymbol{X}_0\boldsymbol{\beta} - \boldsymbol{Q}_{0,\cdot}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})]\}, \end{aligned} \tag{22}$$

where we ignore all normalizing constants and simply view the conditional as a function of the $\boldsymbol{Y}_0$ variable of interest. We finally equate the quadratic part $\boldsymbol{Y}_0^t\text{Var}^{-1}(\boldsymbol{Y}_0|\boldsymbol{Y})\boldsymbol{Y}_0$ and the linear part $\boldsymbol{Y}_0^t\text{Var}^{-1}(\boldsymbol{Y}_0|\boldsymbol{Y})E(\boldsymbol{Y}_0|\boldsymbol{Y})$ of $\boldsymbol{Y}_0$ in the exponent of the Gaussian. We use equation (20) to get

$$\begin{aligned} \text{Var}(\boldsymbol{Y}_0|\boldsymbol{Y}) &= \boldsymbol{Q}_0^{-1} = \boldsymbol{C}_0 - \boldsymbol{C}_{0,\cdot}\boldsymbol{C}^{-1}\boldsymbol{C}_{0,\cdot}^t \\ E(\boldsymbol{Y}_0|\boldsymbol{Y}) &= \boldsymbol{Q}_0^{-1}[\boldsymbol{Q}_0\boldsymbol{X}_0\boldsymbol{\beta} - \boldsymbol{Q}_{0,\cdot}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})] \\ &= \boldsymbol{X}_0\boldsymbol{\beta} + \boldsymbol{C}_{0,\cdot}^t\boldsymbol{C}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}). \end{aligned} \tag{23}$$

Now, we recognize the conditional mean and variance of equation (17) and (18). This exercise indeed shows that the conditional $\boldsymbol{Y}_0$, given $\boldsymbol{Y}$ is Gaussian. It also shows that how the expressions based on precision matrices relate to that of covariance matrices.

In the derivations so far in this section we have treated the $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ parameters as fixed. It is common practice to include the uncertainty of the regression estimator $\hat{\boldsymbol{\beta}}$ into the prediction variance, while $\hat{\boldsymbol{\theta}}$ is just plugged into the covariance matrices. We can include the uncertainty of $\hat{\boldsymbol{\beta}}$ by using the formula for double variance. This gives

$$\begin{aligned}\text{Var}(\boldsymbol{Y}_0|\boldsymbol{Y}) &= \text{Var}_{\hat{\boldsymbol{\beta}}}[E_{\boldsymbol{Y}_0}(\boldsymbol{Y}_0|\boldsymbol{Y},\boldsymbol{\beta})] + E_{\hat{\boldsymbol{\beta}}}[\text{Var}_{\boldsymbol{Y}_0}(\boldsymbol{Y}_0|\boldsymbol{Y},\boldsymbol{\beta})] \\ &= \text{Var}_{\hat{\boldsymbol{\beta}}}[\boldsymbol{C}_{0,\cdot}\boldsymbol{C}^{-1}\boldsymbol{Y} + (\boldsymbol{X}_0 - \boldsymbol{C}_{0,\cdot}\boldsymbol{C}^{-1}\boldsymbol{X})\boldsymbol{\beta}] + E_{\hat{\boldsymbol{\beta}}}[\boldsymbol{C}_0 - \boldsymbol{C}_{0,\cdot}\boldsymbol{C}^{-1}\boldsymbol{C}_{0,\cdot}^t] \\ &= (\boldsymbol{X}_0 - \boldsymbol{C}_{0,\cdot}\boldsymbol{C}^{-1}\boldsymbol{X})\boldsymbol{S}(\boldsymbol{X}_0 - \boldsymbol{C}_{0,\cdot}\boldsymbol{C}^{-1}\boldsymbol{X})^t + \boldsymbol{C}_0 - \boldsymbol{C}_{0,\cdot}\boldsymbol{C}^{-1}\boldsymbol{C}_{0,\cdot}^t.\end{aligned} \tag{24}$$

The prediction variance now gets an additive term compared with equation (23), accounting for not knowing $\boldsymbol{\beta}$ exactly, but rather estimating it from the data. When there is a lot of data, the variance of $\hat{\boldsymbol{\beta}}$, denoted $\boldsymbol{S}$, is small, and the term adds little to the prediction uncertainty.

