CZECH UNIVERSITY OF LIFE SCIENCES PRAGUE
FACULTY OF ENVIRONMENTAL SCIENCES

COMPUTATION ASPECTS OF KRIGING IN CHOSEN ENGINEERING PROBLEMS

Martina Valtrová

April 2009

DECLARATION

I declare that this thesis entitled "Computation Aspects of Kriging in Chosen Engineering Problems" is the result of my own research under guidance of Ing. Petr Máca, Ph.D. and Ing. Matěj Lepš, Ph.D. except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

April 30, 2009 Martina Valtrová

. . . . . .

ACKNOWLEDGMENTS

I wish to express my gratitude and thanks to my supervisors, Ing. Petr Máca, Ph.D. and Ing. Matěj Lepš, Ph.D. for their continuous support and supervising during the whole course of this work. Special thanks are to RNDr. Vladimír Puš, CSc. whose lectures made an invaluable contribution to my education.

Most importantly, I would like to thank Zvenek for his patience, moral support and self-sacrifice.

Not least, the financial support of this work by the research project GAČR 103/07/P554 is gratefully acknowledged.

Abstract

Computation Aspects of Kriging in Chosen Engineering Problems

by Martina Valtrová

In this thesis, two different approaches to Kriging with applications to data are proposed. Kriging is a popular interpolation/approximation method used in different disciplines ranging from the mining and geology, hydrology, meteorology, environmental sciences, global optimization and even structural engineering. The first approach is aimed at Kriging in geostatistics. This access to Kriging is the most widely used. The goal of this method is to predict values of covariance stationarity process at unsampled points with respect to the mean squared error. However, the covariance function is usually not known and has to be estimated. Kriging is the best linear unbiased predictor, i.e. it has the smallest mean squared error among all linear predictors. We also describe what the unbiasedness is and how this condition influences the prediction.

There are many versions of Kriging used in geostatistics. In this study we present the ordinary Kriging, simple Kriging, universal Kriging and the co-Kriging, in detail and inclusive of their mathematical derivation. For geostatistics two or three dimensional data are common. Thus, we show the geostatistical application of Kriging to two dimensional data, and we also remit some problems with estimation of the covariance function or variogram.

The second access to Kriging described in this study is the global optimization approach. Here, Kriging is used as an approximation. Instead of geostatistics, Kriging does not have to be an exact interpolator (does not honor the data). We do not estimate the covariance function or variogram but we have to estimate the correlation function. We apply this approach to less common eight dimensional data using the free Matlab toolbox DACE. We focused mainly on these data. The problems with the compliance of properties of a function which is approximated are also discussed.

Abstrakt

Výpočetní aspekty metody Kriging ve vybraných inženýrských problémech

Martina Valtrová

Tato diplomová práce se zabývá dvěmi rozdílnými přístupy k metodě Kriging. Kriging patří mezi oblíbené interpolační metody. Využití nachází v mnoha různých odvětvích jako je důlní inženýrství, pro jehož účely byla vynalezena, geologii, hydrologii, meteorologii, přírodních vědách a v neposlední řadě i v optimalizaci. Nejprve je představen geostatistický přístup. V geostatistice je Kriging velmi rozšířenou metodou a většina aplikací využívá právě tento přístup. Cílem je predikce hodnot v bodech, kde daný přírodní proces nebyl pozorován. Kriging je tak definován jako nejlepší nestranný lineární odhad, což znamená, že tento odhad má nejmenší kvadratickou chybu mezi všemi ostatními lineárními odhady. Aby byla splněna podmínka nestrannosti, předpokládá se, že odhadovaný proces je kovariačně stacionární, přičemž se vybírá taková kombinace parametrů, která zajišťuje nejmenší rozptyl. Naneštěstí, kovarianční proces je obvykle neznámý a musí být odhadnut.

Přestože existuje mnoho variant metody Kriging, v této práci se podrobně zabýváme pouze jednoduchým Krigingem (simple Kriging), ordinárním Krigingem (ordinary Kriging), univerzálním Krigingem (universal Kriging) a kokrigingem (co-Kriging), přičemž uvádíme i jejich matematické odvození. Dále v této práci popisujeme aplikaci Krigingu na dvou dimenzionální data a zmiňujeme zde problémy, které nastaly při odhadu semivariogramu.

Druhý přístup popisuje Kriging jako optimalizační metodu. Narozdíl od geostatistiky, kde Kriging interpoluje data, v tomto pří stupu aproximuje danou funkci, tudíž ve známých bodech nemusí být odhadovaná hodnota shodná s pozorovanou hodnotou. Také tento typ Krigingu jsme aplikovali na data za použití volně dostupného toolboxu DACE v Matlabu. Jedná se o osmi dimenzionální data, která nejsou příliš běžná při používání této metody. V závěru uvádíme problémy, které se vyskytly při aproximaci a možné směry, jakými by se měl ubírat další výzkum této metody.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION AND HISTORICAL REVIEW

Kriging is an approximation method frequently used in geostatistics, global optimization and statistics. Kriging was originally developed by the South African mining engineer D.G. Krige in the early fifties. In the 1960s the French mathematician G. Matheron gave theoretical foundations to this method [Matheron, 1963]. More recently, Stacks et al. [Sacks et al., 1989], and Jones et al. [Jones et al., 1998] made Kriging well-known in the context of the modeling, and optimization of deterministic functions [Queipo et al., 2005]. Since 1970 Kriging has been intensively adapted, extended, and generalized.

The principle of Kriging is to find the optimal prediction of the covariance stationary process due to the mean squared error. Usually the covariance structure is not known and has to be estimated. This work focuses on the approach in geostatistics and global optimization. The main goal is to present these two approaches in detail with applications to the different data.

In chapter two we introduce some basic statistical concepts which appear throughout this thesis, and which is crucial for right understanding of mathematical background. However, the mentioned statistics which are taken from [Montgomery, 2005], is basic and notorious, therefore the aim of this chapter is just to refresh the knowledge.

The chapter three is concerned with the geostatistical approach to Kriging. The data interpolated by this method have two or three dimensions, therefore they are often called spatial data. These data usually include measurement errors as well. In this chapter we will describe and mathematically derive the ordinary Kriging, universal Kriging and co-Kriging. The ordinary Kriging is presented at first because other types of Kriging can be derived from it. Thus, all types are based on this equation of Kriging predictor:

$$\widehat{Z}(\mathbf{s}_0) = \left(\boldsymbol{\gamma} + \mathbf{1}\frac{(1 - \mathbf{1}'\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma})}{\mathbf{1}'\boldsymbol{\Gamma}^{-1}\mathbf{1}}\right)\boldsymbol{\Gamma}^{-1}\mathbf{Z} \quad . \tag{1.1}$$

In the last theoretical chapter we show the global optimization approach to Kriging.

First of all, we present the model of a linear regression and how can be modified to the general Kriging model:

$$Z(\mathbf{s}) = f(\mathbf{s}) + \epsilon(\mathbf{s}) \quad . \tag{1.2}$$

The Kriging is an approximation method in this approach, i.e. the Kriging predictor does not honor the data. Further, the correlation function and its types will be described with the estimation of the unknown parameters. The data, on which the global optimization is applied, are multidimensional and less common for geostatistical Kriging.

Last but one chapter contains application to data. The two dimensional data are introduced at first. The main attention is paid for the second data set which has eight dimensions. The results of an approximation function are presented to show the applicability of Kriging.

In the last chapter we summarize the Kriging method and also we describe the problems which occurred during this research.

1.1 Notation

For better orientation in the following text, we present used notation.

General Notation

$\mathbf{a}$	Vector
$\mathbf{A}$	Matrix
$\mathbf{a}'$	Transpose of vector
$\mathbf{A}'$	Transpose of matrix
$\mathbf{A}^{-1}$	Inverse of matrix
$\lvert\mathbf{a}\rvert$	Magnitude of vector
$\lvert\mathbf{A}\rvert$	Determinant of matrix A
$\lVert\mathbf{a}\rVert$	Euclidean norm of $\mathbf{a}$
$\hat{A}$ or $\hat{a}$	Predicted value of A or a

Geostatistics Approach to Kriging

Z	Set of random variables or random field
z	Outcome of random variable or a value of random field
$Z \sim N(\mu, \sigma^2)$	Z has the normal distribution with the mean μ and the variance σ^2
$\{z_1, z_2, \ldots, z_n\}$	Set of possible outcomes of random variable
$z_1, z_2, z_3, \ldots$	Actually observed outcomes of random variable or random field
D^p	p-dimensional domain of interest
$\mathbf{s}$	Site of the domain D
$Z(\mathbf{s})$	Random function at location $\mathbf{s}$
$z(\mathbf{s})$	Realization of the random function $Z(\mathbf{s})$
$E[Z(\mathbf{s})]$	Expected value of random function
μ	Mean
$Var[Z(\mathbf{s})]$	Variance of random function
σ^2	Variance
med	Median
$\text{cov}(Z(\mathbf{s}_i), Z(\mathbf{s}_j))$	Covariance
$C(\mathbf{s}_i - \mathbf{s}_j)$ or $C(\mathbf{h})$	Covariance function
$2\gamma(\mathbf{s}_i - \mathbf{s}_j)$ or $2\gamma(\mathbf{h})$	Variogram
$\gamma(\mathbf{s}_i - \mathbf{s}_j)$ or $\gamma(\mathbf{h})$	Semivariogram (sometimes called simply variogram)
$N_{\mathbf{h}}$	Number of pairs separated by the same distance $\mathbf{h}$
c_0	Nugget
c	Sill
a	Range
$\epsilon(\mathbf{s})$	Correlated error process
$\boldsymbol{\lambda}$	Vector of unknown coefficients
MSE or σ_e^2	Mean squared error
ϱ or $\boldsymbol{\varrho}$	Lagrangian multiplier or vector of Lagrangian multipliers

$\boldsymbol{\gamma}$	Vector of variograms such that $\boldsymbol{\gamma} = (\gamma(\mathbf{s}_0 - \mathbf{s}_1), \ldots, \gamma(\mathbf{s}_0 - \mathbf{s}_n))'$
$\boldsymbol{\Gamma}$	Variogram matrix
$\sigma_k^2(\mathbf{s}_0)$	Kriging or predictor variance
$\boldsymbol{\Sigma}$	Matrix of covariance functions
$\boldsymbol{c}$	Vector of covariance function such that $\boldsymbol{c} = (c(\mathbf{s}_0 - \mathbf{s}_1), \ldots, c(\mathbf{s}_0 - \mathbf{s}_n))'$
$C_{YZ}(\mathbf{h})$	Cross-covariance
$2\gamma_{YZ}(\mathbf{h})$ or $\gamma_{YZ}(\mathbf{h})$	Cross-variogram
$Y(\mathbf{s})$	Secondary variable (co-Kriging)
$\boldsymbol{\omega}$	Vector of unknown coefficients belongs to secondary variables

Global Optimization Approach to Kriging

$f(\mathbf{s})$	Linear or nonlinear function of $\mathbf{s}$
$\boldsymbol{\beta}$	Vector of unknown coefficients of regression term
$\boldsymbol{\epsilon}$	Vector of i.i.d. errors
$d(\mathbf{s}_i, \mathbf{s}_j)$	Distance between points $\mathbf{s}_i$ and $\mathbf{s}_j$
θ_h	Unknown parametr which measures the importance of variable $\mathbf{s}_h$
p_h	Unknown parametr which is related to the smoothness of the function in the direction h
$\text{Corr}[\epsilon(\mathbf{s}_i), \epsilon(\mathbf{s}_j)]$ or $\rho[\epsilon(\mathbf{s}_i), \epsilon(\mathbf{s}_j)]$	Correlation between errors
$\mathbf{Z}$	Vector of observed values such that $\mathbf{Z} = (Z_1, \ldots, Z_n)$
$\mathbf{R}$	Correlation matrix
$\mathbf{r}$	Vector or correlation such that $\mathbf{r} = (\text{Corr}[\epsilon(\mathbf{s}_0), \epsilon(\mathbf{s}_1)], \ldots, \text{Corr}[\epsilon(\mathbf{s}_0), \epsilon(\mathbf{s}_n)])$

Chapter 2

SHORT INTRO INTO STATISTICS

First of all some basic statistical concepts will be introduced which are crucial for correct understanding of this work, e.g. expected value, variance, normal distribution etc. These concepts can be found in usual statistics literature, here we follow the description from [Montgomery, 2005].

2.1 Discrete and Continuous Random Variables

In [Bárdossy, 1997] a *random variable* is a function Z which might take different values (outcomes) with given probabilities. An upper case letter Z usually denotes a random variable. A lower case letter, such as z, is used to denote an outcome of a random variable. If the outcomes form a finite or countably infinite set then we speak of a *discrete random variable*. In other words, a variable is discrete if it can have only certain numerical values, with no intermediate values in between, for example a numerical result of throw a die:

$$Z \in \{1, 2, 3, 4, 5, 6\}$$

While a *continuous random variable* might take all possible values of a given interval, for example length:

$$Z \in \langle 0, \infty) \quad .$$

In the notation of discrete variable a distinction is made between the set of possible outcomes which is denoted by $\{z_1, z_2, \ldots, z_n\}$, and the actually observed outcomes $z_1, z_2, z_3, \ldots$

Another important term is the probability distribution. In short, the set of all possible values of a discrete random variable and their probabilities is *probability distribution* described by the *probability function*

$$p(z) = P(Z = z) \quad . \tag{2.1}$$

It means that the probability function gives the probability that a discrete random variable Z will take a particular value z. It is obvious that

$$\sum p(z) = 1 \quad . \tag{2.2}$$

The actual values of a continuous random variable is denoted by $z_1, z_2, z_3, \ldots$ and for all possible values is used the interval $Z \in (z_1, z_2)$. A continuous random variable has no probability function but can be described by the *probability density function*

$$P(z_1 \leq Z \leq z_2) = \int_{z_1}^{z_2} f(z)\,\mathrm{d}z \quad . \tag{2.3}$$

The key property of a function $f(z)$ is

$$\int_{-\infty}^{\infty} f(z)\,\mathrm{d}z = 1 \quad . \tag{2.4}$$

The probability density function is connected to the cumulative distribution function or shortly the *distribution function* that can be defined for both discrete and continuous random variables. It gives the probability that Z takes a value that is lower or equal then z. The distribution function is described for discrete variables as

$$P(Z \leq z_1) = F(z_1) = \sum_{i=-\infty}^{z_1} p(z_i) \tag{2.5}$$

and for continuous variables as

$$P(Z \leq z_1) = F(z_1) = \int_{-\infty}^{z_1} f(z)\,\mathrm{d}z \quad . \tag{2.6}$$

2.2 Parameters of Random Variables

In this section some important parameters of random variables will be introduced. Let $g(Z)$ denote any function of a random variable Z. Then the *expected value* of discrete random variable Z is defined as

$$E[g(Z)] = \sum_{z} g(z)p(z) \tag{2.7}$$

and in case of a continuous variable is defined as

$$E[g(Z)] = \int_{-\infty}^{\infty} g(z)f(z)\,\mathrm{d}z \quad . \tag{2.8}$$

The expected value plays a role of a gravity centre because

$$E[Z - E[Z]] = 0 \tag{2.9}$$

The list of some properties of the expected value follows from which it is obvious that the expected value has a linear behavior:

$$E[Y + Z] = E[Y] + E[Z] \tag{2.10}$$

$$E[cZ] = cE[Z] \tag{2.11}$$

$$E[-Z] = -E[Z] \tag{2.12}$$

$$E[c] = c \tag{2.13}$$

The *mean* of a random variable is

$$\mu = E[Z] \quad . \tag{2.14}$$

For a discrete random variable, this is

$$\mu = \sum_z zp(z) \tag{2.15}$$

and for a continuous random variable it is defined as

$$\mu = \int zf(z)\,\mathrm{d}z \quad . \tag{2.16}$$

Another parametr of random variables is the *variance:*

$$\sigma^2 = Var[Z] = E[(Z - E[Z])^2] = E[Z^2] - E^2[Z] \quad . \tag{2.17}$$

For a discrete random variable the variance can be derived as

$$Var[Z] = \sum_z p(z)z^2 - E^2[Z] \tag{2.18}$$

and for a continuous random variable as

$$Var[Z] = \int_{-\infty}^{\infty} z^2 f(z)\,\mathrm{d}z - E^2[Z] \quad . \tag{2.19}$$

All parameters above are derived for one random variable Z. Next, we will describe parameters which are valid for two variables, e.g. random variables Z and Y. Two random

variables are independent if probabilities taken by a random variable Y are independent on probabilities taken by a random variable Z and oppositely. This independence is measured by the *covariance*:

$$\mathrm{cov}(Y,Z) = E[(Y - E[Y])(Z - E[Z])] \quad . \tag{2.20}$$

It is obvious that

$$\mathrm{cov}(Z,Z) = E[(Z - E[Z])(Z - E[Z])] = E[(Z - E[Z])^2] = var[Z] \quad . \tag{2.21}$$

The four possible cases may occur:

1. $\mathrm{cov}(Y,Z) = 0$ random variables are independent (uncorrelated).

2. $\mathrm{cov}(Y,Z) \neq 0$ random variables are correlated.

3. $\mathrm{cov}(Y,Z) < 0$ random variables are negatively correlated. It means that high values of variable Y can be associated with low values or variable Z.

4. $\mathrm{cov}(Y,Z) > 0$ random variables are positively correlated. Then high values of variable Y can be associated with high values of variable Z.

The covariance is connected with the term *correlation coefficient*. This coefficient is equal to the covariance divided by the standard deviation (= square root of variance) of both variables:

$$\rho_{Y,Z} = \frac{\mathrm{cov}(Y,Z)}{\sigma_Y \sigma_Z} \quad . \tag{2.22}$$

The rank of covariance is

$$-\infty \leq cov(Y,Z) \leq \infty \tag{2.23}$$

in contrast to correlation

$$-1 \leq \rho_{Y,Z} \leq 1 \tag{2.24}$$

As well as in the covariance the four cases may occur:

1. $\rho_{Y,Z} = 0$ random variables Y and Z are uncorrelated.

2. $\rho_{Y,Z} \neq 0$ random variables are correlated.

3. $\rho_{Y,Z} < 0$ random variables are negatively correlated.

4. $\rho_{Y,Z} > 0$ random variables are positively correlated.

Both the covariance and the correlation coefficient are measures of the linear correlation between two random variables but the advantage of the covariance is that it gives information about the magnitude of the data, while the correlation coefficient is normalized with regard to standard deviations. In some cases, it could by also an advantage. However, the covariance and correlation coefficient are influenced by outliers, where the data can be highly correlated but the correlation coefficient is low.

The normal distribution is the most common way how to describe a distribution of random variables. Shortly, it fits on many variables of an empirical nature, obtained from limited and observed data. A random variable Z is normally, or Gaussian, distributed if its probability distribution function is defined as

$$f(z) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(z-\mu)^2\right] \quad . \tag{2.25}$$

where μ is the mean and σ^2 is the variance.

If the random variable Z has the normal distribution with the mean μ and the variance σ^2 it is usual to write $Z \sim N(\mu, \sigma^2)$. The normal distribution $N(0,1)$ is called the standard normal distribution. The normal distribution is crucial due to the central limit theorem, which states that the sum of n independent random variables with a common probability distribution function will approximate a normal distribution when n becomes very large.

Chapter 3

GEOSTATISTICS APPROACH TO KRIGING

3.1 Geostatistics

The apt formulation what geostatistics means is nicely expressed by Caers [Bohling, 2007]:

> "In its broadest sense, geostatistics can be defined as the branch of statistical sciences that studies spatial phenomena and capitalizes on spatial relationships to model possible values of variable(s) at unobserved, unsampled locations."

The name geostatistics has arisen by application of Theory of Regionalized Variables to problems in geology and mining [Clark, 1979]. However, nowadays the geostatistical methods are used in very different disciplines ranging from hydrology, meteorology, environmental sciences, structural engineering etc.

In [Bárdossy, 1997] the main concept of Theory of Regionalized Variables is a *random function*. A random function denoted as Z is a set of random variables corresponding to the points of a spatial region of interest D. Usually D is a two or three dimensional region of the earth's surface. Any site of the domain D is denoted as $\mathbf{s}$, in two dimensions $\mathbf{s} = (x, y)'$ and $\mathbf{s} = (x, y, z)'$ in three dimensions, respectively. The realization of a random function Z is called a *regionalized variable*. Then $z(\mathbf{s})$ is one realization of the random function $Z(\mathbf{s})$ at a location $\mathbf{s}$, in other words $z(\mathbf{s})$ is an observation of some phenomena. $Z(\mathbf{s})$ can be generalized as a multivariate field, i.e. $Z(\mathbf{s})$ represents a suite of variables at a location $\mathbf{s}$. Note that D contains an infinite number of sites but observations may only be taken on a finite collection of sample sites. Let $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n$ denote a such collection of sample sites. $\mathbf{s}_0$ represents unsampled site with value $Z(\mathbf{s}_0)$ which have to be predicted.

3.2 Stationarity

If the geostatistical methods should be optimal then the data have to be stationary. The random function $Z(\mathbf{s})$ is stationary if distribution of this function depends only on the configuration of the points, not on their locations. This form of stationarity is called *strong stationarity*. However, an assumption of strong stationarity is overmuch complex therefore two simpler assumptions are considered. *The second order stationarity* is a weaker form of stationarity and is often used in time series analysis, where the raw data are usually transformed to become stationary. The random function $Z(\mathbf{s})$ is second order stationary if the three following assumptions are satisfied:

1. The expected value does not change over time or position. In other words, the mean is constant:

$$E[Z(\mathbf{s})] = \mu \quad . \tag{3.1}$$

2. The second assumption is similar to the first one, the variance is invariant over time or position:

$$Var[Z(\mathbf{s})] = \sigma^2 \quad . \tag{3.2}$$

3. The covariance function only depends on the difference of arguments, i.e. on the locations (distance and direction) of the pair of sites $\mathbf{s}_i, \mathbf{s}_j \in D$:

$$\mathrm{cov}(Z(\mathbf{s}_i), Z(\mathbf{s}_j)) = C(Z(\mathbf{s}_i), Z(\mathbf{s}_j)) = C(\mathbf{s}_i - \mathbf{s}_j) = C(\mathbf{h}) \quad . \tag{3.3}$$

If the covariance function only depends on the length $\mathbf{h}$ then $C(\mathbf{h}) = C(\|\mathbf{h}\|)$ where $\|\mathbf{h}\|$ denotes Euclidean norm of the vector $\mathbf{h}$. Such random field $Z(\mathbf{s})$ is called *isotropic*. From properties of the covariance mentioned above it is obvious that in case $\mathbf{h} = 0$ following is valid:

$$C(0) = Var[Z(\mathbf{s})] = \sigma^2 \quad . \tag{3.4}$$

Mentioned conditions mean that the random variables corresponding to different locations in a domain D have both the same expected value and the same variance. But the fulfillment of the second condition is rare thus weaker assumption can be formulated.

This assumption is called *intrinsic stationarity*. Two following conditions have to be satisfied:

1. First condition is the same as in the second order stationarity:

$$E[Z(\mathbf{s})] = \mu \quad . \tag{3.5}$$

2. The second one is different and says that the variance of increment depends on the difference of arguments:

$$Var[Z(\mathbf{s}_i) - Z(\mathbf{s}_j)] = E[(Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2] = 2\gamma(\mathbf{h}) \quad . \tag{3.6}$$

The function $2\gamma(\mathbf{h})$ is called *variogram* and $\gamma(\mathbf{h})$ is called *semivariogram*. As well as the covariance function the variogram is isotropic if $2\gamma(\mathbf{h}) = 2\gamma(\|\mathbf{h}\|)$. It is easy to see that the second order stationary process is also intrinsic stationary. However, the reverse is not true.

