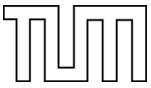


Technische Universität München

Department of Mathematics

Bachelor's Thesis

Kriging methods in spatial statistics

Andreas Lichtenstern

Supervisor: Prof. Dr. Claudia Czado

Advisor: Dipl.-Math.oec. Ulf Schepsmeier

Submission Date: 13.08.2013

I assure the single handed composition of this bachelor's thesis only supported by declared resources.

Garching, August 13, 2013

Zusammenfassung

In vielen angewandten wissenschaftlichen Disziplinen spielt die Prognose unbekannter Werte eine bedeutende Rolle. Da in der Praxis nur einige wenige Stichproben gemacht werden können, müssen die restlichen Werte an Stellen außerhalb der Stichprobenpunkte von den gemessenen Beobachtungen geschätzt werden. Dafür gibt es verschiedene Ansatzmöglichkeiten um geeignete Prognosewerte zu erhalten. In dieser Bachelorarbeit werden wir uns mit "best linear unbiased prediction (BLUP)", also mit optimaler, linearer und erwartungstreuer Schätzung befassen, welches in der Geostatistik auch *Kriging* genannt wird.

Grundsätzlich besteht das Ziel von Kriging aus der Prognose bestimmter Werte eines zugrundeliegenden räumlichen Zufallsprozesses $Z = Z(\mathbf{x})$ mit einem linearen Schätzer, also einem gewichteten Mittel der gemessenen Beobachtungen. Dabei stellt $Z(\mathbf{x})$ für jedes $\mathbf{x}$ in einem geographischen Gebiet eine Zufallsvariable dar. Wir werden uns zum einen mit der Prognose des Mittelwertes von $Z(\mathbf{x})$ über einem Raum und zum anderen mit der Schätzung des Wertes von $Z(\mathbf{x})$ an einem beliebigen Punkt $\mathbf{x}_0$ befassen.

Die generelle Idee hinter Kriging ist dabei, dass die Stichprobenpunkte nahe $\mathbf{x}_0$ eine größere Gewichtung in der Prognose bekommen sollten, um den Schätzwert zu verbessern. Aus diesem Grund stützt sich Kriging auf eine gewisse räumliche Struktur bzw. Abhängigkeit, welche meistens über die Eigenschaften der zweiten Momente der zugrundeliegenden Zufallsfunktion $Z(\mathbf{x})$ modelliert wird, das heißt Variogramm oder Kovarianz. Das Ziel ist es nun, die Gewichte im linearen Schätzer unter Berücksichtigung der Abhängigkeitsstruktur derart zu bestimmen, dass der endgültige Schätzwert unverzerrt ist, und des Weiteren unter allen erwartungstreuen linearen Schätzern minimale Varianz hat. Daraus ergibt sich ein restringiertes Minimierungsproblem, dessen Lösung die "optimalen" Gewichte des linearen Schätzers und damit den Kriging Schätzwert und die minimierte Kriging Varianz eindeutig festlegt.

Wir stellen die genannten Verfahren vor und beweisen deren Eigenschaften. Zum Abschluss werden diese Verfahren zur Vorhersage von Tagestemperaturen in Deutschland illustriert.

Contents

List of Figures

List of Tables

1 Introduction

The problem of obtaining values which are unknown appears and plays a big role in many scientific disciplines. For reasons of economy, there will always be only a limited number of sample points located, where observations are measured. Hence, one has to predict the unknown values at unsampled places of interest from the observed data to obtain their values, or respectively estimates, as well. For this sake, there exist several prediction methods for deriving accurate predictions from the measured observations.

In this thesis we introduce *optimal* or *best linear unbiased prediction (BLUP)*. The French mathematician Georges Matheron (1963) named this method *kriging*, after the South African mining engineer D. G. Krige (1951), as it is still known in spatial statistics today. There, kriging served to improve the precision of predicting the concentration of gold in ore bodies. However, the object "optimal linear prediction" even appeared earlier in literature, as for instance in Wold (1938) or Kolmogorov (1941a). But very much of the credit goes to Matheron for formalizing this technique and for extending the theory.

The general aim of kriging is to predict the value of an underlying random function $Z = Z(\mathbf{x})$ at any arbitrary location of interest $\mathbf{x}_0$, i.e. the value $Z(\mathbf{x}_0)$, from the measured observations $z(\mathbf{x}_i)$ of $Z(\mathbf{x})$ at the $n \in \mathbb{N}$ sample points $\mathbf{x}_i$. For this, let D be some geographical region in $\mathbb{R}^d$, $d \in \mathbb{N}$, which contains all considered points.

The main idea of kriging is that near sample points should get more weight in the prediction to improve the estimate. Thus, kriging relies on the knowledge of some kind of spatial structure, which is modeled via the second-order properties, i.e. variogram or covariance, of the underlying random function $Z(\mathbf{x})$. Further, kriging uses a weighted average of the observations $z(\mathbf{x}_i)$ at the sample points $\mathbf{x}_i$ as estimate. At this point the question arises how to define the "best" or "optimal" weights corresponding to the observed values in the linear predictor. The expressions "best" and "optimal" in our context of prediction with kriging are meant in the sense that the final estimate should be unbiased and then should have minimal error variance among all unbiased linear predictors. These resulting weights will depend on the assumptions on the mean value $\mu(\mathbf{x})$ as well as on the variogram or covariance function of $Z(\mathbf{x})$. Note that we use the term "prediction" instead of "estimation" to clear that we want to predict values of some random quantities, whereas estimation is refered to estimate unknown, but fixed parameters.

The main part of this thesis will be the presentation of four geostatistical kriging methods. First, we introduce *kriging the mean*, which serves to predict the mean value of $Z(\mathbf{x})$ over the spatial domain D. Secondly, we perform *simple kriging* for predicting $Z(\mathbf{x})$ at any arbitrary point $\mathbf{x}_0$, which represents the simplest case of kriging prediction. Then we consider the most frequently used kriging method in practice, *ordinary kriging*. And finally, we present *universal kriging*, which will be our most general considered model in this thesis compared with the previous ones. Hereby a drift in the mean $\mu(\mathbf{x})$ of $Z(\mathbf{x})$ can be taken in the prediction into account to improve the estimate.

Since all above kriging types rely on the same idea, i.e. to derive the best linear unbiased

predictor, the organization of each section treating one method will be similar. First, we state the model assumptions on the mean $\mu(\mathbf{x})$ and on the second-order properties of the underlying random function $Z(\mathbf{x})$. Subsequently, we define the general linear predictor for each kriging method and give the necessary conditions for uniform, i.e. general unbiasedness. These constraints are called *universality conditions* by Matheron (1971). Further we compute the prediction variance. It is defined as the variance of the difference of the linear predictor and the predictand. As mentioned above, our aim is then to minimize this prediction variance subject to the conditions for uniform unbiasedness, since kriging is synonymous for best linear unbiased spatial prediction. In other words, we want to maximize the accuracy of our predictor.

The solution of this resulting constraint minimization problem yields the so-called *kriging equations* for each of the following kriging types. Matheron (1971) called these conditions for achieving minimal prediction variance under unbiasedness constraints *optimality conditions*. Solving these equations will lead us to the "optimal" weights such that we can finally compute the kriging estimate and its corresponding minimized kriging variance at the end.

An closing application paragraph illustrates the performance of every described kriging method. Therefore we will use the *R* package *gstat* (Pebesma 2001) on daily temperature data in Germany.

Last but not least, we give a brief overview and summary of this thesis and say some words about spatio-temporal prediction in the end, see the last part "Summary and Outlook" (p. 85).

2 Mathematical basics

Before beginning with the main topic of this thesis, kriging, we present some mathematical background material. In particular, we need to define random variables, random vectors and their distributions. Afterwards we prepare some properties of definite matrices and of matrices of a block nature. Furthermore, we recall best linear unbiased prediction, which is called kriging in geostatistical literature (Stein 1999) in the last part.

2.1 Probability Theory

Most definitions of this section are taken from Durrett (2010, Chapter 1), the rest can be found for instance in Nguyen and Rogers (1989).

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a *probability space* with *sample space* $\Omega \neq \emptyset$, *σ-field* $\mathcal{F}$ and *probability measure* $\mathbb{P}$. Further let $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ denote the *measurable space* with sample set $\mathbb{R}$ and Borel-σ-algebra $\mathcal{B}(\mathbb{R})$.

Definition 2.1 (Random variable)
A function $X : \Omega \to \mathbb{R}$, $\omega \mapsto X(\omega)$, is called a (real-valued) *random variable* if X is $\mathcal{F}$-$\mathcal{B}(\mathbb{R})$ measurable, i.e.

$$\{X \in A\} := X^{-1}(A) = \{\omega \in \Omega : \ X(\omega) \in A\} \in \mathcal{F}$$

for all $A \in \mathcal{B}(\mathbb{R})$.

Definition 2.2 (Distribution)
For $A \in \mathcal{B}(\mathbb{R})$, set

$$\nu(A) := \mathbb{P}\left(X^{-1}(A)\right) = \mathbb{P}(X \in A),$$

where ν is a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ (*image measure*) and is called the *distribution* of the random variable X.

Definition 2.3 (Random vector)
A function $\mathbf{X} : (\Omega, \mathcal{F}) \to (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, $n \geq 2$, is called a *random vector* if $\mathbf{X}$ is $\mathcal{F}$-$\mathcal{B}(\mathbb{R}^n)$ measurable, i.e.

$$\{\mathbf{X} \in A\} := \mathbf{X}^{-1}(A) = \{\omega \in \Omega : \ \mathbf{X}(\omega) \in A\} \in \mathcal{F}$$

for all $A \in \mathcal{B}(\mathbb{R}^n)$.

Remark 2.4 (Alternative definition of a random vector)
An alternative definition of a random vector can be found in Nguyen and Rogers (1989, Chapter 3), using random variables:

Let $X_1, \ldots, X_n, \ n \in \mathbb{N}$, be random variables on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then $\mathbf{X} := (X_1, \ldots, X_n)^T$ as a mapping from $(\Omega, \mathcal{F})$ to $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ is a random vector, since it satisfies

$$\begin{aligned}\{\mathbf{X} \in A\} &= \{(X_1, \ldots, X_n)^T \in A\} = \{X_1 \in A_1, \ldots, X_n \in A_n\} \\ &= \{X_1 \in A_1\} \cap \ldots \cap \{X_n \in A_n\} \in \mathcal{F}\end{aligned}$$

for all $A = A_1 \times \ldots \times A_n \in \mathcal{B}(\mathbb{R}^n)$. Here, the sets of the form $A_1 \times \ldots \times A_n$ for Borel sets $A_i \in \mathcal{B}(\mathbb{R})$ generate the Borel-σ-algebra $\mathcal{B}(\mathbb{R}^n)$.

Furthermore, we define the multivariate distribution of a random vector following the definition of the distribution of a random variable in one dimension:

Definition 2.5 (Multivariate distribution)
For $A \in \mathcal{B}(\mathbb{R}^n)$, set

$$\nu(A) := \mathbb{P}(\mathbf{X}^{-1}(A)) = \mathbb{P}(\mathbf{X} \in A),$$

where ν is a probability measure on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ (*image measure*) and is called the *multivariate distribution* of the random vector $\mathbf{X}$.

After defining the distribution of a random vector, we want to define its expectation and its covariance. Hence, to guarantee that all expectations and covariances of the variables in the upcoming definitions exist, we consider the space $\mathcal{L}^2 := \{X : \Omega \to \mathbb{R} \text{ random variable: } \mathbb{E}[X^2] < \infty\}$. In the following, we refer particularly to the book by Fahrmeir et al. (1996, Chapter 2), but also to Georgii (2013, Chapter 4).

Definition 2.6 (Expectation vector)
Let $X_1, \ldots, X_n$ be random variables in $\mathcal{L}^2$. The *expectation (vector)* of the random vector $\mathbf{X} = (X_1, \ldots, X_n)^T$ is defined by

$$\mathbb{E}[\mathbf{X}] := (\mathbb{E}[X_1], \ldots, \mathbb{E}[X_n])^T \in \mathbb{R}^n,$$

i.e. the expectation is taken componentwise, such that $\mathbb{E}[\mathbf{X}]_i = \mathbb{E}[X_i]$, $i = 1, \ldots, n$.

