Department of Mathematical Sciences

Bachelor Project

Predictive Analysis of an Attribute Modelled by a Random Field

Filip Schjerven

May 31, 2017

Summary

Contents

List of Figures

1 Introduction

This report is written in coherence with the bachelor project for my studies within the Department of Mathematical Sciences at NTNU. The overall focus of this project is a case that will be handled with tools from spatial statistics.

Spatial statistics refers to the application of statistical methods in a spatial setting. Statistics in itself is an extremely useful mathematical tool to quantify uncertainty. Uncertainty is something that we all face in a day to day basis in our daily lives, but perhaps even more so when viewing the world through scientific research. How can we know that our measurements, and thus how the world is described, are certain? In this sense, Cressie & Wikle refers to statistics as the "Science of Uncertainty". Statistics has many ways of dealing with uncertainty, and one of them is that is interprets uncertainty as a measure of *variability*. Other interpretations could also be used, like *entropy*. By modelling uncertainty via variability, one may model the total variability by the variability in measurements, the model used and due to the uncertainty of parameters in the model [4]. All this happens with the focus on acquiring information from the data that are to be modelled. In spatial statistics, all this happens in a spatial setting, i.e. the data that are to be analyzed via their topological, geometric or geographic attributes.

A typical model thats used in spatial statistics is a *Gaussian Random Field* (GRF). Informally, a GRF is a collection of data that have been given a multivariate Gaussian joint probability function. The random field that it is 'attached' to typically defines the relation between the variables in some quantifiable way. A typical example (that will also be used extensively) is spatial location. As the joint probability function is multivariate Gaussian, all subsets of variables that are included in the GRF is also multivariate Gaussian with their attributes still defined by the relation inherited from the GRF. Due to the many useful properties of a multivariate Gaussian distributed variable, the GRF is a useful way to model data.

In order to model data in a easy to interpret way, a model for GRF is often used to describe the various parts of the model. An example that wil be used in this project is the *Bayes hierarchical model* (BHM). The exact formulation of a BHM is presented later, but summarized it consists of a 1) data model for specifying the data acquirement, 2) a process model for describing the latent process that is typically of interest and lastly any parameter-priors that are used in the Bayesian setting in 3) prior model. By distinguishing between the three, interpretability is made easy and one gets an intuitive way to understand the data acquirement how it relates to the latent process that is often of interest.

As the latent process is of interest, making the correct data acquirement is crucial in order to achieve sufficient information from the data. In spatial statistics this is important, as the data must then be sampled at the correct locations. A set of measurement-locations is called a *sampling design*, or simply *design*, in spatial setting. Choosing the correct sampling design may be crucial for achieving the desired information, and must be evaluated with possible limitations as cost and precision of measurements. Constructing sampling designs happens typically in two settings, *retroactive design* and *proactive design*. A retrospective design will focus on altering the existing design on the basis of new data, while a proactive design will focus on constructing the optimal design on the basis of prior and estimated posterior data.

This paper is organized into 5 sections, including the introduction and conclusion. Section 2 introduces the GRF by going through some basics and relevant properties of random fields, covariance functions and Gaussian processes. A short summary of *Generalized Least Squares* (GLS) and hierarchical models is also part of this section, as it is to be used in the case scenario. The posterior distribution needed for the case scenario in section 4 is also derived. Section 3 introduces the notion of spatial sample design and how it may be related to the model in a quantifiable way. Some examples of path planning criterias is presented along with a short summary of the situation it were used in. Section 4 presents a case that the theory of section 2 and 3 will be applied upon. Input for the case and chosen modelspecification is presented, along with some case-specific procedures for estimating parameters based on the input data. Results follow in the last subsection of section 4, before the conclusion round up the main points of this paper.

Notation

In this paper I'll denote matrices with a capital symbol. Vectors will be denoted by a arrow over its lower-case symbol. An exception will be in the case of a vector of stochastic variables, in that case, a large letter is used. Throughout the paper, only the symbols A, B, Z, Z_a, Y is used for this. Lastly, a matrix $F(\vec{s})$ will be introduced later and used extensively. At some places the notation $Z(\vec{s})/Y(\vec{s})$ will be used in place of $F(\vec{s}) \cdot Z$ / $F(\vec{s}) \cdot Y$. To summarize, with X a matrix,

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots \\ x_{21} & x_{22} & \dots \\ \vdots & \vdots & \ddots \\ & & \end{bmatrix}$$

$\vec{\beta}$ a vector

$$\vec{\beta} = \begin{bmatrix} \beta_0 & \beta_1 & \dots \end{bmatrix}$$

where x_{ii} and β_i are values $\forall$ i, and Y as a vector of stochastic variables

$$Y = \begin{bmatrix} Y_0 & Y_1 & \dots \end{bmatrix}$$

where Y_i is a stochastic variable $\forall$ i

2 Spatial Gaussian Random Fields

2.1 Random Fields

A random field is a set of random variables $\vec{Y}$ that has a distribution function

$$F(Y(s_1) \leq y_1, \ldots, Y(s_n) \leq y_n) \tag{1}$$

for any number for any number n, and for any combination of coordinates s_1, ..., s_n. A random field has typically some attribute that connects the random variables in $\vec{Y}$. Some examples of such attributes may be physical location, placement in a graph or ordering in time.

Depending on the properties of the distribution function and the setting in which it is defined, we have many different types of random fields, e.g. Markov Random Fields, Gibbs Random Fields and Conditional Random Fields, all of which may again be used with different attributes connecting the variables of the field.

For this particular project, the main focus will revolve around a random field of type *Gaussian Random Field* (GRF). A GRF is set up so that the variables $\vec{Y}$ is governed by a *Gaussian process*. In addition, the Gaussian process- variance will be assumed to be stationary, meaning that the Gaussian probability distributions of the process is invariant to translations in time or space.

2.2 Covariance functions

For a GRF with equal variance for all associated variables, two of them being X and $\vec{Y}$, we have that the covariance is defined by:

$$\begin{aligned} \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} &= \frac{Cov(X,Y)}{\sigma^2} = Corr(X,Y) \\ &\implies Cov(X,Y) = \sigma^2 Corr(X,Y) \end{aligned}$$

with σ^2 being the variance parameter of the covariance function. So the covariance of two variables are linear with their pairwise correlation.

Evidently, by defining a covariance function to use for a Gaussian process, one achieves a covariance with fewer parameters to estimate or set. This simplification is desireable and acceptable for many different models, as we now are down to estimate or set what is typically a couple of parameters compared to estimating $\frac{(n-1)\cdot(n-2)}{2}$ covariances for a Gaussian process with n variables. In addition, by using a defined covariance function one is then guaranteed that the resulting covariance matrix is a *positive definite matrix* [1]. This further implies the existence of a inverse of the covariance matrix, that's positive definite as well.

In a spatial GRF setting as this case will use, the associated variables will be linked with a spatial location. The main link will be the Euclidean distance between points, denoted as $d = |s_X - s_Y|$ where s_A is the spatial location for variable A. Combine this with the fact that we have

[1] Let $M \in \mathcal{R}^n \times \mathcal{R}^n$ with $M = M^T$ be a positive definite matrix. Then, for any $\vec{z} \in \mathcal{R}^n$ we have $\vec{z}^T M \vec{z} > 0$.

the Gaussian process as time-stationary, the covariance function simplifies to a single-argument function of d:

$$Cov(X,Y) = \sigma^2 Corr(X,Y) = \sigma^2 C(|s_X - s_Y|) = \sigma^2 C(d) \tag{2}$$

This means the covariance function of our model is *isotropic*, relying solely on distance between variables for its calculation. Whether this simplification is reasonable comes down to model specification. However, in a spatial setting where one is modelling geophysical attributes that only change significantly in time windows of thousands of years, there is a strong argument for defining our covariance in this form. However, covariance functions may also be defined in convenient forms in the case of non-stationarity. In Springer one may find a decent review of covariance functions and methods that allows these to be defined within a non-stationary setting [5].

For the choice of covariance functions there are many possibilities. A simple function is the exponential correlation function:

$$C(d) = \sigma^2 \cdot \exp(-\frac{d}{\tau})$$

where d is the Euclidean distance between two stochastic points X and Y in the GRF. τ is a *range* parameter, describing the intensity in which the distance should affect the correlation. For easy interpretation with the exponential covariancefunction, the distance $d = 3\tau$ is of interest as the resulting correlation is $\exp(-3) \approx 0.05$. The exponential covariance function is derived as a special case of the family of *Mátern covariance functions*. The Mátern covariance functions is defined in a stationary form, with distance d, range parameter τ and degrees of freedom ν, as:

$$C_\nu(d) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{d}{\tau}\right)^\nu K_\nu\left(\sqrt{2\nu}\frac{d}{\tau}\right) \tag{3}$$

where K_ν is the modified Bessel function of the second kind. The covariance functions obtained from the Mátern family makes sample paths of an associated Gaussian process become $\nu - 1$ differentiable. The equation in (3) is seemingly a complicated affair to evaluate, but it is reducable to convenient forms for many choices of ν. Three of them are:

$$\begin{aligned}
\nu = \frac{1}{2}: &\quad C_{\frac{1}{2}}(d) = \sigma^2 \exp(-\frac{d}{\tau}) \\
\nu = \frac{3}{2}: &\quad C_{\frac{3}{2}}(d) = \sigma^2\left(1 + \frac{\sqrt{3}d}{\tau}\right)\exp(-\frac{\sqrt{3}d}{\tau}) \\
\nu = \frac{5}{2}: &\quad C_{\frac{5}{2}}(d) = \sigma^2\left(1 + \frac{\sqrt{5}d}{\tau} + \frac{5d^2}{3\tau^2}\right)\exp(-\frac{\sqrt{5}d}{\tau})
\end{aligned} \tag{4}$$

These covariancefunctions are presented in figure (1). The exponential covariance function is recognized as the Mátern covariance function with d.o.f. $\nu = \frac{1}{2}$. This again implies that we will not have differentiable sample paths when taken from a Gaussian process using this covariance function, something that's illustrated in the figure (2).