3.3 Covariance Function and Variogram

The covariance function $C(\mathbf{h})$ is valid if and only if it is positive semi-definite. A matrix $\mathbf{C}$ is positive definite if for every nonzero vector $\mathbf{a}$ the following expression is true [Shewchuk, 1994]:

$$\mathbf{a^T C\, a} \geq 0 \quad .$$

In other words, the covariance function $C(\mathbf{h})$ is positive definite if for all finite sets of sites $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n$ and all finite sets of constants $a_1, a_2, \ldots, a_n$ a following statement is valid [Cressie, 1993]:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j\, C(\mathbf{s}_i - \mathbf{s}_j) \geq 0 \quad . \tag{3.7}$$

On the contrary, the function $2\gamma(\mathbf{h})$ is valid if and only if it is conditionally negative definite [Cressie, 1993]:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j\, 2\gamma(\mathbf{s}_i - \mathbf{s}_j) \leq 0 \quad , \tag{3.8}$$

but sets of constants $a_1, a_2, \ldots, a_n$ are restricted by condition

$$\sum_{i=1}^{n} a_i = 0 \quad . \tag{3.9}$$

If the second order stationarity is valid the relationship between covariance function and variogram can be derived

$$\begin{aligned} 2\gamma(\mathbf{h}) = E[(Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2] = E[((Z(\mathbf{s}_i) - \mu) - (Z(\mathbf{s}_j) - \mu))^2] = \\ = E[(Z(\mathbf{s}_i) - \mu)^2] - 2E[(Z(\mathbf{s}_i) - \mu) - (Z(\mathbf{s}_j) - \mu)] + E[(Z(\mathbf{s}_j) - \mu)^2] = \\ = Var[Z(\mathbf{s}_i)] + Var[Z(\mathbf{s}_j)] - 2E[(Z(\mathbf{s}_i) - \mu) - (Z(\mathbf{s}_j) - \mu)] = \\ = 2C(\mathbf{0}) - 2C(\mathbf{h}) \end{aligned}$$

Consequently:

$$\gamma(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h}) \quad . \tag{3.10}$$

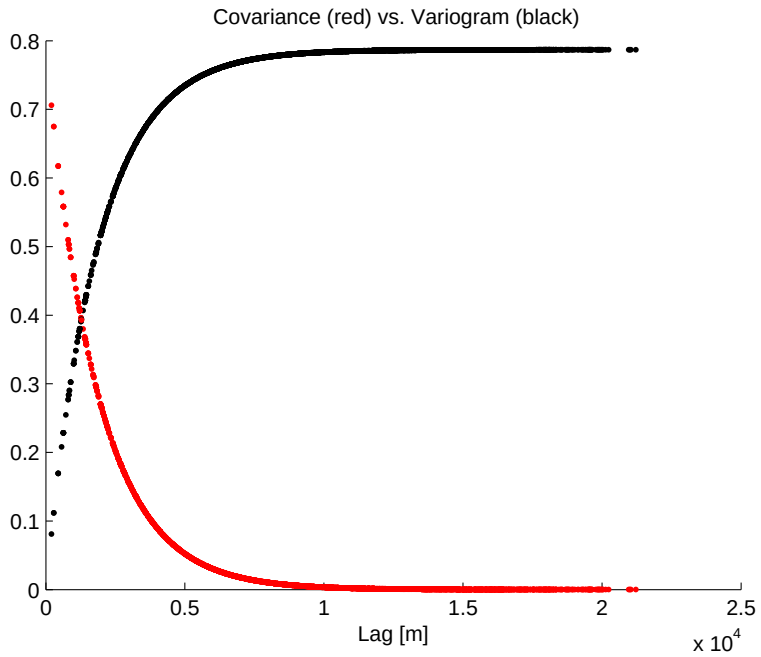

Figure 3.1: The relationship between the covariance function and the semivariogram (or simply variogram).

Covariance function is measure of the similarity of the points but variance is measure of the dissimilarity. Both covariance function and variogram play crucial role in geostatistics. However, the main problem of practical applications is that variogram neither covariance is not known and has to be estimated. First of all it is necessary to estimate the experimental covariance function and variogram based on Matheron's formulas [Cressie, 1993]:

$$\widehat{C}(\mathbf{h}) = \frac{1}{N_{\mathbf{h}}} \sum_{N_{\mathbf{h}}} (Z(\mathbf{s}_i) - \widehat{\mu})(Z(\mathbf{s}_j) - \widehat{\mu}) \quad , \tag{3.11}$$

where the $\widehat{C}(\mathbf{h})$ is estimated covariance function. The sum is over all pairs separated by the same distance $\mathbf{h}$, $N_{\mathbf{h}}$ is the number of such pairs and $\widehat{\mu}$ is the estimation of a mean computed according to

$$\widehat{\mu} = \frac{1}{n} \sum_{i=1}^{n} Z(\mathbf{s}_i) \quad . \tag{3.12}$$

And formula for the variogram reads [Michálek, 2001]:

$$2\widehat{\gamma}(\mathbf{h}) = \frac{1}{N_{\mathbf{h}}} \sum_{N_{\mathbf{h}}} (Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2 \quad . \tag{3.13}$$

In geostatistics the variogram is preferred against the covariance function. There are more reasons for that. It is obvious from intrinsic stationarity that if the covariance function exists then the variogram also exists. In some cases the covariance function cannot be defined but the variogram still exists. Another reason is stated in [Bárdossy, 1997]: "The covariance function is defined using the value of the expectation, while the variogram does not depend on this value. This is an advantage because slight trends do not influence the variogram severely, in contrast to the covariance function where through the improper estimation of the mean these effects are more severe." Owing to these reasons, the attention will be paid to the variogram rather then covariance function in the following text.

3.4 The Experimental Variogram

The variogram defined as (2.13) is estimated on the basis of measured (observed) data. Because the variogram is sensitive to data outliers, Cressie and Hawkins suggested two robust formulas for estimation of a variogram [Michálek, 2001]:

$$2\widehat{\gamma}(\mathbf{h}) = \frac{(\frac{1}{N_{\mathbf{h}}} \sum_{N_{\mathbf{h}}} |Z(\mathbf{s}_i) - Z(\mathbf{s}_j)|^{\frac{1}{2}})^4}{0.457 + \frac{0.494}{N_{\mathbf{h}}}} \quad . \tag{3.14}$$

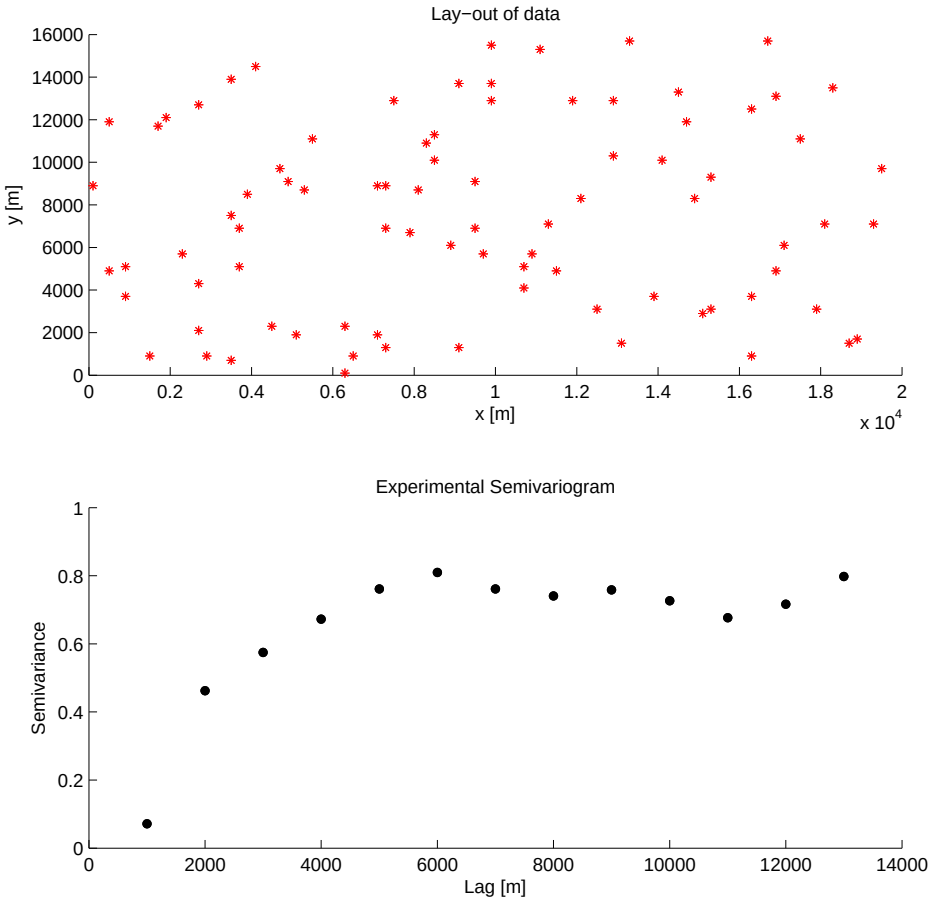

Figure 3.2: A Lay-out of data and the corresponding experimental semivariogram for lag 1000 m.

The second formula is based on a median:

$$2\widehat{\gamma}(\mathbf{h}) = \frac{(med\{|Z(\mathbf{s}_i) - Z(\mathbf{s}_j)|^{\frac{1}{2}} : (\mathbf{s}_i - \mathbf{s}_j) \in N_{\mathbf{h}}\})^4}{0.457} \quad . \tag{3.15}$$

It is proved that these robust formulas give better estimation then a Matheron's formula (2.13). Variograms (2.13), (2.14) and (2.15) are called *experimental variograms* and can be applied only if data points are regularly spaced. In the case of irregularly spaced data the summation could be carried out over the pairs fulfilling [Bárdossy, 1997]:

$$|\mathbf{s}_i - \mathbf{s}_j| - |\mathbf{h}| \leq \varepsilon \tag{3.16}$$

$$angle(\mathbf{s}_i - \mathbf{s}_j, \mathbf{h}) \leq \delta \tag{3.17}$$

where ε is a certain allowed difference in the length and δ is a certain allowed difference in the angle.

3.5 The Theoretical Variogram

The problem is that experimental variograms do not fulfill the condition (2.8), therefore models of *theoretical variograms* have to be introduced. However, it is necessary to explain some important terms at first.

3.5.1 Nugget Effect

The distance $\mathbf{h}$ is also called *lag*. Apparently the value of the semivariogram at lag equals to zero should be also zero:

$$\gamma(\mathbf{h}) = 0, \quad \text{for } \mathbf{h} \to 0 \tag{3.18}$$

If the semivariogram value is meaningfully different from zero then this value has been called the *nugget effect*:

$$\gamma(\mathbf{h}) \to c_0 > 0, \quad \text{for } \mathbf{h} \to 0 \tag{3.19}$$

It is mentioned in [Cressie, 1993] that this is caused by a microscale variation (variation at spatial scales shorter than separation of the sample sites) that makes a discontinuity at the origin, and also a measurement error.

3.5.2 Sill

It is supposed for very distant points that corresponding random variables are independent due to the second order stationarity. If variables $Z(\mathbf{s}_i)$ and $Z(\mathbf{s}_j)$ are independent then $C(\mathbf{h}) = 0$ and according to (2.10) and (2.4)

$$\gamma(\mathbf{h}) = C(\mathbf{0}) = \sigma^2 \quad . \tag{3.20}$$

It means that after a certain distance the semivariogram becomes constant. The semivariogram value c at which the semivariogram levels off is called the *sill* [Bohling, 2007]. It also means that if the condition of the second order stationarity is not satisfied, just the intrinsic stationarity is met, then the semivariogram does not have sill and it can show an unlimited increase. The nugget effect can be included also in the sill.

3.5.3 Range

Range is a certain lag distance at which the semivariogram achieve the sill. In other words, the range is the distance which separates the correlated and the uncorrelated random variables [Bárdossy, 1997].

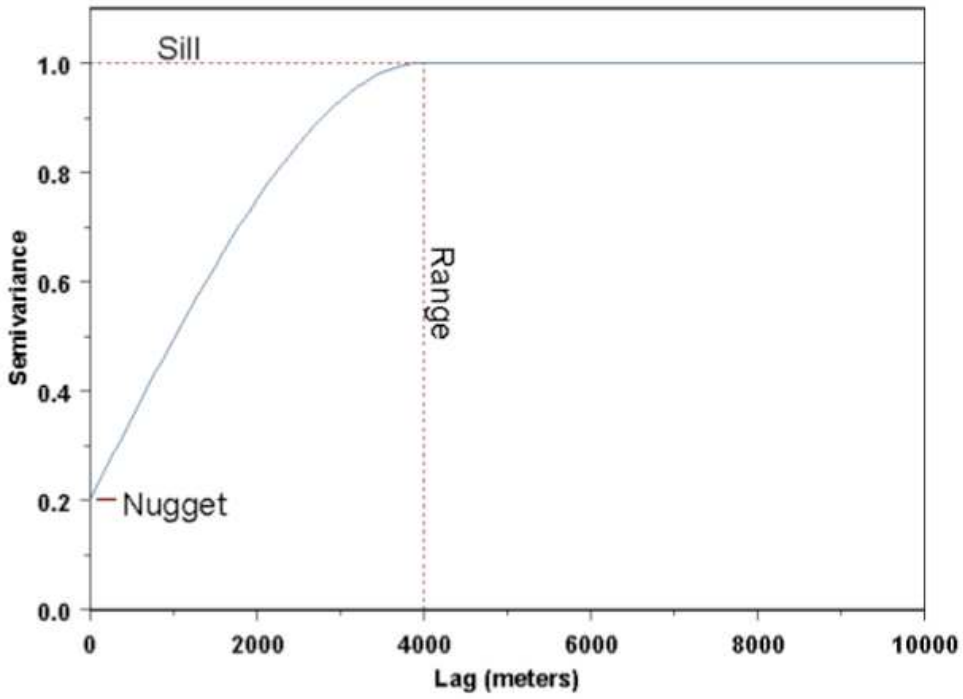

Figure 3.3: The nugget effect, the sill and the range are illustrated on the variogram.

3.5.4 Models of Theoretical Variogram

Theoretical semivariograms can be divided into two groups, with the sill and without the sill. As mentioned above, $\mathbf{h}$ represents lag distance, c_0 is the nugget effect and c represents the sill. Main semivariograms with the sill are [Cressie, 1993]:

1. Nugget:

$$\widehat{\gamma}(\mathbf{h}) = \begin{cases} 0 & \text{if } \mathbf{h} = 0 \\ c_0 & \text{if } \mathbf{h} > 0 \end{cases} \quad . \tag{3.21}$$

2. Spherical:

$$\widehat{\gamma}(\mathbf{h}) = \begin{cases} c_0 + c_s \left(\frac{3\,\mathbf{h}}{2\,a_s} - \frac{1\,\mathbf{h}^3}{2\,a_s^3} \right) & \text{if } \mathbf{h} \leq a_s \\ c_0 + c_s & \text{if } \mathbf{h} > a_s \end{cases} \quad , \tag{3.22}$$

where the sill is $C = c_0 + c_s$ and the range is a_s.

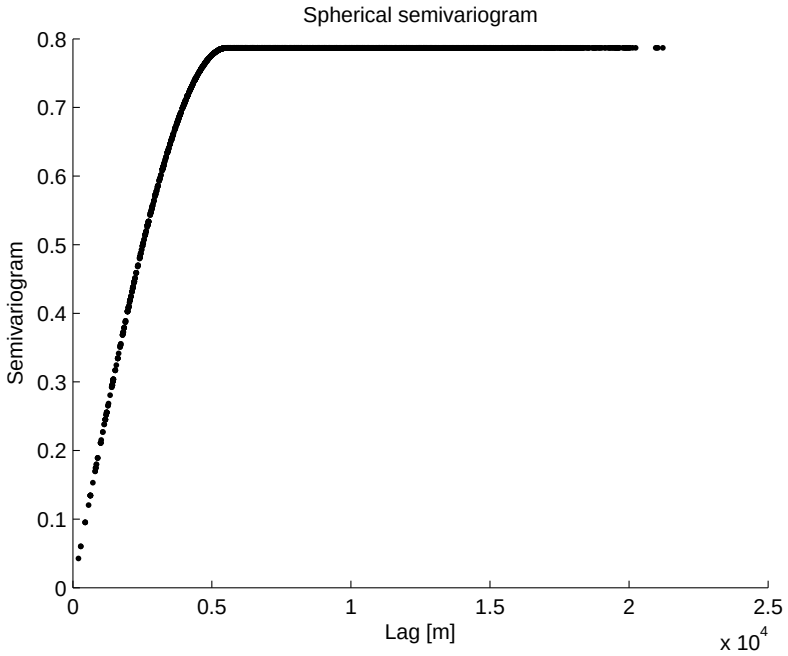

Figure 3.4: The Spherical Semivariogram.

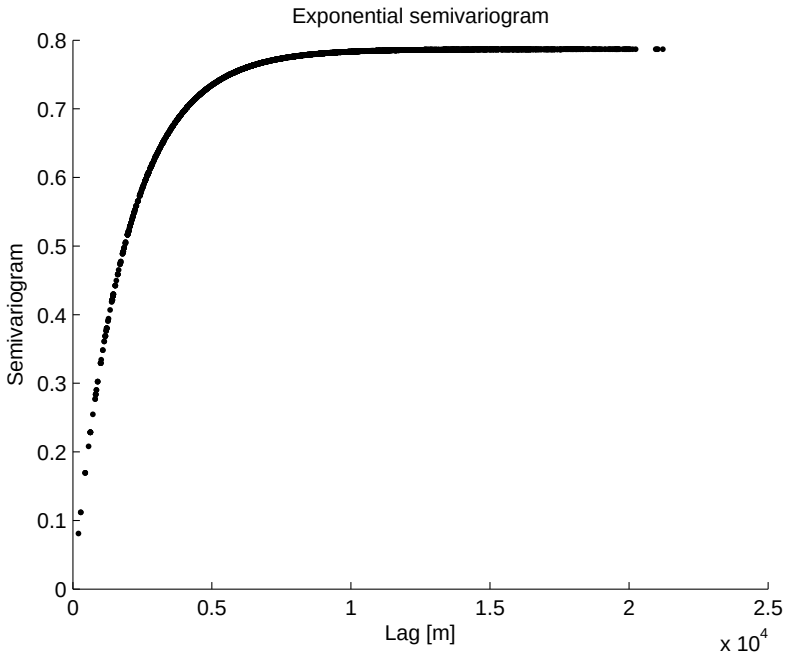

Figure 3.5: The Exponential Semivariogram.

3. Exponential:

$$\widehat{\gamma}(\mathbf{h}) = c_0 + c_e \left(1 - e^{\frac{-3\mathbf{h}}{a_e}}\right) \quad , \tag{3.23}$$

where $c_0 + c_e$ is the sill and a_e is the range.

4. Gaussian:

$$\widehat{\gamma}(\mathbf{h}) = c_0 + c_g \left(1 - e^{-\frac{3\mathbf{h}^2}{a_g^2}}\right) \quad , \tag{3.24}$$

where a_g is the range of spatial correlation and $c_0 + c_g$ is the sill.

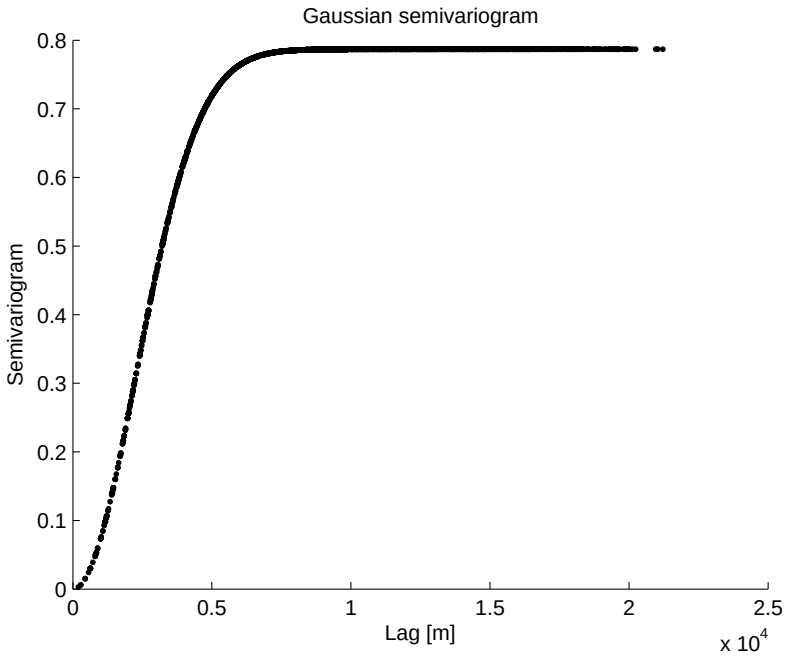

Figure 3.6: The Gaussian Semivariogram.

The semivariogram without the sill:

5. Power:

$$\widehat{\gamma}(\mathbf{h}) = c_0 + a_p \mathbf{h}^p \quad \text{for } 0 < p < 2 \quad , \tag{3.25}$$

where a_p represents the range. In case of $p = 1$ the semivariogram is linear. This semivariogram is often used. If the semivariogram has no sill then the variance of the process is infinite.

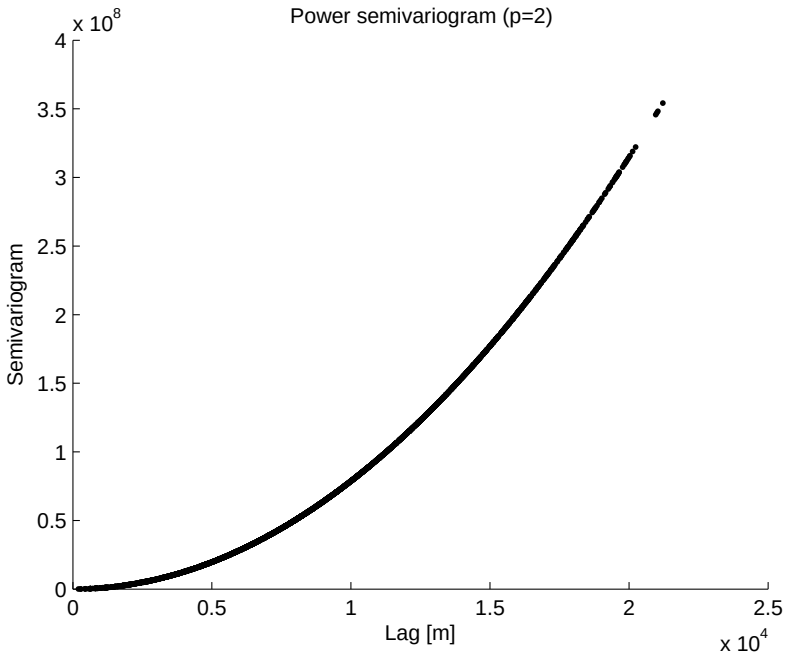

Figure 3.7: The Power Semivariogram.