Definition 2.7 (Covariance matrix)
Let $X_1, \ldots, X_n \in \mathcal{L}^2$. The symmetric *covariance matrix* of the random vector $\mathbf{X} = (X_1, \ldots, X_n)^T$ is defined by

$$\Sigma := Cov(\mathbf{X}) := \mathbb{E}\left[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T\right]$$

$$= \begin{pmatrix} Var(X_1) & Cov(X_1, X_2) & \cdots & Cov(X_1, X_n) \\ Cov(X_2, X_1) & Var(X_2) & \cdots & Cov(X_2, X_n) \\ \vdots & \vdots & \cdots & \vdots \\ Cov(X_n, X_1) & Cov(X_n, X_2) & \cdots & Var(X_n) \end{pmatrix},$$

i.e. $\Sigma_{i,j} := Cov(X_i, X_j)$ with

- $Cov(X_i, X_j) := \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])] = \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\,\mathbb{E}[X_j]$ and
- $Var(X_i) := Cov(X_i, X_i)$.

Proposition 2.8 (Properties of the covariance)
Let X, Y, X_i, $Y_i \in \mathcal{L}^2$ and let a_i, b_i, $c \in \mathbb{R}$, $i = 1, \ldots, n$.
The covariance satisfies the following properties:

(i) Symmetry: $Cov(X, Y) = Cov(Y, X)$

(ii) Bilinearity: $\sum_{i=1}^n a_i X_i \in \mathcal{L}^2$ and $\sum_{j=1}^n b_j Y_j \in \mathcal{L}^2$ with $Cov\left(\sum_{i=1}^n a_i X_i, \sum_{j=1}^n b_j Y_j\right) = \sum_{i=1}^n \sum_{j=1}^n a_i b_j Cov(X_i, Y_j)$

(iii) Constants: $Cov(X, Y + c) = Cov(X, Y),\ Cov(X, c) = 0$

(iv) The covariance matrix $\Sigma \in \mathbb{R}^{n\times n}$ of $\mathbf{X} = (X_1, \ldots, X_n)^T$ is *positive semidefinite*, i.e. $\forall \mathbf{v} = (v_1, \ldots, v_n)^T \in \mathbb{R}^n$:

$$\mathbf{v}^T \Sigma \mathbf{v} = \sum_{i=1}^{n} \sum_{j=1}^{n} v_i \Sigma_{i,j} v_j \geq 0.$$

Proof:
The statements (i)-(iii) in Proposition 2.8 simply follow by inserting the definition of the covariance.

Hence, we give only the proof of the last property, the positive semidefiniteness of the symmetric covariance matrix Σ, since this will occur later in this thesis. For this, let $\mathbf{v} = (v_1, \ldots, v_n)^T \in \mathbb{R}^n$ be given and define the random variable $Z := \sum_{i=1}^{n} v_i X_i$. It follows:

$$\mathbf{v}^T \Sigma \mathbf{v} = \sum_{i=1}^{n} \sum_{j=1}^{n} v_i \Sigma_{i,j} v_j$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} v_i v_j Cov(X_i, X_j) = Cov\left(\sum_{i=1}^{n} v_i X_i, \sum_{j=1}^{n} v_j X_j \right)$$

$$= Cov(Z, Z) = \mathbb{E}\left[(Z - \mathbb{E}[Z])^2 \right] \geq 0$$

□

In this thesis we will always assume the variance of a linear combination of random variables to be strictly positive, i.e. $Var(\sum_{i=1}^{n} v_i X_i) > 0$, and not only nonnegative (cf. Proposition 2.8 (iv)). This assumption makes sense as in the case that the variance equals zero, the sum $\sum_{i=1}^{n} v_i X_i$ would be almost surely equal to a constant, i.e. to its own expectation. This follows from $0 = Var(Z) = \mathbb{E}\left[\underbrace{(Z - \mathbb{E}[Z])^2}_{\geq 0} \right]$ and hence $Z = \mathbb{E}[Z]$ almost surely for any random variable Z. This fact leads us to the following intuitive assumption:

Assumption 2.9 (Nondegenerate covariance matrix)
We assume the covariance matrix Σ to be *nondegenerate*, i.e. to be strictly *positive definite*, such that

$$\mathbf{v}^T \Sigma \mathbf{v} > 0$$

$\forall \mathbf{v} \in \mathbb{R}^n,\ \mathbf{v} \neq \mathbf{0}$.

2.2 Definite and block matrices

For our later calculations, we will need some properties of the covariance matrix Σ, which we supposed to be positive definite. Hence, we state some useful equivalences about positive definite matrices in the following proposition. And afterwards we give a formula how the determinant of a block matrix can be computed (see Fahrmeir et al. 1996, pp. 807–808, 815–819).

Proposition 2.10 (Characterization positive definite matrices)
Let $M \in \mathbb{R}^n,\ n \in \mathbb{N}$, be a symmetric and positive semidefinite matrix.
Then, the following statements are equivalent:

(i) M is positive definite.

(ii) All eigenvalues of M are strictly positive.

(iii) M is invertible, i.e. $\det(M) \neq 0$.

(iv) M^{-1} is symmetric and positive definite.

Hence we conclude that the covariance matrix Σ of some random vector is invertible and its inverse Σ^{-1} is also positive definite.

Proposition 2.11 (Determinant of block matrix)
Let $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{m\times n}$ and $D \in \mathbb{R}^{m\times m}$ for $m, n \in \mathbb{N}$. Further let the matrix A be invertible.
Then, the determinant of the block matrix $\begin{pmatrix} A & B \\ C & D \end{pmatrix}$ can be written as

$$\det \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \det(A) \det\left(D - CA^{-1}B\right).$$

Proof:
This fact simply follows from the invertibility of the matrix A and the decomposition of the block matrix:

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} A & 0 \\ C & 1 \end{pmatrix} \begin{pmatrix} 1 & A^{-1}B \\ 0 & D - CA^{-1}B \end{pmatrix}.$$

Hence, the formula in Proposition 2.11 holds due to the multiplicativity property of the determinant:

$$\begin{aligned} \det \begin{pmatrix} A & B \\ C & D \end{pmatrix} &= \det \begin{pmatrix} A & 0 \\ C & 1 \end{pmatrix} \det \begin{pmatrix} 1 & A^{-1}B \\ 0 & D - CA^{-1}B \end{pmatrix} \\ &= \det \begin{pmatrix} A^T & C^T \\ 0 & 1 \end{pmatrix} \det \begin{pmatrix} 1 & A^{-1}B \\ 0 & D - CA^{-1}B \end{pmatrix} \\ &= \det(A^T) \det\left(D - CA^{-1}B\right) = \det(A) \det\left(D - CA^{-1}B\right). \end{aligned}$$

□

2.3 Linear prediction

The last point which we want to consider, is linear prediction, following Stein (1999) in Chapter 1. Since kriging is used as a synonym for best linear unbiased prediction in spatial statistics, we want to define what is actually meant by this expression.

For this reason, suppose a random field $Z = Z(\mathbf{x})$ with $\mathbf{x} \in D \subseteq \mathbb{R}^d$ for some $d \in \mathbb{N}$, where $Z(\mathbf{x})$ is a random variable for each $\mathbf{x} \in D$. We observe this random function at the $n \in \mathbb{N}$ distinct sample points $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Further, let $\mathbf{Z} := (Z(\mathbf{x}_1), \ldots, Z(\mathbf{x}_n))^T \in \mathbb{R}^n$ denote the random vector providing the random function $Z(\mathbf{x})$ evaluated at the sample locations, i.e. the random variables $Z(\mathbf{x}_i)$. We wish to predict the value of the random variable $Z(\mathbf{x})$ at any arbitrary point $\mathbf{x}_0 \in D$.

Stein (1999) defined the linear predictor of this value as the linear combination of some constant and of the random function $Z(\mathbf{x})$ evaluated at the samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$, weighted with some unknown coefficients:

Definition 2.12 (Linear predictor)
The *linear predictor* $Z^*(\mathbf{x}_0)$ of the value of $Z(\mathbf{x})$ at $\mathbf{x}_0$ is defined by

$$Z^*(\mathbf{x}_0) := \lambda_0 + \boldsymbol{\omega}^T \mathbf{Z} = \lambda_0 + \sum_{i=1}^{n} \omega_i Z(\mathbf{x}_i),$$

with constant $\lambda_0 \in \mathbb{R}$ and weight vector $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)^T \in \mathbb{R}^n$.

Definition 2.13 (Unbiased predictor)
The predictor $Z^*(\mathbf{x}_0)$ of $Z(\mathbf{x}_0)$ is called *unbiased* if its expected error is zero:

$$\mathbb{E}[Z(\mathbf{x}_0) - Z^*(\mathbf{x}_0)] = 0.$$

Note that to ensure this unbiasedness for any choice of $\boldsymbol{\omega}$, we infer that the constant λ_0 has to be chosen such that $\lambda_0 = \mathbb{E}[Z(\mathbf{x}_0)] - \boldsymbol{\omega}^T \mathbb{E}[\mathbf{Z}] = \mu_0 - \boldsymbol{\omega}^T \boldsymbol{\mu} = \mu_0 - \sum_{i=1}^{n} \omega_i \mu_i$, where μ_0 denotes the expected value of $Z(\mathbf{x}_0)$, $\boldsymbol{\mu} := \mathbb{E}[\mathbf{Z}]$ and $\mu_i := \mathbb{E}[Z(\mathbf{x}_i)]$ the *ith* component of the mean vector $\boldsymbol{\mu}$.

Definition 2.14 (Best linear unbiased predictor (BLUP))
The linear predictor $Z^*(\mathbf{x}_0)$ of $Z(\mathbf{x}_0)$ is called *best linear unbiased predictor (BLUP)*, if it is unbiased and has minimal prediction variance among all linear unbiased predictors.

Notice that minimizing the prediction variance $Var(Z(\mathbf{x}_0) - Z^*(\mathbf{x}_0))$ of an unbiased predictor is identical with minimizing the *mean squared error*

$$mse(Z^*(\mathbf{x}_0)) := \mathbb{E}\left[(Z(\mathbf{x}_0) - Z^*(\mathbf{x}_0))^2\right]$$

of the predictor $Z^*(\mathbf{x}_0)$, since the squared *bias* $\left(\underbrace{\mathbb{E}[(Z(\mathbf{x}_0) - Z^*(\mathbf{x}_0))]}_{bias=0}\right)^2 = 0.$

Finally, since finding the best linear unbiased predictor, i.e. finding the "best" weights

$\omega_1, \ldots, \omega_n$, is equal with finding the minimum of the prediction variance subject to the unbiasedness condition of the linear predictor, we want to present a theorem by Rao (1973, p. 60). We will apply this theorem later in the kriging methods, where minimizing the prediction variance turns out to be finding the minimum of a quadratic form subject to some linear constraints.

Theorem 2.15 (Minimum of definite quadratic form with constraints)
Let A be a positive definite $m \times m$ matrix, B a $m \times k$ matrix, and U be a k-vector. Denote by S^- any generalized inverse of $B^T A^{-1} B$. Then

$$\inf_{B^T X = U} X^T A X = U^T S^- U,$$

where X is a column vector and the infimum is attained at $X_* = A^{-1} B S^- U$.

Since Theorem 2.15 makes use of the object generalized inverse of a matrix, we do not want to omit its definition, given by Rao (1973, p. 24):

Definition 2.16 (Generalized inverse)
Consider an $m \times n$ matrix A of any rank. A *generalized inverse* (or a *g-inverse*) of A is a $n \times m$ matrix, denoted by A^-, such that $X = A^- Y$ is a solution of the equation $AX = Y$ for any $Y \in \mathcal{M}(A)$.

Here, $\mathcal{M}(A)$ stands for the space spanned by A, i.e. the smallest subspace containing A.

Remark 2.17 (Characterization of generalized inverse)
Rao (1973, pp. 24–25) observed that

$$A^- \text{ is a generalized inverse of } A \Leftrightarrow AA^-A = A.$$

Note that for an invertible matrix A, a possible generalized inverse is simply given by its own inverse matrix, i.e. $A^- = A^{-1}$, since $AA^{-1}A = A$.

Rao (1973) also mentioned that this generalized inverse A^- always exists, but is not necessarily unique.

Arriving at this point, we can finish with our preparation of the needed mathematical basics. But just one step before beginning with the main part of this thesis, we want to give some basic notation. We will use them in most of the following sections whenever the expressions make sense, i.e. if they are well-defined. For their definitions and some description see Table 2.1 below. We will always consider an underlying random function $Z(\mathbf{x})$ for $\mathbf{x}$ in a spatial domain $D \subseteq \mathbb{R}^d$ for some $d \in \mathbb{N}$. We observe this function on the $n \in \mathbb{N}$ sample points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and obtain their values $z(\mathbf{x}_1), \ldots, z(\mathbf{x}_n)$. Our object of interest is to predict the value of $Z(\mathbf{x})$ at any arbitrary and mostly unsampled location of interest, denoted by $\mathbf{x}_0$.