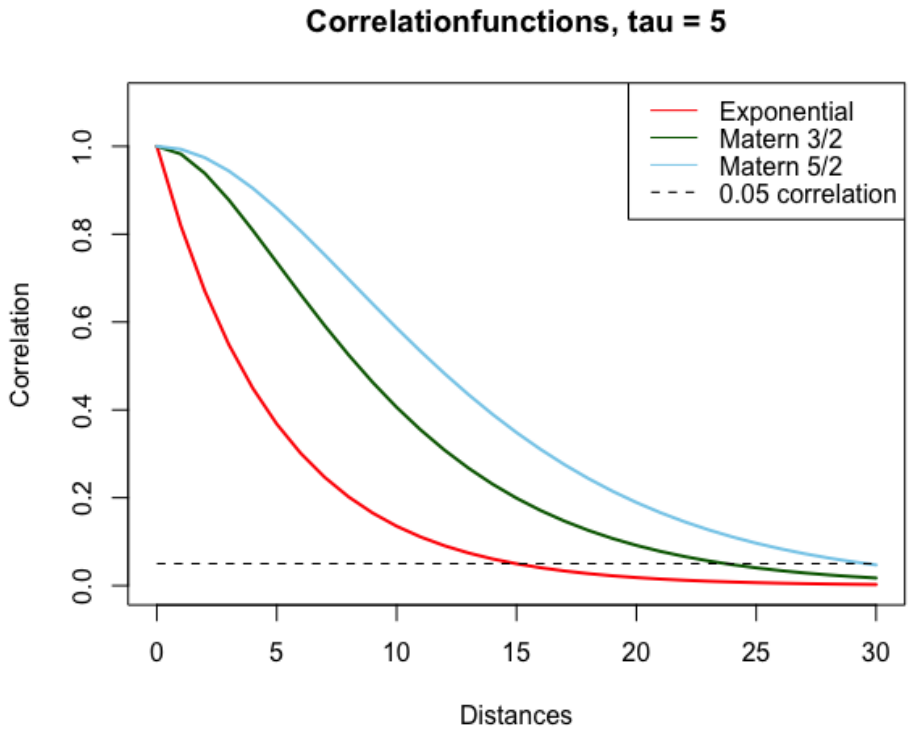

Figure 1: Correlation with $\tau = 5$ as function of distance.

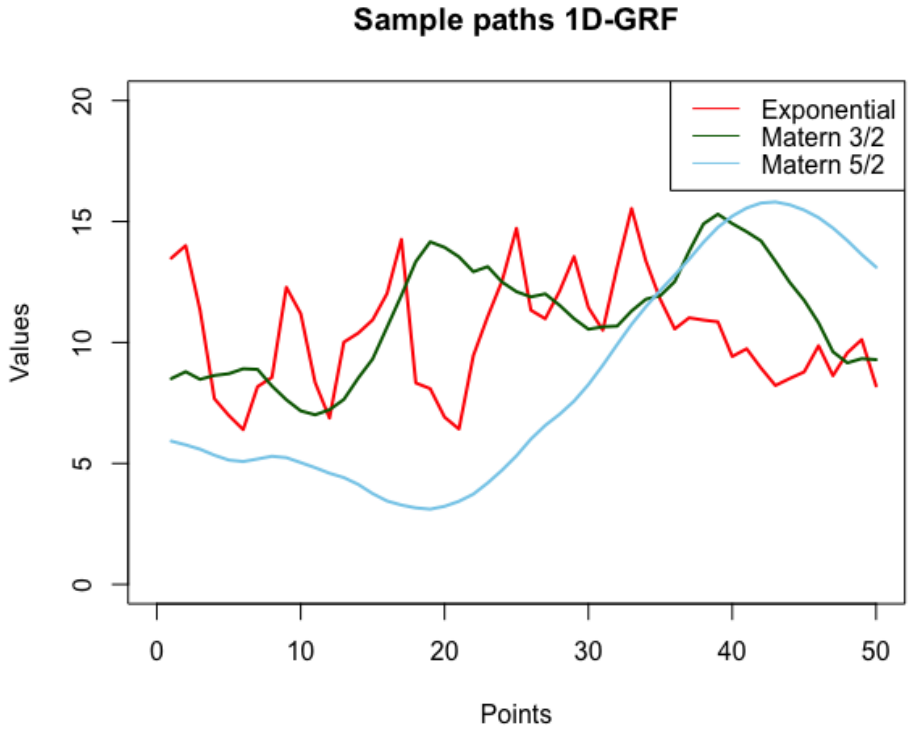

Figure 2: Sample paths from a 1D-GRF, mean $\mu = 10$, variance $\sigma^2 = 10$ and range $\phi = 5$.

2.3 Gaussian processes

Associated with every GRF is a underlying Gaussian process detailing the distributions of the variables defining the field. The fundamental characterization of such a process is that all finitie-dimensional joint distributions for the variables of the process is multivariate normally distributed (MVN), and especially that each variable is normally distributed. Being distributed as a MVN, a Gaussian process is completely determined by its mean and covariance functions [2]. Thus, the mean and covariance functions of a process are often a focus point in dealing with a Gaussian process.

Amongst the many favourable properties of MVN distributions, we have that the best predictor for an unobserved variable in a Gaussian process is linear function of the observed variables. The predictor being a linear function of MVN distributed variables, the predictor as well is MVN distributed with an adjusted mean and covariance.

More precisely, if $\begin{bmatrix}\vec{A}\\ \vec{B}\end{bmatrix}$ is a vector of n stochastic variables associated with a Gaussian process, meaning $\begin{bmatrix}\vec{A}\\ \vec{B}\end{bmatrix}$ has a distribution on the form of $\mathcal{MVN}(\begin{bmatrix}\mu_{\vec{A}}\\ \mu_{\vec{B}}\end{bmatrix}, \begin{bmatrix}\Sigma_{\vec{A}\vec{A}} & \Sigma_{\vec{A}\vec{B}}\\ \Sigma_{\vec{B}\vec{A}} & \Sigma_{\vec{B}\vec{B}}\end{bmatrix})$, where $\Sigma_{\vec{A}\vec{B}} = \Sigma^T_{\vec{B}\vec{A}}$ denotes the covariance between the variables in $\vec{A}$ versus the variables in $\vec{B}$. Given $\vec{B}$, we have exact results for the best linear predictor. $P(\vec{A}|\vec{B} = \vec{b})$ will then be MVN, with corrected mean $\mu_{\vec{A}|\vec{B}}$ and variance $\Sigma_{\vec{A}|\vec{B}}$ as:

$$\mu_{\vec{A}|\vec{B}} = \mu_{\vec{A}} + \Sigma_{\vec{A}\vec{B}} \cdot \Sigma^{-1}_{\vec{B}\vec{B}}(\vec{b} - \mu_{\vec{B}}) \tag{5}$$

$$\Sigma_{\vec{A}|\vec{B}} = \Sigma_{\vec{A}\vec{A}} - \Sigma_{\vec{A}\vec{B}} \cdot \Sigma^{-1}_{\vec{B}\vec{B}} \cdot \Sigma_{\vec{B}\vec{A}} \tag{6}$$

In an informal setting, one may interpret this as a linear smoothing of the parameters for the MVN distribution of $\vec{A}$. Especially it's worth noting that since the covariance matrix $\Sigma_{\vec{B}\vec{B}}$ is positive definite, its inverse is as well, meaning that the diagonal entries of $\Sigma_{\vec{A}\vec{B}} \cdot \Sigma^{-1}_{\vec{B}\vec{B}} \cdot \Sigma_{\vec{B}\vec{A}}$ are all greater or equal to zero. As the variance of each variable in $P(\vec{A}|\vec{B} = \vec{b})$ are the diagonal elements of $\Sigma_{\vec{A}|\vec{B}}$, we have that the elementwise variance after conditioning on $\vec{B}$ is lower or equal than with no conditioning:

$$\Sigma_{\vec{A}|\vec{B}} = \Sigma_{\vec{A}\vec{A}} - \Sigma_{\vec{A}\vec{B}} \cdot \Sigma^{-1}_{\vec{B}\vec{B}} \cdot \Sigma_{\vec{B}\vec{A}} \implies \Sigma_{\vec{A}|\vec{B}}[i,\ i] \leq \Sigma_{\vec{A}\vec{A}}[i,\ i] \quad i = 1, \ldots, n. \tag{7}$$

where $\Sigma_{\vec{A}\vec{A}}[i,\ i]$ denotes the entry at index (i, i) in the matrix. As the matrices $\Sigma_{\vec{A}\vec{B}}$ and $\Sigma^{-1}_{\vec{B}\vec{B}}$ consists solely of positive real numbers, the equality will only hold in the case that there are no correlation between variable i of $\vec{A}$ and all the variables in $\vec{B}$. Intuitively, this makes sense as then the variable $\vec{A}_i$ have no relation to the variables in $\vec{B}$.