3.6 Estimation of Suitable Semivariogram

The appropriate estimation of a semivariogram is obtained by fitting a theoretical curve to the experimental one see Fig. 3.8. In practical geostatistics the most extended method is "by eye." It is provided by plotting the experimental variogram and finding a linear combination of theoretical models whose plot is the most similar to the experimental one [Bárdossy, 1997].

This method has the advantage that it can detect errors of the data set, e.g. the intrinsic stationarity is not valid, there are trends in the data set, inhomogeneities etc. However, this method is disadvantageous because it is too subjective. Hence, the common practice is to use other methods, for example nonlinear least squared approach or the maximum likelihood method which will be discussed later.

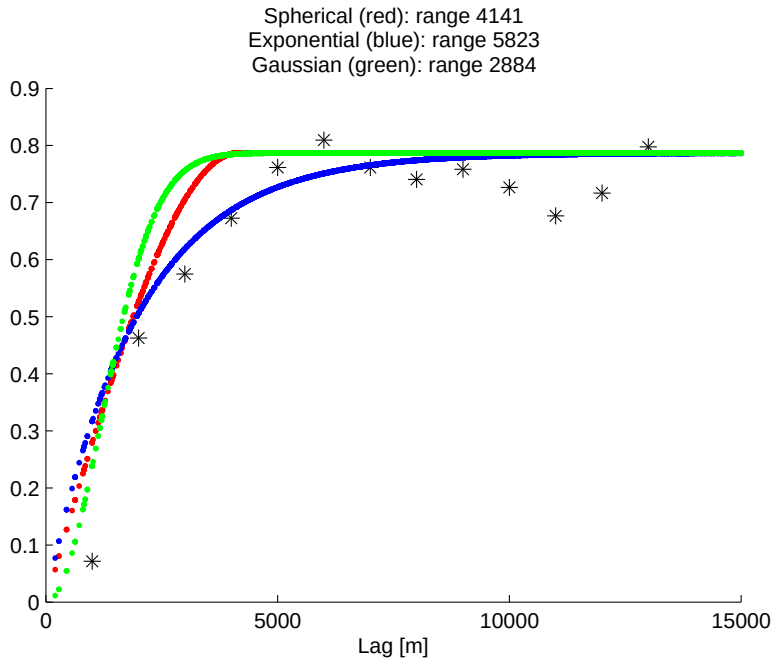

Figure 3.8: An example of fitting the curves of theoretical variograms to a curve of the experimental variogram.

3.7 Kriging

As mentioned earlier, the Kriging is the geostatistical method of predicting values at unsampled points. The semivariogram or variogram plays the crucial role in Kriging therefore the theory of variogram was discussed above. Kriging is similar to other interpolation methods, e.g. splines and radial basis functions. For more details about the relationship between Kriging and splines see [Watson, 1984], and the relationship between Kriging and radial basis function can be found in [Hornby and Horowitz, 1999].

There are many versions of Kriging. We introduce the ordinary Kriging at first because most of the other Kriging techniques can be derived from it. Further in this section we will discuss the universal Kriging and co-Kriging. The other types of Kriging, e.g. indicator Kriging, soft Kriging, fixed rank Kriging etc., can be found, e.g. in [Cressie, 1993], [Cressie and Johannesson, 2008], or [Bárdossy, 1997].

3.7.1 Ordinary Kriging

Let assume the model:

$$Z(\mathbf{s}) = \mu(\mathbf{s}) + \epsilon(\mathbf{s}) \quad . \tag{3.26}$$

The term $\mu(\mathbf{s})$ is the deterministic part of Z which describes *large-scale variability* of a process Z. The term $\epsilon(\mathbf{s})$ is a correlated error process and it describes *smooth-scale variability, micro-scale variability* and measurement errors [Michálek, 2001]. It is supposed that this term has following properties:

1.
$$E[\epsilon(\mathbf{s})] = 0 \tag{3.27}$$

2.
$$Var[\epsilon(\mathbf{s})] = \sigma^2 \tag{3.28}$$

   The variance does not depend on spatial location $\mathbf{s} \in D$.

3.
$$C(\mathbf{s}_i - \mathbf{s}_j) = \mathrm{cov}(\epsilon(\mathbf{s}_i), \epsilon(\mathbf{s}_j)) \tag{3.29}$$

   The covariance function only depends on the difference in the locations, i.e. the distance and direction of the pairs of sites $\mathbf{s}_i, \mathbf{s}_j \in D$.

If all these presumptions are satisfied then following is valid:

1.
$$E[Z(\mathbf{s})] = \mu(\mathbf{s}) \tag{3.30}$$

2.
$$Var[Z(\mathbf{s})] = \sigma^2 \tag{3.31}$$

3.

$$C(\mathbf{s}_i - \mathbf{s}_j) = \text{cov}(Z(\mathbf{s}_i), Z(\mathbf{s}_j)) \tag{3.32}$$

We also assume that the covariance function $C(\mathbf{h})$ or semivariogram $\gamma(\mathbf{h})$ is known.

Then the goal of Kriging is to find *the best linear unbiased predictor* (BLUP) or *the best linear unbiased estimator* (BLUE) in the point $\mathbf{s}_0$:

$$\widehat{Z}(\mathbf{s}_0) = \sum_{i=1}^{n} \lambda_i Z(\mathbf{s}_i) \quad , \tag{3.33}$$

where terms λ_i's are chosen to minimize the mean squared error:

$$\text{MSE} = E[(\widehat{Z}(\mathbf{s}_0) - Z(\mathbf{s}_0))^2] = E\left[(Z(\mathbf{s}_0) - \sum_{i=1}^{n} \lambda_i Z(\mathbf{s}_i))^2\right] \quad . \tag{3.34}$$

If the prediction $\widehat{Z}$ in the unsampled point $\mathbf{s}_0$ should be unbiased then the following condition has to be fulfilled:

$$E[(\widehat{Z}(\mathbf{s}_0)] = E[(Z(\mathbf{s}_0)] \quad , \tag{3.35}$$

then according to (3.30)

$$\begin{aligned} E[\widehat{Z}(\mathbf{s}_0)] &= E\left[\sum_{i=1}^{n} \lambda_i Z(\mathbf{s}_i)\right] \\ &= \sum_{i=1}^{n} \lambda_i E[Z(\mathbf{s}_i)] \\ &= \mu(\mathbf{s}) \sum_{i=1}^{n} \lambda_i \end{aligned}$$

Hence, the prediction $\widehat{Z}(\mathbf{s}_0)$ is unbiased for $Z(\mathbf{s}_0)$ if and only if

$$\sum_{i=1}^{n} \lambda_i = 1 \quad . \tag{3.36}$$

The mean squared error of predictor $\widehat{Z}(\mathbf{s}_0)$ is:

$$\text{MSE}(\mathbf{s}_0) = E\left[(\widehat{Z}(\mathbf{s}_0) - Z(\mathbf{s}_0))^2\right] = \sigma_e^2 \quad , \tag{3.37}$$

or equivalently:

$$\begin{aligned} E\left[(\widehat{Z}(\mathbf{s}_0) - Z(\mathbf{s}_0))^2\right] &= Var[\widehat{Z}(\mathbf{s}_0) - Z(\mathbf{s}_0)] \\ &= Var[\widehat{Z}(\mathbf{s}_0)] + Var[Z(\mathbf{s}_0)] - 2\text{cov}[\widehat{Z}(\mathbf{s}_0), Z(\mathbf{s}_0)] \\ &= Var\left[\sum_{i=1}^{n} \lambda_i Z(\mathbf{s}_i)\right] + Var[Z(\mathbf{s}_0)] - 2\text{cov}\left[\sum_{i=1}^{n} \lambda_i Z(\mathbf{s}_i), Z(\mathbf{s}_0)\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i\lambda_j C(\mathbf{s}_i - \mathbf{s}_j) + \sigma^2 - 2\sum_{i=1}^{n} \lambda_i C(\mathbf{s}_i - \mathbf{s}_0) \quad . \end{aligned}$$

With respect to condition (3.10), we can rewrite these equations as a function of the variogram:

$$\begin{aligned} \text{MSE}(\mathbf{s}_0) &= \sigma^2 + \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i\lambda_j[\sigma^2 - \gamma(\mathbf{s}_i - \mathbf{s}_j)] - 2\sum_{i=1}^{n} \lambda_i[\sigma^2 - \gamma(\mathbf{s}_i - \mathbf{s}_0)] \\ &= \sigma^2 + \sigma^2 - \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i\lambda_j\gamma(\mathbf{s}_i - \mathbf{s}_j) - 2\sigma^2 + 2\sum_{i=1}^{n} \lambda_i\gamma(\mathbf{s}_i - \mathbf{s}_0) \\ &= -\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i\lambda_j\gamma(\mathbf{s}_i - \mathbf{s}_j) + 2\sum_{i=1}^{n} \lambda_i\gamma(\mathbf{s}_i - \mathbf{s}_0) \quad . \end{aligned}$$

To obtain the vector of coefficients $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_n)'$ by minimizing the mean squared error subjected to the constraint (3.36) we have to apply the method of Lagrange multipliers:

$$Q(\boldsymbol{\lambda}, \varrho) = -\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i\lambda_j\gamma(\mathbf{s}_i - \mathbf{s}_j) + 2\sum_{i=1}^{n} \lambda_i\gamma(\mathbf{s}_i - \mathbf{s}_0) + 2\varrho\left(1 - \sum_{i=1}^{n} \lambda_i\right) \quad . \tag{3.38}$$

By taking derivatives of the function $Q(\boldsymbol{\lambda}, \varrho)$ with respect to the $\boldsymbol{\lambda}$ and ϱ one obtains:

$$\frac{\partial}{\partial\lambda_i}Q(\boldsymbol{\lambda}, \varrho) = -2\sum_{i=1}^{n} \lambda_i\gamma(\mathbf{s}_i - \mathbf{s}_j) + 2\gamma(\mathbf{s}_i - \mathbf{s}_0) - 2\varrho \tag{3.39}$$

$$\frac{\partial}{\partial\varrho}Q(\boldsymbol{\lambda}, \varrho) = 2\left(1 - \sum_{i=1}^{n} \lambda_i\right) \tag{3.40}$$

Kriging equations can be obtained by setting each of the expressions above equal to zero, and by re-arranging terms:

$$\sum_{i=1}^{n} \lambda_i\gamma(\mathbf{s}_i - \mathbf{s}_j) + \varrho = \gamma(\mathbf{s}_o - \mathbf{s}_i) \tag{3.41}$$

$$\sum_{i=1}^{n} \lambda_i = 1 \tag{3.42}$$

The both kriging equations can be written in the matrix notation which is more transparent [Cressie, 1993]:

$$\underbrace{\begin{pmatrix} 0 & \gamma(\mathbf{s}_1-\mathbf{s}_2) & \gamma(\mathbf{s}_1-\mathbf{s}_3) & \cdots & \gamma(\mathbf{s}_1-\mathbf{s}_n) & 1 \\ \gamma(\mathbf{s}_2-\mathbf{s}_1) & 0 & \gamma(\mathbf{s}_2-\mathbf{s}_3) & \cdots & \gamma(\mathbf{s}_2-\mathbf{s}_n) & 1 \\ \gamma(\mathbf{s}_3-\mathbf{s}_1) & \gamma(\mathbf{s}_3-\mathbf{s}_2) & 0 & \cdots & \gamma(\mathbf{s}_3-\mathbf{s}_n) & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma(\mathbf{s}_n-\mathbf{s}_1) & \gamma(\mathbf{s}_n-\mathbf{s}_2) & \gamma(\mathbf{s}_n-\mathbf{s}_3) & \cdots & 0 & 1 \\ 1 & 1 & 1 & \cdots & 1 & 0 \end{pmatrix}}_{\boldsymbol{\Gamma_O}} \underbrace{\begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \vdots \\ \lambda_n \\ \varrho \end{pmatrix}}_{\boldsymbol{\lambda_O}} = \underbrace{\begin{pmatrix} \gamma(\mathbf{s}_0-\mathbf{s}_1) \\ \gamma(\mathbf{s}_0-\mathbf{s}_2) \\ \gamma(\mathbf{s}_0-\mathbf{s}_3) \\ \vdots \\ \gamma(\mathbf{s}_0-\mathbf{s}_n) \\ 1 \end{pmatrix}}_{\boldsymbol{\gamma_O}}$$

Shortly:

$$\boldsymbol{\lambda_O} = \boldsymbol{\Gamma_O^{-1}}\boldsymbol{\gamma_O} \quad , \tag{3.43}$$

where $\boldsymbol{\Gamma_O}$ is a symmetric $(n+1)\times(n+1)$ matrix. From (3.43), the coefficients $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_n)'$ are given by [Cressie, 1993]:

$$\boldsymbol{\lambda}' = \left(\boldsymbol{\gamma} + \mathbf{1}\frac{(1 - \mathbf{1}'\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma})}{\mathbf{1}'\boldsymbol{\Gamma}^{-1}\mathbf{1}}\right)' \boldsymbol{\Gamma}^{-1} \quad , \tag{3.44}$$

and

$$\varrho = -\frac{(1 - \mathbf{1}'\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma})}{\mathbf{1}'\boldsymbol{\Gamma}^{-1}\mathbf{1}} \quad , \tag{3.45}$$

where $\boldsymbol{\gamma} = (\gamma(\mathbf{s}_0-\mathbf{s}_1), \ldots, \gamma(\mathbf{s}_0-\mathbf{s}_n))'$, $\boldsymbol{\Gamma}$ is the $n \times n$ matrix whose (i,j) th element is $\gamma(\mathbf{s}_i - \mathbf{s}_j)$ and $\mathbf{1}$ is $1 \times n$ vector of ones.

The *Kriging (prediction) variance* is the minimized mean-squared prediction error (3.37):

$$\begin{aligned} \sigma_k^2(\mathbf{s}_0) &= \boldsymbol{\lambda}_O'\boldsymbol{\gamma}_O = \textstyle\sum_{i=1}^n \lambda_i \gamma(\mathbf{s}_0 - \mathbf{s}_i) + \varrho \\ &= \boldsymbol{\gamma}'\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma} - \frac{\left(\mathbf{1}'\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma}_{-1}\right)^2}{\mathbf{1}'\boldsymbol{\Gamma}^{-1}\mathbf{1}} \quad . \end{aligned} \tag{3.46}$$

For the sake of completeness, we introduce also a matrix notation of Kriging in terms of the covariance function [Cressie, 1993]:

$$\underbrace{\begin{pmatrix} C(\mathbf{0}) & C(\mathbf{s}_1-\mathbf{s}_2) & C(\mathbf{s}_1-\mathbf{s}_3) & \cdots & C(\mathbf{s}_1-\mathbf{s}_n) & 1 \\ C(\mathbf{s}_2-\mathbf{s}_1) & C(\mathbf{0}) & C(\mathbf{s}_2-\mathbf{s}_3) & \cdots & C(\mathbf{s}_2-\mathbf{s}_n) & 1 \\ C(\mathbf{s}_3-\mathbf{s}_1) & C(\mathbf{s}_3-\mathbf{s}_2) & C(\mathbf{0}) & \cdots & C(\mathbf{s}_3-\mathbf{s}_n) & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ C(\mathbf{s}_n-\mathbf{s}_1) & C(\mathbf{s}_n-\mathbf{s}_2) & C(\mathbf{s}_n-\mathbf{s}_3) & \cdots & C(\mathbf{0}) & 1 \\ 1 & 1 & 1 & \cdots & 1 & 0 \end{pmatrix}}_{\boldsymbol{\Sigma_O}} \underbrace{\begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \vdots \\ \lambda_n \\ \varrho \end{pmatrix}}_{\boldsymbol{\lambda_O}} = \underbrace{\begin{pmatrix} C(\mathbf{s}_0-\mathbf{s}_1) \\ C(\mathbf{s}_0-\mathbf{s}_2) \\ C(\mathbf{s}_0-\mathbf{s}_3) \\ \vdots \\ C(\mathbf{s}_0-\mathbf{s}_n) \\ 1 \end{pmatrix}}_{\boldsymbol{c_O}}$$

That is:

$$\boldsymbol{\lambda_O} = \boldsymbol{\Sigma_O^{-1}} \boldsymbol{c_O} \quad , \tag{3.47}$$

where $\boldsymbol{\Sigma_O}$ is a symmetric $(n+1) \times (n+1)$ matrix. As well as in the case of variogram, the coefficients $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_n)'$ are given by:

$$\boldsymbol{\lambda}' = \left(\boldsymbol{c} + \mathbf{1} \frac{(1 - \mathbf{1}'\boldsymbol{\Sigma}^{-1}\boldsymbol{c})}{\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}} \right)' \boldsymbol{\Sigma}^{-1} \quad , \tag{3.48}$$

and

$$\varrho = -\frac{(1 - \mathbf{1}'\boldsymbol{\Gamma}^{-1}\boldsymbol{c})}{\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}} \quad , \tag{3.49}$$

where $\boldsymbol{c} = (C(\mathbf{s}_0 - \mathbf{s}_1), \ldots, C(\mathbf{s}_0 - \mathbf{s}_n))'$, $\boldsymbol{\Sigma}$ is the $n \times n$ matrix whose (i,j) th element is $C(\mathbf{s}_i - \mathbf{s}_j)$ and $\mathbf{1}$ is $1 \times n$ vector of ones. These equations require the assumption of the second order stationarity.

Finally, the Kriging variance is defined as:

$$\sigma_k^2(\mathbf{s}_0) = C(\mathbf{0}) - \boldsymbol{\lambda}'\boldsymbol{c} + \varrho \quad . \tag{3.50}$$

3.7.2 Simple Kriging

The simple Kriging is the simplest type. In the case of the simple Kriging it is assumed that $\mu(\mathbf{s})$ is known constant mean μ over the whole domain D and the coefficients do not add up to one. A distinction between simple and ordinary Kriging can be seen when all n observation have the same value z. With ordinary Kriging the estimate at the unobserved location is also z, whereas with simple Kriging the estimate at the unobserved location is a linear combination of z and the specified model mean μ.

3.7.3 Universal Kriging

Universal Kriging is similar to the ordinary Kriging. The differences are in the deterministic term $\mu(\mathbf{s})$. Let assume that the mean value $\mu(\mathbf{s})$ of a process (or random function) Z is the linear combination of known functions $x_1(\mathbf{s}), x_2(\mathbf{s}), \ldots, x_p(\mathbf{s})$, $s \in D$. The coefficients of this linear combination are unknown parameters $\beta_0, \beta_1, \ldots, \beta_n$. These parameters have to be estimated. Thus, it is supposed:

$$\mu(\mathbf{s}) = \sum_{j=0}^{p} \beta_j x_j(\mathbf{s}) \quad . \tag{3.51}$$

The best linear unbiased predictor in the point $\mathbf{s}_0$ is identical to ordinary Kriging:

$$\widehat{Z}(\mathbf{s}_0) = \sum_{i=1}^{n} \lambda_i Z(\mathbf{s}_i) \quad , \tag{3.52}$$

where terms λ_i's are chosen to minimize the mean squared error:

$$\text{MSE} = E[(\widehat{Z}(\mathbf{s}_0) - Z(\mathbf{s}_0))^2] = E\left[(Z(\mathbf{s}_0) - \sum_{i=1}^{n} \lambda_i Z(\mathbf{s}_i))^2\right] \quad . \tag{3.53}$$

If the prediction $\widehat{Z}$ in the unsampled point $\mathbf{s}_0$ should be unbiased then the following condition has to be fulfilled:

$$E[(\widehat{Z}(\mathbf{s}_0)] = E[(Z(\mathbf{s}_0)] \quad , \tag{3.54}$$

for all $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$. The expected value of predictor is:

$$E[(\widehat{Z}(\mathbf{s}_0)] = \beta_0 + \beta_1 x_1(\mathbf{s}_0) + \beta_2 x_2(\mathbf{s}_0) + \cdots + \beta_p x_p(\mathbf{s}_0) \quad , \tag{3.55}$$

then

$$\begin{aligned}
E[\widehat{Z}(\mathbf{s}_0)] &= E\left[\sum_{i=1}^{n} \lambda_i Z(\mathbf{s}_i)\right] \\
&= \sum_{i=1}^{n} \lambda_i E[Z(\mathbf{s}_i)] \\
&= \sum_{i=1}^{n} \lambda_i (\beta_0 + \beta_1 x_1(\mathbf{s}_i) + \beta_2 x_2(\mathbf{s}_i) + \cdots + \beta_p x_p(\mathbf{s}_i)) \\
&= \beta_0 \sum_{i=1}^{n} \lambda_i + \beta_1 \sum_{i=1}^{n} \lambda_i x_1(\mathbf{s}_i) + \beta_2 \sum_{i=1}^{n} \lambda_i x_2(\mathbf{s}_i) + \cdots + \beta_p \sum_{i=1}^{n} \lambda_i x_p(\mathbf{s}_i) \quad .
\end{aligned}$$

So the prediction $\widehat{Z}(\mathbf{s}_0)$ is unbiased for $Z(\mathbf{s}_0)$ if and only if

$$\sum_{i=1}^{n} \lambda_i = 1 \quad , \tag{3.56}$$

$$\sum_{i=1}^{n} \lambda_i x_j(\mathbf{s}_i) = x_j(\mathbf{s}_0) \quad \text{for} \quad j = 1, 2, \ldots, p \quad . \tag{3.57}$$

The condition (3.54) also can be expressed as [Michálek, 2001]:

$$\mu(\mathbf{s}_0) = \sum_{j=0}^{p} \beta_j x_j(\mathbf{s}_0) = \sum_{i=1}^{n} \lambda_i \sum_{j=0}^{p} \beta_j x_j(\mathbf{s}_i) \quad . \tag{3.58}$$

In matrix notation it is:

$$\mathbf{x}'\boldsymbol{\beta} = \boldsymbol{\lambda}'\mathbf{X}\boldsymbol{\beta} \quad , \tag{3.59}$$

$$\underbrace{\begin{pmatrix} (x_1(\mathbf{s}_0) \\ x_2(\mathbf{s}_0) \\ \vdots \\ x_p(\mathbf{s}_0) \end{pmatrix}'}_{\mathbf{x}} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}}_{\boldsymbol{\beta}} = \underbrace{\begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{pmatrix}'}_{\boldsymbol{\lambda}} \underbrace{\begin{pmatrix} x_1(\mathbf{s}_1) & \cdots & x_p(\mathbf{s}_1) \\ x_1(\mathbf{s}_2) & \cdots & x_p(\mathbf{s}_2) \\ \vdots & & \vdots \\ x_1(\mathbf{s}_n) & \cdots & x_p(\mathbf{s}_n) \end{pmatrix}}_{\mathbf{X}} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}}_{\boldsymbol{\beta}}$$

Thus, we can rewrite the model (3.26) as:

$$Z(\mathbf{s}_0) = \mu(\mathbf{s}_0) + \epsilon(\mathbf{s}_0) = \mathbf{x}'\boldsymbol{\beta} + \epsilon(\mathbf{s}_0) \quad , \tag{3.60}$$

and the predictor is:

$$\widehat{Z}(\mathbf{s}_0) = \sum_{i=1}^{n} \lambda_i Z(\mathbf{s}_i) = \boldsymbol{\lambda}'\mathbf{Z} = \boldsymbol{\lambda}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\lambda}'\boldsymbol{\epsilon} \quad . \tag{3.61}$$

where $\mathbf{Z} = (Z(\mathbf{s}_1), Z(\mathbf{s}_2), \ldots, Z(\mathbf{s}_n))'$ is a vector of observed values in points $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n$. And $\boldsymbol{\epsilon} = (\epsilon(\mathbf{s}_1), \epsilon(\mathbf{s}_2), \ldots, \epsilon(\mathbf{s}_n))'$ is vector of corresponding random errors.