Symbol	Dimension	Definition	Description
$\mathbf{Z}$	$\mathbb{R}^n$	$\mathbf{Z}_i := Z(\mathbf{x}_i)$	Random vector at samples
$\mathbf{z}$	$\mathbb{R}^n$	$\mathbf{z}_i := z(\mathbf{x}_i)$	Observation vector
Σ	$\mathbb{R}^{n \times n}$	$\Sigma_{i,j} := Cov(Z(\mathbf{x}_i), Z(\mathbf{x}_j))$	Covariance matrix of $\mathbf{Z}$
$\mathbf{c}_0$	$\mathbb{R}^n$	$(\mathbf{c}_0)_i := Cov(Z(\mathbf{x}_i), Z(\mathbf{x}_0))$	Covariances between samples and location of interest
$\mathbb{1}$	$\mathbb{R}^n$	$\mathbb{1} := (1, \ldots, 1)^T$	Vector containing only ones

Table 2.1: Basic notation

3 Data set

At the end of each following section, we want to apply our theoretical results to some real data. We will investigate how the different kriging methods perform and how they are implemented in the statistical software *R*.

Therefore, we will always take exemplarily the two dates 2010/11/28 and 2012/06/09 during this thesis. Our complete data set contains 78 weather stations in Germany, where the mean temperature is measured in °C. To perform our prediction, we will use the first 54 stations for model fitting and the last 24 stations as test data, i.e. for comparison of the prediction estimates with their corresponding measured temperature values. The coordinates of each station are given by its latitude, longitude and elevation, i.e. its height given in meters above sea level. For an overview see Table 3.1, where the first ten stations and their corresponding values are printed.

	Station	Abbreviation	Latitude	Longitude	Elevation
1	Angermünde	ange	53.03	13.99	54.00
2	Aue	auee	50.59	12.72	387.00
3	Berleburg, Bad-Stünzel	brle	50.98	8.37	610.00
4	Berlin-Buch	brln	52.63	13.50	60.00
5	Bonn-Roleber	bonn	50.74	7.19	159.00
6	Braunschweig	brau	52.29	10.45	81.20
7	Bremen	brem	53.05	8.80	4.00
8	Cottbus	cott	51.78	14.32	69.00
9	Cuxhaven	cuxh	53.87	8.71	5.00
10	Dresden-Hosterwitz	dres	51.02	13.85	114.00

Table 3.1: First 10 weather stations included in our data set

In the following sections, we will always compare our predicted temperature estimates for the last 24 weather stations with the ”true”, i.e. measured ones, where the prediction is based on the fitted models of the first 54 basic stations.

Additionally, we will perform a grid of latitude and longitude of Germany. Then, for graphical representation, we will plot our predicted results and their corresponding prediction variances in the map of Germany, as it is drawn in Figure 3.1. For comparison, Figure 3.2 shows the additional weather stations, where our 54 basic stations are labeled as ”+”.

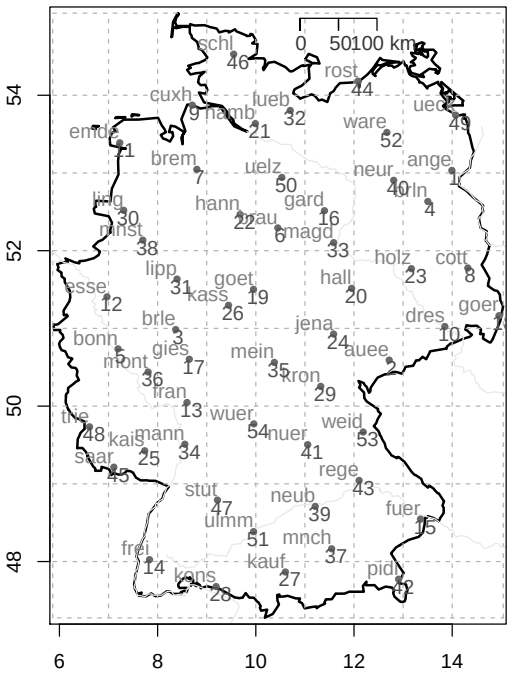


Figure 3.1: 54 basic weather stations for model fitting

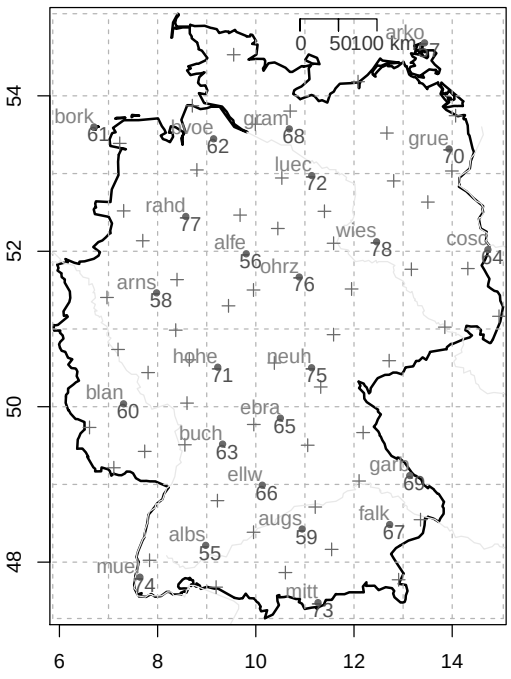


Figure 3.2: Additional 24 weather stations used as test data; "+" labels the 54 stations of Figure 3.1

4 The Variogram

The variogram as a geostatistical method is a convenient tool for the analysis of spatial data and builds the basis for kriging (Webster and Oliver 2007, p. 65), which we will discuss in the following sections and is the main topic of this thesis.

The idea of the variogram relies on the assumption that the spatial relation of two sample points does not depend on their absolute geographical location itself, but only on their relative location (Wackernagel 2003). Thus, our problem of interest is to find a measure for the spatial dependence, given n distinct sample points $(\mathbf{x}_i)_{i=1,\dots,n}$ in a spatial domain D, where the observed values $z(\mathbf{x}_i)$ are modeled as realizations of real-valued random variables $Z(\mathbf{x}_i)$ of a random function $Z = Z(\mathbf{x})$, $\mathbf{x} \in D$. This measure we will later call *variogram function* according to Matheron (1962) and Cressie (1993). Most common in practice are the cases, where $d = 1, 2$ or 3.

Hence, the aim of this section is to derive, i.e. estimate, a suitable variogram function from the underlying observed data, which we can use in our kriging methods afterwards. For this reason we have to go forward according to the following steps, as for instance presented in Wackernagel (2003, pp. 45–61) or Cressie (1993, pp. 29–104):

(i) Draw the so-called *variogram cloud* by plotting the dissimilarities of two location points against their lag distance $\mathbf{h}$.

(ii) Construct the *experimental variogram* by grouping similar lags $\mathbf{h}$.

(iii) Fit the experimental variogram with a *parametric variogram model function* by choosing a suitable variogram model and estimating the corresponding parameters, e.g. by a least squares fit.

At the end we can use the estimated variogram function in our prediction of temperature values at unsampled locations, since kriging requires the knowledge of a variogram or covariance function.

Details on the three steps above are given in the following, but first we want to introduce the theory behind, the *theoretical variogram*. It will restrict the set of all valid variogram functions in the estimation in point (iii), due to the consequence of some of its properties.

Note that there are plenty of scientific books which cover the subject variogram and among them, Georges Matheron (1962, 1963, 1971) was one of the first who introduced it. Other popular examples are the books by Cressie (1993), Journel and Huijbregts (1978), Wackernagel (2003), Webster and Oliver (2007) and Kitanidis (1997), who based their theory on the work of Matheron.

4.1 The theoretical variogram

Our first aim is to find a function to measure the spatial relation of the random function $Z(\mathbf{x})$, i.e. how two different random variables of $Z(\mathbf{x})$ influence each other. In theory,

this is usually done by introducing a quantity as defined in Definition 4.2 in the following, which uses the variance of the increments $Z(\mathbf{x}+\mathbf{h})-Z(\mathbf{x})$ for $\mathbf{x}$, $\mathbf{x}+\mathbf{h} \in D$ and separating vector $\mathbf{h}$. The increments display the variation in space of $Z(\mathbf{x})$ at $\mathbf{x}$ and the variance acts as a measure for the average spread of the values. Matheron (1962) named this quantity *theoretical variogram*, although it has even appeared earlier, e.g. in Kolmogorov (1941b) or Matérn (1960).

Now, to guarantee that our future definition of the variogram function is well-defined, we assume the underlying random function $Z(\mathbf{x})$ to be *intrinsically stationary*, or $Z(\mathbf{x})$ is to satisfy the *intrinsic hypothesis* respectively, as defined by Matheron (1971, p. 53) and Wackernagel (2003, pp. 50–51):

Definition 4.1 (Intrinsic stationarity)
$Z(\mathbf{x})$ is *intrinsically stationary of order two* if for the increments $Z(\mathbf{x}+\mathbf{h})-Z(\mathbf{x})$ it holds:

(i) The mean $\mu(\mathbf{h})$ of the increments is translation invariant in D and equals zero, no matter where $\mathbf{h}$ is located in D, i.e. $\mathbb{E}[Z(\mathbf{x}+\mathbf{h})-Z(\mathbf{x})] = \mu(\mathbf{h}) = 0 \; \forall \mathbf{x}, \mathbf{x}+\mathbf{h} \in D$. In other words, $Z(\mathbf{x})$ has a constant mean.

(ii) The variance of the increments is finite and its value only depends on the separating vector $\mathbf{h}$ in the domain, but not on its position in D, i.e.

$$Var(Z(\mathbf{x}+\mathbf{h}) - Z(\mathbf{x})) < \infty \; \forall \mathbf{x}, \mathbf{x}+\mathbf{h} \in D$$

and is only a function of $\mathbf{h}$.

Definition 4.2 (Theoretical variogram)
Matheron (1962) and Cressie (1993, p. 58) defined the *theoretical variogram* $2\gamma(\mathbf{h})$ as the function

$$2\gamma(\mathbf{h}) := Var\left(Z(\mathbf{x}+\mathbf{h}) - Z(\mathbf{x})\right)$$

for $\mathbf{x}$, $\mathbf{x}+\mathbf{h} \in D$ and lag $\mathbf{h} = (\mathbf{x}+\mathbf{h}) - \mathbf{x}$.

In the literature, for instance in Wackernagel (2003) or Webster and Oliver (2007), $\gamma(\mathbf{h})$ is called *semivariogram*, *semivariance* or even synonymously *variogram*, too. For notational convenience, we will use these terms simultaneously for $\gamma(\mathbf{h})$ from now on.

Furthermore, notice that by assuming intrinsic stationarity of $Z(\mathbf{x})$, the variogram is indeed well-defined due to the previous stationarity conditions in Definition 4.1 (i) and (ii), i.e. it is finite and does not depend on the explicit location of $\mathbf{x}$ in the spatial domain D, but only on $\mathbf{h}$. For any intrinsically random function $Z(\mathbf{x})$ we also infer that

$$2\gamma(\mathbf{h}) := Var\left(Z(\mathbf{x}+\mathbf{h}) - Z(\mathbf{x})\right) = \mathbb{E}\left[\left(Z(\mathbf{x}+\mathbf{h}) - Z(\mathbf{x})\right)^2\right]$$

But according to Cressie (1993, p. 69), we should keep in mind that in fact, the variogram function, whenever it makes sense for other more general random processes (except intrinsically stationary), should be defined as the variance of the increments $Var(Z(\mathbf{x}+\mathbf{h}) - Z(\mathbf{x}))$ and not as the expectation of the squared increments $\mathbb{E}\left[\left(Z(\mathbf{x}+\mathbf{h}) - Z(\mathbf{x})\right)^2\right]$ of

$Z(\mathbf{x})$, since for a general, nonconstant first moment function $\mu(\mathbf{x})$ of $Z(\mathbf{x})$, i.e. $\mathbb{E}[Z(\mathbf{x})] = \mu(\mathbf{x})$, it follows

$$2\gamma(\mathbf{h}) := Var(Z(\mathbf{x}+\mathbf{h}) - Z(\mathbf{x})) = \mathbb{E}\left[(Z(\mathbf{x}+\mathbf{h}) - Z(\mathbf{x}))^2\right] - (\mathbb{E}[Z(\mathbf{x}+\mathbf{h}) - Z(\mathbf{x})])^2$$

$$= \mathbb{E}\left[(Z(\mathbf{x}+\mathbf{h}) - Z(\mathbf{x}))^2\right] - \underbrace{(\mu(\mathbf{x}+\mathbf{h}) - \mu(\mathbf{x}))^2}_{\neq 0} \neq \mathbb{E}\left[(Z(\mathbf{x}+\mathbf{h}) - Z(\mathbf{x}))^2\right].$$