As an example, three predictions using these results have been plotted in figures (3), (4) and (5). The actual processes shown in red are Gaussian processes in a 1D-GRF with spatial points $s \in \{1, \ldots, 50\}$. A common set of parameters are set with mean $\mu = 10$, variance $\sigma^2 = 10$ and range $\tau = 5$. Data were sampled in all cases as noise-free at location $s \in \{8, 25, 43\}$. The difference is which covariancefunction that's used to simulate the process in the three figures.

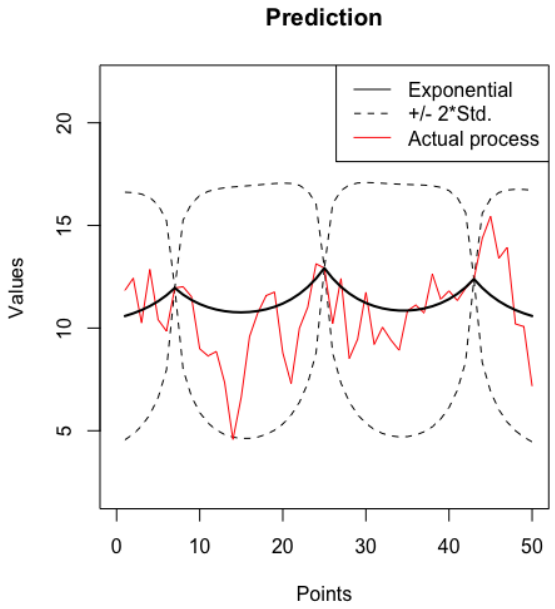

Figure 3: Predictions with an Exponential covariancefunction

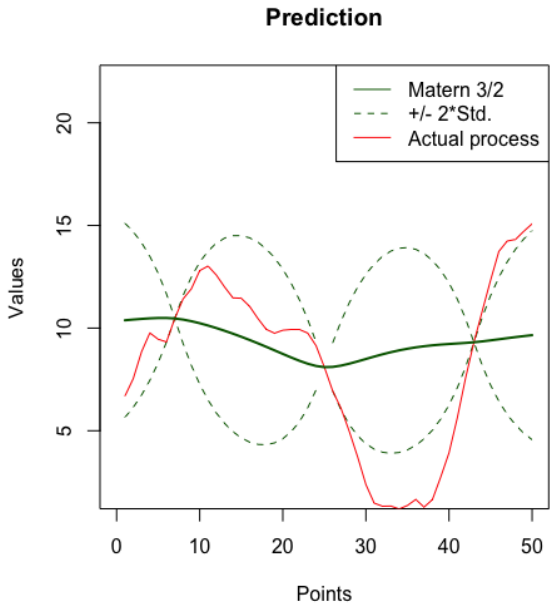

Figure 4: Predictions with Matern covariancefunction, d.o.f. $\nu = \frac{3}{2}$

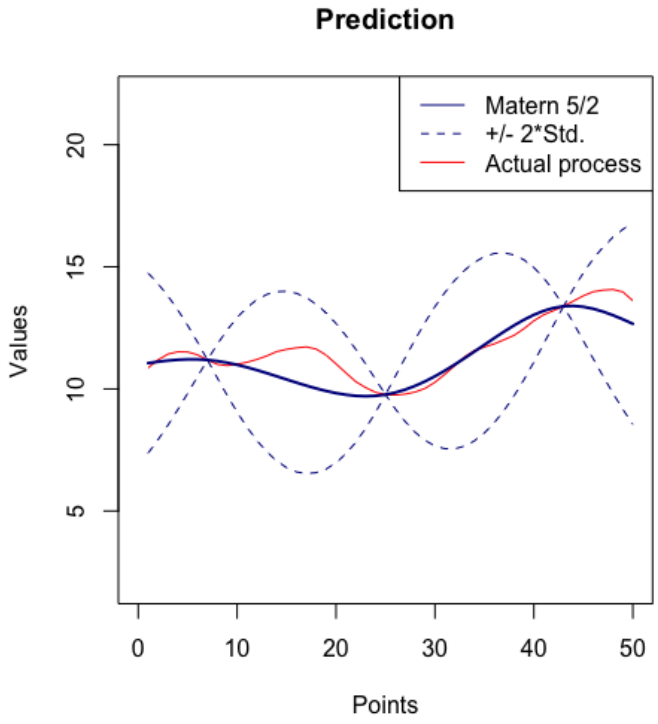

Figure 5: Predictions with Matern covariancefunction, d.o.f. $\nu = \frac{5}{2}$

Evidently from the easy to achieve posterior distribution, when dealing with a Gaussian process it would be favourable to have well-defined expressions for its mean and covariance. Regarding the process-covariance, covariancefunctions has been mentioned as a possible solution. In this project results will be presented for all the covariancefunctions defined in equation (4).

2.4 Generalized least squares

The trendfunction is another matter. As we are unaware of any data in the grid that may be related to the trend of the GRF model, the better option that is left is to estimate the trend based on the curent data aqcuisition. The trendfunction is defined as a polynomial function of spatial location on the form $\mu(\vec{s}) = X(\vec{s})\vec{\beta}$, where X is a design matrix corresponding to the locations $\vec{s}$ and $\vec{\beta}$ is vector of coefficients. By doing so we may utilize results from the theory of *Generalized Least Squares* (GLS) in the estimation of trend. One particular advantage of using GLS is that we are guaranteed by the Gauss-Markov Theorem that the GLS estimates is the *Best Linear Unbiased Estimator* (BLUE) for the vector of coefficients $\vec{\beta}$.

One of two caveats with using GLS estimates is that one still need to specify the polynomial trendfunction to fit the data, meaning there's still a selection process. As the trendfunction is to specified by spatial location for this case, one must face the second caveat. The GLS estimates are the BLUE only for the data and their associated spatial location used in the estimation, giving no guarantee for the prediction of data not part of the estimates. By this, the trendfunction will be selected to be a polynomial of lower order with simple interaction between easting- and northing-coordinates, i.e. for a variable with coordinates (s_E, s_N):

$$X(s_E, s_N)\vec{\beta} = \beta_0 + s_E \cdot \beta_1 + s_N \cdot \beta_2 + s_E \cdot s_N \cdot \beta_4 \tag{8}$$

This has been chosen as an prediction far away from the sample data will typically be over- or underpredicted by an unacceptable amount using a higher-order polynomial. This is due to the covariates of said polynomial will be fitted to the data that has been sampled, which in this case will be done mainly by a specified design of low regularity. In sampling designs with high regularity over the field, this is effect is handled as then the estimation will try to adapt to grid points that has a regular distance between each other. An example of this issue is shown in figures (6) and (7), where a proposed sampling design is used two different polynomial distributions are fitted.

When the samples are drawn from a distribution with known correlation matrix $\Omega = \sigma^2 \cdot W$, with σ^2 as a constant finite variance parameter and W the correlation matrix, the estimates of $\vec{\beta}$ converge in distribution to a MVN, or more precisely

$$\hat{\vec{\beta}} \xrightarrow{d} \mathcal{MVN}\big(\ \vec{\beta},\ (X^T\Omega^{-1}X)^{-1}\big) \tag{9}$$

where X denotes the associated design matrix of the samples used to estimate the trend. In figures(6) and (7) the GLS estimation has been used on the data acquired at the sample path indicated by the black line. The data is from the case data, modelled with an Exponential covariancefunction. A GLS estimate was fitted with both a simple linear formula of equation (8) and a quadratic formula of equation (8) $+s_E^2 \cdot \beta_5 + s_N^2 \cdot \beta_6$. The effect of overestimation in the

fitted modell is not so obvious as the grid size is fairly low, but you can see that the quadratic formula's variance increases rapidly as we move towards the outer edges, away from the sampled data.

2.5 Hierarchical model

Apart from specifying the design of the GRF, it's preferable to clearly relate the GRF and its attributes to the data acquisition. One may also specify any noise in the data acquisition in a precise matter. One common way of specifying this relation is to describe the data acquisition and gaussian process in a hierarchical model. An advantage of doing so is the lack of restriction that the model presents on the parameters of the model. In the case of Bayesian modelling, we also have the advantage of clearly connecting the different levels through the updating posterior distributions that are of interest, encapsulating the uncertainty of the different levels in an easy to interpret way [5].