Under the assumptions we have made, the prediction distribution is Gaussian. A 95 % prediction interval for the response at site $\boldsymbol{s}_0$ is obtained by

$$\left\{E(Y(\boldsymbol{s}_0)|\boldsymbol{Y}) - 1.96\sqrt{\text{Var}(Y(\boldsymbol{s}_0)|\boldsymbol{Y})}, E(Y(\boldsymbol{s}_0)|\boldsymbol{Y}) + 1.96\sqrt{\text{Var}(Y(\boldsymbol{s}_0)|\boldsymbol{Y})}\right\}. \tag{25}$$

5 Examples

5.1 Synthetic data

We define a spatial domain $\mathcal{D} = [(0,1) \times (0,1)]$, and generate a spatial design with $n = 500$ sites. The selected design is of a regular + random infill type. Figure 1 shows the observation sites (dots) along with some randomly located prediction sites (crosses). By generating 500 replicates of data, each of size $n = 500$, we can compare the maximum likelihood estimates with the true values used to generate the data.

We use covariate $x^t(\boldsymbol{s}) = (1, s_1)$, where s_1 is the east coordinate of the site $\boldsymbol{s}$. The true regression parameters are set to $\beta = (-1,1)^t$. We use a Matern (3/2) covariance model $\Sigma(h) = \sigma^2(1+\phi h)\exp(-\phi h)$. We use a parameterization with $\theta_1 = \log(\sigma^{-2})$, $\theta_2 = \log(\phi)$, $\theta_3 = \log\tau^{-2}$. The true parameters are set to $\boldsymbol{\theta} = (0, 2.5, 2)$.

Figure 2 shows a histogram of the estimated parameters for the 500 replicates of data. The covariance parameters $\hat{\theta}$ are to the left, while the regression parameters $\hat{\beta}$ are to the right. The true values are indicated by dots along the first axis. Note that the true values are captured almost symmetrically by the histogram. Asymptotically, as the number of data $n \to \infty$, the variability in the histogram should reflect the asymptotic variance given by the inverse Hessian of the likelihood. For the covariance parameters $\boldsymbol{\theta}$, the diagonal of $[d^2l/d\boldsymbol{\theta}^2]^{-1}$ is 0.5, 0.4 and 0.3 for the three components, and this matches the histograms quite well, indicating that the asymptotic result is not so farfetched when $n = 500$. If we repeated this exercise for a larger n, the histograms would become narrower.

Figure 3 displays the predictions along with the true values generated at the prediction sites. This is done for every replicate, and the number of dots should be $50 \times 500 = 25.000$. We can display this only because we generate the data ourselves. In a real application it is common to

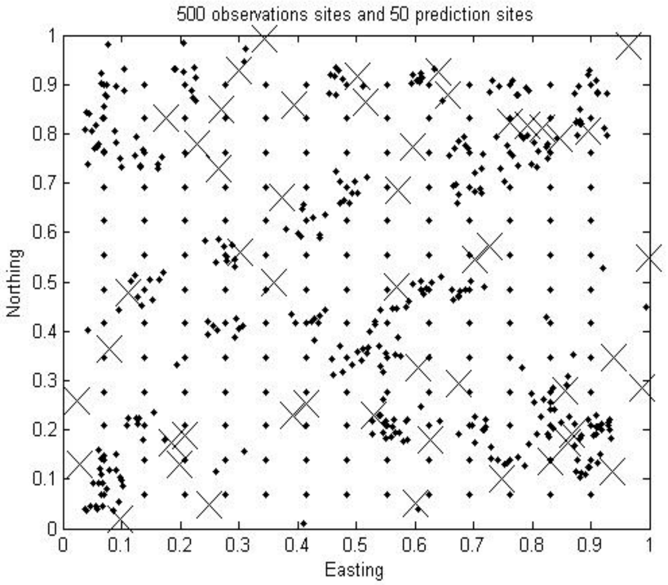

Figure 1: Design with 500 observation sites (dots) and 50 prediction sites (crosses).

use a hold-out set of the data, and then use the remaining data to predict the hold-out set. In this way one can study the performance of predictions, and compare different methods of predicting. Here, the correlation between the predictions and the true values is quite high. The average mean square prediction error (MSPE) is 0.25. Based on the prediction and the prediction variance, we can compute a 95 percent coverage interval, at all sites, for every replicate. We can check if the true data is within this interval. For this example the data is within the interval about 94 percent of the time, indicating a very accurate prediction coverage.