In the following, after defining the theoretical variogram, we want to present some useful, important and characterizing properties of the variogram $\gamma(\mathbf{h})$. First of all, one very practical feature is represented in the equivalence with a covariance function $C(\mathbf{h})$ of $Z(\mathbf{x})$. But just one step before, we need to strengthen the stationarity assumption on $Z(\mathbf{x})$ to guarantee the existence of the covariance, since intrinsic stationarity does only imply the existence of the variogram, but not a finite covariance in general. Therefore, in accordance with Matheron (1971, p. 52) and Wackernagel (2003, p. 52), we assume *second-order stationarity* of the random function $Z(\mathbf{x})$, $\mathbf{x} \in D$, which is also called *weak stationarity* or even *hypothesis of stationarity of the first two moments*, i.e. mean μ and covariance $C(\mathbf{h})$:

Definition 4.3 (Second-order stationarity)
$Z(\mathbf{x})$ is *second-order stationary* with mean μ and covariance function $C(\mathbf{h})$ if

(i) the mean $\mu \in \mathbb{R}$ of $Z(\mathbf{x})$ is constant, i.e. $\mathbb{E}[Z(\mathbf{x})] = \mu(\mathbf{x}) = \mu \ \forall \mathbf{x} \in D$ and

(ii) the covariance function $C(\mathbf{h})$ only depends on the separating vector $\mathbf{h}$ of the two inserted locations, i.e. $\forall \mathbf{x}, \mathbf{x}+\mathbf{h} \in D$:

$$C(\mathbf{h}) := Cov(Z(\mathbf{x}), Z(\mathbf{x}+\mathbf{h})) = \mathbb{E}\left[Z(\mathbf{x})Z(\mathbf{x}+\mathbf{h})\right] - \mathbb{E}\left[Z(\mathbf{x})\right]\mathbb{E}\left[Z(\mathbf{x}+\mathbf{h})\right]$$

$$= \mathbb{E}[Z(\mathbf{x})Z(\mathbf{x}+\mathbf{h})] - \mu^2.$$

This implies that the covariance function $C(\mathbf{h})$ is bounded (Matheron 1971, p. 53) with

$$|C(\mathbf{h})| \leq C(\mathbf{0}) = Var(Z(\mathbf{x})) \ \forall \mathbf{x} \in D,$$

since $0 \leq Var(Z(\mathbf{x}+\mathbf{h}) \pm Z(\mathbf{x})) = Var(Z(\mathbf{x}+\mathbf{h})) \pm 2Cov(Z(\mathbf{x}+\mathbf{h}), Z(\mathbf{x})) + Var(Z(\mathbf{x})) = 2C(\mathbf{0}) \pm 2C(\mathbf{h})$ (see Cressie 1993, p. 67).

Remark 4.4

(i) In many textbooks, e.g. in Cressie (1993, p. 53), $C(\mathbf{h})$ is often called *covariogram*. Webster and Oliver (2007, p. 53) even called it *autocovariance function* because it displays the covariance of $Z(\mathbf{x})$ with itself and hence describes the relation between the values of $Z(\mathbf{x})$ for changing lag $\mathbf{h}$.

(ii) Further note that the intrinsic stationarity of $Z(\mathbf{x})$ in Definition 4.1 is more general than the second-order stationarity in Definition 4.3, since any second-order stationary random process is automatically intrinsically stationary, i.e. the set of all second-order stationary random functions is a subset of the set of all intrinsically stationary functions. But in general, the reversal is not true. Hence, a variogram

function, which requires e.g. only intrinsic stationarity, could exist even if there is no covariance function, wich requires e.g. the stronger assumption of second-order stationarity (cf. Wackernagel 2003, p. 52; Cressie 1993, p. 67). An example of a random process being intrinsically, but not second-order stationary can be found in Haskard (2007, pp. 8-9), where a "discrete Brownian motion", a symmetric random walk on $\mathbb{Z}$ starting at 0, is presented.
Journel and Huijbregts (1978, p. 12) also noticed that the intrinsic hypothesis of $Z(\mathbf{x})$ is just the second-order stationarity of its increments $Z(\mathbf{x}+\mathbf{h}) - Z(\mathbf{x})$.

We want to go further and present an important proposition, which can be found in nearly all books about variogram functions, for instance in Matheron (1971, p. 53) and Wackernagel (2003, p. 52), where it is often called equivalence of variogram and covariance function:

Proposition 4.5 (Equivalence of variogram and covariance function)

(i) If $Z(\mathbf{x})$ is second-order stationary, i.e. there exists a covariance function $C(\mathbf{h})$ of $Z(\mathbf{x})$, then a variogram function $\gamma(\mathbf{h})$ can be deduced from $C(\mathbf{h})$ according to the formula
$$\gamma(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h}).$$

(ii) If $Z(\mathbf{x})$ is intrinsically stationary with a bounded variogram $\gamma(\mathbf{h})$, i.e. there is a finite value $\gamma(\infty) := \lim_{|\mathbf{h}|\to\infty} \gamma(\mathbf{h}) < \infty$, which denotes the lowest upper bound of an increasing variogram function, then a covariance function $C(\mathbf{h})$ can be specified as
$$C(\mathbf{h}) = \gamma(\infty) - \gamma(\mathbf{h}).$$

(iii) For second-order stationary processes $Z(\mathbf{x})$, both properties (i) and (ii) hold, and the variogram and the covariogram are said to be equivalent.

Proof:
Kitanidis (1997, p. 52) stated the proof of this proposition:

(i) Since the second moment $\mathbb{E}\left[Z(\mathbf{x})^2\right] = C(\mathbf{0}) + \mu^2$ of $Z(\mathbf{x})$ is constant and hence independent of $\mathbf{x}$ for all $\mathbf{x}$ in D, it follows for all $\mathbf{x}, \mathbf{x}+\mathbf{h} \in D$:
$$\gamma(\mathbf{h}) = \frac{1}{2}\mathbb{E}\left[(Z(\mathbf{x}+\mathbf{h}) - Z(\mathbf{x}))^2\right] = \frac{1}{2}\mathbb{E}\left[Z(\mathbf{x}+\mathbf{h})^2\right] - \mathbb{E}[Z(\mathbf{x}+\mathbf{h})Z(\mathbf{x})] + \frac{1}{2}\mathbb{E}\left[Z(\mathbf{x})^2\right]$$
$$= \mathbb{E}\left[Z(\mathbf{x})^2\right] - \mathbb{E}[Z(\mathbf{x}+\mathbf{h})Z(\mathbf{x})] = \left(\mathbb{E}\left[Z(\mathbf{x})^2\right] - \mu^2\right) - \left(\mathbb{E}[Z(\mathbf{x}+\mathbf{h})Z(\mathbf{x})] - \mu^2\right)$$
$$= C(\mathbf{0}) - C(\mathbf{h}) < \infty.$$

(ii) If the variogram is bounded, then similar to (i), we can write the covariance function as
$$C(\mathbf{h}) = C(\mathbf{0}) - \gamma(\mathbf{h}) = \gamma(\infty) - \gamma(\mathbf{h}) < \infty.$$

(iii) The last part simply follows from (i) and (ii), since second-order stationarity of $Z(\mathbf{x})$ infers intrinsic stationarity and the existence of a bounded variogram.

□

Remark 4.6

(i) The proposition shows the equivalence of the variogram with its corresponding covariogram function in the case of a bounded variogram, e.g. if the underlying random function $Z(\mathbf{x})$ is second-order stationary. In general, the reverse statement (ii) in Proposition 4.5 is not true because unbounded variogram functions do not have corresponding covariance functions in general (see Wackernagel 2003, p. 52; Remark 4.4 (ii)).

(ii) The proposition also implies that a graph of the semivariogram $\gamma(\mathbf{h})$ plotted versus the absolute value, i.e. the Euclidean norm $|\mathbf{h}|$, of the lag $\mathbf{h}$, is simply the mirror image of the corresponding covariance function $C(\mathbf{h})$ about a line parallel to the $|\mathbf{h}|$-coordinate (Webster and Oliver 2007, p. 55).

(iii) Cressie (1993, p. 67) also stated that if $Z(\mathbf{x})$ is second-order stationary and if $C(\mathbf{h}) \to 0$ as $|\mathbf{h}| \to \infty$, then $\gamma(\mathbf{h})$ converges to $C(\mathbf{0})$, i.e. $\gamma(\mathbf{h}) \to C(\mathbf{0})$ as $|\mathbf{h}| \to \infty$ due to the stationarity criterion. The value $C(\mathbf{0})$, which is equal to the variance of $Z(\mathbf{x})$, is called the sill (see Definition 4.9).

Afterwards we want to show some more basic properties of the theoretical variogram, which are provided in the next proposition and are stated by Matheron (1971, pp. 54–56), but can also be found exemplarily in Wackernagel (2003, pp. 51–55). These properties will restrict the choice of the underlying variogram in the estimation later:

Proposition 4.7 (Properties of the variogram function)
Let $Z(\mathbf{x})$ be intrinsically stationary. The variogram function $\gamma(\mathbf{h})$ satisfies the following five conditions:

(i) $\gamma(\mathbf{0}) = 0$

(ii) $\gamma(\mathbf{h}) \geq 0$

(iii) $\gamma(-\mathbf{h}) = \gamma(\mathbf{h})$

(iv) The variogram grows slower than $|\mathbf{h}|^2$ as $|\mathbf{h}| \to \infty$, i.e.

$$\lim_{|\mathbf{h}| \to \infty} \frac{\gamma(\mathbf{h})}{|\mathbf{h}|^2} = 0.$$

(v) The variogram is a *conditionally negative semidefinite* function, i.e. for any finite sequence of points $(\mathbf{x}_i)_{i=1,\dots,n}$ and for any finite sequence of real numbers $(\omega_i)_{i=1,\dots,n}$ such that $\sum_{i=1}^{n} \omega_i = 0$, it holds

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \omega_i \omega_j \gamma(\mathbf{x}_i - \mathbf{x}_j) \leq 0.$$

Proof:
The parts of the proof of this proposition are presented in most books dealing with variogram functions, for instance in Wackernagel (2003, pp. 51–55) or Matheron (1971, pp. 54–56).

(i) The value at the origin of the variogram is zero by definition, since the variance of a constant equals zero

$$\gamma(\mathbf{0}) = \frac{1}{2} Var\left(Z(\mathbf{x}+\mathbf{0}) - Z(\mathbf{x})\right) = 0.$$

(ii) The variogram is nonnegative, since the variance of some random variables cannot take negative values.

(iii) Additionally, the theoretical variogram is also an even function, i.e.

$$\gamma(-\mathbf{h}) = \frac{1}{2} Var\left(Z(\mathbf{x}-\mathbf{h}) - Z(\mathbf{x})\right) = \frac{1}{2} Var\left(Z(\mathbf{x}) - Z(\mathbf{x}+\mathbf{h})\right) = \gamma(\mathbf{h}),$$

due to the invariance for any translation of $\mathbf{h}$ in the domain D.

(iv) Further, we proof the behavior of $\gamma(\mathbf{h})$ at infinity by contradiction. Therefore we assume

$$\lim_{|\mathbf{h}|\to\infty} \frac{\gamma(\mathbf{h})}{|\mathbf{h}|^2} \neq 0,$$

i.e. the variogram grows at least as fast as the square of the lag, and it follows:

$$0 \neq \lim_{|\mathbf{h}|\to\infty} \frac{\gamma(\mathbf{h})}{|\mathbf{h}|^2} = \lim_{|\mathbf{h}|\to\infty} \frac{1}{2} \frac{\mathbb{E}\left[\left(Z(\mathbf{x}+\mathbf{h}) - Z(\mathbf{x})\right)^2\right]}{|\mathbf{h}|^2}$$

$$= \lim_{|\mathbf{h}|\to\infty} \frac{1}{2} \mathbb{E}\left[\left(\frac{Z(\mathbf{x}+\mathbf{h}) - Z(\mathbf{x})}{|\mathbf{h}|}\right)^2\right] \geq \lim_{|\mathbf{h}|\to\infty} \frac{1}{2} \left(\mathbb{E}\left[\frac{Z(\mathbf{x}+\mathbf{h}) - Z(\mathbf{x})}{|\mathbf{h}|}\right]\right)^2 \geq 0,$$

which follows by applying Jensen's inequality for the convex function $\varphi : \mathbb{R} \to \mathbb{R}$, $y \mapsto \varphi(y) := y^2$, or simply by the formula $\mathbb{E}\left[X^2\right] \geq \left(\mathbb{E}[X]\right)^2$ for any random variable X.

$$\Rightarrow \lim_{|\mathbf{h}|\to\infty} \left(\mathbb{E}\left[\frac{Z(\mathbf{x}+\mathbf{h}) - Z(\mathbf{x})}{|\mathbf{h}|}\right]\right)^2 > 0.$$

Hence, we obtain

$$\lim_{|\mathbf{h}|\to\infty} \mathbb{E}\left[\frac{Z(\mathbf{x}+\mathbf{h}) - Z(\mathbf{x})}{|\mathbf{h}|}\right] \neq 0,$$

which implies

$$\lim_{|\mathbf{h}|\to\infty} \mu(\mathbf{h}) = \lim_{|\mathbf{h}|\to\infty} \mathbb{E}\left[Z(\mathbf{x}+\mathbf{h}) - Z(\mathbf{x})\right] \neq 0.$$

This is a contradiction to the assumption that the drift $\mu(\mathbf{h})$ equals zero.