A hierarchical model divides the spatial design model into three parts:

- Data model - Describes the relation between process and datasampling
- Process model - Describes the model properties of the GRF
- Prior model - Priors used in the case of Bayesian modelling of parameters

An example with fixed parameters, i.e. without a "Prior model", is given here. Denoting $\vec{Z}$ as the sampled data, $\vec{Y}$ as the underlying GP and parameters $\mu, \tau, \sigma^2, \phi$:

$$\begin{aligned} \textbf{Data model: } & \vec{Z}(\vec{s}) \sim \mathcal{MVN}\big(F(\vec{s}) \cdot \vec{Y}, \phi^2 I_{n\times n}\big) \\ \textbf{Process model: } & \vec{Y} \sim \mathcal{MVN}\big(\vec{\mu},\ C_{\vec{Y}}(D, \sigma^2, \tau)\big) \end{aligned}$$

where $\vec{s} \in \mathcal{R}^n \times \mathcal{R}^2$ is the matrix of coordinates for the n variables in question, D denotes a symmetric $n \times n$ matrix of distances between the points in the field, F is a matrix indicating sampling of data and $C_{\vec{Y}}(\dots)$ the chosen covariance function. In this case ρ indicates the covariance that arises due to the prosess, whilst ϵ may be interpreted as possible noise that arises when the data is acquired. In this paper, ϵ will be treated as a fixed constant and not a stochastic variable within the Bayesian framework. A motivation for this is that ϵ models the knwon uncertainty of measurement-equipment.

By using the linearity property of Gaussian variables, we may write the models as linear combinations of their means and covariances, i.e.

$$\begin{aligned} \textbf{Data model: } & \vec{Z}(\vec{s}) = F(\vec{s}) \cdot \vec{Y} + \vec{\epsilon}, \quad \vec{\epsilon} \sim \mathcal{MVN}\big(\vec{0}, \phi^2 I(\vec{s})\big) \\ \textbf{Process model: } & \vec{Y} = \vec{\mu} + \rho, \quad \rho \sim \mathcal{MVN}\big(\vec{0},\ C_Y(D, \sigma^2, \tau)\big) \end{aligned} \tag{10}$$

This is a unique property of models with a GRF and Gaussian distributed noise and is of no difference compared to the first model. Sample paths for a model of this type was model is shown in (2), where $F(\vec{s})$ is set to be $I_{n\times n}$ for the n variables in the field, i.e. perfect sampling of the field.

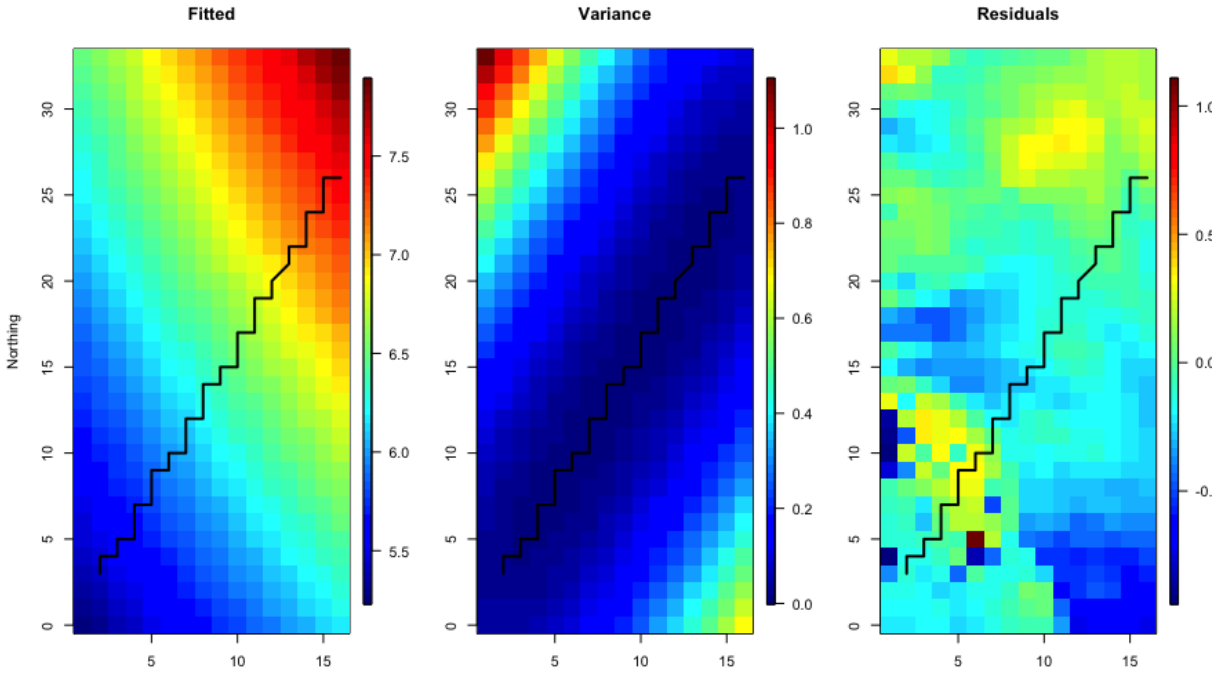

Figure 6: GLS fit by perfect data from the black line. Linear model with interaction.

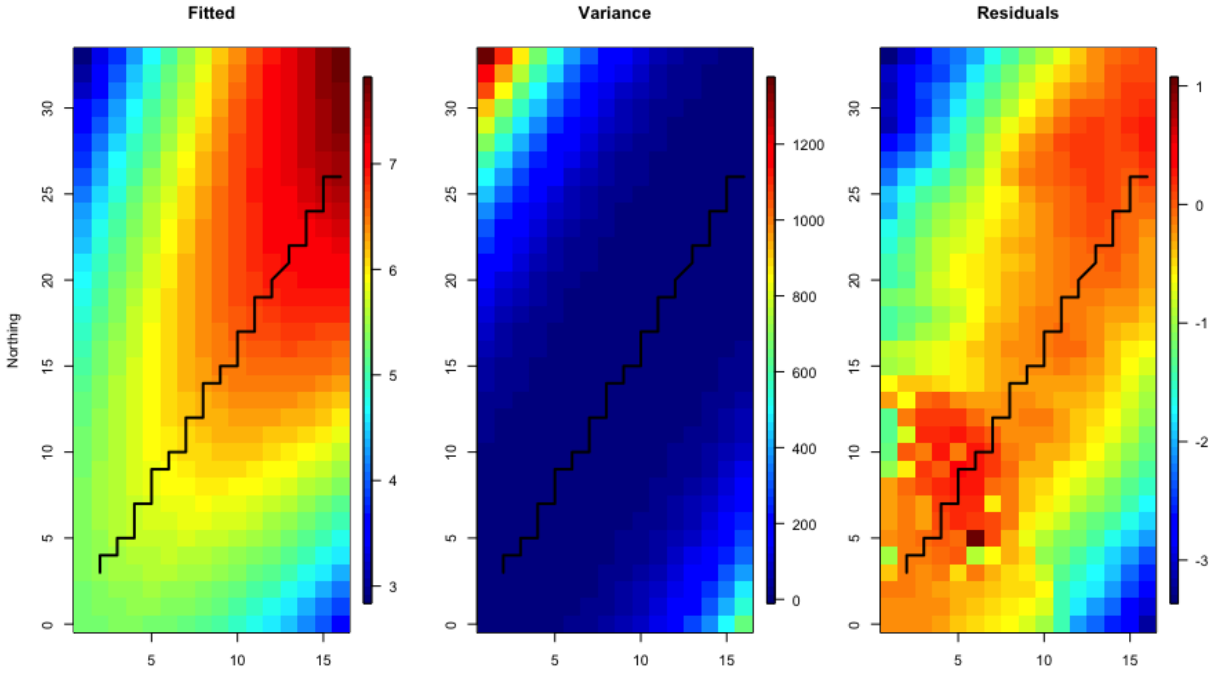

Figure 7: GLS fit by perfect data from the black line. Quadratic model with interaction.

2.5.1 Deriving the posterior GRF

The ultimate goal of the analysis is to provide a result for the posterior distribution of the underlying Gaussian process, given some sample data, that is $P\big(\vec{Y}|\vec{Z}(\vec{s_0})\big)$ where $\vec{s}$ denotes the whole coordinates of the whole field and $\vec{s_0}$ denotes the coordinates of samples obtained. This distribution is derived as

$$\begin{aligned} P\big(\vec{Y}|\vec{Z}(\vec{s_0})\big) &= \int\int P\big(\vec{Y}, \sigma^2, \tau|\ \vec{Z}(\vec{s_0})\big)\ d\sigma^2\ d\tau \\ &= \int\int P\big(\vec{Y}|\vec{Z}(\vec{s_0}), \sigma^2, \tau\big) \cdot P\big(\sigma^2, \tau|\vec{Z}(\vec{s_0})\big)\ d\sigma^2\ d\tau \end{aligned}$$

Further, we use Bayes Theorem

$$\begin{aligned} P\big(\sigma^2, \tau|\vec{Z}(\vec{s_0})\big) &\propto P\big(\vec{Z}(\vec{s_0})|\sigma^2, \tau\big) \cdot P\big(\sigma^2, \tau\big) \\ &\propto P\big(\vec{Z}(\vec{s_0})|\sigma^2, \tau\big) \cdot P\big(\sigma^2\big) \cdot P\big(\tau\big) \\ \implies P\big(\sigma^2, \tau|\vec{Z}(\vec{s_0})\big) &= \frac{P\big(\vec{Z}(\vec{s_0})|\sigma^2, \tau\big) \cdot P\big(\sigma^2\big) \cdot P\big(\tau\big)}{\int\int P\big(\vec{Z}(\vec{s_0})|\sigma^2, \tau\big) \cdot P\big(\sigma^2\big) \cdot P\big(\tau\big) d\sigma^2 d\tau} \end{aligned} \tag{11}$$