The mean squared error of a predictor $\widehat{Z}(\mathbf{s}_0)$ is similar to the MSE in the case of ordinary Kriging:

$$\mathrm{MSE}(\mathbf{s}_0) = E\left[(\widehat{Z}(\mathbf{s}_0) - Z(\mathbf{s}_0))^2\right] = \sigma_e^2 \quad , \tag{3.62}$$

or equivalently:

$$\begin{aligned}
\mathrm{MSE}(\mathbf{s}_0) &= \sigma^2 + \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j[\sigma^2 - \gamma(\mathbf{s}_i - \mathbf{s}_j)] - 2\sum_{i=1}^{n}\lambda_i[\sigma^2 - \gamma(\mathbf{s}_i - \mathbf{s}_0)] \\
&= \sigma^2 + \sigma^2 - \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j\gamma(\mathbf{s}_i - \mathbf{s}_j) - 2\sigma^2 + 2\sum_{i=1}^{n}\lambda_i\gamma(\mathbf{s}_i - \mathbf{s}_0) \\
&= -\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j\gamma(\mathbf{s}_i - \mathbf{s}_j) + 2\sum_{i=1}^{n}\lambda_i\gamma(\mathbf{s}_i - \mathbf{s}_0) \quad .
\end{aligned}$$

To obtain the vector of coefficients $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_n)'$ by minimizing the mean squared error subjected to constraints (3.56, 3.57) we have to apply the method of Lagrange multipliers:

$$\begin{aligned}
Q(\boldsymbol{\lambda}, \boldsymbol{\varrho}) = &- \textstyle\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j\gamma(\mathbf{s}_i - \mathbf{s}_j) + 2\sum_{i=1}^{n}\lambda_i\gamma(\mathbf{s}_i - \mathbf{s}_0) \\
&+ 2\varrho_0\left(1 - \textstyle\sum_{i=1}^{n}\lambda_i\right) + 2\sum_{j=1}^{p}\varrho_j\left(x_j(\mathbf{s}_0) - \sum_{i=1}^{n}\lambda_i x_j(\mathbf{s}_i)\right) \quad .
\end{aligned} \tag{3.63}$$

By taking derivatives of the function $Q(\boldsymbol{\lambda}, \boldsymbol{\varrho})$ with respect to the $\boldsymbol{\lambda}$ and $\boldsymbol{\varrho}$ we get:

$$\frac{\partial}{\partial\lambda_k}Q(\boldsymbol{\lambda}, \boldsymbol{\varrho}) = -2\sum_{i=1}^{n}\lambda_i\gamma(\mathbf{s}_i - \mathbf{s}_k) + 2\gamma(\mathbf{s}_k - \mathbf{s}_0) - 2\varrho_0 - 2\sum_{j=1}^{p}\varrho_j x_j(\mathbf{s}_k) \quad , \tag{3.64}$$

for $k = 1, 2, \ldots, n$,

$$\frac{\partial}{\partial\varrho_0}Q(\boldsymbol{\lambda}, \boldsymbol{\varrho}) = 2\left(1 - \sum_{i=1}^{n}\lambda_i\right) \tag{3.65}$$

$$\frac{\partial}{\partial\varrho_k}Q(\boldsymbol{\lambda}, \boldsymbol{\varrho}) = 2\left(x_k(\mathbf{s}_0) - \sum_{i=1}^{n}\lambda_i x_k(\mathbf{s}_i)\right) \quad \text{for} \quad k = 1, \ldots, p \quad . \tag{3.66}$$

Kriging equations can be obtained by setting each of the expressions above equal to zero, and by re-arranging terms:

$$\sum_{i=1}^{n}\lambda_i\gamma(\mathbf{s}_i - \mathbf{s}_j) + \varrho_0 + \sum_{j=1}^{p}\varrho_j x_j(\mathbf{s}_k) = \gamma(\mathbf{s}_o - \mathbf{s}_k) \quad \text{for} \quad k = 1, \ldots, n \quad , \tag{3.67}$$

$$\sum_{i=1}^{n}\lambda_i = 1 \quad , \tag{3.68}$$

$$\sum_{i=1}^{n}\lambda_i x_k(\mathbf{s}_i) = x_k(\mathbf{s}_o) \quad \text{for} \quad k = 1, \ldots, p \quad . \tag{3.69}$$

We can rewrite equations in matrix notation:

$$\underbrace{\begin{pmatrix} 0 & \gamma(\mathbf{s}_1-\mathbf{s}_2) & \cdots & \gamma(\mathbf{s}_1-\mathbf{s}_n) & 1 & x_1(\mathbf{s}_1) & \cdots & x_p(\mathbf{s}_1) \\ \gamma(\mathbf{s}_2-\mathbf{s}_1) & 0 & \cdots & \gamma(\mathbf{s}_2-\mathbf{s}_n) & 1 & x_1(\mathbf{s}_2) & \cdots & x_p(\mathbf{s}_2) \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & & \vdots \\ \gamma(\mathbf{s}_n-\mathbf{s}_1) & \gamma(\mathbf{s}_n-\mathbf{s}_2) & \cdots & 0 & 1 & x_1(\mathbf{s}_n) & \cdots & x_p(\mathbf{s}_n) \\ 1 & 1 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ x_1(\mathbf{s}_1) & x_2(\mathbf{s}_2) & \cdots & x_n(\mathbf{s}_n) & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ x_p(\mathbf{s}_1) & x_p(\mathbf{s}_2) & \cdots & x_p(\mathbf{s}_n) & 0 & 0 & \cdots & 0 \end{pmatrix}}_{\boldsymbol{\Gamma_U}} \underbrace{\begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ \varrho_0 \\ \varrho_1 \\ \vdots \\ \varrho_p \end{pmatrix}}_{\boldsymbol{\lambda_U}} =$$

$$= \underbrace{\begin{pmatrix} \gamma(\mathbf{s}_0-\mathbf{s}_1) \\ \gamma(\mathbf{s}_0-\mathbf{s}_2) \\ \vdots \\ \gamma(\mathbf{s}_0-\mathbf{s}_n) \\ 1 \\ x_1(\mathbf{s}_0) \\ \vdots \\ x_p(\mathbf{s}_0) \end{pmatrix}}_{\boldsymbol{\gamma_U}}$$

or easily:

$$\boldsymbol{\lambda_U} = \boldsymbol{\Gamma_U^{-1}\gamma_U} \quad , \tag{3.70}$$

where $\boldsymbol{\Gamma_U}$ is a symmetric $(n+p+1)\times(n+p+1)$ matrix. Thus, the coefficients $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_n)'$ are given by [Cressie, 1993]:

$$\boldsymbol{\lambda}' = \left(\boldsymbol{\gamma} + \mathbf{X}(\mathbf{X}'\boldsymbol{\Gamma}^{-1}\mathbf{X})^{-1}(\mathbf{x} - \mathbf{X}'\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma})'\boldsymbol{\Gamma}^{-1}\right) \quad , \tag{3.71}$$

and

$$\varrho = -(\mathbf{x} - \mathbf{X}'\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma})'(\mathbf{X}'\boldsymbol{\Gamma}^{-1}\mathbf{X})^{-1} \quad , \tag{3.72}$$

where $\boldsymbol{\gamma} = (\gamma(\mathbf{s}_0-\mathbf{s}_1), \ldots, \gamma(\mathbf{s}_0-\mathbf{s}_n))'$, $\boldsymbol{\Gamma}$ is the $n\times n$ matrix whose (i,j) th element is $\gamma(\mathbf{s}_i-\mathbf{s}_j)$ and $\mathbf{X}$ is an $n\times(p+1)$ matrix whose (i,j) th element is $x_{j-1}(\mathbf{s}_i)$.

The prediction variance is:

$$\begin{aligned}\sigma_k^2(\mathbf{s}_0) &= \boldsymbol{\lambda}_U'\boldsymbol{\gamma}_U = \sum_{i=1}^{n}\lambda_i\gamma(\mathbf{s}_0-\mathbf{s}_i)+\varrho_0+\sum_{j=1}^{p}\varrho_j x_j(\mathbf{s}_0)\\ &= \boldsymbol{\gamma}'\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma}-\left(\mathbf{x}-\mathbf{X}'\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma}\right)^{-1}\left(\mathbf{x}-\mathbf{X}'\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma}\right) \quad .\end{aligned} \tag{3.73}$$

We also release a matrix notation of universal Kriging in terms of the covariance function [Cressie, 1993]:

$$\underbrace{\begin{pmatrix} C(\mathbf{0}) & C(\mathbf{s}_1-\mathbf{s}_2) & \cdots & C(\mathbf{s}_1-\mathbf{s}_n) & 1 & x_1(\mathbf{s}_1) & \cdots & x_p(\mathbf{s}_1)\\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & & \vdots\\ C(\mathbf{s}_n-\mathbf{s}_1) & C(\mathbf{s}_n-\mathbf{s}_2) & \cdots & C(\mathbf{0}) & 1 & x_1(\mathbf{s}_n) & \cdots & x_p(\mathbf{s}_n)\\ 1 & 1 & \cdots & 1 & 0 & 0 & \cdots & 0\\ x_1(\mathbf{s}_1) & x_2(\mathbf{s}_2) & \cdots & x_n(\mathbf{s}_n) & 0 & 0 & \cdots & 0\\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots\\ x_p(\mathbf{s}_1) & x_p(\mathbf{s}_2) & \cdots & x_p(\mathbf{s}_n) & 0 & 0 & \cdots & 0 \end{pmatrix}}_{\boldsymbol{\Sigma_U}} \underbrace{\begin{pmatrix}\lambda_1\\ \vdots\\ \lambda_n\\ \varrho_0\\ \varrho_1\\ \vdots\\ \varrho_p\end{pmatrix}}_{\boldsymbol{\lambda_U}} =$$

$$= \underbrace{\begin{pmatrix} C(\mathbf{s}_0-\mathbf{s}_1)\\ \vdots\\ C(\mathbf{s}_0-\mathbf{s}_n)\\ 1\\ x_1(\mathbf{s}_0)\\ \vdots\\ x_p(\mathbf{s}_0)\end{pmatrix}}_{\boldsymbol{c_U}}$$

That is:

$$\boldsymbol{\lambda_U} = \boldsymbol{\Sigma_U^{-1}}\boldsymbol{c_U} \quad , \tag{3.74}$$

where $\boldsymbol{\Sigma_U}$ is a symmetric $(n+p+1)\times(n+p+1)$ matrix. Thus, the coefficients $\boldsymbol{\lambda} = (\lambda_1,\lambda_2,\ldots,\lambda_n)'$ are given by [Cressie, 1993]:

$$\boldsymbol{\lambda}' = \left(\mathbf{c}+\mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}(\mathbf{x}-\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{c}\right)'\boldsymbol{\Sigma}^{-1} \quad , \tag{3.75}$$

and

$$\varrho = (\mathbf{x}-\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{c})'(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1} \quad , \tag{3.76}$$

where $\mathbf{c} = (C(\mathbf{s}_0 - \mathbf{s}_1), \ldots, C(\mathbf{s}_0 - \mathbf{s}_n))'$, $\mathbf{\Sigma}$ is the $n \times n$ matrix whose (i,j) th element is $C(\mathbf{s}_i - \mathbf{s}_j)$ and $\mathbf{X}$ is an $n \times (p+1)$ matrix whose (i,j) th element is $x_{j-1}(\mathbf{s}_i)$.

The prediction variance is:

$$\begin{aligned} \sigma_k^2(\mathbf{s}_0) &= C(\mathbf{0}) + \textstyle\sum_{i=1}^{n} \lambda_i C(\mathbf{s}_0 - \mathbf{s}_i) + \varrho_0 + \sum_{j=1}^{p} \varrho_j x_j(\mathbf{s}_0) \\ &= C(\mathbf{0}) - \mathbf{c}'\mathbf{\Sigma}^{-1}\mathbf{c} + \left(\mathbf{x} - \mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{c}\right)' \left(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X}\right)^{-1} \left(\mathbf{x} - \mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{c}\right) \quad . \end{aligned} \tag{3.77}$$

3.7.4 Co-Kriging

Before we describe the co-Kriging we have to introduce two new terms. The optimization of the linear co-Kriging estimate requires modeling of the *cross-covariance* or *cross-variogram* between pairs of variable at location pairs. Thus, the cross-covariance is formulated as:

$$C_{YZ}(\mathbf{h}) = \frac{1}{N_{\mathbf{h}}} \sum_{h_{ij}=h} (Y(\mathbf{s}_i) - \widehat{\mu}_Y)(Y(\mathbf{s}_j) - \widehat{\mu}_Y)(Z(\mathbf{s}_i) - \widehat{\mu}_Z)(Z(\mathbf{s}_j) - \widehat{\mu}_Z) \quad , \tag{3.78}$$

and the cross-variogram si defined as:

$$2\gamma_{YZ}(\mathbf{h}) = \frac{1}{N_{\mathbf{h}}} \sum_{h_{ij}=h} (Y(\mathbf{s}_i) - Y(\mathbf{s}_j))(Z(\mathbf{s}_i) - Z(\mathbf{s}_j)) \quad , \tag{3.79}$$

where $N_{\mathbf{h}}$ is number of pairs whose locations are separated by vector $\mathbf{h}$. Index $h_{ij} = \text{location}[Y(\mathbf{s}_i)] - \text{location}[Y(\mathbf{s}_j)] = \text{location}[Z(\mathbf{s}_i)] - \text{location}[Z(\mathbf{s}_j)]$, where $\text{location}[Y(\mathbf{s}_i)]$ is location where $Y(\mathbf{s}_i)$ is calculated, and similarly for other values [Memarsadeghi, 2004].

It is clear from equations (3.78, 3.79) that we are dealing with values of two random variables from different distributions. The first variable $Z(\mathbf{s})$ is called primary variable (or variable of interest) and $Y(\mathbf{s})$ is secondary variable.

Co-Kriging is defined as a multivariate version of Kriging as well as the universal Kriging. The best linear unbiased predictor of $Z(\mathbf{s}_0)$ is estimated as a linear combination of both the variable of interest $Z(\mathbf{s})$ and the secondary variable $Y(\mathbf{s})$:

$$\widehat{Z}(\mathbf{s}_0) = \sum_{i=1}^{n} \lambda_i Z(\mathbf{s}_i) + \sum_{j=1}^{m} \omega_j Y(\mathbf{s}_j) \quad , \tag{3.80}$$

where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_n)'$ and $\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots, \omega_m)'$ are vectors of unknown coefficients that have to be estimated. Note that auxiliary information does not need to be collected at

the same data points as the variable of interest [Memarsadeghi, 2004]. These coefficients are chosen to minimize the mean squared error:

$$\text{MSE} = E[(\widehat{Z}(\mathbf{s}_0) - Z(\mathbf{s}_0))^2] = E\left[(Z(\mathbf{s}_0) - \sum_{i=1}^{n} \lambda_i Z(\mathbf{s}_i) - \sum_{j=1}^{m} \omega_j Y(\mathbf{s}_j))^2\right] \quad . \tag{3.81}$$

As well as previous types of Kriging, the predictor $\hat{Z}$ at unsampled point $\mathbf{s}_0$ is unbiased for $Z(\mathbf{s}_0)$ if:

$$E[(\widehat{Z}(\mathbf{s}_0)] = E[(Z(\mathbf{s}_0)] \quad , \tag{3.82}$$

and using property of the expected value (2.10):

$$\begin{aligned} E[\widehat{Z}(\mathbf{s}_0)] &= E\left[\sum_{i=1}^{n} \lambda_i Z(\mathbf{s}_i) + \sum_{j=1}^{m} \omega_j Y(\mathbf{s}_j)\right] \\ &= \sum_{i=1}^{n} \lambda_i E[Z(\mathbf{s}_i)] + \sum_{j=1}^{m} \omega_j E[Y(\mathbf{s}_j)] \\ &= \mu_Z \sum_{i=1}^{n} \lambda_i + \mu_Y \sum_{j=1}^{m} \omega_j \end{aligned}$$

So constraints on coefficients are:

$$\sum_{i=1}^{n} \lambda_i = 1 \quad , \tag{3.83}$$

$$\sum_{j=1}^{m} \omega_j = 0 \quad . \tag{3.84}$$

The mean squared error of predictor $\widehat{Z}(\mathbf{s}_0)$ is:

$$\text{MSE}(\mathbf{s}_0) = E\left[(\widehat{Z}(\mathbf{s}_0) - Z(\mathbf{s}_0))^2\right] = \sigma_e^2 \quad , \tag{3.85}$$

and by [Memarsadeghi, 2004] follows:

$$
\begin{aligned}
E\left[(\widehat{Z}(\mathbf{s}_0) - Z(\mathbf{s}_0))^2\right] &= Var[\widehat{Z}(\mathbf{s}_0) - Z(\mathbf{s}_0)] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j \mathrm{cov}(Z(\mathbf{s}_i), Z(\mathbf{s}_j)) + \sum_{i=1}^{m}\sum_{j=1}^{m}\omega_i\omega_j \mathrm{cov}(Y(\mathbf{s}_i), Y(\mathbf{s}_j)) \\
&+ 2\sum_{i=1}^{n}\sum_{j=1}^{m}\lambda_i\omega_j \mathrm{cov}(Z(\mathbf{s}_i), Y(\mathbf{s}_j)) - 2\sum_{i=1}^{n}\lambda_i \mathrm{cov}(Z(\mathbf{s}_0), Z(\mathbf{s}_i)) \\
&- 2\sum_{j=1}^{m}\omega_j \mathrm{cov}(Z(\mathbf{s}_0), Y(\mathbf{s}_j)) + \mathrm{cov}(Z(\mathbf{s}_0), Z(\mathbf{s}_0)) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j C_Z(\mathbf{s}_i - \mathbf{s}_j) + \sum_{i=1}^{m}\sum_{j=1}^{m}\omega_i\omega_j C_Y(\mathbf{s}_i - \mathbf{s}_j) \\
&+ 2\sum_{i=1}^{n}\sum_{j=1}^{m}\lambda_i\omega_j C_{ZY}(\mathbf{s}_i - \mathbf{s}_j) - 2\sum_{i=1}^{n}\lambda_i C_Z(\mathbf{s}_0 - \mathbf{s}_i) \\
&- 2\sum_{j=1}^{m}\omega_j C_{ZY}(\mathbf{s}_0 - \mathbf{s}_j) + \sigma^2 \quad .
\end{aligned}
$$

We can expressed previous equations as a function of the variogram:

$$
\begin{aligned}
E\left[(\widehat{Z}(\mathbf{s}_0) - Z(\mathbf{s}_0))^2\right] &= Var[\widehat{Z}(\mathbf{s}_0) - Z(\mathbf{s}_0)] \\
&= -\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j \gamma_Z(\mathbf{s}_i - \mathbf{s}_j) - \sum_{i=1}^{m}\sum_{j=1}^{m}\omega_i\omega_j \gamma_Y(\mathbf{s}_i - \mathbf{s}_j) \\
&- 2\sum_{i=1}^{n}\sum_{j=1}^{m}\lambda_i\omega_j \gamma_{ZY}(\mathbf{s}_i - \mathbf{s}_j) + 2\sum_{i=1}^{n}\lambda_i \gamma_Z(\mathbf{s}_0 - \mathbf{s}_i) \\
&+ 2\sum_{j=1}^{m}\omega_j \gamma_{ZY}(\mathbf{s}_0 - \mathbf{s}_j) \quad .
\end{aligned}
$$

To obtain vectors of coefficients $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_n)$ and $\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots, \omega_m)$ to minimize the mean squared error subjected to the constraint of unbiasedness we have to again apply the method of Lagrange multipliers:

$$
\begin{aligned}
Q(\boldsymbol{\lambda}, \boldsymbol{\omega}, \boldsymbol{\varrho}) &= -\textstyle\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j \gamma_Z(\mathbf{s}_i - \mathbf{s}_j) - \sum_{i=1}^{m}\sum_{j=1}^{m}\omega_i\omega_j \gamma_Y(\mathbf{s}_i - \mathbf{s}_j) \\
&- 2\textstyle\sum_{i=1}^{n}\sum_{j=1}^{m}\lambda_i\omega_j \gamma_{ZY}(\mathbf{s}_i - \mathbf{s}_j) + 2\sum_{i=1}^{n}\lambda_i \gamma_Z(\mathbf{s}_0 - \mathbf{s}_i) \qquad (3.86) \\
&+ 2\textstyle\sum_{j=1}^{m}\omega_j \gamma_{ZY}(\mathbf{s}_0 - \mathbf{s}_j) + 2\varrho_Z\left(\sum_{i=1}^{n}\lambda_i - 1\right) + 2\varrho_Y\left(\sum_{j=1}^{m}\omega_j\right) \quad .
\end{aligned}
$$

By taking derivatives of the function Q($\boldsymbol{\lambda}$,$\boldsymbol{\omega}$,$\boldsymbol{\varrho}$) with respect to the $\boldsymbol{\lambda}$, $\boldsymbol{\omega}$ and $\boldsymbol{\varrho}$ one obtains:

$$\begin{aligned}\frac{\partial}{\partial\lambda_j}Q(\boldsymbol{\lambda},\boldsymbol{\omega},\boldsymbol{\varrho}) = & - 2\sum_{i=1}^{n}\lambda_i\gamma_Z(\mathbf{s}_i-\mathbf{s}_j)+2\sum_{i=1}^{m}\omega_i\gamma_{ZY}(\mathbf{s}_i-\mathbf{s}_j) \\ & - 2\gamma_Z(\mathbf{s}_j-\mathbf{s}_0)+2\varrho_Z \quad ,\end{aligned} \tag{3.87}$$

for $j = 1, 2, \ldots, n$,

$$\begin{aligned}\frac{\partial}{\partial\omega_j}Q(\boldsymbol{\lambda},\boldsymbol{\omega},\boldsymbol{\varrho}) = & - 2\sum_{i=1}^{n}\lambda_i\gamma_{ZY}(\mathbf{s}_i-\mathbf{s}_j)+2\sum_{i=1}^{m}\omega_i\gamma_{Y}(\mathbf{s}_i-\mathbf{s}_j) \\ & - 2\gamma_{ZY}(\mathbf{s}_j-\mathbf{s}_0)+2\varrho_Y \quad ,\end{aligned} \tag{3.88}$$

for $j = 1, 2, \ldots, m$,

$$\frac{\partial}{\partial\varrho_Z}Q(\boldsymbol{\lambda},\boldsymbol{\omega},\boldsymbol{\varrho}) = 2\varrho_Z\left(\sum_{i=1}^{n}\lambda_i-1\right) \tag{3.89}$$

$$\frac{\partial}{\partial\varrho_Y}Q(\boldsymbol{\lambda},\boldsymbol{\omega},\boldsymbol{\varrho}) = 2\varrho_Y\left(\sum_{j=1}^{m}\omega_j\right) \quad . \tag{3.90}$$