(v) Finally, in most books, this last part of the proposition is proved assuming second-order stationarity of $Z(\mathbf{x})$, i.e. the existence of a covariance function $C(\mathbf{h})$ (e.g. see Matheron 1971). But Cressie (1993, pp. 86–87) gives a much nicer proof assuming only intrinsic stationarity, i.e. weaker assumptions. For this reason, we present the proof by Cressie (1993) at this point:

Let $\boldsymbol{\omega} := (\omega_1, \ldots, \omega_n)^T \in \mathbb{R}^n$ be given, such that $\sum_{i=1}^n \omega_i = 0$. Since

$$-\frac{1}{2}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}\omega_i\omega_j \left(Z(\mathbf{x}_i) - Z(\mathbf{x}_j)\right)^2\right]$$

$$= -\frac{1}{2}\left[\sum_{i=1}^{n}\omega_i \left(Z(\mathbf{x}_i)\right)^2 \underbrace{\sum_{j=1}^{n}\omega_j}_{=0} - 2\sum_{i=1}^{n}\sum_{j=1}^{n}\omega_i\omega_j Z(\mathbf{x}_i)Z(\mathbf{x}_j) + \sum_{j=1}^{n}\omega_j \left(Z(\mathbf{x}_j)\right)^2 \underbrace{\sum_{i=1}^{n}\omega_i}_{=0}\right]$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n}\omega_i\omega_j Z(\mathbf{x}_i)Z(\mathbf{x}_j) = \left(\sum_{i=1}^{n}\omega_i Z(\mathbf{x}_i)\right)^2,$$

it follows by taking expectations:

$$\sum_{i=1}^{n}\sum_{i=1}^{n}\omega_i\omega_j\gamma(\mathbf{x}_i - \mathbf{x}_j) = \sum_{i=1}^{n}\sum_{j=1}^{n}\omega_i\omega_j \mathbb{E}\left[\frac{\left(Z(\mathbf{x}_i) - Z(\mathbf{x}_j)\right)^2}{2}\right]$$

$$= \mathbb{E}\left[\frac{1}{2}(\sum_{i=1}^{n}\sum_{j=1}^{n}\omega_i\omega_j \left(Z(\mathbf{x}_i) - Z(\mathbf{x}_j)\right)^2)\right] = \underbrace{-\mathbb{E}\left[\left(\sum_{i=1}^{n}\omega_i Z(\mathbf{x}_i)\right)^2\right]}_{\geq 0} \leq 0.$$

□

4.2 Variogram cloud

After the theoretical part about the variogram, we want to show the way how an underlying variogram function can be deduced from our data points $\mathbf{x}_i$ in the geographical region D with observations $z(\mathbf{x}_i)$, $i = 1, \ldots, n$, to get a measure for the dependence in the spatial space. As in practice, unfortunately, the "real", "truth" underlying variogram behind is not known, we have to estimate it.

As a first step, Wackernagel (2003, p. 45) as well as Webster and Oliver (2007, p. 65) introduced a measure for dissimilarity $\gamma^*_{i,j}$ of two sample points $\mathbf{x}_i$ and $\mathbf{x}_j$, which can be computed by the half of the squared difference between the observed values $z(\mathbf{x}_i)$ and $z(\mathbf{x}_j)$ at these points, i.e.

$$\gamma^*_{i,j} := \frac{\left(z(\mathbf{x}_i) - z(\mathbf{x}_j)\right)^2}{2}.$$

Further, Wackernagel (2003) supposed the dissimilarity γ^* to depend only on the separating vector $\mathbf{h}$ of the sample points $\mathbf{x}_i$ and $\mathbf{x}_i + \mathbf{h}$, then

$$\gamma^*(\mathbf{h}) := \frac{\left(z(\mathbf{x}_i + \mathbf{h}) - z(\mathbf{x}_i)\right)^2}{2}.$$

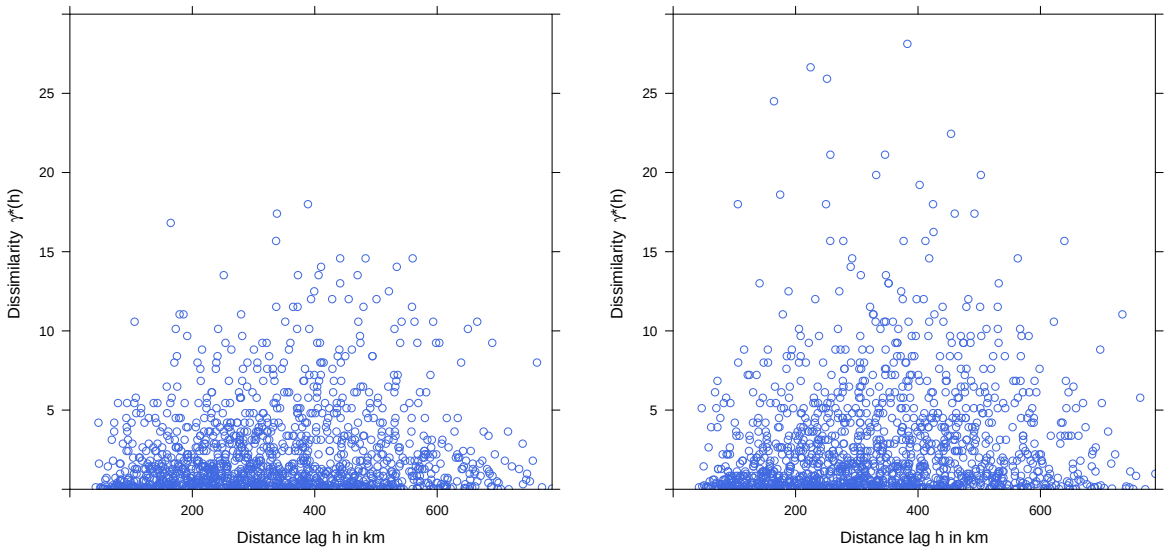
Obviously, this dissimilarity is symmetric with respect to $\mathbf{h}$ as a squared function, i.e. $\gamma^*(\mathbf{h}) = \gamma^*(-\mathbf{h})$.

For graphical representation, the resulting dissimilarities $\gamma^*(\mathbf{h})$ are plotted against the Euclidean distances $|\mathbf{h}|$ of the spatial separation vectors $\mathbf{h}$. This plot, or scatter diagram, of the dissimilarities against the lag distances, which takes any of the $\frac{n(n-1)}{2}$ pairs of samples into account, is called the *variogram cloud* by Wackernagel (2003, p. 46). It contains the information about the spatial structure of the sample and gives a first idea of the relationship between two points in D. Therefore, Wackernagel (2003) described the variogram cloud itself as a "powerful tool" for the analysis of spatial data and also Cressie (1993, pp. 40–41) characterized it as a "useful diagnostic tool".

Note that in most cases, the dissimilarity function $\gamma^*(\mathbf{h})$ is increasing as near sample points tend to have more similar values (Wackernagel 2003, p. 46). Below, we show examplarily the first lines of the values of the variogram cloud given the mean temperature data of 2010/11/28 and 2012/06/09, as it occurs using the function *variogram()* in the *R* package *gstat*. An illustrative example of the plotted variogram cloud of the data of 2010/11/28 is given on Figure 4.1a), while Figure 4.1b) illustrates the variogram cloud of 2012/06/09.

```
> #Create gstat objects:
> g1<-gstat(g=NULL,id="temp1",formula=temp1~1,locations=~longkm+latkm,
+ data=data1)

data:
temp1 : formula = temp1~1 ; data dim = 54 x 1
~longkm + latkm

> g2<-gstat(g=NULL,id="temp2",formula=temp2~1,locations=~longkm+latkm,
+ data=data2)

data:
temp2 : formula = temp2~1 ; data dim = 54 x 1
~longkm + latkm

> #Variogram cloud:
> vcloud1<-variogram(object=g1,cutoff=Inf,cloud=TRUE)
> vcloud2<-variogram(object=g2,cutoff=Inf,cloud=TRUE)

> head(vcloud1) #2010/11/28

       dist gamma dir.hor dir.ver    id left right
1 285.96629 0.605       0       0 temp1    2     1
2 453.94377 3.380       0       0 temp1    3     1
3 306.54582 1.125       0       0 temp1    3     2
4  56.11389 0.005       0       0 temp1    4     1
5 233.74657 0.500       0       0 temp1    4     2
6 402.62469 3.125       0       0 temp1    4     3
```

```
> head(vcloud2) #2012/06/09

       dist  gamma dir.hor dir.ver    id left right
1 285.96629  1.125       0       0 temp2    2     1
2 453.94377 22.445       0       0 temp2    3     1
3 306.54582 13.520       0       0 temp2    3     2
4  56.11389  0.125       0       0 temp2    4     1
5 233.74657  0.500       0       0 temp2    4     2
6 402.62469 19.220       0       0 temp2    4     3
```



(a) 2010/11/28 (b) 2012/06/09

Figure 4.1: Variogram clouds of the temperature data of 2010/11/28 and 2012/06/09 in Germany

4.3 The experimental variogram

Subsequently, since there could exist more than only one dissimilarity value for some distance lags $\mathbf{h}$, and since we will always have only finitely many sample points in practice and hence most lags $\mathbf{h}$ will be without any observation and thus still without dissimilarity values $\gamma^*(\mathbf{h})$, we must go on and find a solution to these two problems (Webster and Oliver 2007, pp. 77–79). Following Matheron (1962), we define the *classical estimator*, or also known as *method-of-moment estimator*, as

$$\gamma^*(\mathbf{h}) := \frac{1}{2\,|N(\mathbf{h})|} \sum_{N(\mathbf{h})} \left(z(\mathbf{x}_i) - z(\mathbf{x}_j)\right)^2,$$

with $N(\mathbf{h}) = \{(\mathbf{x}_i, \mathbf{x}_j) : \ \mathbf{x}_i - \mathbf{x}_j = \mathbf{h} \text{ for } i, j = 1, \ldots, n\}$ the set of all pairs of points with lag $\mathbf{h}$ and $|N(\mathbf{h})|$ the number of pairs in $N(\mathbf{h})$.

Cressie (1993) noticed that the symmetry property remains, since $\gamma^*(\mathbf{h}) = \gamma^*(-\mathbf{h})$, although $N(\mathbf{h}) \neq N(-\mathbf{h})$. Further advantages of this estimator are its unbiasedness and that it is not necessary to estimate the mean μ of $Z(\mathbf{x})$. But unfortunately, it is also sensitive to outliers due to the square of the differences (Webster and Oliver 2007, p. 113; Cressie 1993, pp. 40, 69). However, the problem of most distances being without a value still remains.

For this reason, Wackernagel (2003, p. 47) grouped the separation vectors into K vector classes $H_k,\ k = 1, \ldots, K$ with $K \in \mathbb{N}$, i.e. lag intervals, such that the union $\bigcup_{k=1}^{K} H_k$ covers all linking vectors $\mathbf{h}$ up to the maximum distance $\max_{i,j=1,\ldots,n} |\mathbf{x}_i - \mathbf{x}_j|$ in the sample.