This implies that

$$P\big(\vec{Y}|\vec{Z}(\vec{s_0})\big) = \int\int P\big(\vec{Y}|\vec{Z}(\vec{s_0}), \sigma^2, \tau\big) \cdot \frac{P\big(\vec{Z}(\vec{s_0})|\sigma^2, \tau\big) \cdot P\big(\sigma^2\big) \cdot P\big(\tau\big)}{\int\int P\big(\vec{Z}(\vec{s_0})|\sigma^2, \tau\big) \cdot P\big(\sigma^2\big) \cdot P\big(\tau\big) d\sigma^2 d\tau}\ d\sigma^2\ d\tau \tag{12}$$

An exact analytical result for the posterior found in both (11) and (12) may be found under the with the appropriate priors, but to acommodate any type of prior from a probability distribution, the integrals is approximated by a simple numerical procedure. This is done by discretizing the domain of σ^2 and τ and summing over the discretizied values with some weight Δ for each discretization to ensure normalization, giving an approximation as

$$P\big(\vec{Y}|\vec{Z}(\vec{s_0})\big) \approx \frac{1}{\Pi} \sum_{\tau_i}^{n_\tau} \sum_{\sigma_j^2}^{n_\sigma} P\big(\vec{Y}|\vec{Z}(\vec{s_0}), \sigma_j^2, \tau_i\big) \cdot P\big(\vec{Z}(\vec{s_0})|\sigma_j^2, \tau_i\big) \cdot P\big(\sigma_j^2\big) \cdot P\big(\tau_i\big) \cdot \Delta_{ij}$$

With $\Pi = \sum_{\tau_i}^{n_\tau} \sum_{\sigma_j^2}^{n_\sigma} P\big(\vec{Z}(\vec{s_0})|\sigma_j^2, \tau_i\big) \cdot P\big(\sigma_j^2\big) \cdot P\big(\tau_i\big) \cdot \Delta_{ij}$ as the normalizing constant.

For this integration to be performed, four distributions need to be defined: $P\big(\sigma_j^2\big)$ and $P\big(\tau_i\big)$ are priors given by the model, $P\big(\vec{Z}(\vec{s_0})|\sigma_j^2, \tau_i\big)$ is known MVN by the model specification and lastly we have that $P\big(\vec{Y}|\vec{Z}(\vec{s_0}), \sigma_j^2, \tau_i\big)$ is MVN by the results presented in section (2.3) with $\vec{Y}$ as A and $\vec{Z}(\vec{s_0})$ as B.

When performing the numerical integration, it is noted that it's the mean and variance of

$P\big(\sigma^2, \tau | \vec{Z}(\vec{s_0})\big)$ that's of relevance for the predictions. By similar arguments as above, one gets:

$$\begin{aligned}
\mathbf{E}\big(\vec{Y}|\vec{Z}(\vec{s_0})\big) &\approx \frac{1}{\Pi}\sum_{\tau_i}^{n_\tau}\sum_{\sigma_j^2}^{n_\sigma}\Delta_{ij}\cdot\mathbf{E}\big(\vec{Y}|\vec{Z}(\vec{s_0}),\sigma_j^2,\tau_i\big)\cdot P\big(\vec{Z}(\vec{s_0})|\sigma_j^2,\tau_i\big)\cdot P\big(\sigma_j^2\big)\cdot P\big(\tau_i\big) \\
\mathbf{Var}\big(\vec{Y}|\vec{Z}(\vec{s_0})\big) &\approx \frac{1}{\Pi}\sum_{\tau_i}^{n_\tau}\sum_{\sigma_j^2}^{n_\sigma}\Delta_{ij}\cdot\mathbf{Var}\big(\vec{Y}|\vec{Z}(\vec{s_0}),\sigma_j^2,\tau_i\big)\cdot P\big(\vec{Z}(\vec{s_0})|\sigma_j^2,\tau_i\big)\cdot P\big(\sigma_j^2\big)\cdot P\big(\tau_i\big)
\end{aligned} \tag{13}$$

with Π defined as before. The expressions of $\mathbf{E}\big(\vec{Y}|\vec{Z}(\vec{s_0}),\sigma^2,\tau\big)$ and $\mathbf{Var}\big(\vec{Y}|\vec{Z}(\vec{s_0}),\sigma^2,\tau\big)$ is given in (5) and (6).

3 Spatial Design

3.1 Sampling design

In spatial prediction, the choice of which spatial locations, or spatial design, to sample data from becomes critical. In the model setting, this means defining how our matrix $F(\vec{s})$ mentioned in (**??**) is constructed. In order to achieve the best design, one would have to choose in cohesion with the model that is defined for the data situation and as to what criteria the model is trying to satisfy.

There are two common situations when producing sampling designs. Firstly we have the *retrospective design*, where one assesses the sampling design after obtaining information about the attributes of interest. A goal in this setting might be to reduce the number of sampling locations without a significant loss in the amount or quality of information. This problem is exemplified by a want to economize measuring-stations while still achieving satisfactory measurements from the remaining locations.

The second approach is that of *proactive design*. In this form, the overall goal is to design a set of sampling locations in advance of the data acquisitioning that align with some criteria set for the project. The criteria is most often set on a applicationbased setting, indicating that it doesn't exist any general solution to produce the optimal design. Typically in a predictive setting, criterias that revolve around minimizing the variance or predicitve error of a design or maximizing the entropy is common. Two examples of criterias relating to the variance are found in equations (14) and (15).

The setting in which this criteria was used was the desire to find the optimal path through a graph that were modelled as a random field [6]. The criteria was to minimize the *Expected Mean Squared Error* (EMSE), that they calculated as

$$\text{EMSE} = \frac{1}{n}[tr(\Sigma_p) - tr(\Sigma_o)] \tag{14}$$

where Σ_p denotes the posterior covariancematrix and Σ_o denotes the prior covariancematrix.

For equation (15), the setting was retroactive design planning, were the ultimate goal was to remove as many measurement-locations in a random field without significant loss in measurement amount or quality [3]. This criteria was adhered to by minimizing the equation

$$E(\vec{\nu}) = \int_A E_{Y|\theta_0}[Var(S(x)|Y)]dx \tag{15}$$

where $S(x)$ denotes a function of the proposed reduced sampling design, Y known data and θ_0 the true parameter values.

To adhere to the criteria that's set, the design need not be solved within the workframe as stated by the model ones uses. Diggle & Lophaven [3] refers to some purely geometrical approaches that may work well for finding adequate designs. In this project a satisfiable design will be estimated on the basis of a combination of predefined design proposals, i.e. the project will use focus on proactive design. The criteria that will be used in this project is minimizing the spatially

averaged standard deviation of the proposed sample design, i.e.

$$I = \sum_{Z_a} \sum_{i=1}^{n} \sqrt{Var(Y_i|Z_a, Z)} P(Z_a|Z). \tag{16}$$

where Z_a denotes the proposed design, Z current data that has been acquired and Y_i as the latent process at the location to be predicted.

3.2 Constructing designs in Bayesian hierarchical models

With a Bayesian hierarchical model, one has to make some alterations to the calculation. In our setting, with prior distributions for σ^2 and τ, we need to evaluate the integral that arises from

$$P(Z_a|Z) = \int \int P(Z_a|Z, \sigma^2, \tau) d\sigma^2 d\tau$$

In stead of evaluating this double-integral, Monte Carlo sampling of $P(Z_a|Z)$ is used to estimate I. To generate a sample Z_a, a realization of both σ^2 and τ is drawn from their respective distributions. With these, a sample Z_a^* is drawn from $P(Z_a|Z, \sigma^2, \tau)$, and evaluating the criteria as

$$I = \sum_{Z_a^*} \sum_{i=1}^{n} \sqrt{Var(Y_i|Z_a^*, Z)}.$$

Performing this Monte Carlo sampling an adequate amount of times, we obtain a result for our criteria.

4 Case

4.1 Case introduction

4.1.1 Introduction and case specification

In this case I've been presented with the task of performing a predictive analysis of an attribute that may be modelled by a random field. The scenario is that there is a desire in knowing the temperature for a specific physical location, and one would like to perform predictive analysis with the following goals in mind:

- Perform satisfactory accurate predictions
- Minimize data acquirement needed

In order to ensure that these goals will be met, the analysis of the case will be set up within the framework of a hierarchical model as described in (10). Within this setting, suitable criterias must be set so that the analysis may be performed in a way that easily relates to these goals. The criterias set for this case will be to minimize equation (16) where the proposed paths focus on covering areas where the prediction variance is high. Informally we may say that we are seeking for the most uncertainty-reducing paths of data sampling.

4.1.2 Input

The only information that we're presented with beforehand is a variogram computed on the basis of simulation model output based on weather forecasts, oceanographic models and satelite data. This is however reckoned as slightly unreliable data and must be handled accordingly. The area in which the analysis will be based on is given as a list of xy-coordinates with each coordinate being given a z-variable denoting the numerical value to the attribute of interest. There is a total of 29116 coordinates, ordered into a rectangular grid with size 116 in x-direction and size 251 in y-direction. The directions will informally be called easting and northing when plotting. The variogram is presented in (8) along with the original data which we desire to predict.