Kriging equations can be obtained by setting each of the expressions above equal to zero, and by re-arranging terms:

$$\sum_{i=1}^{n}\lambda_i\gamma_Z(\mathbf{s}_i-\mathbf{s}_j)+\sum_{i=1}^{m}\omega_i\gamma_{ZY}(\mathbf{s}_i-\mathbf{s}_j)+\varrho_Z=\gamma_Z(\mathbf{s}_j-\mathbf{s}_0) \quad , \tag{3.91}$$

for $j = 1, 2, \ldots, n$,

$$\sum_{i=1}^{n}\lambda_i\gamma_{ZY}(\mathbf{s}_i-\mathbf{s}_j)+\sum_{i=1}^{m}\omega_i\gamma_{Y}(\mathbf{s}_i-\mathbf{s}_j)+\varrho_Y=\gamma_{ZY}(\mathbf{s}_j-\mathbf{s}_0) \quad , \tag{3.92}$$

for $j = 1, 2, \ldots, m$,

$$\sum_{i=1}^{n}\lambda_i = 1 \tag{3.93}$$

$$\sum_{i=1}^{m}\omega_i = 0 \quad . \tag{3.94}$$

Next, we can express previous equations in matrix notation:

$$\underbrace{\begin{pmatrix} \gamma_Z(\mathbf{s}_1-\mathbf{s}_1) & \cdots & \gamma_Z(\mathbf{s}_1-\mathbf{s}_n) & \gamma_{ZY}(\mathbf{s}_1-\mathbf{s}_1) & \cdots & \gamma_{ZY}(\mathbf{s}_1-\mathbf{s}_m) & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \gamma_Z(\mathbf{s}_n-\mathbf{s}_1) & \cdots & \gamma_Z(\mathbf{s}_n-\mathbf{s}_n) & \gamma_{ZY}(\mathbf{s}_n-\mathbf{s}_1) & \cdots & \gamma_{ZY}(\mathbf{s}_n-\mathbf{s}_m) & 1 & 0 \\ \gamma_{YZ}(\mathbf{s}_1-\mathbf{s}_1) & \cdots & \gamma_{YZ}(\mathbf{s}_1-\mathbf{s}_n) & \gamma_Y(\mathbf{s}_1-\mathbf{s}_1) & \cdots & \gamma_Y(\mathbf{s}_1-\mathbf{s}_m) & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \gamma_{YZ}(\mathbf{s}_m-\mathbf{s}_1) & \cdots & \gamma_{YZ}(\mathbf{s}_m-\mathbf{s}_n) & \gamma_Y(\mathbf{s}_1-\mathbf{s}_1) & \cdots & \gamma_Y(\mathbf{s}_m-\mathbf{s}_m) & 1 & 0 \\ 1 & \cdots & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 & 0 & 0 \end{pmatrix}}_{\boldsymbol{\Gamma_C}} \underbrace{\begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ \omega_1 \\ \vdots \\ \omega_n \\ \varrho_Z \\ \varrho_Y \end{pmatrix}}_{\boldsymbol{\lambda_C}} =$$

$$= \underbrace{\begin{pmatrix} \gamma_Z(\mathbf{s}_0-\mathbf{s}_1) \\ \vdots \\ \gamma_Z(\mathbf{s}_0-\mathbf{s}_n) \\ \gamma_{ZY}(\mathbf{s}_0-\mathbf{s}_1) \\ \gamma_{ZY}(\mathbf{s}_0-\mathbf{s}_m) \\ 1 \\ 0 \end{pmatrix}}_{\boldsymbol{\gamma_C}}$$

Or shortly:

$$\boldsymbol{\lambda_C} = \boldsymbol{\Gamma_C^{-1}}\boldsymbol{\gamma_C} \quad . \tag{3.95}$$

The matrix of covariance functions $\boldsymbol{\Sigma_C}$ is provided by replacing the variogram $\gamma(\cdot)$ by the covariance function $C(\cdot)$.

Finally, the prediction variance can be computed from:

$$\sigma_k^2(\mathbf{s}_0) = -\sum_{i=1}^{n} \lambda_i \gamma_{ZY}(\mathbf{s}_0-\mathbf{s}_i) - \sum_{i=1}^{m} \omega_i \gamma_{ZY}(\mathbf{s}_0-\mathbf{s}_i) + \varrho_Z + \varrho_Y \quad . \tag{3.96}$$

3.7.5 Properties of Kriging

At the end of this chapter we summarize some properties of Kriging which can be seen from above-mentioned equations.

1. By construction, Kriging is unbiased:

$$E[\widehat{Z}(\mathbf{s}_0)] = Z(\mathbf{s}_0) \quad .$$

2. By construction, the Kriging predictor is the best linear unbiased predictor. It means that it has the smallest mean squared error among all linear predictors.

3. Kriging is an exact interpolator, see Fig. 3.9. For a sampled site $\mathbf{s}_i$, the Kriging predictor is:

$$\widehat{Z}(\mathbf{s}_i) = Z(\mathbf{s}_i) \quad ,$$

and the corresponding estimation variance is:

$$\sigma_k^2(\mathbf{s}_i) = 0 \quad .$$

Nevertheless, Kriging can be an approximation, too. The distinction between the method presented here and an approximation is whether the fitted/interpolated function goes exactly through all the input data points (interpolation), or whether it allows measurement errors to be specified and then "smooths" to get a statistically better predictor that does not generally go through the data points (does not "honor the data"). Such access can be found, e.g. in [Krivoruchko, 2001] or in [Kleijnen, 2007].

4. The Kriging predictor is continuous if there is no nugget effect:

$$\lim_{h \to 0} \widehat{Z}(\mathbf{s}_0 + \mathbf{h}) = \widehat{Z}(\mathbf{s}_0) \quad .$$

If there is a nugget effect, then the Kriging predictor is continuous everywhere except at the data points where there are discontinuities.

5. Kriging coefficients $\boldsymbol{\lambda}$ sum up to 1, but they can also be negative.

6. Furthermore, Kriging coefficients $\boldsymbol{\lambda}$ are not influenced by the measurement values. If the same configuration appears at two different locations the Kriging weights $\boldsymbol{\lambda}$ will be the same, abstractedly from the measured values [Bárdossy, 1997].

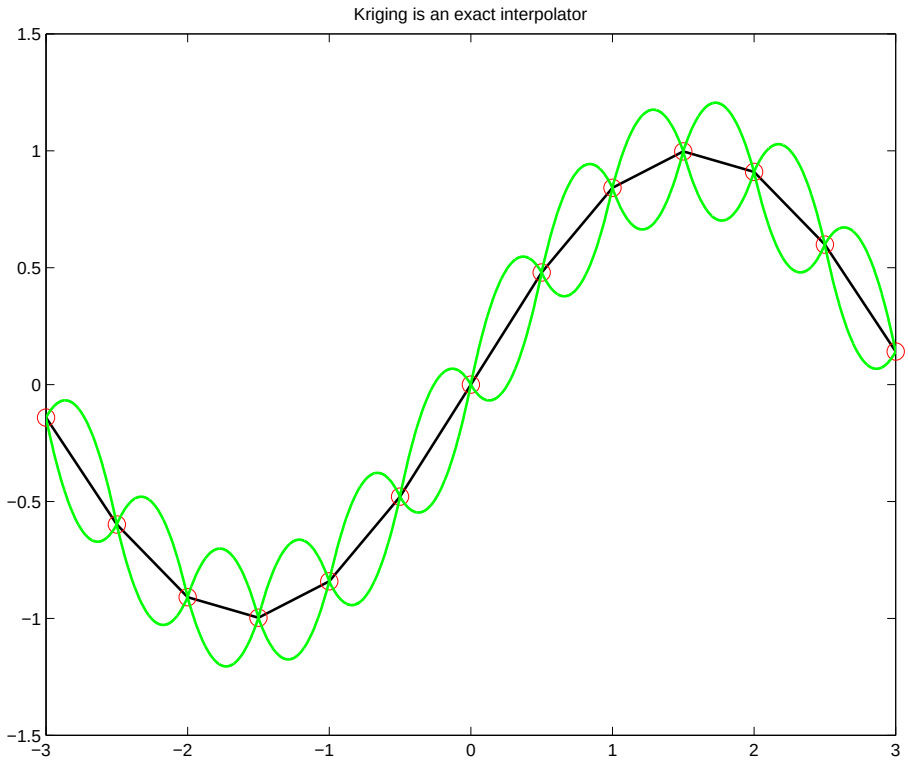

Figure 3.9: Kriging is an exact interpolator. Red circles are sampled points, black line is predicted value at unsampled points and green lines bound the confidence range.

7. It is typical for Kriging weights that distant points get lower weights if the further measurements are available. This property is named a screening effect.

8. It is obvious from equations presented above that the computation of Kriging weights $\boldsymbol{\lambda}$ requires inverting of the variogram matrix (or the matrix of covariance function) which is time consuming if a number of observations is very large. The time required to solve a dense $n \times n$ matrix grows at order of n^3. The problem of a large number of observations can be solved by application of sparse matrix techniques, see [Barry and Pace, 1997]. Alternatively, the special type of Kriging can be used: for instance fixed rank Kriging, that is adapted to large data set, see e.g. [Cressie and Johannesson, 2008].

Chapter 4

GLOBAL OPTIMIZATION APPROACH TO KRIGING

In this chapter the global optimization approach to Kriging is presented. We are inspired by the work of [Jones et al., 1998] and [Jones, 2001]. They took the stochastic process model (Kriging) used in statistics and applied it to global optimization. We start with showing how the Kriging can be derived as a modification to linear regression. The notation used in the previous chapter is preserved with less but necessary differences.

Let assume a evaluated function of k variables at n points. The sampled point i is denoted by $\mathbf{s}_i = (\mathbf{s}_{i\,1}, \mathbf{s}_{i\,2}, \ldots, \mathbf{s}_{i\,n})'$ and a function value at the same point by $Z_i = Z(\mathbf{s}_i)$, for $i = 1, 2, \ldots, n$. The model of linear regression fitting a response of given data is provided by:

$$Z(\mathbf{s}_i) = \sum_h \beta_h f_h(\mathbf{s}_i) + \epsilon_i \qquad \text{for } i = 1, 2, \ldots, n \quad , \tag{4.1}$$

where each $f_h(\mathbf{s})$ is a linear or nonlinear function of $\mathbf{s}$. The term $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_h)'$ is a vector of unknown coefficients which have to be estimated, and the $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)'$ are independent error terms with normal distribution. Among others, it means that these terms have mean equals to zero and variance equals to σ^2:

$$E[\boldsymbol{\epsilon}] = 0 \quad , \tag{4.2}$$

$$Var[\boldsymbol{\epsilon}] = \sigma^2 \quad . \tag{4.3}$$

When we are applying linear regression to a computer code, there is a problem that the functional form of regression terms is often unknown. Another problem is that the assumption of an independent error is wrong. Because the code is deterministic, any lack of fit will be entirely modeling error, not measurement error or noise [Jones et al., 1998]. From this cognisance ensues that we may write ϵ_i as $\epsilon(\mathbf{s}_i)$, and that if the points $\mathbf{s}_i$ and $\mathbf{s}_j$ are close together, then the errors $\epsilon(\mathbf{s}_i)$ and $\epsilon(\mathbf{s}_j)$ should be close, too. It all means that the assumption of correlated errors is more reasonable. Similar to geostatistics the

correlation between errors is related to the distance between the corresponding points. We compute the distance by the help of formula [Jones et al., 1998]:

$$d(\mathbf{s}_i, \mathbf{s}_j) = \sum_{h=1}^{k} \theta_h \, |\mathbf{s}_{i\,h} - \mathbf{s}_{j\,h}|^{p_h} \quad , \tag{4.4}$$

where the parameter $\theta_h \geq 0$ is a measure of the importance of the variable $\mathbf{s}_h$. A larger θ is related to a more active variable, and the correlation descends more rapidly with the change in s. The parametr $p_h \in \{1, 2\}$ is related to the smoothness of the function in the direction h. A higher value of p_h makes a function more smoother.

Originally [Jones et al., 1998], the distance formula presented above leads to the definition of the correlation between errors:

$$\mathrm{Corr}[\epsilon(\mathbf{s}_i), \epsilon(\mathbf{s}_j)] = \exp[-d(\mathbf{s}_i - \mathbf{s}_j)] \quad . \tag{4.5}$$

Then, the correlation can be denoted for any function ρ as:

$$\mathrm{Corr}[\epsilon(\mathbf{s}_i), \epsilon(\mathbf{s}_j)] = \rho[\epsilon(\mathbf{s}_i), \epsilon(\mathbf{s}_j)] \quad , \tag{4.6}$$

where ρ is a correlation function selected by user, see next page.

The Kriging model can be seen as a combination of a global model plus a localized 'deviation' [Jin, 2005]:

$$Z(\mathbf{s}) = f(\mathbf{s}) + \epsilon(\mathbf{s}) \quad , \tag{4.7}$$

where $f(\mathbf{s})$ is a known function of $\mathbf{s}$ as a global model of the original function, and $\epsilon(\mathbf{s})$ is a Gaussian random function with zero mean and non-zero covariance that represents a localized deviation from the global model. Usually the regression term $f(\mathbf{s})$ is a polynomial function but in [Jones et al., 1998] is stated that under circumstances mentioned above, we can replace the regression term by the constant term. Then, we get the ordinary Kriging model:

$$Z(\mathbf{s}_i) = \mu + \epsilon(\mathbf{s}_i) \qquad \text{for} \quad i = 1, 2, \ldots, n \quad . \tag{4.8}$$

The parameters μ and variance of error terms σ^2 is unknown and have to be estimated. The estimation of these parameters is not direct and should be connected with estimation of the correlation parameters θ_h and p_h. As is mentioned in [Jones et al., 1998] and

[Giunta and Watson, 1998], it is common to call this model the DACE stochastic process model [Sacks et al., 1989], where DACE is abbreviated from Design and Analysis of Computer Experiments.

The model (4.6) has $2k+2$ parameters, namely $\mu, \sigma^2, \theta_1, \theta_2, \ldots, \theta_k$ and $p_1, p_2, \ldots, p_k$. These parameters are estimated by the method of maximum likelihood. Here, the likelihood function is:

$$\frac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}|\mathbf{R}|^{1/2}} \exp\left[-\frac{(\mathbf{Z}-\mathbf{1}\mu)'\mathbf{R}^{-1}(\mathbf{Z}-\mathbf{1}\mu)}{2\sigma^2}\right] \quad , \tag{4.9}$$

where $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_n)'$ denotes the vector of observed values, $\mathbf{R}$ is the correlation matrix $n \times n$ whose $(i,j)\,th$ element is $\mathrm{Corr}[\epsilon(\mathbf{s}_i), \epsilon(\mathbf{s}_j)]$. And finally, $\mathbf{1}$ denotes $1 \times n$ vector of ones.

By maximization of this function we get the parameters $\theta_1, \theta_2, \ldots, \theta_k$ and $p_1, p_2, \ldots, p_k$ which are hidden in correlation matrix $\mathbf{R}$. Next, the two remaining parameters can be estimated:

$$\widehat{\mu} = \frac{\mathbf{1}'\mathbf{R}^{-1}\mathbf{Z}}{\mathbf{1}'\mathbf{R}^{-1}\mathbf{1}} \quad , \tag{4.10}$$

$$\widehat{\sigma}^2 = \frac{(\mathbf{Z}-\mathbf{1}\widehat{\mu})'\mathbf{R}^{-1}(\mathbf{Z}-\mathbf{1}\widehat{\mu})}{n} \quad . \tag{4.11}$$

The concentrated likelihood function can be obtained by substituting the equations (4.10) and (4.11) to the likelihood function (4.9). Then the concentrated likelihood function will depend on parameters $\theta_1, \theta_2, \ldots, \theta_k$ and $p_1, p_2, \ldots, p_k$ only. Now the Kriging is essentially a generalized least squares (GLS) [1] model with a constant term as a set of regressors and a special correlation matrix that depends upon distances between sampled points [Jones et al., 1998]. When the errors are correlated then the GLS estimates of μ and σ^2 are more efficient then ordinary least squares (OLS) estimates.

If we get all needed parameters we can predict the value $Z(\mathbf{s}_0)$. The best linear unbiased predictor of $\widehat{Z}(\mathbf{s}_0)$ is:

$$\widehat{Z}(\mathbf{s}_0) = \widehat{\mu} + \mathbf{r}'\mathbf{R}^{-1}(\mathbf{Z}-\mathbf{1}\widehat{\mu}) \quad , \tag{4.12}$$

where $\mathbf{r}$ is $1 \times n$ vector of correlations between the error terms at unsampled point $\mathbf{s}_0$ and the error terms at sampled points. The i th element $r_i = \mathrm{Corr}[\epsilon(\mathbf{s}_0), \epsilon(\mathbf{s}_i)]$. If there is no correlation, i.e. $\mathbf{r} = 0$, then $\widehat{Z}(\mathbf{s}_0) = \widehat{\mu}$ and we again obtain ordinary Kriging.

[1] For details how the predictor of GLS looks like, see Appendix A

As well as in the case of Kriging used in geostatistics, the predictor variance is defined by:

$$\sigma_k^2(\mathbf{s}_0) = \sigma^2 \left[1 - \mathbf{r}'\mathbf{R}^{-1}\mathbf{r} + \frac{(1 - \mathbf{1}'\mathbf{R}^{-1}\mathbf{r})^2}{\mathbf{1}'\mathbf{R}^{-1}\mathbf{1}} \right] \quad . \tag{4.13}$$

The term $-\mathbf{r}'\mathbf{R}^{-1}\mathbf{r}$ represents the reduction in a prediction error with respect to the fact that $\mathbf{s}_0$ is correlated with the sampled points [Jones et al., 1998]. And the term $\frac{(1-\mathbf{1}'\mathbf{R}^{-1}\mathbf{r})^2}{\mathbf{1}'\mathbf{R}^{-1}\mathbf{1}}$ reflects the uncertainty that follows from unknowing the $\widehat{\mu}$ exactly.

4.1 Types of Correlation Function

The correlation function $\rho[\epsilon(\mathbf{s}_i), \epsilon(\mathbf{s}_j)]$ is specified by the user. Depending on the choice of a correlation function, Kriging can either 'honor the data', providing an exact interpolation of the data, or 'smooth the data', providing an inexact interpolation [Simpson et al., 2001]. We assume that the correlation function depends only on the distance between points. Thus, we can write:

$$h_h(i,j) = |\mathbf{s}_{i\,h} - \mathbf{s}_{j\,h}| \qquad \text{for} \quad h = 1, \ldots, k;\ i, j = 1, 2, \ldots, n \quad , \tag{4.14}$$

so the $\rho[\epsilon(\mathbf{s}_i), \epsilon(\mathbf{s}_j)]$ reduces to $\rho[h_h(i,j)]$. Then matrix $\mathbf{R}$ reads:

$$\mathbf{R} = \begin{pmatrix} 1 & \rho(\mathbf{s}_1 - \mathbf{s}_2) & \rho(\mathbf{s}_1 - \mathbf{s}_3) & \cdots & \rho(\mathbf{s}_1 - \mathbf{s}_n) \\ \rho(\mathbf{s}_2 - \mathbf{s}_1) & 1 & \rho(\mathbf{s}_2 - \mathbf{s}_3) & \cdots & \rho(\mathbf{s}_2 - \mathbf{s}_n) \\ \rho(\mathbf{s}_3 - \mathbf{s}_1) & \rho(\mathbf{s}_3 - \mathbf{s}_2) & 1 & \cdots & \rho(\mathbf{s}_3 - \mathbf{s}_n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho(\mathbf{s}_n - \mathbf{s}_1) & \rho(\mathbf{s}_n - \mathbf{s}_2) & \rho(\mathbf{s}_n - \mathbf{s}_3) & \cdots & 1 \end{pmatrix} \quad ,$$

and the vector $\mathbf{r}$ can be rewritten as:

$$\mathbf{r} = \begin{pmatrix} \rho(\mathbf{s}_0 - \mathbf{s}_1) \\ \rho(\mathbf{s}_0 - \mathbf{s}_2) \\ \rho(\mathbf{s}_0 - \mathbf{s}_3) \\ \vdots \\ \rho(\mathbf{s}_0 - \mathbf{s}_n) \end{pmatrix} \quad .$$

The most usual form put on the correlation function is

$$\rho(\mathbf{h}) = \prod_{h=1}^{k} \rho(h_h) \quad , \tag{4.15}$$

where $\rho[h]$ is a user selected correlation function. The most frequently used types of functions according to [Kleijnen, 2008] are:

1.

$$\rho[h] = \exp\left[-\theta_h h_h^{p_h}\right] \quad , \tag{4.16}$$

2. which for $p_h = 1$ simplifies into the exponential correlation function:

$$\rho[h] = \exp(-\theta h) \quad , \tag{4.17}$$

3. or in case of $p_h = 2$ the traditional Gaussian correlation function is obtained:

$$\rho[h] = \exp\left(-\theta h^2\right) \quad . \tag{4.18}$$

4. Other possibility is a linear correlation function:

$$\rho[h] = \max(1 - \theta h, 0) \quad , \tag{4.19}$$

5. or inspect Tab. 5.1 for other popular functions.

Chapter 5

APPLICATION ON EXPERIMENTAL DATA

Our research in this work is focused on multidimensional data which has more then a three dimensions. Nevertheless, we also show application of Kriging to a spatial data for the sake of completeness. Another application of Kriging to the spatial data is introduced in the graduation thesis by Maxim Bernstein [Bernstein, 2006]. His work was aimed at an interpolation of rainfall on the area of Prague, see again [Bernstein, 2006] for more details.

5.1 Two Dimensional Data

The spatial data are frequently used in geostatistics. The data can be multivariate but the domain D is usually limited to two or three dimensions. The estimation of the variogram is pivotal for prediction values at unsampled points. There the free available data which can be downloaded from http://people.ku.edu/~gbohling/BoiseGeostat are used. The data consist of 85 entries of vertically averaged porosity values in percents. The eighty five porosity values are measured at points distributed throughout the domain which is approximately 20 km in east-west extent and 16 km north-south. The porosity range from 12 % to 17 %, see Fig. 5.1. In this figure we can see that the values of porosity are irregularly spaced. Therefore, we used the equation (3.17) for including allowed difference ε in length:

$$|\mathbf{s}_i - \mathbf{s}_j| - |\mathbf{h}| \leq \varepsilon \tag{5.1}$$

The experimental variogram was computed from (3.13):

$$2\widehat{\gamma}(\mathbf{h}) = \frac{1}{N_{\mathbf{h}}} \sum_{N_{\mathbf{h}}} (Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2 \quad . \tag{5.2}$$

Different lags $h = \{500, 1000, 2000, 3000\}$ has been tested with two lag tolerances $\varepsilon = \{0, 500\}$:

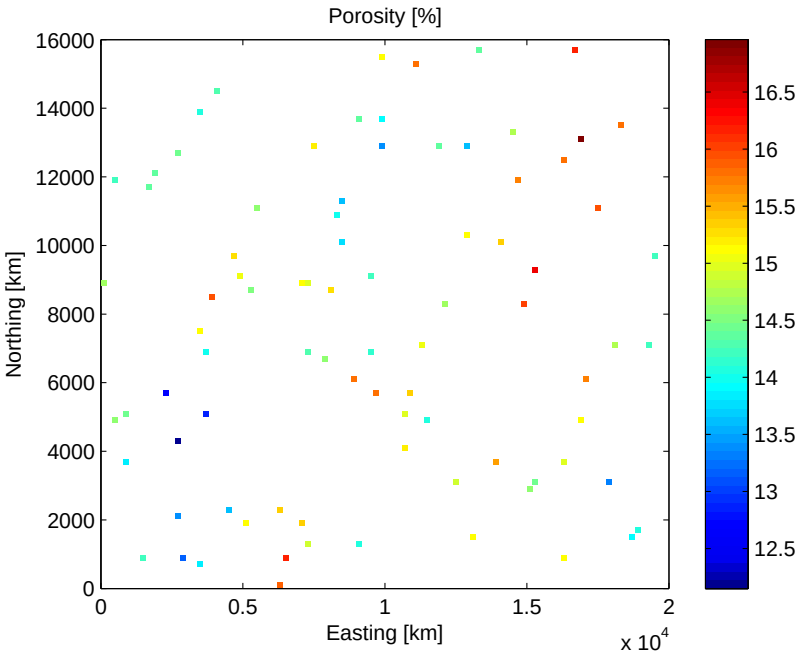

Figure 5.1: Scatter plot of locations where the porosity was measured.