Therefore, following Wackernagel (2003), we can determine the average dissimilarity $\gamma^*(H_k)$ corresponding to the vector class H_k, and hence we get an estimate of the dissimilarity value for all lags, by computing the average of dissimilarities $\gamma^*(\mathbf{h})$ for all point pairs with linking vector $\mathbf{h}$ belonging to vector class H_k, such that

$$\gamma^*(H_k) := \frac{1}{2\,|N(H_k)|} \sum_{N(H_k)} \left(z(\mathbf{x}_i) - z(\mathbf{x}_j)\right)^2,\ k \in \mathbb{N},$$

where $N(H_k) = \{(\mathbf{x}_i, \mathbf{x}_j) :\ \mathbf{x}_i - \mathbf{x}_j \in H_k \text{ for } i, j = 1, \ldots, n\}$ denotes the set of all pairs of points with separation vector in H_k and $|N(H_k)|$ the number of distinct elements in $N(H_k)$.

These average dissimilarities $\gamma^*(H_k)$ of the vector classes H_k form the *experimental variogram* (Wackernagel 2003, p. 47), which is in literature often called *empirical*, *estimated*, *sample variogram* or even *semivariance*, too (Webster and Oliver 2007, p. 60).

Note that the resulting empirical variogram strongly depends on the choice of the vector classes H_k and that this explicit choice also depends on the underlying problem. In literature, there exist two common ways to define these classes, which are presented in Cressie (1993, pp. 60–64) and Haskard (2007, pp. 9–10, 16):

(i) The vector sets H_k only depend on the Euclidean distance $|\mathbf{h}| = |\mathbf{x} - \mathbf{y}|$ between the points $\mathbf{x}$ and $\mathbf{y}$ in D. Hence, the empirical variogram is also a function only of the Euclidean norm $|\mathbf{h}|$ of $\mathbf{h}$, i.e. $\gamma^*(\mathbf{h}) = \gamma_0^*(|\mathbf{h}|)$, and is called *isotropic*.

(ii) The second way is to take the direction of the lag $\mathbf{h}$ in addition to the distance into account, e.g. by dividing the interval of angle $[0, \pi)$ into $m \in \mathbb{N}$ intervals, e.g. for $m = 4$: $[0, \frac{\pi}{4})$, $[\frac{\pi}{4}, \frac{\pi}{2})$, $[\frac{\pi}{2}, \frac{3\pi}{4})$ and $[\frac{3\pi}{4}, \pi)$ in the situation of a two-dimensional domain D. In this case the sample variogram is also a function of both, the distance and the direction, i.e. angle between the two inserted points, and is called *anisotropic* or even *directional variogram*. Anisotropies appear for instance when the underlying process $Z(\mathbf{x})$ in the vertical direction is different from its behavior in the horizontal direction (cf. Webster and Oliver 2007, p. 59).

Remark 4.8

(i) Since the final variogram estimator is still sensitive to outliers (Webster and Oliver 2007, p. 113), Cressie (1993) presented a robust estimator together with Hawkins;

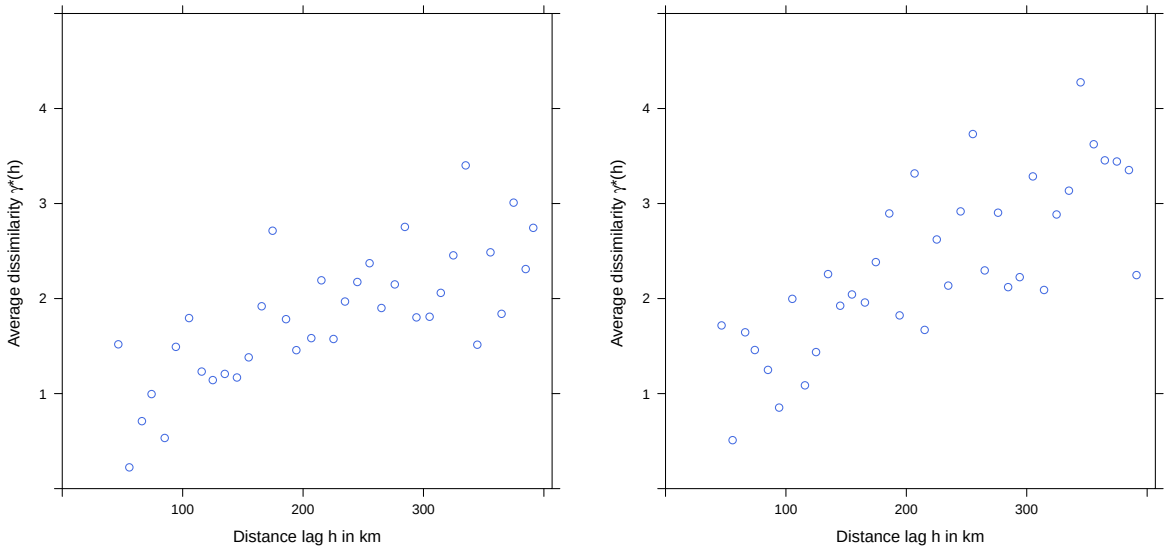
for more details see the original literature Cressie and Hawkins (1980) or Cressie (1993, p. 40, 74–76).

(ii) In practice, disjoint and equidistant vector classes are prefered and usually the experimental variogram is computed using lag vectors $\mathbf{h}$ of length up to a distance of half the diameter of the region, since for larger distances, the empirical variogram becomes more and more unreliable (Wackernagel 2003, p. 47; Haskard 2007, p. 19).

A concrete example for our data set of mean temperatures in Germany is given on Figure 4.2, where the isotropic experimental variogram is obtained from the variogram cloud by subdividing the distance vectors into the classes $H_k := \{\mathbf{h} \in \mathbb{R}^2 : \ (k-1) \cdot 10\text{km} \leq |\mathbf{h}| < k \cdot 10\text{km}\}$ with $K = 79$ the maximum distance in the sample divided by the width of each vector class, which is set to 10km, i.e. $H_1 = \{\mathbf{h} \in \mathbb{R}^2 : \ 0\text{km} \leq |\mathbf{h}| < 10\text{km}\}$, $H_2 = \{\mathbf{h} \in \mathbb{R}^2 : \ 10\text{km} \leq |\mathbf{h}| < 20\text{km}\}$ and so on. For instance, assuming a linear distance of 56 kilometers between $\mathbf{x}$ =Munich and $\mathbf{y}$ =Augsburg, their lag $\mathbf{h} = \mathbf{x} - \mathbf{y}$ would be an element of H_6.

Analogously to the variogram cloud, we show the first lines of the empirical variogram obtained from the function *variogram()* in the *R* package *gstat*. In the plot, the value of the empirical variogram for each H_k is printed as a blue dot at the averaged distance.

```
> #Empirical variogram:
> #Cutoff=half of maximum distance in variogram cloud
> #Vector classes have width of 10km
>
> vemp1<-variogram(object=g1,cutoff=max(vcloud1$dist)/2,width=10)
> vemp2<-variogram(object=g2,cutoff=max(vcloud2$dist)/2,width=10)

> head(vemp1) #2010/11/28

  np     dist     gamma dir.hor dir.ver    id
1  4 46.54447 1.5187500       0       0 temp1
2  8 55.71406 0.2237500       0       0 temp1
3  7 66.15195 0.7100000       0       0 temp1
4 18 74.15924 0.9950000       0       0 temp1
5 19 85.08439 0.5331579       0       0 temp1
6 20 94.36247 1.4920000       0       0 temp1

> head(vemp2) #2012/06/09

  np     dist    gamma dir.hor dir.ver    id
1  4 46.54447 1.717500       0       0 temp2
2  8 55.71406 0.510000       0       0 temp2
3  7 66.15195 1.645000       0       0 temp2
4 18 74.15924 1.458611       0       0 temp2
5 19 85.08439 1.249474       0       0 temp2
6 20 94.36247 0.852750       0       0 temp2
```



(a) 2010/11/28 (b) 2012/06/09

Figure 4.2: Empirical variograms of the temperature data of 2010/11/28 and 2012/06/09 in Germany

4.4 Fitting the experimental variogram

The experimental variogram $\gamma^*(\mathbf{h})$ provides a first estimate of the assumed underlying theoretical variogram $\gamma(\mathbf{h})$, which can be used to characterize the spatial structure and is needed for our future kriging methods. But at this point, Cressie (1993, pp. 89–90) argued that we cannot use this estimator directly, since we have to take care about several limitations on the variogram function, which are summarized in Proposition 4.7. In particular, Cressie (1993, p. 90) infered that the conditionally definiteness cannot be removed, since otherwise it can happen that the prediction variances in kriging will turn negative.

This implies that we have to fit a variogram function to the empirical variogram, i.e. replace it by a theoretical variogram. But unfortunately, by the same argument, not every arbitrary function is authorized to be a variogram function; a suitable, valid function is needed.

For instance, a first approach could be fitting a more or less arbitrary function to the empirical variogram by a least squares fit. But given the resulting function, checking all important conditions of the variogram would be not really efficient, since it can be very difficult to verify e.g. the conditionally definiteness or the growth condition of the variogram function. Therefore, to avoid this disadvantage and to save time, there exist a few valid *parametric variogram models* - parametric, since they depend on parameters - satisfying these conditions and which are often used in applied geostatistics (Webster and Oliver 2007, p. 82).

Hence, our idea is to search for a suitable variogram function within all of these valid parametric variogram models, i.e. we automatically guarantee that all important conditions of the variogram are satisfied, which provides the best fit to our data (see Cressie 1993, p. 90; Webster and Oliver 2007, p. 290).

In summary, for fitting a valid variogram function to the empirical variogram, we conclude that first the parameters of the parametric variogram model functions have to be estimated and then we can select the best fitting model (cf. Webster and Oliver 2007, pp. 101–102, 290).

For estimating these parameters, there exist a few methods. Cressie (1993, pp. 91–96) introduced estimation by ordinary or weighted least squares, which we will focus on, but alternatively also discussed maximum likelihood and restricted maximum likelihood (REML) estimation, minimum norm quadratic (MINQ) estimation and generalized-least-squares (GLS) fitting. In the case of a least squares fit, the function with the least sum of squares is chosen, since it seems to be the closest to our data.

Webster and Oliver (2007) also mentioned that the behavior of the fitted variogram at the origin is important, i.e. for very small distances, where the variogram can be differentiable, continuous but not differentiable or even discontinuous, and the behavior for large distances (beyond the range), where one has to decide whether the variogram is bounded or not. Further details and interpretation are presented in the books by Matheron (1971, 1989) and Wackernagel (2003, pp. 48–49, 115–117).

4.5 Parametric variogram models

As mentioned above, so-called parametric variogram models are used to fit the experimental variogram, since they provide the required properties. Therefore, before giving some examples of valid variogram model families, we want to introduce the three most common parameters *nugget*, *sill* and *range*, defined in accordance with Matheron (1962) and Cressie (1993, pp. 59, 67–68, 130–131):

Definition 4.9 (Variogram parameters)

(i) Nugget:
If the empirical variogram is discontinous at the origin, i.e. $\gamma(\mathbf{h}) \to c_0 > 0$ as $|\mathbf{h}| \to 0$, then the height of the jump c_0 is called the nugget, or nugget effect respectively, representing the value which could be caused by measurement error or some microscale variation.

(ii) Sill:
The value $\gamma(\infty) = \lim_{|\mathbf{h}|\to\infty} \gamma(\mathbf{h})$ is called the sill.

(iii) Range:
The distance at which the semivariogram $\gamma(\mathbf{h})$ exceeds the sill value for the first time is called the range.

An illustrative example for the parameters is given on Figure 4.3.

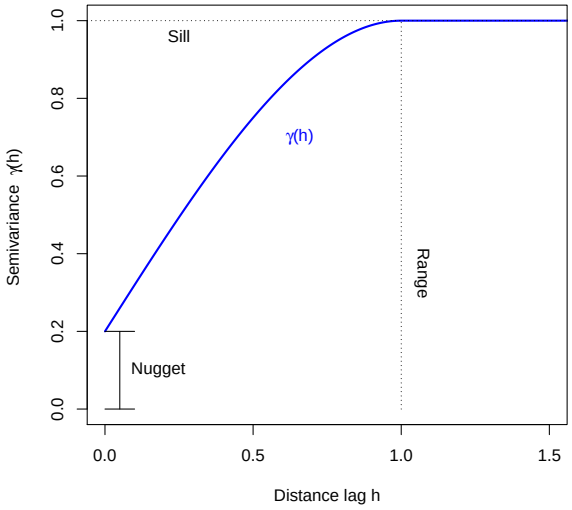


Figure 4.3: Variogram parameters nugget, sill and range

Remark 4.10

- Note that in practice, sometimes the range is defined as the distance at which the semivariogram achieves about 95% of its sill value, called *practical* or *effective range* (Wackernagel 2003, p. 57). This makes sense, since for some variogram functions, the sill can only be reached asymptotically and never in a finite distance. For instance see the Exponential or Gaussian model in Definition 4.11.