4.1.3 Model specification and starting point of analysis

As a necessary preliminary for the predictive analysis the temperature attribute of all the coordinates is modelled as a stochastic variable. Attaining a Gaussian process to describe trend and variability of model, the grid is viewed as a GRF. The connection between variables in this model will revolve around spatial location within the grid, where the trendfunction will be estimated by GLS regression on the sampled data. The covariance will be described by an isotropic, where results from all three covariancefunctions of (4) will be presented individually.
The model process is fitted into the hierarchical model (HM) design described in (2.5). With Y denoting the temperature variables at locations $\vec{s}$, that's set to be distributed according to the

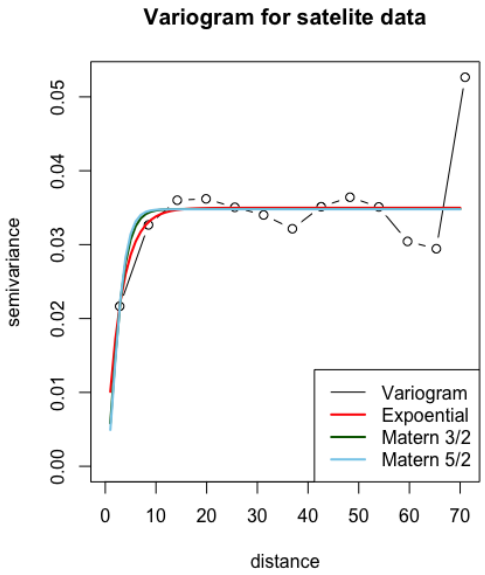

Figure 8: Variogram given as input to the case.

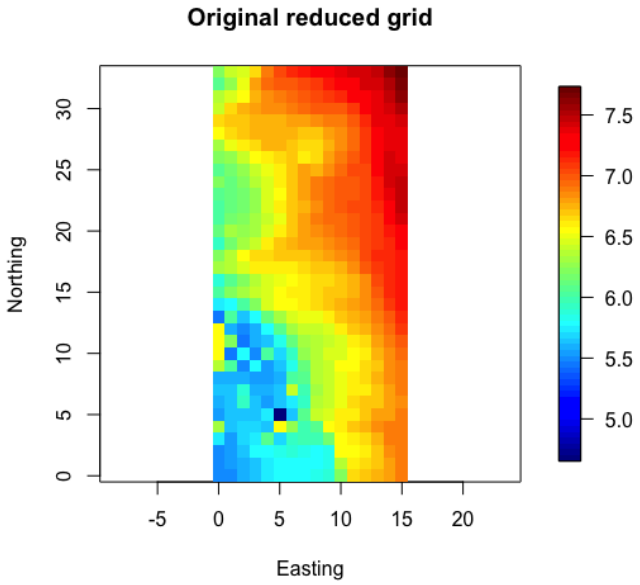

Figure 9: Actual data that are to be predicted on a reduced grid from the original.

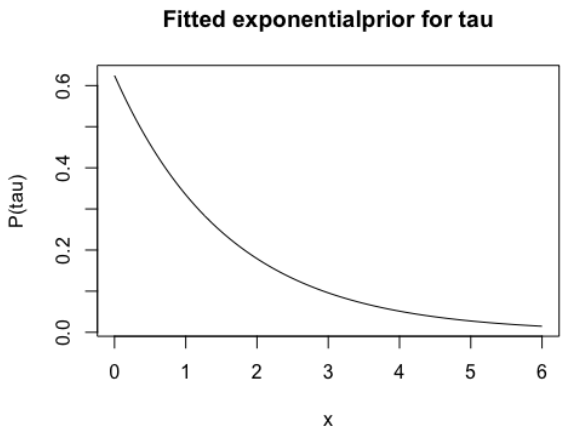

Figure 10: Fitted prior distribution for τ.

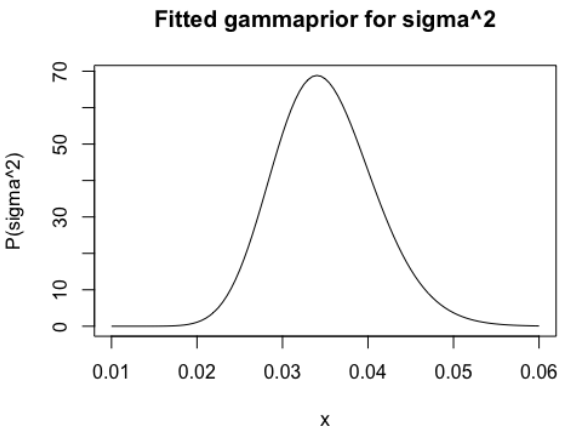

Figure 11: Fitted prior distribution for σ^2

Gaussian process, the hierarchical model follows as

$$\begin{aligned}
\textbf{Data model:}\ & \vec{Z}(\vec{s}) = \vec{Y}(\vec{s}) + \vec{\epsilon}(\vec{s}), \quad \vec{\epsilon}(\vec{s}) \sim \mathcal{MVN}\big(\vec{0}, \phi^2 I(\vec{s})\big) \\
\textbf{Process model:}\ & \vec{Y} = \vec{\mu} + \rho(\vec{s}), \quad \rho(\vec{s}) \sim \mathcal{MVN}\big(\vec{0},\ C_Y(D(\vec{s}), \sigma^2, \tau)\big) \\
\textbf{Prior model:}\ & \sigma^2 \sim \text{Gamma}(\alpha, \kappa), \quad \tau \sim \text{Exp}(\lambda)
\end{aligned}$$

The intensity of the sampling noise, ϕ^2, is the only one of the parameters that will be fixed. The motivation for doing so is that sampling noise may be viewed as more related to the equipment being used than the analysis in which it is included. Compared to the other parameters, one may have much more information about this parameter, as it may be stated by the equipment manufacturer or be tested separately in controlled environments. For the results, $\phi^2 = 0.2$ has been used.

With the trend $\vec{\mu}$ is to be estimated by GLS on the sampled data eventually the data model is redefined as

$$\begin{aligned}
\textbf{Data model:}\ & \vec{Z}(\vec{s}) = X(\vec{s})\hat{\vec{\beta}} + \epsilon(\vec{s}) + \rho(\vec{s}) \\
\implies\ & \vec{Z}(\vec{s}) \sim \mathcal{MVN}\big(X(\vec{s})\hat{\vec{\beta}},\ X(\vec{s})\Sigma_{\hat{\beta}}X(\vec{s})^T + C_Y(D(\vec{s}), \sigma^2, \tau) + \phi^2 I(\vec{s})\big)
\end{aligned} \tag{17}$$

where $\text{Var}\hat{\beta} = \Sigma_{\hat{\beta}} = \left(X(\vec{s_0})^T \cdot \big(C_Y(D(\vec{s_0}), \sigma^2, \tau) + \phi^2 I(\vec{s_0})\big)^{-1} \cdot X(\vec{s_0})\right)^{-1}$.

4.2 Analysis

4.2.1 Choosing prior-parameters

As a consequence of defining our model with prior-distributions for σ^2 and τ, we must somehow estimate or choose parameters for their distributions. In order to do so, the variogram given as a part of the case preliminary has been utilized to estimate the GRF parameters of σ^2 and τ. Given the three covariancefunctions in 4, the one may try to fit the the parameters of the chosen covariancefunction to the line. Denoting the discrete variogram values at points k as γ_k and the fitted covariancefunction at the same points as $\hat{\gamma}_k(\sigma^2, \tau)$, this is done by minimizing the associated lossfunction

$$Loss(\sigma^2, \tau) = \sum_k \left(\frac{\gamma_k - \hat{\gamma}_k(\sigma^2, \tau)}{\gamma_k}\right)^2 \tag{18}$$

Notice that we only have two parameters to be estimated for each covariancefunction as the nugget effect has been assumed to be negligible, resulting in more stable computation of fit. This minimization has been performed using the *variofit* function in the *geoR* package found in the programming environement R. Obviously, the covariance functions are exponentially decaying as a function of distance with the three chosen in 4. In order to visually compare the fit, we use

the relation between the variogram of a isotropic process and its covariancefunction that is

$$\begin{aligned} 2\gamma(\mathbf{s},\mathbf{t}) &= Var(Y(\mathbf{s}) - Y(\mathbf{t})) \\ &= Var(Y(\mathbf{s})) + Var(Y(\mathbf{t})) - 2Cov(Y(\mathbf{s}), Y(\mathbf{t})) \\ \implies \gamma(\mathbf{s},\mathbf{t}) &= \sigma^2 - Cov(\|\mathbf{s}-\mathbf{t}\|) \end{aligned} \tag{19}$$

The variogram is then overlain with fitted covariancefunctions, adapted by equation (19), from 4 in figure (8).