5.2 Precipitation data

We analyze a precipitation data set from middle Norway, Sept-Oct 2006. The number of days with rainfall is registered at 92 monitoring stations in the region (Figure 4). In this Fall of 2006 Norway experienced record high precipitation. For instance, Bergen registered rainfall 90 days in a row. Our goal here is to use the data from middle Norway to estimate parameters and predict at a regular grid of prediction sites covering the domain. Note that the data are counts $\{0,\ldots,61\}$, but since the average is about 40, and none of the data are extreme, we use a Gaussian model. We use a constant mean β in the modeling, and a Matern (3/2) covariance function $C(h) = \sigma^2(1 + \phi h)\exp(-\phi h) + I(h=0)\tau^2$.

The Fisher-scoring algorithm is initiated by an a priori guess. It converges within a few iterations. After 5 iterations the difference $||\boldsymbol{\beta}(5) - \boldsymbol{\beta}(4)|| = 0.0004$ and $||\boldsymbol{\theta}(5) - \boldsymbol{\theta}(4)|| = 0.0007$. The final estimates are $\hat{\beta} = 41$ and $\hat{\boldsymbol{\theta}} = (-2.66, 1.17, -3.07)$, which corresponds to $\sigma^2 = 3.8^2$, $\phi = 3.2$, and $\tau^2 = 4.6^2$.

Figure 5 shows the prediction and prediction standard deviations. Note that the number of days with rain in this south-central part is predicted to 35, and a 95 % prediction interval is about

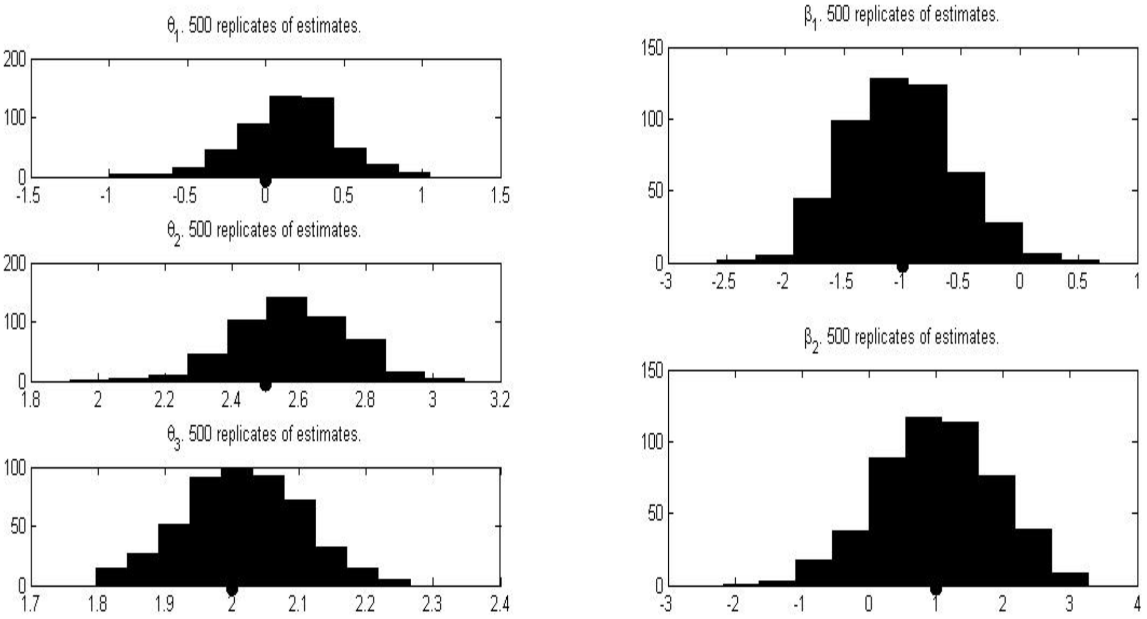

Figure 2: Left) Parameter estimates of the covariance parameters for 500 replicates of data. Right) Parameter estimates of the regression parameters for 500 replicates of data.

$(25, 45)$, based on the standard deviations (Figure 5 Right)). For the north-western parts the prediction is 47, with 95 % prediction interval about $(37, 57)$. Depending on the direction of wind it rains in the eastern or western part, but the central south region seems to have relatively few days of rain.