- For distances beyond the range, the corresponding random variables are said to be uncorrelated because its associated covariogram equals zero (in case of a bounded variogram) (Journel and Huijbregts 1978, p. 13).

Since in the experience, bounded variation is more common than unbounded variation (Webster and Oliver 2007, p. 84), we introduce the most common and most frequently used bounded, or also called stationary or transitional, isotropic and valid variogram model families, which can be found for instance in Webster and Oliver (2007, pp. 82–95) and Wackernagel (2003, pp. 57–61, 334–336).

These are the *Nugget-effect model*, the *Bounded linear model*, the *Spherical model*, the *Exponential model*, the *Gaussian model* and the *Matérn class*, defined in the sense of Webster and Oliver (2007, p. 94), which is a more general class containing variogram functions.

Definition 4.11 (Parametric variogram models)
Let $\gamma_{a,b}(\mathbf{h})$ denote the variogram function, $C_{a,b}(\mathbf{h})$ the corresponding covariance function with lag $\mathbf{h}$ and $a, b > 0$ the parameters of each model, where a represents the range parameter and b the sill value.

(i) Nugget-effect model:

$$\gamma_{a,b}^{nug}(\mathbf{h}) := \begin{cases} 0, \text{ if } |\mathbf{h}| = 0 \\ \\ b, \text{ otherwise,} \end{cases}$$

$$C_{a,b}^{nug}(\mathbf{h}) := \begin{cases} b, \text{ if } |\mathbf{h}| = 0 \\ \\ 0, \text{ otherwise.} \end{cases}$$

(ii) Bounded linear model:

$$\gamma_{a,b}^{lin}(\mathbf{h}) := \begin{cases} b\left(\frac{|\mathbf{h}|}{a}\right), \text{ if } 0 \leq |\mathbf{h}| \leq a \\ \\ b, \text{ otherwise,} \end{cases}$$

$$C_{a,b}^{lin}(\mathbf{h}) := \begin{cases} b\left(1 - \frac{|\mathbf{h}|}{a}\right), \text{ if } 0 \leq |\mathbf{h}| \leq a \\ \\ 0, \text{ otherwise.} \end{cases}$$

(iii) Spherical model:

$$\gamma_{a,b}^{sph}(\mathbf{h}) := \begin{cases} b\left(\frac{3}{2}\frac{|\mathbf{h}|}{a} - \frac{1}{2}\left(\frac{|\mathbf{h}|}{a}\right)^3\right), \text{ if } 0 \leq |\mathbf{h}| \leq a \\ \\ b, \text{ otherwise,} \end{cases}$$

$$C_{a,b}^{sph}(\mathbf{h}) := \begin{cases} b\left(1 - \frac{3}{2}\frac{|\mathbf{h}|}{a} + \frac{1}{2}\left(\frac{|\mathbf{h}|}{a}\right)^3\right), \text{ if } 0 \leq |\mathbf{h}| \leq a \\ \\ 0, \text{ otherwise.} \end{cases}$$

(iv) Exponential model:

$$\gamma_{a,b}^{exp}(\mathbf{h}) := b\left(1 - \exp\left(-\frac{|\mathbf{h}|}{a}\right)\right) \text{ for } |\mathbf{h}| \geq 0,$$

$$C_{a,b}^{exp}(\mathbf{h}) := b\exp\left(-\frac{|\mathbf{h}|}{a}\right) \text{ for } |\mathbf{h}| \geq 0.$$

(v) Gaussian model:

$$\gamma_{a,b}^{gau}(\mathbf{h}) := b\left(1 - \exp\left(-\frac{|\mathbf{h}|^2}{a^2}\right)\right) \text{ for } |\mathbf{h}| \geq 0,$$

$$C_{a,b}^{gau}(\mathbf{h}) := b\exp\left(-\frac{|\mathbf{h}|^2}{a^2}\right) \text{ for } |\mathbf{h}| \geq 0.$$

(vi) Matérn model class:

$$\gamma^{mat}_{a,b,\nu}(\mathbf{h}) := b\left[1 - \frac{1}{2^{\nu-1}\Gamma(\nu)}\left(\frac{|\mathbf{h}|}{a}\right)^{\nu} K_\nu\left(\frac{|\mathbf{h}|}{a}\right)\right] \text{ for } |\mathbf{h}| \geq 0,$$

$$C^{mat}_{a,b,\nu}(\mathbf{h}) := b\left[\frac{1}{2^{\nu-1}\Gamma(\nu)}\left(\frac{|\mathbf{h}|}{a}\right)^{\nu} K_\nu\left(\frac{|\mathbf{h}|}{a}\right)\right] \text{ for } |\mathbf{h}| \geq 0,$$

with *smoothness parameter* ν varying from 0 to ∞, gamma function $\Gamma(\cdot)$ and modified Bessel function $K_\nu(\cdot)$.

The equivalence of the variogram and the covariance function simply follows from Proposition 4.5 (Equivalence of variogram and covariance function) and the boundedness of the variogram of all models (Webster and Oliver 2007, pp. 84–95).

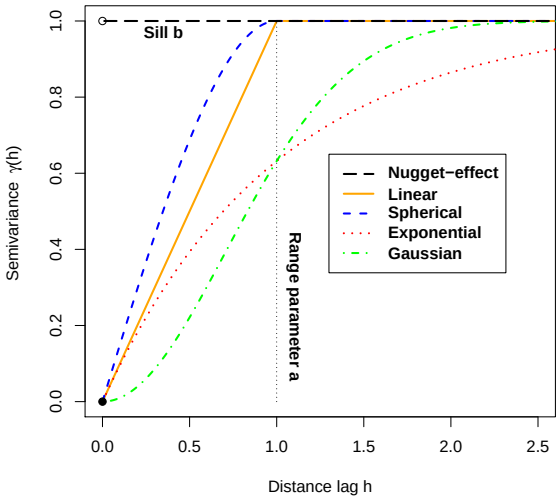
Wackernagel (2003, pp. 57–58) noted that in contrast to the Linear and the Spherical model, which reach the specified sill value b exactly at the finite range a, i.e. $\gamma^{lin}_{a,b}(a) = \gamma^{sph}_{a,b}(a) = b$, the Exponential and the Gaussian model approach the sill asymptotically for $|\mathbf{h}| \to \infty$, which has a non zero covariance as consequence. In this case, the practical range equals approximately $3a$ for the Exponential and $\sqrt{3}a$ for the Gaussian variogram model, i.e. $\gamma^{exp}_{a,b}(3a) \approx 0.95b$ and $\gamma^{gau}_{a,b}(\sqrt{3}a) \approx 0.95b$ (cf. Webster and Oliver 2007, pp. 88–93; Pebesma 2001, p. 38).

Figure 4.4 shows the variogram and the corresponding covariogram functions of the models (i)-(v) in Definition 4.11 for parameters $a = b = 1$. The range parameter a determines how fast the sill value, the variance of the process, is achieved; the smaller the value of a, the faster the sill b is reached, assuming b to be constant. Hence, the corresponding variogram function will increase faster. The sill b also decides about the shape of the variogram, since a higher value of b, with constant range parameter a, also infers a more increasing variogram function. Analogously to these considerations, the same applies to the corresponding covariance functions, which will decrease faster for increasing sill or decreasing range parameter. The influence of varying values of a and b on the variogram is shown on Figure 4.5.

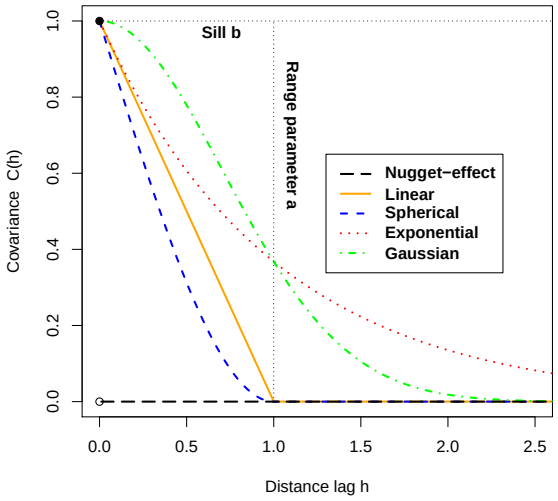
Figure 4.6 shows the variogram and covariogram functions of the Matérn class for different values of ν. The changing behavior of the variogram and covariance functions from rough ($\nu = 0.1$) to smooth ($\nu = 5.0$) for varying ν is a characterizing feature of the Matérn class. More details on the interpretation of ν are presented in Remark 4.12 below.

Remark 4.12

- Definition 4.11 (i)-(v) agrees with the definition of the parametric variogram models *Nug*, *Lin*, *Sph*, *Exp* and *Gau* in the *R* package *gstat* (Pebesma 2001, pp. 36–39).
- The definition of the Matérn model class, which is named after the work of Bertil Matérn by Stein (1999), see for instance Matérn (1960), coincides with the variogram model *Mat* in *R*. But there also exists another parameterization of this model, given by Stein (1999, pp. 48–51), which is also implemented in *R* via the model *Ste*.

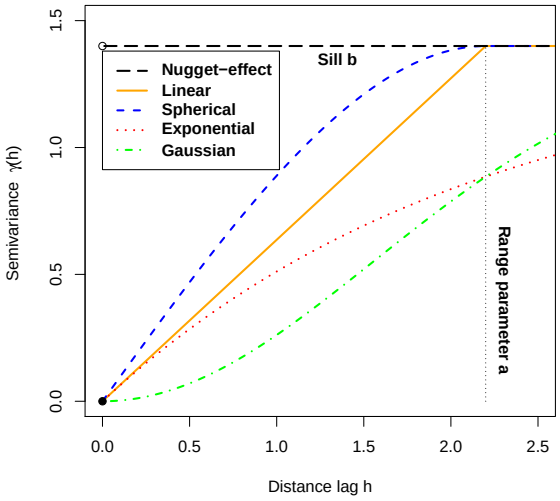


(a) Variogram functions

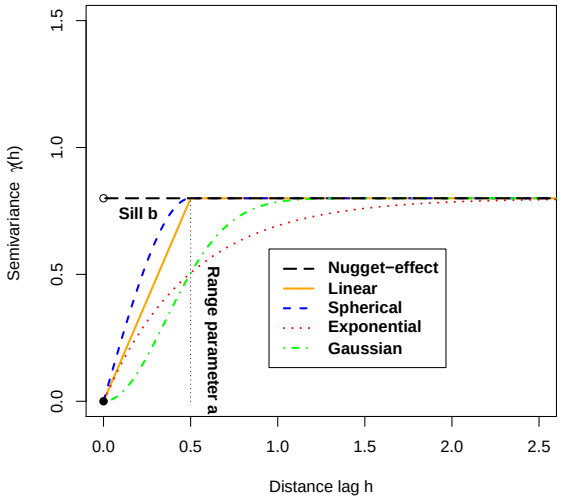


(b) Covariance functions

Figure 4.4: Variogram and covariance functions with range parameter $a = 1$ and sill $b = 1$



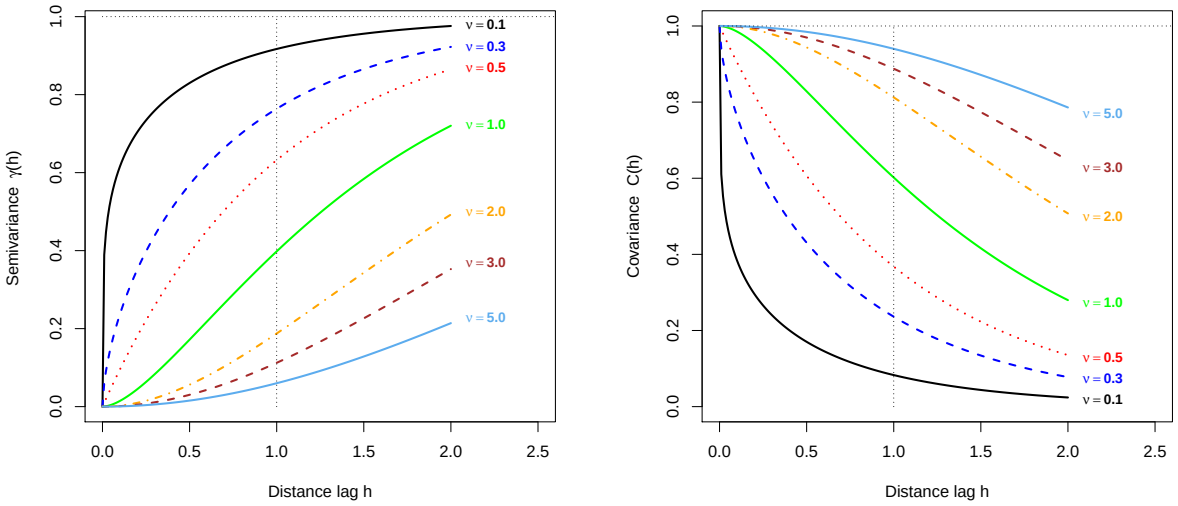
(a) Variogram functions with $a = 0.5$ and $b = 0.8$



(b) Variogram functions with $a = 2.2$ and $b = 1.4$

Figure 4.5: Variogram functions for varying range parameter a and sill b

The Matérn class is a generalization of several other variogram model functions, e.g. it includes the Exponential model for $\nu = 0.5$ and a variogram model called *Whittle's model* for $\nu = 1$. The parameter ν is called *smoothness parameter*, since it is the crucial value which decides about the smoothness of the variogram function. $\nu \approx 0$ is related to a very rough and the value $\nu = \infty$ to a very smooth behavior of the corresponding variogram function (Webster and Oliver 2007, p. 94). In the



(a) Matérn variogram functions

(b) Matérn covariance functions

Figure 4.6: Variogram and covariance functions of the Matérn class with range parameter $a = 1$, sill $b = 1$ and varying ν

cases where ν equals $m + \frac{1}{2}$ for a nonnegative integer m, i.e. if ν is half-integer, the Matérn variogram and covariogram functions can be expressed in a simpler form: an exponential times a polynomial of order m. The interested reader may look at Rasmussen and Williams (2006, pp. 84–85) for further details and examples.