Denoting the estimated values as $\hat{\sigma}^2$ and $\hat{\tau}$, one could have performed predictioning straight away. However, in an attempt to reduce the uncertainty in using this variogram, $\hat{\sigma}^2$ and $\hat{\tau}$ is used to fit parameters for the priors. This has been done by using a *method of moments* inspired approach. As we have a single parameter prior for τ, its mean and variance is estimated by

$$\hat{\lambda} = \frac{1}{\hat{\tau}} \tag{20}$$

With σ^2 given a gamma-prior, one needs to estimate two parameters. Denoting $\sum_k (\gamma_k - \hat{\sigma}^2)^2$ as s_*^2, the following approximation is used to set parameters α and β:

$$\mathbf{E}(\sigma^2) = \frac{\alpha}{\beta} \approx \hat{\sigma}^2, \quad \mathbf{Var}(\sigma^2) = \frac{\alpha}{\beta^2} \approx s_*^2 \implies \hat{\alpha} = \frac{(\hat{\sigma}^2)^2}{s_*^2}, \quad \hat{\beta} = \frac{\hat{\sigma}^2}{s_*^2} \tag{21}$$

The resulting prior-distributions are presented in figures (10) and (11).

4.2.2 Discretizing the prior domains

As the analysis relies on a numerical summation over the prior domains, a strategy must be defined to discretize the domains into n_{σ^2} and n_τ divisions used the summation 13. In this case a method of *stratified sampling* has been used. Given size-input n_{σ^2}, $(0, \inf)$ is subdivided into n_{σ^2} divisions, each part corresponding to a $\frac{1}{n_{\sigma^2}}$ probability mass of the prior. Random samples are simulated until minimum one sample within each division is obtained. The same procedure is performed for n_τ.

The joined domain of τ and σ^2 is then considered a grid, with each cell corresponding to samples (τ_i, σ_j^2) having a probability mass of $\Delta_{ij} = \frac{1}{n_{\sigma^2} \cdot n_\tau}$ due to independence of priors. With both the normalizing constaHowever, may also be definedt and the summation for predictioThis progame discretization, all Δ_{ij} cancels as they are independent of (τ_i, σ_j^2) and are therefore omitted from the calculations. A disadvantage is obviously the problem of simulating random samples for all the divisions if the number of divisions become high. However, n_τ and n_{σ^2} is kept relatively low in this case.

4.2.3 Results

Due to time-limitations, only a select predefined designs were used to produce results. The designs are presented in figure (12).

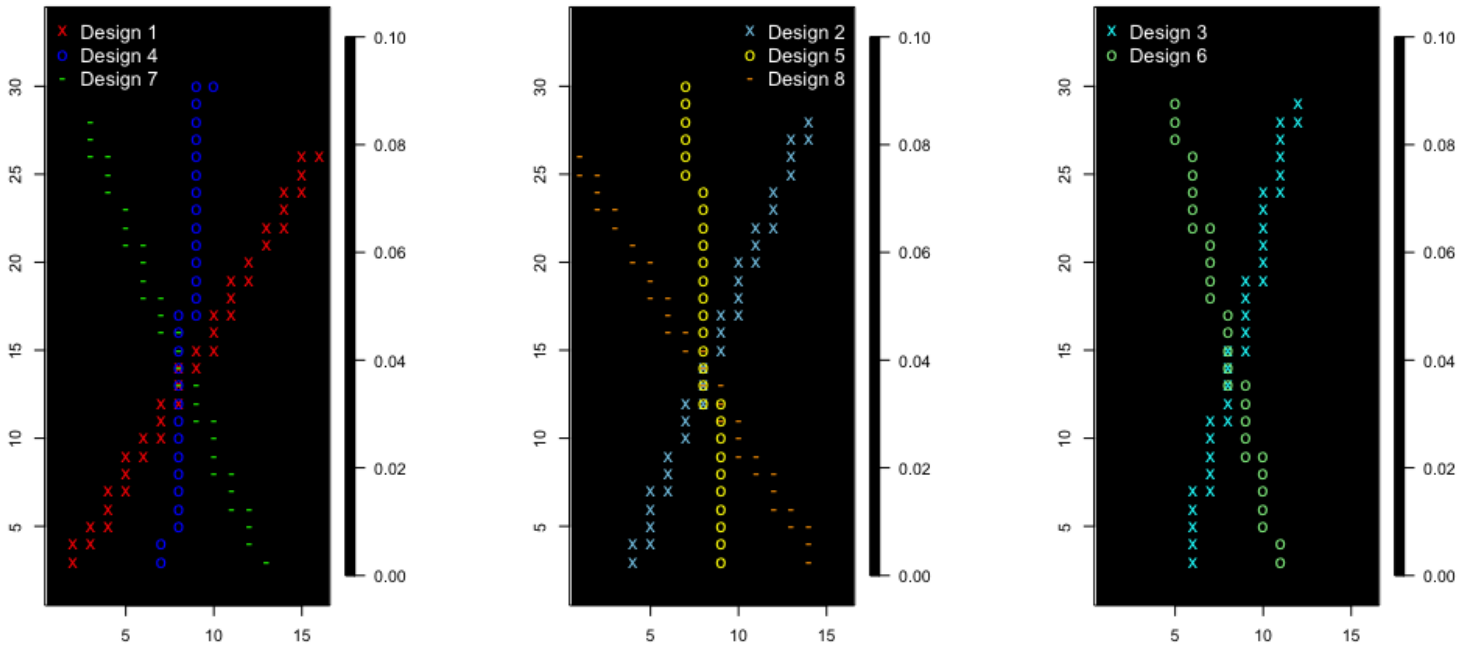

Figure 12: Sampling designs used to produce results.

As the model is reliant on estimation of trend, an initial data sample has to be obtained. To estimate the spatially averaged predictive standard deviation, each design presented in figure (12) was used as an initial design. Samples was obtained from this initial design, and possible new designs for further predictive analysis of the area was the remaining sample designs.

Using the procedure described in subsection (**??**), results are presented in table (1). Each number signifies the estimated spatially averaged standard deviation from equation (16) by the various combinations. The design on top of the table signifies the initial path with realized values Z. The design at the left signifies the proposed added path in which we Monte Carlo simulate realizations Z_a to use in equation (16).

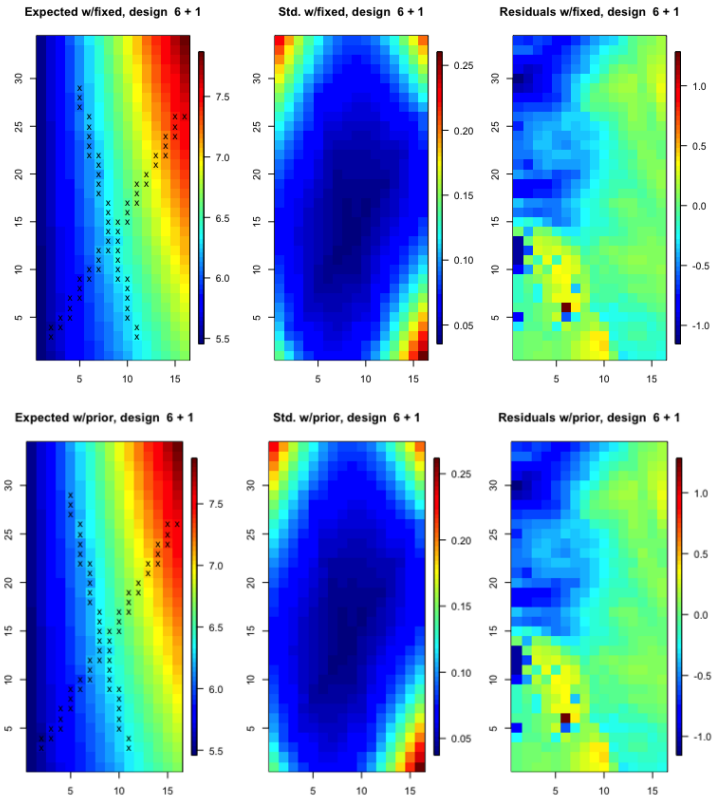

Figure 13: Prediction results for combined designs 6 + 1

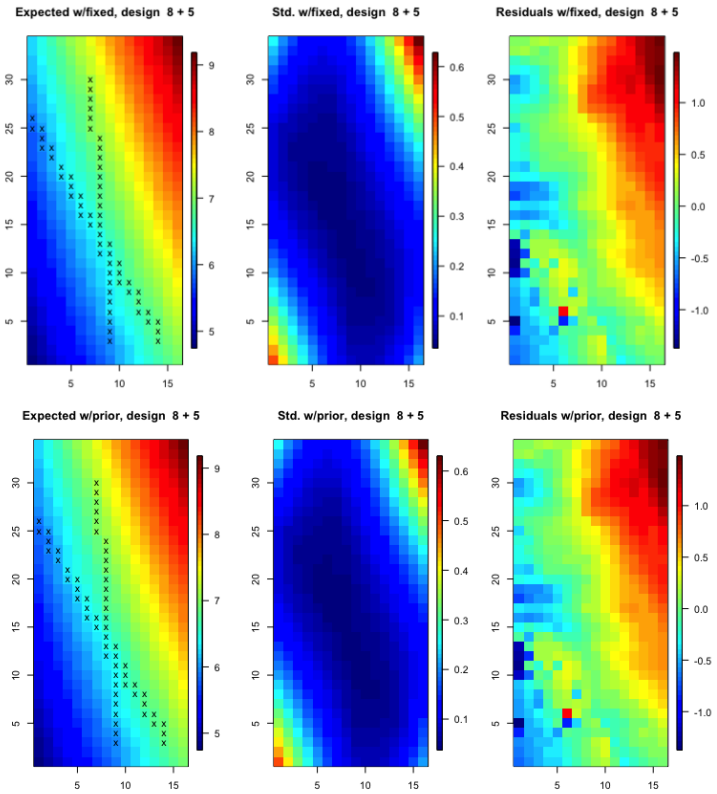

Figure 14: Prediction results for combined designs 8 + 5

In these simulations, the measurement noise was set as $\phi^2 = 0.64$ and the Matern covariance function with $\nu = \frac{5}{2}$ was used. The prior discretization was performed with $n_{\sigma^2} = n_\tau = 10$ and yielded the following prior values:

$$\begin{aligned}\sigma^2 \in &\{0.02580851, 0.02883705, 0.03074835, 0.03233533, 0.03380639\\ &0.03527218, 0.03683043, 0.03861879, 0.04093687, 0.04503625\}\\ \tau \in &\{0.02811741, 0.08908788, 0.15769846, 0.23614194, 0.32771585,\\ &0.43771735, 0.57548019, 0.75992392, 1.03994281, 1.64216826\}\end{aligned}$$

Lastly, a total of 50 Monte Carlo samples was used to evaluate the criteria of m, spatially averaged standard deviation, for now denoted as SASD.