5.3 Mining data

This dataset contains a particular mineral content in a mine in Norway. The number of data is $n = 1466$, and the observation sites are made along boreholes. The locations are in (east, north, depth), denoted $\boldsymbol{s}_i = (s_{i1}, s_{i2}, s_{i3})$, $i = 1, \dots, 1466$. In Figure 6 Left) we display the three dimensional locations of the measurements. The data are log transformed to obtain near-Gaussian variables. We want to estimate parameters and predict the mineral content at unobserved locations.

The Fisher-scoring algorithm is initiated at values indicated by the variogram displayed in Figure 6 Right). Based on this variogram it appears as if the nugget is about $\tau^2 = 0.03$, the sill is about $\tau^2 + \sigma^2 = 0.1$, and the effective correlation range is about 100 m. In the optimization, the difference in the covariance parameters $||\boldsymbol{\theta}(j) - \boldsymbol{\theta}(j-1)||$ at subsequent steps goes like: 0.0796, 0.0060, 0.0019, 0.0007, 0.0002, 0.0001. At the last iteration we have $\hat{\theta} = (2.6, -2.9, 3.6)$, which corresponds to $\sigma^2 = 0.27^2$, effective range about 60m and $\tau^2 = 0.17^2$. The convergence speed for the scalar regression parameter is about the same.

In Figure 7 we show the prediction and prediction standard deviation at depth 79m. This is the average depth of the geological formation. Some zones show a much higher level (1.5) of the mineral content, while others have a poorer quality (0.7). In a real application these levels of

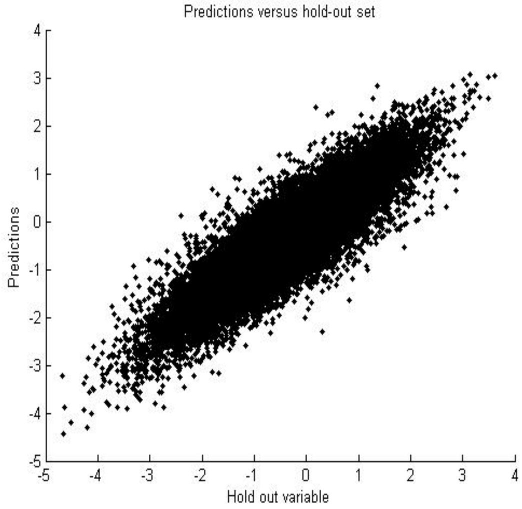

Figure 3: Predictions (second axis) plotted against the true data (first axis) at the same observation sites. The number of prediction sites is 50, and the plot displays all of them for every replicate $1, \dots, 500$ of data.

content must be transformed to the attributes required for decision making. The prediction standard deviation ranges from 1 to 2. Note that we can have a relatively high uncertainty at locations where there appears to be data (see central domain with standard deviation near 2). This is an effect of the $3D$ measurement locations. The depth of the observations (dots) can be quite different from the 79m we display here.

6 References

Banerjee, S., Carlin, B.K. and Gelfand, A.E. (2004). Hierarchical modeling and analysis for spatial data. Chapman & Hall.

Cressie, N. (1993). Statistics for spatial data. Wiley.

Diggle, P.J. and Ribeiro, P.J. (2007). Model-based geostatistics. Springer.

Gaetan, C. and Guyon, X. (2009). Spatial statistics and modeling. Springer.

Le, N.D. and Zidek, J.V., (2006). Statistical analysis of environmental space-time processes. Springer.

Schabenberger, O. and Gotway, C.A. (2004). Statistical methods for spatial data analysis. Chapman & Hall.

Stein, M.L. (1999). Interpolation of spatial data: Some theory for Kriging. Springer.