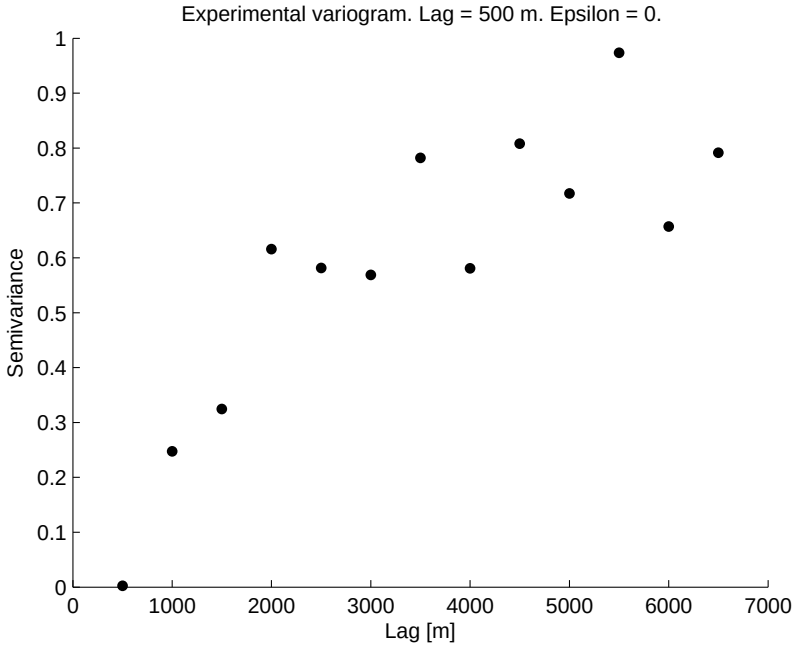

Figure 5.2: Experimental variogram with $h = 500$ and lag tolerance $\varepsilon = 0$.

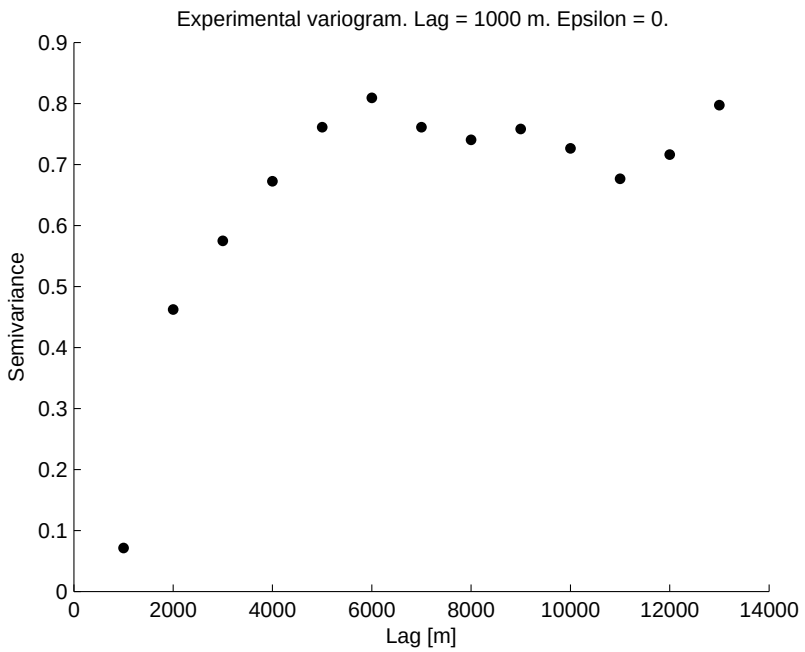

Figure 5.3: Experimental variogram with $h = 1000$ and lag tolerance $\varepsilon = 0$.

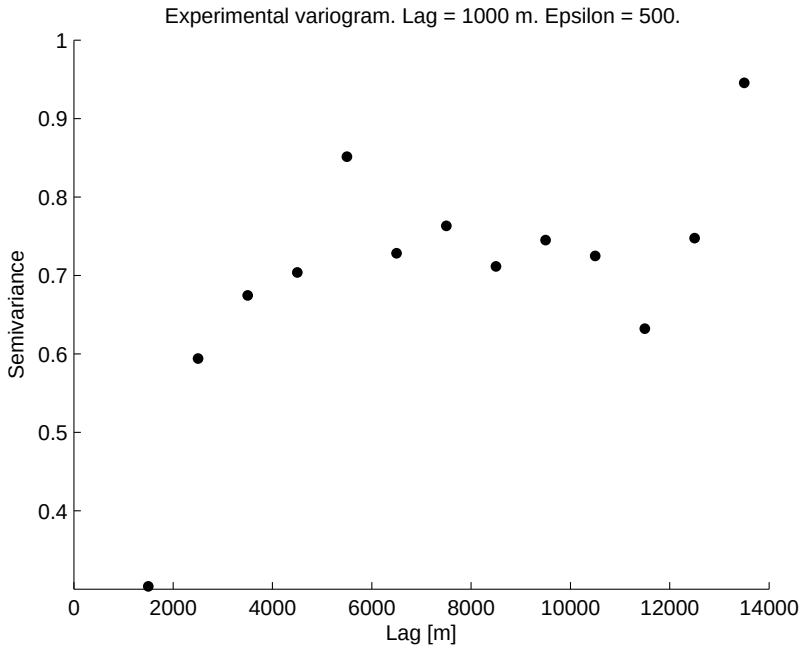

Figure 5.4: Experimental variogram with $h = 1000$ and lag tolerance $\varepsilon = 500$.

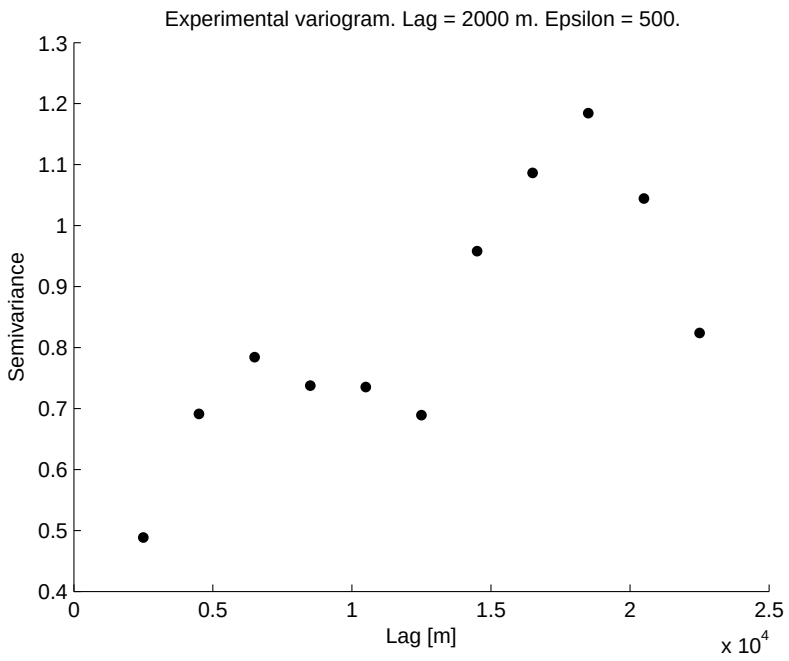

Figure 5.5: Experimental variogram with $h = 2000$ and lag tolerance $\varepsilon = 500$.

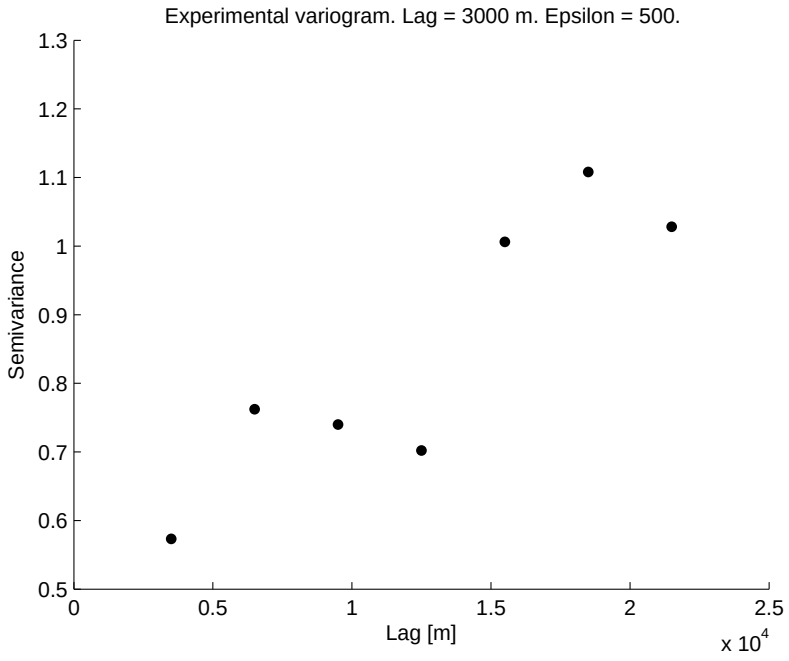

Figure 5.6: Experimental variogram with $h = 3000$ and lag tolerance $\varepsilon = 500$.

We decided o use the experimental variogram with lag $h = 1000$ and lag tolerance $\varepsilon = 500$, because the corresponding variogram seems to be the most appropriate for tuning. Then, the curve of theoretical variogram was fitted to the experimental one. The method 'by eye' was applied. The sill was fixed at 0.78 and the range was changing in an aim to get the best fit. Following figures show chosen examples of several possible fits:

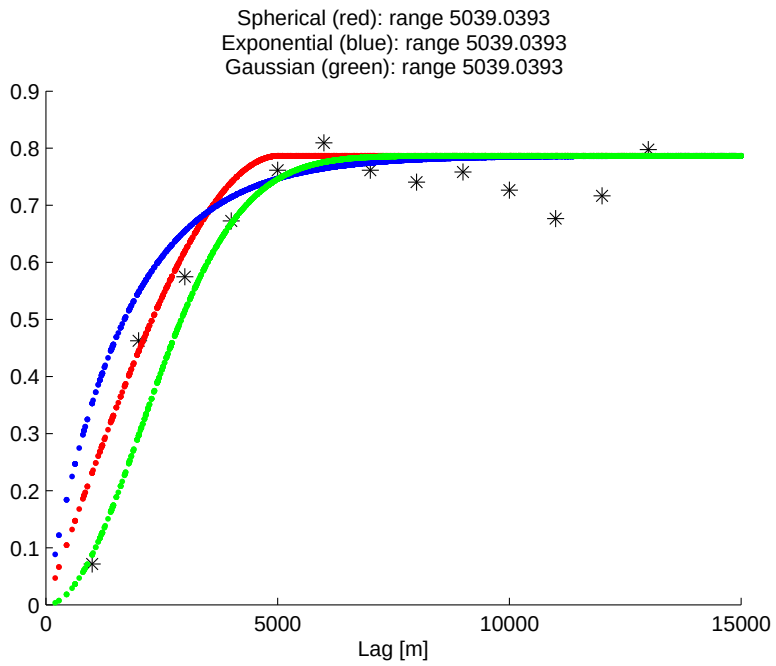

Figure 5.7: Three theoretical variograms with a range $a = 5039$.

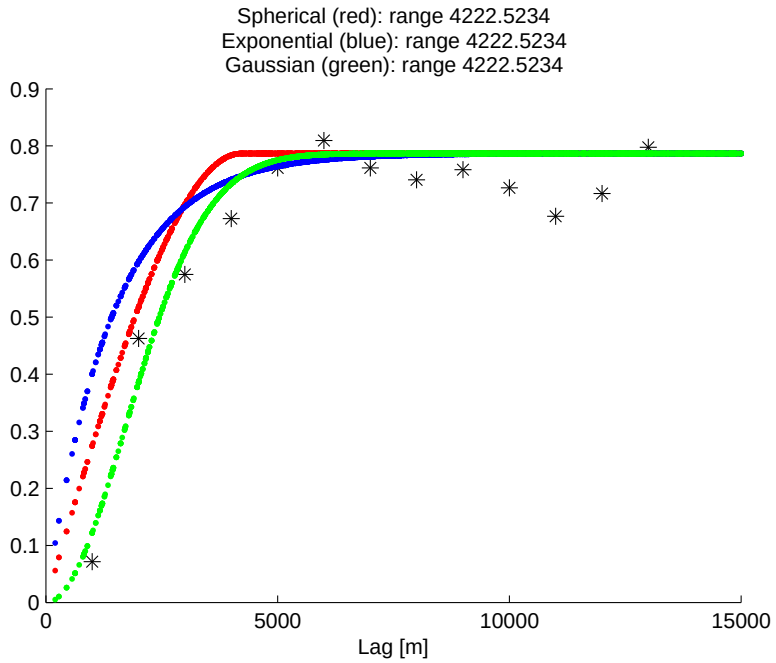

Figure 5.8: Three theoretical variograms with a range $a = 4223$.

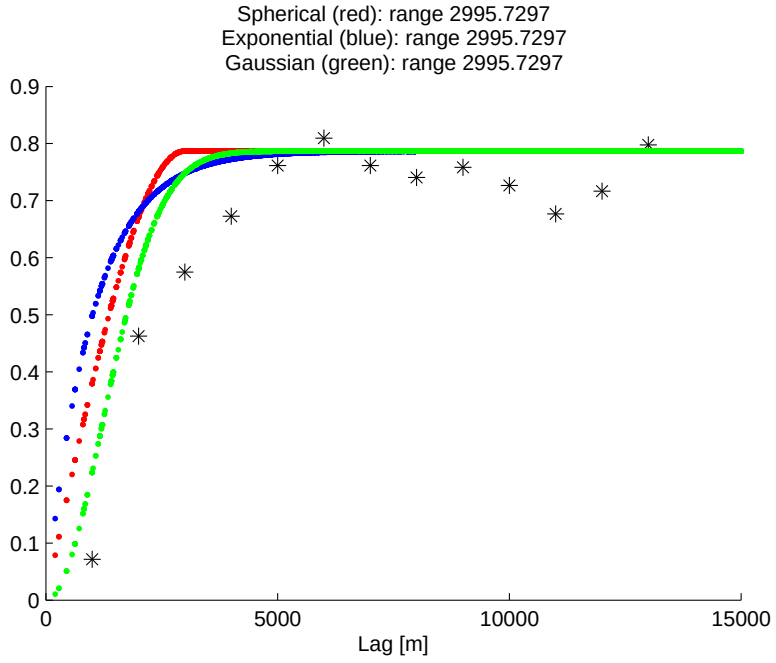

Figure 5.9: Three theoretical variograms with a range $a = 2996$.

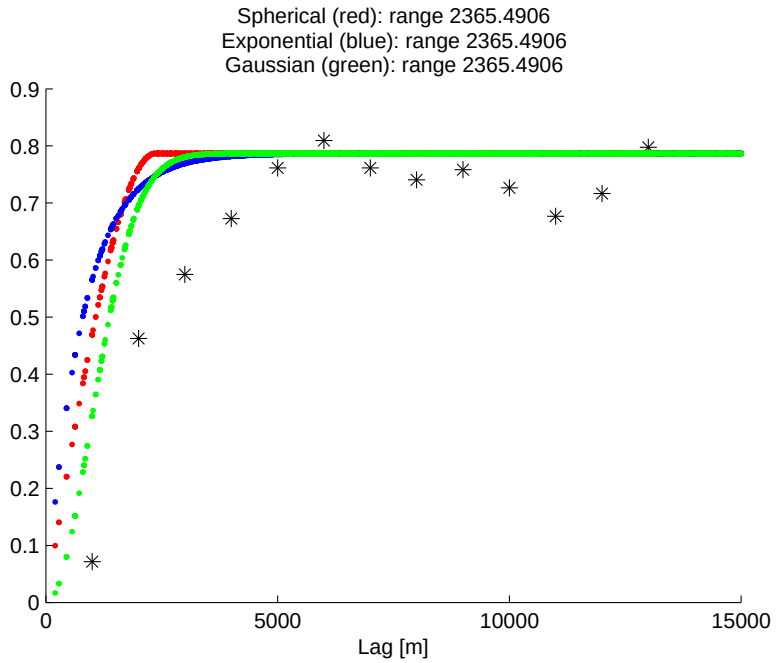

Figure 5.10: Three theoretical variograms with a range $a = 2365$.

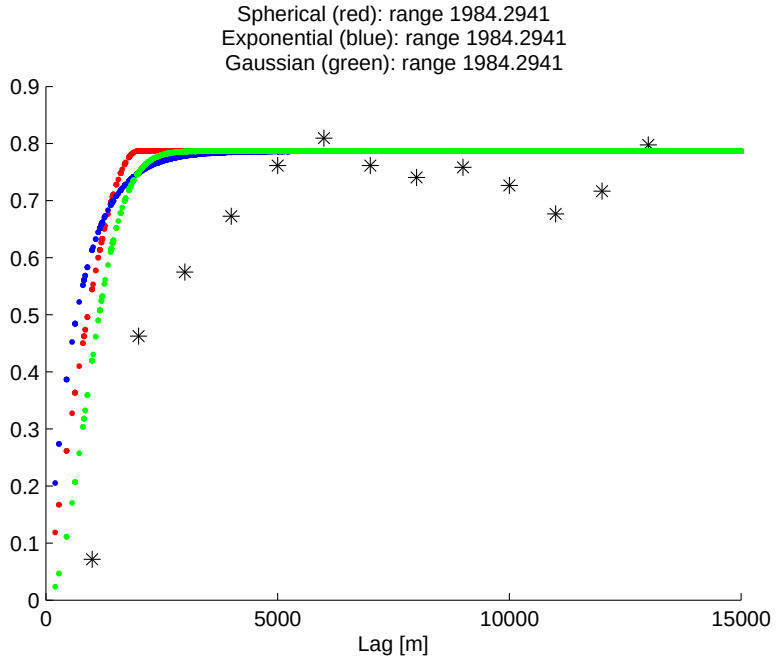

Figure 5.11: Three theoretical variograms with a range $a = 1984$.

After carefully study of all pictures, we have decided to use a model with a spherical semivariogram, no nugget and a range $a = 4223$. Then a regular grid of unsampled points from $19\,500 \times 19\,500$ area with spacing of 100 m has been generated. Thus, we have to predict 38 416 values of porosity. The simple Kriging was used for prediction. The map of porosity is a result, see Fig. 5.12, 5.13 and 5.14:

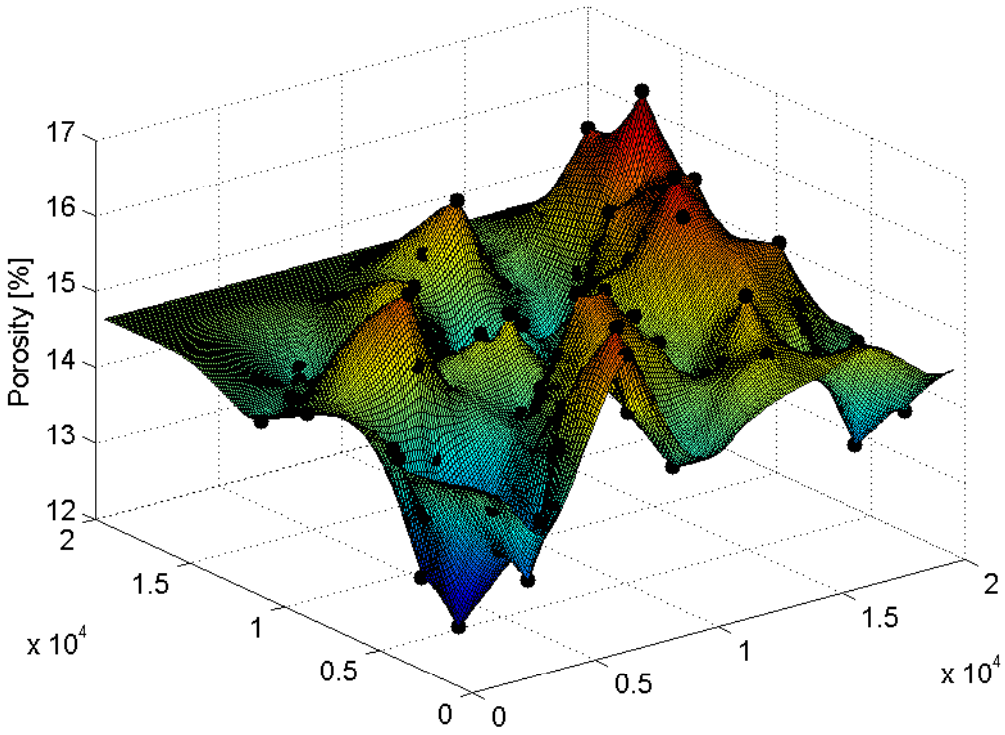

Figure 5.12: Surface plot of predicted values of porosity. Black points represent values at sampled points.

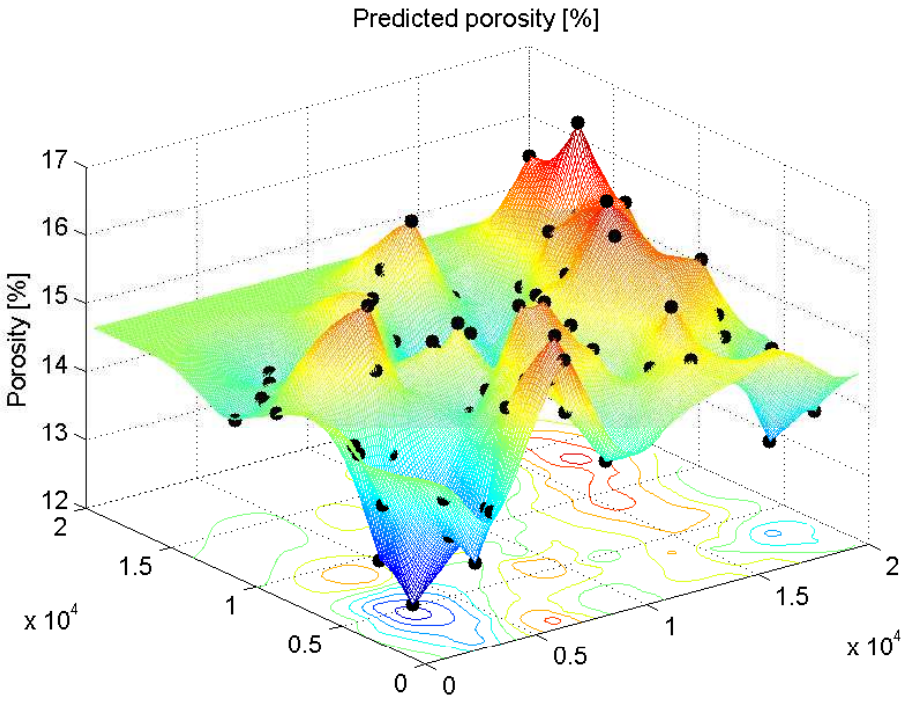

Figure 5.13: Surface plot with the contour map of predicted values of porosity. Black points represent values at sampled points.