Table 1: SASD with prior parameters

$Z_a \setminus Z$	Design 1	Design 2	Design 3	Design 4	Design 5	Design 6	Design 7	Design 8
Design 1	-	548.0111	490.2626	472.2716	**465.4389**	**461.4315**	**458.3172**	**457.1136**
Design 2	548.0420	-	549.6023	490.1057	473.6499	465.6733	460.4351	458.5897
Design 3	490.2403	549.4966	-	547.7782	495.0345	474.8508	464.7677	461.2044
Design 4	472.3915	490.0712	547.8146	-	580.9237	498.8230	473.9019	466.1729
Design 5	465.5245	473.6227	495.0963	580.9661	-	573.4793	495.3815	476.5188
Design 6	461.4052	465.4790	474.7950	498.7048	573.4010	-	555.6704	495.8328
Design 7	458.3496	460.5030	464.8691	473.9674	495.4149	555.7863	-	559.4372
Design 8	**457.1382**	**458.4126**	**461.1775**	**466.1601**	476.5174	495.9166	559.3891	-

Evidently from the results, design 1 or design 8 are the ones prefered by our criteria, although only marginally in some cases. Based on the geometry of the design paths used in this analysis, this doesn't come as any surprise. As a note, the preferred addition to the existing path is seemingly the path that covered most of the area not included by the initial path. To assess the value of integrating the prior parameters out from the posterior distribution, the natural comparison are the same results with fixed parameters. In these simulations $\sigma^2 = 0.03486239$ and $\tau = 0.5481692$, which were the optimized values found by equation (18) on the input-variogram. the results are found in table (2).

Table 2: SASD with fixed parameters

$Z_a \setminus Z$	Design 1	Design 2	Design 3	Design 4	Design 5	Design 6	Design 7	Design 8
Design 1	-	548.3832	490.2182	471.9795	**464.9614**	**460.7763**	**457.5650**	**456.2909**
Design 2	548.3832	-	549.8854	490.0902	473.3898	465.1057	459.8781	457.8017
Design 3	490.2182	549.8854	-	548.2868	495.0969	474.4967	464.3427	460.5580
Design 4	471.9795	490.0902	548.2868	-	581.2475	498.6920	473.6160	465.6364
Design 5	464.9614	473.3898	495.0969	581.2475	-	573.8424	495.4848	476.2670
Design 6	460.7763	465.1057	474.4967	498.6920	573.8424	-	555.9909	495.9254
Design 7	457.5650	459.8781	464.3427	473.6160	495.4848	555.9909	-	559.8999
Design 8	**456.2909**	**457.8017**	**460.5580**	**465.6364**	476.2670	495.9254	559.8999	-

We see that the results coincide almost exactly as the simulations performed with prior dis-

tributions for the parameters σ^2 and τ. As the results are similar for both simulations in terms of path design combinations, a measure was needed to see if the predictions produces different predictive results. Nooot

4.2.4 Notes on computational cost

An important aspect to consider when assessing these two methods used is the computational cost of using these methods.

Integrating the parameters out by a numerical procedure is time-consuming due to the many matrix operations that are demanded in both GLS-estimation and evaluation of the multivariate Gaussian distribution. A quick analysis of the required calculations yields a running-time of $\mathcal{O}(n_\tau \cdot n_{\sigma^2} \cdot n^3_{gridsize} \cdot n_{MonteCarlo})$ for the model using prior parameters. Assuming a rectangular grid with dimensions $n_x \cdot n_y$ we get the running time as $\mathcal{O}(n_\tau \cdot n_\tau \cdot n_x^3 \cdot n_y^3 \cdot n_{MonteCarlo})$ versus a $\mathcal{O}(n_x^3 \cdot n_y^3 \cdot n_{MonteCarlo})$ for fixed parameters. Although the procedure for fixed parameter is quicker by some constantfactors, it would still be unacceptable in most cases when n_{grid_size} becomes large. Based on the time used to derive the results in tables (1) and (1), I would characterize them as *extremely slow* in the model with prior parameter and *slow* despite the relatively small size of the gridded area.

4.3 Discussion

By the results obtained in tables (1) and (2), combined with ..., there's strong arguments for saying that the two models don't yield significant different results. When viewing the results, the effect of estimating the trend and how it affects the covariance of the latent process. As the trend has to be estimated, we introduce more variability into the model, perhaps in such a way that the GP-covariance "drowns" in the covariance introduced by estimating the trend by GLS. Evidently, this may be the case as the variogram gives a very low estimated σ^2, and is obviously the case when looking away from the sample paths as the standard deviation of examples such as figures (**??**) and (**??**) are much higher in some areas. The end result may be that the parameter σ^2 becomes of little importance.

Another possible explanation as to why the prior parameter model and fixed parameter model yielded the same results may be the fact that they were based on a variogram that is in essence a very accurate calculation. In that sense, assigning a prior model on the basis of this variogram could be interpreted as attaining too much variability to something that has in essence very little variability.

The correlation parameter τ achieves a somewhat decent estimate, with an estimated correlation of ≈ 0.07 at distance 3 in the Matern covariancefunction with $\nu = \frac{5}{2}$. However, this means that the covariance matrix of the latent GP is in reality quite sparse. This could be a motivation for redefining the model in some way to exploit this fact, as the cubic term in the computational running-time arises from the many inverses and determinants that has to be computed from the covariance matrix of the model.

Looking back at the results, one has to evaluate the simulations that were performed as well.

The discretization of the priorgrid for σ^2 and τ could be increased, the measurement noise could be varied, other covariancefunctions should have been tried out and especially the number of Monte Carlo samples obtained in the evaluation of the criteria. This was all neglected due to a combination of the computational cost of performing multiple simulations and how much time that was available. A possible tool would be to use parallellization of the computations as each Monte Carlo sample may be performed independently of any other. However, this doesn't solve the underlying problem as it won't change the immense running time of the calculations.

Another aspect that should have been altered in simulations is the number and variety of sample paths that were tested. At the very least more varied paths should have been tested, as to perhaps achieve more informative results. Given more time and effort, some sort of "path designer" could have been programmed, so as to construct proposals for paths that were probable to adhere to the criteria in equation (16).

5 Conclusion

The results underlines the notion that using a prior model and using fixed parameter is of no difference. Evidently from the pictures in ..., the two methods also yield the same predictions. However, the quality of simulations are somewhat poor if they are to be used in such a comparison as we don't have too much theory to aid us in separating the two procedures. The final conclusion based on the results is that the optimal paths are the combinations that cover the most of the field, but again, the simulations was possible performed with too few sample paths to compare.

With respect to the criterias presented in section (**??**) 4, the fixed parameter procedure would be preferred with the main argument being the difference in computational cost. As described, this yielded a quicker calculation with the same results and is therefore preferred.

Recommende further work would be to possible redefine the model in such a way to exploit the sparseness of the covariancematrix. As the case is presented with a variogram, one could possible redefine the model as a Gaussian Random Markov Field. One could also utilize INLA as an approximation technique, as the latent variables are modelled by a Gaussian process in both cases.

References

[1] Eidsvik, Jo and Mukerji, Tapan & Bhattacharjya, Debarun (2015)
Value of Information in the Earth Sciences - Integrating Spatial Modeling and Decision Analysis,
Cambridge University Press, 2015.

[2] Rasmussen, C. E. & Williams, C. K. I.
Gaussian Processes for Machine Learning,
the MIT Press, 2006.

[3] Diggle, Peter & Lophaven, Søren
Bayesian Geostatistical Design,
Scandinavian Journal of Statistics, volume 33 (2006): 53-64, print.

[4] Cressie, Noel & Wikle, Christopher K.
Statistics for Spatio-Temporal Data,
John Wiley & Sons (2011).

[5] Le, Nhu D. & Zidek, James V.
Statistical Analysis of Environmental Space-Time Processes,
Springer Science+Business Media Inc. (2006).

[6] Binney, Jonathan & Krause, Andreas & Sukhatme, Gaurav S.
Optimizing waypoints for monitoring spatiotemporal phenomena,
The International Journal of Robotics Research (2013), print.