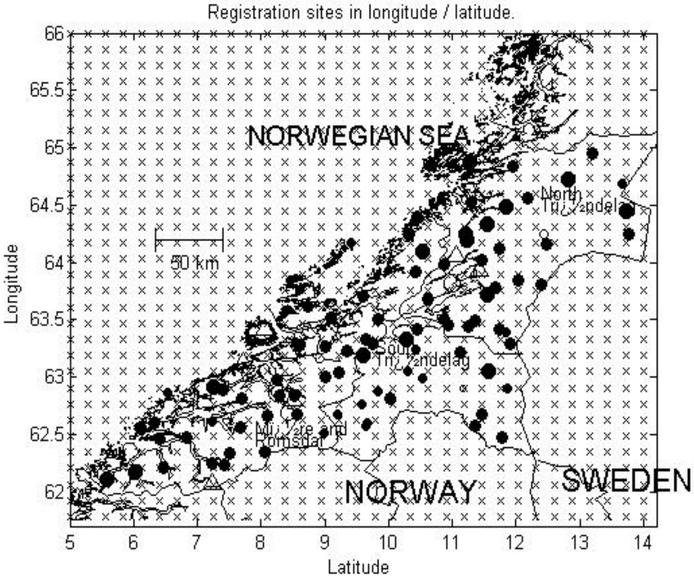

Figure 4: Registration sites (dots) for precipitation in the middle Norway. Prediction sites are on a regular grid (crosses) covering the domain.

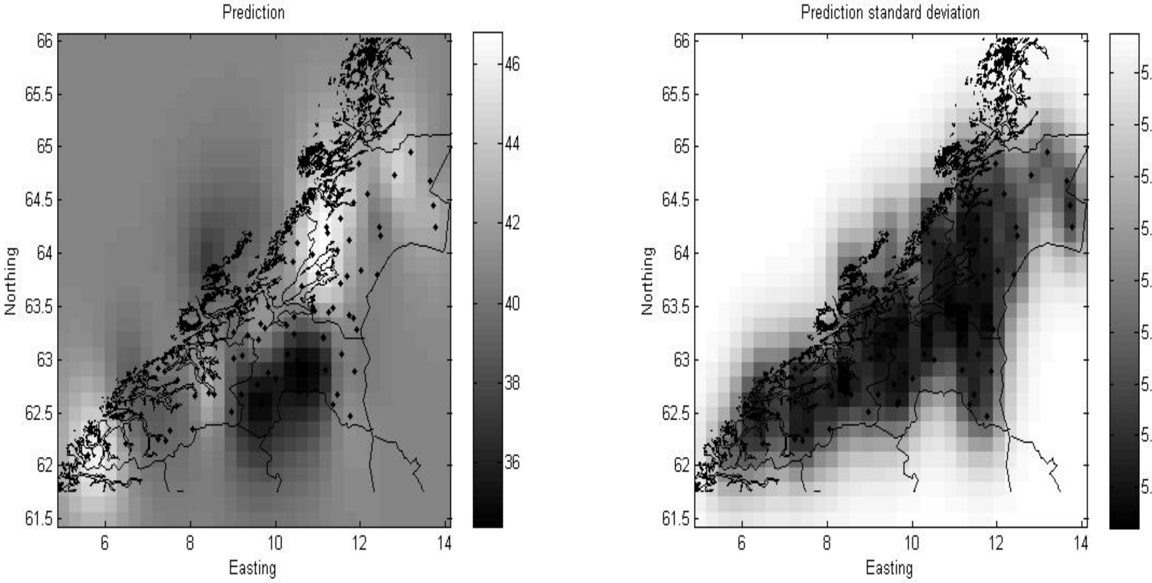

Figure 5: Left) Kriging prediction of the number of days with rain. Right) Kriging standard deviation for the number of days with rain.