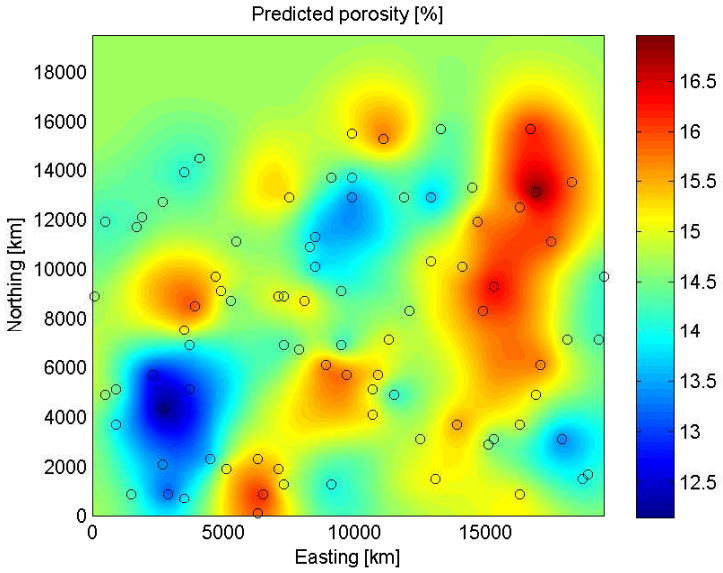

Figure 5.14: Predicted values of porosity. Circles denote sampled points.

5.2 Kriging Approximation in Cement Paste Experimental Performance

This section shows an engineering application utilizing approximation as well as interpolation properties of Kriging. This research is a part of a project aimed at a development of new concrete compositions. This section was also published in [Valtrová* and Lepš, 2009] as a conference paper.

5.2.1 Introduction

Cement paste is a fundamental scale from which concrete inherits majority of its properties. Experimental results show considerable scatter in the elastic response of cement paste samples, however, virtual testing in a computer allows testing the influence of input parameters on resulting macroscopic response [Šmilauer, 2006]. Last year, a combination of CEMHYD3D model with homogenization processes was employed as a basis for an optimization [Šmilauer et al., 2008], where the Young modulus and heat of hydration appear as objective functions. Question arises, whether results from the optimization of the virtual model can be trusted. Our proposed solution is based on a so-called robust optimization [H.-G. Beyer, 2007] where some selected distance to existing experimental results is employed as the robustness measure. Hence, our goal is to create the closest approximation to available experimental data and to provide estimation of the quality of that approximation.

In this contribution, we demonstrate that Kriging/DACE [Sacks et al., 1989]), approximation is far away from the surface that is expected to describe physical process underneath. Therefore, nonlinear optimization of the maximum monotonicity is presented. Overall, dozen of combinations of regression and correlation parts have been tested. Unfortunately, the importance of a proper regression part is more crucial than presented in the optimization literature [Jones, 2001]. We have found a regression part that almost ideally describes the physical problem, however, the strict monotonicity has not been preserved.

5.2.2 Methods

Since the response in terms of mixture parameters is non-linear, the Kriging [Jones, 2001] approximation in the space of hydration heat of available real measurements seems as natural choice. Generally, Kriging predictor is composed of a regression and interpolation part that constitutes the nonlinear surface among available data [Lophaven et al., 2002]:

$$\widehat{\boldsymbol{Z}}(\mathbf{s}_0) = f(\mathbf{s}_0)'\boldsymbol{\beta} + \mathbf{r}(\mathbf{s}_0)'\mathbf{R}^{-1}\widehat{\phi} \quad , \tag{5.3}$$

where $vekf(\mathbf{s})$ is an a-priori selected set of basis functions creating the response surface and $\mathbf{r}(\mathbf{s}_0)$ is the correlation term between an unsampled point $\mathbf{s}_0$ and known points $\mathbf{s}_i$, $i = 1, \ldots, n : \mathbf{r}(\mathbf{s}_0) = [R(\boldsymbol{\theta}; \mathbf{s}_1; \mathbf{s}_0), \ldots, R(\boldsymbol{\theta}; \mathbf{s}_n; \mathbf{s}_0)]'$, where R is a-priori selected correlation function with unknown coefficients $\boldsymbol{\theta}$, see later. The regression part is solved by a generalized least squares solution

$$\boldsymbol{\beta} = (\mathbf{X}'\mathbf{R}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z} \quad , \tag{5.4}$$

where $\mathbf{X}$ is a matrix containing $f(\mathbf{s})$ evaluated at known sites $\mathbf{s}_i$, $\mathbf{R}$ stems for correlation among $\mathbf{s}_i$ using again the correlation function R and $\mathbf{Z}$ are known values of Z_i at $\mathbf{s}_i$. The Kriging part then interpolates the residual leading to the system of linear equations

$$\widehat{\phi} = \mathbf{R}^{-1}(\mathbf{Z} - \mathbf{X}\widehat{\boldsymbol{\beta}}) \quad . \tag{5.5}$$

The use of such metamodel for optimization purposes is less demanding on the regression part since an interpolation is dominant and hence, the constant regression part usually suffices. Then, the correlation function is traditionally selected to obtain a positive-definite system of equations, mainly restricted to the form

$$R(\boldsymbol{\theta}, \mathbf{s}, \mathbf{s}_0) = \prod_{h=1}^{n} R_h(\boldsymbol{\theta}, \mathbf{s}_{i\,h} - \mathbf{s}_{j\,h}) \quad . \tag{5.6}$$

In our case, a free Matlab toolbox DACE [Lophaven et al., 2002] is utilized providing seven correlation functions, where five of them are presented in this work, see Tab. 5.1.

Note that at this point we still do not know the tuning/shape parameters $\boldsymbol{\theta}$. Their functionality is twofold: they express the anisotropy among dimensions and also determine the shape of the metamodel in the vicinity of given samples. Traditionally, these

Name	$R(\boldsymbol{\theta}; d_h)$,	$d_h = \mathbf{s}_{i\,h} - \mathbf{s}_{j\,h}$
EXP	$exp(-\boldsymbol{\theta}_h \lvert d_h \rvert)$	
GAUSS	$exp(-\boldsymbol{\theta}_h d_h^2)$	
LIN	$max\{0, 1 - \boldsymbol{\theta}_h \lvert d_h \rvert\}$	
SPHERICAL	$1 - 1.5\xi_h + 0.5\xi_h^3$,	$\xi_h = min\{1, \boldsymbol{\theta}_h \lvert d_h \rvert\}$
SPLINE	$1 - 15\xi_h + 30\xi_h^3$,	$for \quad 0 \leq \xi_h \leq 0.2$
	$1.25(1 - \xi_h)^3$,	$for \quad 0.2 \leq \xi_h \leq 1$
	0,	$for \quad \xi_h > 1, \quad \xi_h = \boldsymbol{\theta}_h \lvert d_h \rvert$

Table 5.1: Correlation functions

parameters are found a-posteriori by minimizing an expected mean squared error (MSE), which leads to the constrained nonlinear optimization problem. See e.g. [Jones, 2001] for discussion how to efficiently solve this problem without re-calculation of $\boldsymbol{\beta}$ and $\widehat{\phi}$ for this new $\boldsymbol{\theta}$.

Designation	**C3S**	**C2S**	**C3A**	**C4AF**	**Gyp.**	**w/c**	**Fin.**
	Mass %	Mass %	Mass %	Mass %	Vol %	-	m^2/kg
Aalborg white	0.666	0.238	0.034	0.004	0.036	0.400	390
Principallo	0.554	0.184	0.082	0.091	0.051	0.375	530
BAM Fontana	0.492	0.243	0.090	0.076	0.0652	0.300	380
Hua	0.688	0.075	0.081	0.092	0.04	0.420	400
Robeyst	0.634	0.084	0.074	0.100	0.05	0.500	390
Smolik_Litos	0.612	0.126	0.070	0.100	0.05	0.500	306
Tamtsia early	0.465	0.246	0.104	0.083	0.05	0.500	340

Table 5.2: Composition of experimental measurements

5.2.3 Fitting of experimental data

Particular application is shown on experimental data see Tab. 5.2 and Tab. 5.3, obtained from seven sources, consecutively from top: Data measured at CTU by TAMAir

Designation	**Hydration heat (time in h, heat in J/g of cement)**			
Aalborg white	24	48	168	[h]
	170.3	234	327	[J/g]
Principallo	9.42	80.24	400.00	[h]
	63.388	323.247	377.466	[J/g]
BAM Fontana	10.01	144.03	310.69	[h]
	159.2624	295.6692	322.3247	[J/g]
Hua	24.00	168.00	600.00	[h]
	233.4	317.25	339.8	[J/g]
Robeyst	14.66	45.79	140.99	[h]
	94.07	238.623	348.757	[J/g]
Smolik_Litos	10.01	19.19	261.78	[h]
	59.9159	329.2083	466.1589	[J/g]
Tamtsia early	18.00	24.00	102.00	[h]
	279.4076	307.3484	447.0522	[J/g]

Table 5.3: Experimental results of hydration heat

isothermal calorimeter, from [Principigallo et al., 2003], data from private communication and determined from evaporable water content and assumed potential hydration heat 480 J/g, from [Hua et al., 1995], [Robeyst et al., 2007], data measured at CTU by TAMAir isothermal calorimeter and finally, from [Tamtsia et al., 2004] assuming potential heat 500 J/g. In Tab. 5.2 cement chemical composition, gypsum content, w/c and fineness are presented. Measurements of hydration heat at three different times recalculated for the same reference temperature of 20 °C are shown in Tab. 5.3.

First, almost linear dependency within the input data caused by volume unity and hence the resulting ill-conditioning is solved by Principal Component Analysis (PCA) by transforming inputs into the space of principal directions and removing the direction with the smallest eigenvalue, for more details about Principal Component Analysis see [Shlens, 2003] or [Smith, 2002] Therefore, our approximation is a real function (hydration heat) of seven inputs – time plus seven original inputs transformed with PCA to the six dimensions. Next, several combinations of regression and correlation functions have been tested, see Fig. 5.15 and Fig. 5.16. Horizontal axes are for time and vertical axes for hydration heat. Note that zero point [0 h, 0 J/g] has been added to enforce a physically reasonable start of the heat-time relationship.

There are two main requirements on the approximation. We need an interpolation of experimental data to precisely describe the behavior in the vicinity of existing experiments and oppositely, the best possible description of the trend in extrapolation. This is of great importance since there is low number of available data and the range of parameters covered is usually also small. The deficiency of created metamodels for extrapolation purposes is clearly visible from Fig. 5.15 and Fig. 5.16. Whenever the metamodel is far away from given data, the prediction is approaching the mean trend. This means that in distant extrapolation we would obtain a flat surface in the case of a constant regression term and a linear surface in a linear case.

We have tried a dozen of combinations of regression descriptions and correlation functions and finally, a combination of an exponential correlation function, a linear regression of mixture parameters and an exponential regression term $(1 - e^{-T})$ for time T gives reasonable regression output, see Fig. 5.17. In Fig. 5.18, the result for optimized weights $\boldsymbol{\theta}$ with respect to the minimal MSE is presented. Since the curve of hydration heat his-

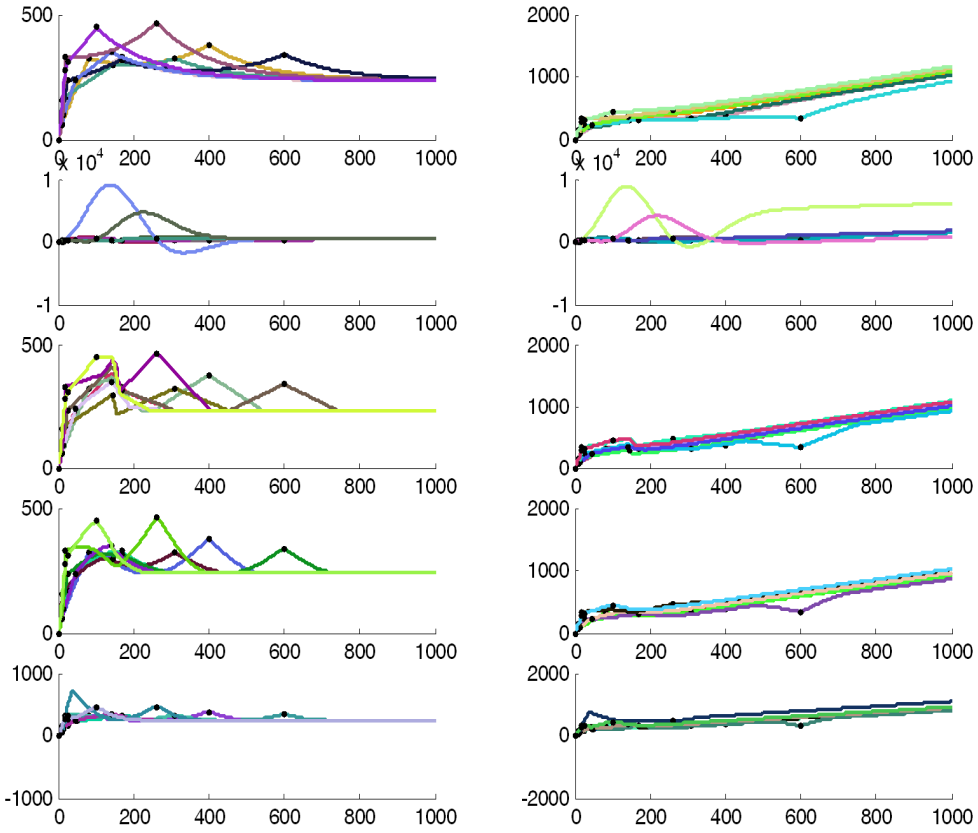

Figure 5.15: Cuts of approximations for **not optimized weights $\boldsymbol{\theta}$**: Constant regression term (left column), linear regression term (right column) and (from top) five correlation functions.

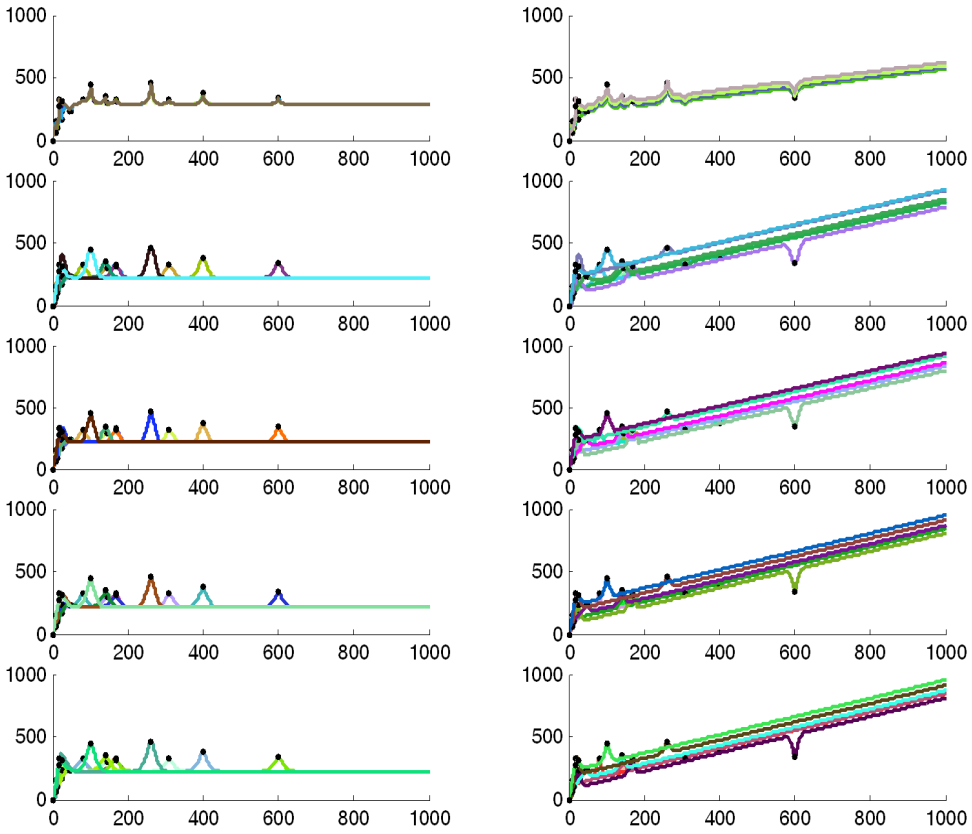

Figure 5.16: Cuts of approximations for **optimized weights $\boldsymbol{\theta}$ for minimal MSE**: Constant regression term (left column), linear regression term (right column) and (from top) five correlation functions.

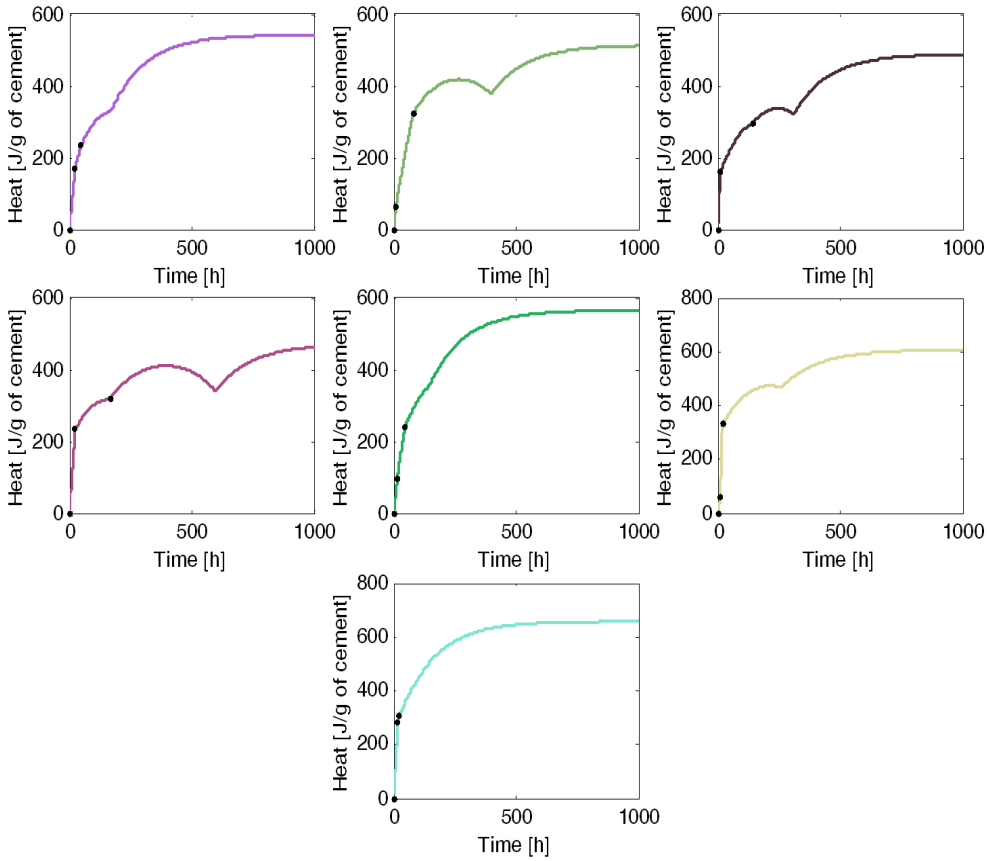

Figure 5.17: Cut of approximation through experiments using exponential correlation function, linear term of composition and exponential regression term in time for **not optimized weights $\boldsymbol{\theta}$**.

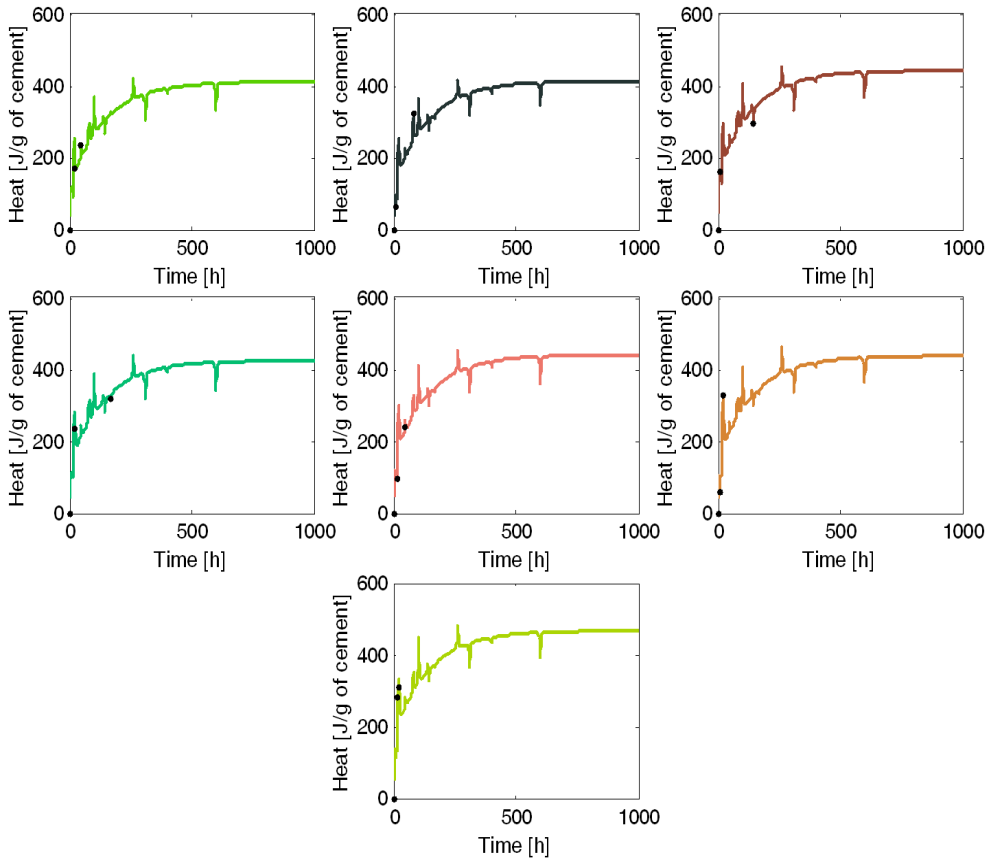

Figure 5.18: Cut of approximation through experiments using exponential correlation function, linear term of composition and exponential regression term in time for **optimized weights θ for minimal MSE**.

tory should be (from physical principles) monotonous, the traditional MSE minimization is replaced by minimization of a negative (numerical) derivative of a resulting curve in the time direction. As an optimization algorithm, the Quasi-Newton line-search method available in Matlab Optimization toolbox was used. The optimization algorithm ran 6.5 minutes on AMD Turion MT-37 notebook processor with more than 700 evaluations of the metamodel. The resulting curves are presented in Fig. 5.19. The approximation that almost ideally describes the physical problem has been found, however, the strict monotonicity has not been preserved, see again Fig. 5.19.

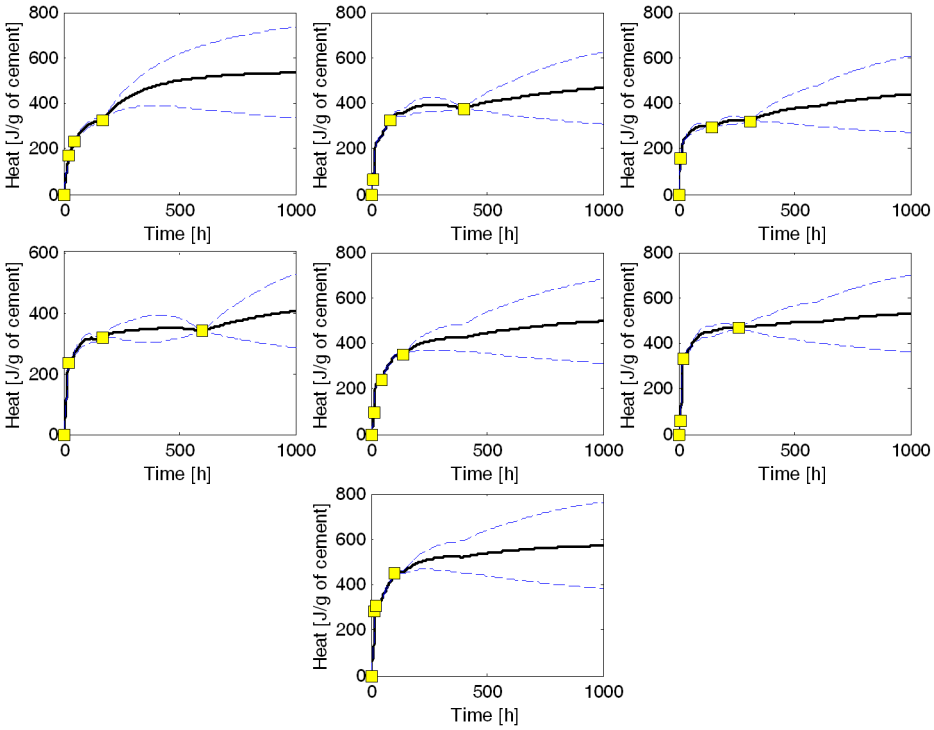

Figure 5.19: Cut of approximation through experiments with expected mean (black continuous line) and MSE bounds (blue dashed lines) for **optimized weights $\boldsymbol{\theta}$ for maximal monotonicity**.

Chapter 6

CONCLUSION

The two different approaches to Kriging was presented. Namely, the geostatistical and global optimization derivations have been shown. We also applied both Kriging methods to distinct data sets to show pros and cons of both approaches.

Chapter three was concerned with the geostatistical approach to Kriging. In this method the determination of a variogram (or semivariogram) is crucial. However, the determination of a variogram is not easy task. The usual practice of estimating the variogram is the method "by eye". It has the advantage that many errors of data can be uncovered. However, such fitting of a variogram requires very experienced curve fitter and enhanced knowledge of the data structure. The method of least squares can be also used to fit the variogram. In the case of the least squares, data errors are not uncovered. Furthermore, errors (deviation of a theoretical and experimental variogram) are supposed to be independent. However, this assumption is rarely met. From this summary it is obvious that the fitting of a variogram is more an art of an experienced curve fitter than the strict science.

In the chapter four a quite different access to Kriging was shown. Namely, the global optimization approach to Kriging was presented. This approach is approximation rather a interpolating method. The correlation function plays the crucial role instead of the variogram which is used in geostatistics. Nevertheless, the difficulties connected with the selection of the right model of a correlation function are similar to difficulties of the choice of a variogram model. The most common technique of estimating of a correlation function is maximizing the likelihood function. The main problem is that a less attention is paid to the regression term which is often taken as a constant. But analysts often know that the function, which has to be approximate, has a certain properties, e.g. monotonicity. However, Kriging do not preserve these properties. The searching for a proper model function is exhausting and time consuming, but the result is not guaranteed.

There is also the problem with large data sets. For example the free Matlab toolbox DACE is not able to compute a data set which has more than thousand points. Last but not least, another questions arise here. Kriging is defined as the best linear predictor. Another restriction in this definition is hidden: why should be the estimation computed only from the linear combinations of values? Why should be the minimum of variance the only one optimization criteria? It is clear that further research is needed. Nowadays, the new techniques are developed which use the nonlinear estimation and different optimization criteria. In the case of searching for a proper model, the Genetic Programming can be used to solve this task using a burden of computational power, see e.g. [Streeter, 2001] for more details.

BIBLIOGRAPHY

[Bárdossy, 1997] Bárdossy, A. (1997). Introduction to geostatistics. Technical report, Institute for Hydraulic Engineering and Water Resources Management, University of Stuttgart.

[Barry and Pace, 1997] Barry, R. P. and Pace, R. K. (1997). Kriging with large data sets using sparse matrix techniques. *Communications in statistics, simulation and computation*, 26(2):619–629.

[Bernstein, 2006] Bernstein, M. (2006). Interpolace srážkových dat na území prahy. Master's thesis, Česká zemědělská univerzita v Praze.

[Bohling, 2007] Bohling, G. (2007). Introduction to geostatistics. Boise State University.

[Clark, 1979] Clark, I. (1979). *Practical Geostatistics*. Applied Science Publishers LTD.

[Cressie and Johannesson, 2008] Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *Royal Statistical Society*, 70:209–226.

[Cressie, 1993] Cressie, N. A. C. (1993). *Statistics for spatial data revised edition*. A Wiley-Interscience publication.

[Giunta and Watson, 1998] Giunta, A. A. and Watson, L. T. (1998). A comparison of approximation modeling techniques: Polynomial versus interpolating models. Technical report, MAD Center for Advanced Vehicles, Virginia Polytechnic Institute and State University.

[H.-G. Beyer, 2007] H.-G. Beyer, B. S. (2007). Robust optimization – a comprehensive survey. *Comput. Methods Appl. Mech. Engrg.*, 196:3190–3218.

[Hornby and Horowitz, 1999] Hornby, P. and Horowitz, F. G. (1999). Closely related alternatives to kriging interpolation. Technical report, CSIRO Division of Exploration and Mining, Nedlands, WA 6009, Australia.

[Hua et al., 1995] Hua, C., Acker, P., and Ehrlacher, A. (1995). Analyses and models of the autogenous shrinkage of hardening cement paste. I. modelling at macroscopic scale. 25(7):1457–1468.

[Jin, 2005] Jin, Y. (2005). A comprehensive survey of fitness approximation in evolutionary computation. *Soft Computing*, 9:3–12.

[Jones, 2001] Jones, D. R. (2001). A taxonomy of global optimization methods based on response surfaces. *Journal of Global Optimization*, 21:345–383.

[Jones et al., 1998] Jones, D. R., Schonlau, M., and Welch, W. J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13:455–492.

[Kleijnen, 2007] Kleijnen, J. P. C. (2007). Kriging metamodeling in simulation: a review. Technical report, Department of Information Management, Tilburg University, Netherlands.

[Kleijnen, 2008] Kleijnen, J. P. C. (2008). *Design and Analysis of Simulation Experiments*. Springer.

[Krivoruchko, 2001] Krivoruchko, K. (2001). Using linear and non-linear kriging interpolation to produce probability maps. In *Annual Conference of the International Association for Mathematical Geology*.

[Lophaven et al., 2002] Lophaven, S. N., Nielsen, H. B., and Sondergaard, J. (2002). Aspects of the matlab toolbox dace. Technical Report IMM-REP-2002-13, Informatics and Mathematical Modelling, Technical University of Danemark. 44 pages, `http://www.imm.dtu.dk/~hbn/publ/TR0213.ps`.

[Matheron, 1963] Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, 58:1246–1266.

[Memarsadeghi, 2004] Memarsadeghi, N. (2004). Cokriging interpolation. Technical report, Computer Science Department of University of Maryland at College Park.

[Michálek, 2001] Michálek, J. (2001). Dvě poznámky o metodě kriging. *Robust 2000*, pages 189–197.

[Montgomery, 2005] Montgomery, D. C. (2005). *Design and Analysis of Experiments*. John Wiley and Sons, Inc., sixth edition edition.

[Principigallo et al., 2003] Principigallo, A., Lura, P., van Breugel, K., and Levita, G. (2003). Early development of properties in a cement paste: A numerical and experimental study. 33(7):1013 – 1020.

[Queipo et al., 2005] Queipo, N. V., Haftka, R. T., Shyy, W., Goel, T., Vaidyanathan, R., and Tucker, P. K. (2005). Surrogate-based analysis and optimization. *Progress in Aerospace Sciences*, 41:1–28.

[Robeyst et al., 2007] Robeyst, N., Gruyaert, E., and Belie, N. D. (2007). *Advances in Construction Materials 2007: Ultrasonic and calorimetric measurements on fresh concrete with blast-furnace slag*, chapter VI, pages 497–504. Springer Berlin Heidelberg.

[Sacks et al., 1989] Sacks, J., Welch, W., Mitchell, T., and Wynn, H. (1989). Design and analysis of computer experiments. *Statistical Science*, 4(4):409–435.

[Shewchuk, 1994] Shewchuk, J. R. (1994). An introduction to the conjugate gradient method without the agonizing pain. Technical report, School of Computer Science, Carnegie Mellon University, Pittsburgh.

[Shlens, 2003] Shlens, J. (2003). A tutorial on principal component analysis.

[Simpson et al., 2001] Simpson, T. W., Peplinski, J. D., Koch, P. N., and Allen, J. K. (2001). Metamodels for computer-based engineering design: Survey and recommendations. *Engineering with Computers*, 17:129–150.

[Smith, 2002] Smith, L. I. (2002). A tutorial on principal components analysis.

[Streeter, 2001] Streeter, M. J. (2001). Automated discovery of numerical approximation formulae via genetic programming. Master's thesis, Worcester Polytechnic Institute.

[Tamtsia et al., 2004] Tamtsia, B., Beaudoin, J., and Marchand, J. (2004). The early age short-term creep of hardening cement paste: Load-induced hydration effects. *Cement and Concrete Composites*, 26:481–489.

[Valtrová* and Lepš, 2009] Valtrová*, M. and Lepš, M. (2009). Kriging approximation in cement paste experimental performance. In *Proceedings of Engineering mechanics 2009*. stav teoretick a aplikovan mechaniky AVR.

[Šmilauer, 2006] Šmilauer, V. (2006). *Elastic properties of hydrating cement paste determined from hydration models*. PhD thesis, CTU in Prague. `http://mech.fsv.cvut.cz/~smilauer/`.

[Šmilauer et al., 2008] Šmilauer, V., Lepš, M., and Vitingerová, Z. (2008). Multiobjective optimization of cement paste performance. In *Engineering Mechanics 2008*, pages 961–967. Ústav termomechaniky AV ČR.

[Watson, 1984] Watson, G. S. (1984). Smoothing and interpolation by kriging and with splines. *Mathematical Geology*, 16(6):601–615.

Appendix A

THE LEAST SQUARES PREDICTOR

The purpose of this chapter is to gather information related to the Least Squares Predictor used throughout the text. In linear regression the Ordinary Least Squares (OLS) Estimator for $\boldsymbol{\beta}$ is used:

$$\widehat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z} \quad , \tag{A.1}$$

and

$$Var[\widehat{\boldsymbol{\beta}}_{OLS}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \quad . \tag{A.2}$$

$\boldsymbol{\beta}$ is the vector of unknown coefficients which have to be estimated:

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \quad .$$

Matrix $\mathbf{X}$ looks like this:

$$\mathbf{X} = \begin{pmatrix} 1 & x_1(\mathbf{s}_1) & x_2(\mathbf{s}_1) & \cdots & x_p(\mathbf{s}_1) \\ 1 & x_1(\mathbf{s}_2) & x_2(\mathbf{s}_2) & \cdots & x_p(\mathbf{s}_2) \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_1(\mathbf{s}_n) & x_2(\mathbf{s}_n) & \cdots & x_p(\mathbf{s}_n) \end{pmatrix} \quad ,$$

and the vector of observed value $\mathbf{Z}$ is defined as:

$$\boldsymbol{Z} = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{pmatrix} \quad .$$

Finally, the term $\mathbf{\Sigma}$ is a covariance matrix:

$$\mathbf{\Sigma} = \begin{pmatrix} C(\mathbf{0}) & C(\mathbf{s}_1 - \mathbf{s}_2) & C(\mathbf{s}_1 - \mathbf{s}_3) & \cdots & C(\mathbf{s}_1 - \mathbf{s}_n) \\ C(\mathbf{s}_2 - \mathbf{s}_1) & C(\mathbf{0}) & C(\mathbf{s}_2 - \mathbf{s}_3) & \cdots & C(\mathbf{s}_2 - \mathbf{s}_n) \\ C(\mathbf{s}_3 - \mathbf{s}_1) & C(\mathbf{s}_3 - \mathbf{s}_2) & C(\mathbf{0}) & \cdots & C(\mathbf{s}_3 - \mathbf{s}_n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ C(\mathbf{s}_n - \mathbf{s}_1) & C(\mathbf{s}_n - \mathbf{s}_2) & C(\mathbf{s}_n - \mathbf{s}_3) & \cdots & C(\mathbf{0}) \end{pmatrix}$$

In Kriging the General Least Squares (GLS) Estimator for $\boldsymbol{\beta}$ is used:

$$\widehat{\boldsymbol{\beta}}_{GLS} = (\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{Z} \quad , \tag{A.3}$$

and

$$Var[\widehat{\boldsymbol{\beta}}_{GLS}] = (\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1} \quad . \tag{A.4}$$

Appendix B

MATLAB CODES

B.1 Experimental Semivariogram

```
% function [V,A,C,sill] = esemivario (X,t,r1,r2,nl)
% Function computes experimental semivariogram
%
% Input:
%       X  - Matrix of location of sampled points (x, y)
%       t  - Vector of value at sampled points
%       r1 - Minimum range
%       r2 - Maximum range
%       nl - Number of lags
%
% Output:
%       V    - Distance matrix
%       A    - Experimental semivariogram
%       C    - Vector of lags
%       sill - Variance of data

function [V,A,C,sill] = esemivario (X,t,r1,r2,nl)

% Distance Matrix
[n nn] = size(X);
V = zeros(n,n);
for i = 1:n
    for j = 1:n
        dott = 0;
        for k = 1:nn
            dott = dott + (X(i,k)-X(j,k))^2;
        end
        V(i,j) = dott^0.5;
    end
end

% ----- Experimental Semivariogram ------------------------
sill = var(t);
A = zeros(nl,1);
d = r2 - r1;
for l = 1:nl
```

```
37      [r,c] = find((V >= ((l-1)*d + r1) & (V < ((l-1)*d + r2))));
38      if length(r) >= 1
39          dott = 0;
40          for k = 1:length(r)
41              dott = dott + (t(r(k)) - t(c(k)))^2;
42          end
43          A(l,1) = dott /(2*length(r));
44      end
45  end
46
47  % Vector of lags
48  C = (r2:d:(nl+1)*d)';
49
50  % --------- Plot semivariogram ----------------------
51  figure(1)
52  subplot(2,1,1)
53  hold on
54  for i = 1:n
55      plot(X(i,1),X(i,2),'*red','MarkerSize',5)
56  end
57  title('Lay-out of data');
58  xlabel('x [m]');
59  ylabel('y [m]');
60
61  subplot(2,1,2)
62  hold on
63  for i = 1:nl
64      if (A(i,1) ~= 0)
65          plot(C(i,1),A(i,1),'.k','MarkerSize',15);
66      end
67  end
68  title('Experimental Semivariogram')
69  xlabel('Lag [m]');
70  ylabel('Semivariance');
71  set(gcf,'PaperSize',[19.75 18.00],'PaperPositionMode','auto')
72  hold off
73  print -dpdf E_Semivario.pdf
74  close all
```

B.2 Theoretical Semivariograms

```
% function tsemivario (sill,V,A,C,sillS,sillE,sillG,nr,krok)
% Function creates theoretical semivariograms
% The range can be find out to call fucntion recursive
%
% Input:
%       sill  - Variance of data
%       V     - Distance matrix
%       A     - Vector of experimental semivariogram
%       C     - Vector of lags
%       sillS - Practical range for spherical s.
%       sillE - Practical range for exponential s.
%       sillG - Practical range for gaussian s.
%       nr    - Number of calling
%       step  - Step for correction of sill
%

function tsemivario(sill,V,A,C,sillS,sillE,sillG,nr,step)
if nr >= 1

    % -------- theoretical semivariogram (Spherical)----------------------
    S = A - (sillS - nr*step/10)*sill;
    if (S(1) > 0)
        [r] = find(S < 0);
    else
        [r] = find(S > 0);
    end

    rangeS = C(r(1)-1) + ((abs(S(r(1)-1))*(C(r(1)) - C(r(1)-1)))...
             /(abs(S(r(1)-1))+abs(S(r(1)))));

    % matrix of spherical theoretical semivariogram
    [n] = size(V);
    AS = zeros(n);

    for i = 1:n(1)
        for j = 1:n(2)
            if (i == j)
                AS(i,j) = sill;
            else
                if (V(i,j) <= rangeS)
                    AS(i,j) = sill*(1.5*(V(i,j)/rangeS)...
                              - 0.5*(V(i,j)/rangeS)^3);
                else
                    AS(i,j) = sill;
                end
```

```
46              end
47          end
48      end
49
50
51      % -------- theoretical semivariogram (Exponential)----------------------
52      E = A - (sillE - nr*step/10)*sill;
53      if (E(1) > 0)
54          [r] = find(E < 0);
55      else
56          [r] = find(E > 0);
57      end
58
59      rangeE = C(r(1)-1) + ((abs(E(r(1)-1))*(C(r(1)) - C(r(1)-1)))...
60               /(abs(E(r(1)-1))+abs(E(r(1)))));
61
62      % matrix of exponential theoretical semivariogram
63      AE = zeros(n);
64
65      for i = 1:n(1)
66          for j = 1:n(2)
67              if (i == j)
68                  AE(i,j) = sill;
69              else
70                  AE(i,j) = sill*(1-exp(-3*V(i,j)/rangeE));
71              end
72          end
73      end
74
75
76      % -------- theoretical semivariogram (Gaussian)----------------------
77      G = A - (sillG - nr*step/10)*sill;
78      if (G(1) > 0)
79          [r] = find(G < 0);
80      else
81          [r] = find(G > 0);
82      end
83
84      rangeG = C(r(1)-1) + ((abs(G(r(1)-1))*(C(r(1)) - C(r(1)-1)))...
85               /(abs(G(r(1)-1))+abs(G(r(1)))));
86
87      % matrix of gaussian theoretical semivariogram
88      AG = zeros(n);
89
90      for i = 1:n(1)
91          for j = 1:n(2)
92              if (i == j)
```

```
93                  AG(i,j) = sill;
94              else
95                  AG(i,j) = sill*(1-exp(-3*(V(i,j)^2)/(rangeG^2)));
96              end
97          end
98      end
99
100     % --------- plot semivariograms ----------------------
101     % plot experimental semivariogram
102     figure(nr);
103     hold on
104     for i = 1:length(A)
105         if (A(i,1) ~= 0)
106             plot(C(i,1),A(i,1),'*black','MarkerSize',5);
107         end
108     end
109
110     % plot Spherical theoretical semivariogram
111      for i = 1:(n(1))
112         for j = (i+1):n(2)
113             plot(V(i,j),AS(i,j),'.red','MarkerSize',4)
114         end
115     end
116
117     % plot Exponencial s.
118     for i = 1:(n(1))
119         for j = (i+1):n(2)
120             plot(V(i,j),AE(i,j),'.blue','MarkerSize',4)
121         end
122     end
123
124     % plot Gaussian s.
125     for i = 1:(n(1))
126         for j = (i+1):n(2)
127             plot(V(i,j),AG(i,j),'.green','MarkerSize',4)
128         end
129     end
130
131     title({['Spherical (red): range ',num2str(rangeS)];
132            ['Exponential (blue): range ',num2str(rangeE)];
133            ['Gaussian (green): range ',num2str(rangeG)]});
134     xlabel('Lag [m]');
135     print ('-dpdf',['Power ',num2str(nr),'.pdf'])
136     close all
137
138     tsemivario(sill,V,A,C,sillS,sillE,sillG,nr-1,step);
139 end
```

B.3 Simple Kriging in Geostatistics

```
function [K,k,lam,tp,v] = etkriging (sill,X,x,t,range,typ)
% Simple Kriging
%
% Input:
%       sill  - Variance
%       X     - Matrix of Sampled points (x, y)
%       x     - Matrix of Unsampled points (x, y)
%       t     - Vector of values at sampled points
%       range - Range
%       typ   - Type of semivarogram: s = spherical
%                                     e = exponential
%                                     g = gaussian
%
% Output:
%       K   - Covariance matrix
%       k   - Covariance matrix between sampled and unsampled points
%       lam - Vector of unknown coefficients
%       tp  - Predicted values at unsampled points



% Distance matrix
[n nn] = size(X);
V = zeros(n);
for i = 1:n
    for j = 1:n
        dott = 0;
        for k = 1:nn
            dott = dott + (X(i,k)-X(j,k))^2;
        end
        V(i,j) = dott^0.5;
    end
end

% Dinstance matrix between points of sampled and unsampled points
[r s] = size(x);
v = ones(r,n);
for j = 1:r
    for i = 1:n
        dott = 0;
        for k = 1:s
            dott = dott + (x(j,k)-X(i,k))^2;
        end
        v(j,i) = dott^0.5;
    end
```

```
46  end
47
48  [n] = size(V);
49  [m] = size(v);
50  K = zeros(n(1),n(2));
51  k = zeros(m(1),m(2));
52
53  % spherical semi
54  if (typ == 's')
55      for i = 1:n(1)
56          for j = 1:n(2)
57              if (V(i,j) <= range)
58                  K(i,j) = sill*(1 - (1.5*(V(i,j)/range) - 0.5*(V(i,j)/range)^3));
59              else
60                  K(i,j) = 0;
61              end
62          end
63      end
64
65
66      for i = 1:m(1)
67          for j = 1:m(2)
68              if (v(i,j) <= range)
69                      k(i,j) = sill*(1 - (1.5*(v(i,j)/range)...
70                               - 0.5*(v(i,j)/range)^3));
71                  else
72                      k(i,j) = 0;
73              end
74          end
75      end
76
77
78  % exponential semi
79  elseif (typ == 'e')
80      for i = 1:n(1)
81          for j = 1:n(2)
82              K(i,j) = sill*(1 - (1-exp(-3*V(i,j)/range)));
83          end
84      end
85
86      for i = 1:m(1)
87          for j = 1:m(2)
88              k(i,j) = sill*(1 - (1-exp(-3*v(i,j)/range)));
89          end
90      end
91
92  % gaussian semi
```

```
93  elseif (typ == 'g')
94      for i = 1:n(1)
95          for j = 1:n(2)
96              K(i,j) = sill*(1-(1-exp(-3*(V(i,j)^2)/(range^2))));
97          end
98      end
99
100     for i = 1:m(1)
101         for j = 1:m(2)
102             k(i,j) = sill*(1-(1-exp(-3*(v(i,j)^2)/(range^2))));
103         end
104     end
105 end
106
107 k = k';
108
109 % Vector of coefficients
110 lam = inv(K)*k;
111
112 % Mean of all data
113 mu = mean(t);
114
115 % Residuals from mean
116 Ra = t - mu;
117
118 R = lam'*Ra;
119
120 % Predicted values
121 tp = R + mean(t);
```