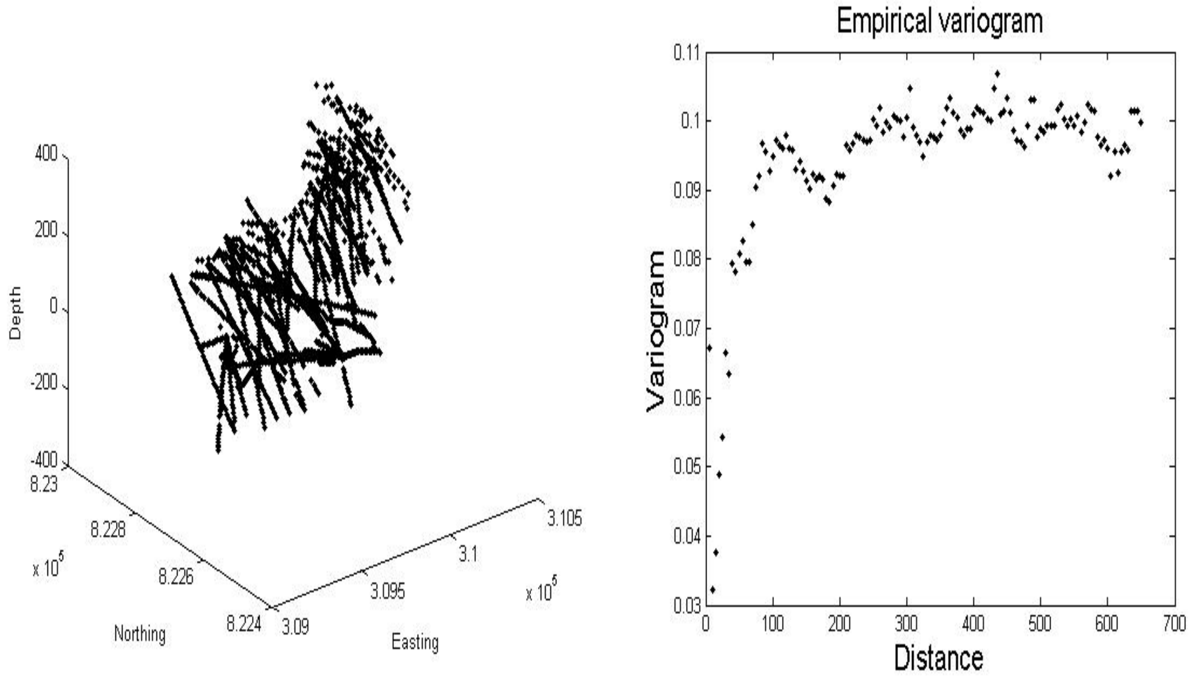

Figure 6: Left) Observation sites in a mine in Norway. The number of data is $n = 1466$ collected in several boreholes. Right) Empirical estimate of the variogram constructed by bins every 5m.

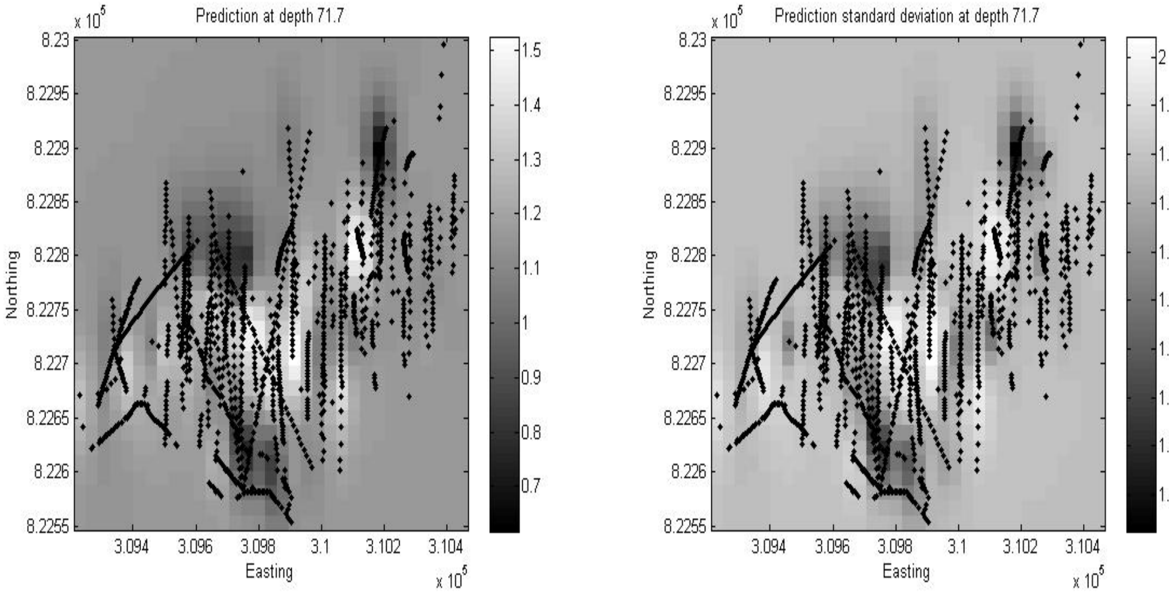

Figure 7: Left) Predictions of mineral content in a mine in Norway. Right) Prediction standard deviation of mineral content in a mine in Norway.