In practice, in the case of many data, the variogram appears more complex. Webster and Oliver (2007, p. 95) argued that it is common to combine some "simple" variogram models from Definition 4.11 for achieving a better fit. We profit from the fact that the combination of some conditionally negative semidefinite functions is again a conditionally negative semidefinite function. The most common combination is a nugget component added to another model, e.g. Webster and Oliver (2007) stated the exponential variogram $\gamma_{a,b}^{exp}(\mathbf{h})$ including a nugget c_0, i.e. $\gamma_{a,b,c_0}^{exp}(\mathbf{h}) := c_0 + \gamma_{a,b}^{exp}(\mathbf{h})$. In this situation of the existence of a nugget component c_0, the sill becomes the sum of b and c_0, $b + c_0$, and the value b is called *partial sill* (Cressie 1988).

There are variations of the models and their variogram and covariance functions as definded above, too. For instance, as shown in Bohling (2005, pp. 15–16), in some cases there exist shifted models. E.g. the Exponential and the Gaussian model are shifted at $\mathbf{h}$ with a factor such that they have the effective range a.

In literature, there exist various other bounded isotropic models as the *circular* or *Whittle's model*, or even unbounded isotropic models as the *logarithmic model* or the *models in* $|\mathbf{h}|^\theta$ for $0 < \theta < 2$, too. These and other models can be found in Journel and Huijbregts (1978, pp. 161–170) or Webster and Oliver (2007, pp. 82–95), just to name a few references. Journel and Huijbregts (1978, pp. 175–183) also talked about models for

anisotropy.

For interpretation of each model in practical application, we refer to the books by Wackernagel (2003) in Chapter 8, Webster and Oliver (2007, pp. 84–95) and to Bohling (2005, pp. 15–17).

Finally, we conclude that for choosing the most suitable variogram model, given the empirical variogram, the parameters a and b of each model have to be estimated first, e.g. by an ordinary or by a weighted least squares method, and then the best fitting model with the lowest sum of squares is chosen.

In *R* this could be done by using the function *fit.variogram()* in the package *gstat*, which provides a least squares fit. It contains nonweighted as well as weighted least squares, which can be selected by the argument *fit.method*. For instance, *fit.method=6* gives nonweighted and *fit.method=1* weighted least squares with weights $N(H_k)$ (see Pebesma 2001, p. 42). For comparison one could apply the *R* function *nls()* for nonlinear least squares, which is part of the standard package *stats*, and uses a Gauss-Newton algorithm as default algorithm. In most cases, this ends in nearly the same estimated values $\hat{a}$ for the range and $\hat{b}$ for the sill. It also contains an argument *weight*, where the weights for weighted least squares can be set.

Note that the *gstat* function *variogram()* also includes a command *cressie*. If this attribute is set to *cressie=TRUE*, then Cressie's robust variogram estimator in Remark 4.8 (i) is used instead of the classical method-of-moments estimator.

Further note that *fit.method=6* infers nonweighted least squares, i.e. all vector classes get the same weight 1 independent of their number of elements. *fit.method=1* implies weighted least squares with weights *np* from the experimental variogram, which counts the number of sample pairs in each vector class.

For our data set we infer that in both cases, the Matérn model fitted with nonweighted least squares (*fit.method=6*) provides the "best" fit, since their residuals are the lowest. Below one can find an overview of the parametric variogram models with fitted parameters and their sum of squares, where Table 4.1 and Table 4.2 refer to 2010/11/28 and Table 4.3 and Table 4.4 to 2010/06/09.

	Linear	Spherical	Exponential	Gaussian	Matérn
Range	270.02	405.44	190.74	180.15	215.44
Nugget	0.55	0.50	0.08	0.72	0.00
Partial Sill	1.74	1.94	2.73	1.67	2.87
Sum of squares	7.35	7.12	7.12	7.16	7.12
ν	0.00	0.00	0.00	0.00	0.44

Table 4.1: Parameters from weighted least squares (fit.method=1) of the temperature data of 2010/11/28 in Germany

	Linear	Spherical	Exponential	Gaussian	Matérn
Range	273.05	419.79	260.59	194.28	102.62
Nugget	0.51	0.48	0.29	0.75	0.65
Partial Sill	1.80	2.00	2.85	1.68	2.03
Sum of squares	7.32	7.11	7.08	7.11	7.05
ν	0.00	0.00	0.00	0.00	1.49

Table 4.2: Parameters from ordinary least squares (fit.method=6) of the temperature data of 2010/11/28 in Germany

	Linear	Spherical	Exponential	Gaussian	Matérn
Range	264.76	537.18	445.73	237.07	2013.70
Nugget	0.53	0.69	0.52	1.11	0.01
Partial Sill	2.48	3.03	4.96	2.36	10.22
Sum of squares	11.46	10.81	10.85	10.73	10.92
ν	0.00	0.00	0.00	0.00	0.31

Table 4.3: Parameters from weighted least squares (fit.method=1) of the temperature data of 2012/06/09 in Germany

	Linear	Spherical	Exponential	Gaussian	Matérn
Range	278.59	434.63	345.76	216.07	72.23
Nugget	0.71	0.66	0.54	1.08	1.03
Partial Sill	2.30	2.61	4.07	2.20	2.44
Sum of squares	11.36	10.68	10.74	10.66	10.63
ν	0.00	0.00	0.00	0.00	2.98

Table 4.4: Parameters from ordinary least squares (fit.method=6) of the temperature data of 2012/06/09 in Germany

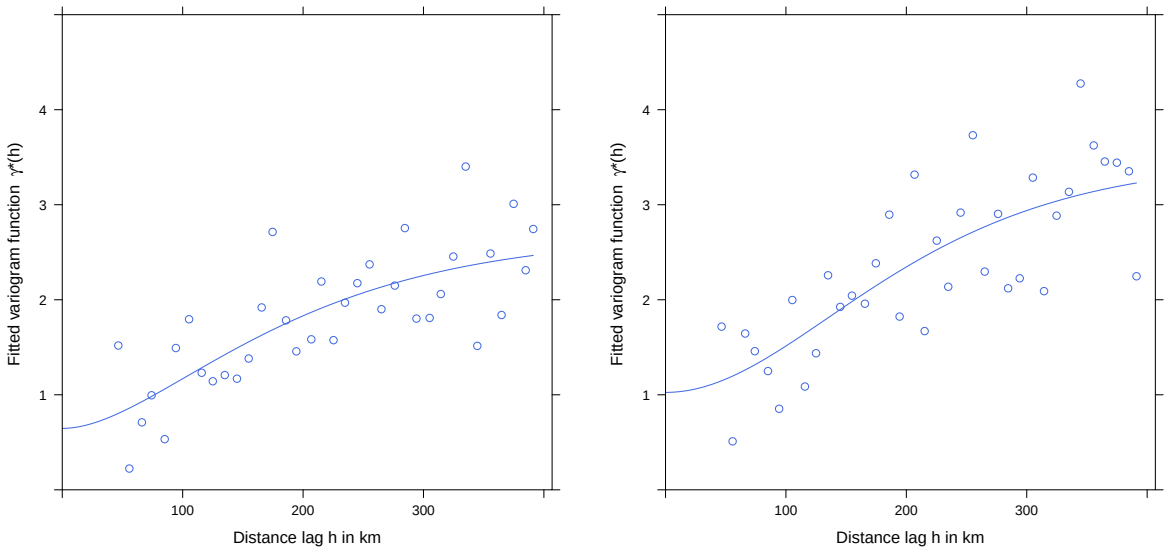
Hence, by taking the models with the lowest sum of squares, we obtain our valid variogram functions, which are indeed Matérn model functions in both cases:

```
> #Best fitting variogram model for 2010/11/28:
> vfit1<-fit.variogram(object=vemp1, model=vgm(psill=1, model="Mat",
+ range=100, nugget=1, kappa=1.49), fit.sills=TRUE, fit.ranges=TRUE,
+ fit.method=6)

  model     psill    range kappa
1   Nug 0.6464684   0.0000  0.00
2   Mat 2.0341866 102.6184  1.49

> attributes(vfit1)$SSErr #sum of squares

[1] 7.054155

> #Best fitting variogram model for 2012/06/09:
> vfit2<-fit.variogram(object=vemp2, model=vgm(psill=1, model="Mat",
+ range=100, nugget=1, kappa=2.98), fit.sills=TRUE, fit.ranges=TRUE,
+ fit.method=6)

  model    psill    range kappa
1   Nug 1.025047  0.00000  0.00
2   Mat 2.442959 72.23161  2.98

> attributes(vfit2)$SSErr #sum of squares

[1] 10.62915
```

Both estimated parametric variogram model functions are fitted to the empirical variogram, which can be seen on Figure 4.7.

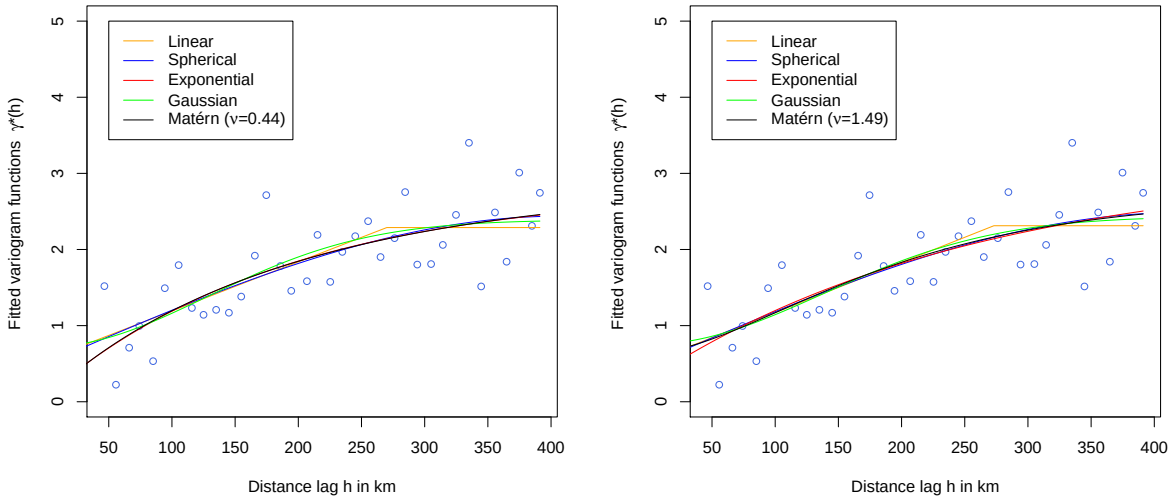
The other fitted variogram functions are printed below, for 2010/11/28 see Figure 4.8 and for 2012/06/09 see Figure 4.9. The left pictures contain the variogram functions corresponding to the estimation from weighted least squares (*fit.method=1*) and the right pictures to the estimation from nonweighted least squares (*fit.method=6*).

Finally, we can finish with our preparation for kriging, since we now have valid variogram functions, namely *vfit1* and *vfit2*, of our data set at hand.



(a) 2010/11/28, ν=1.49 (b) 2012/06/09, ν=2.98

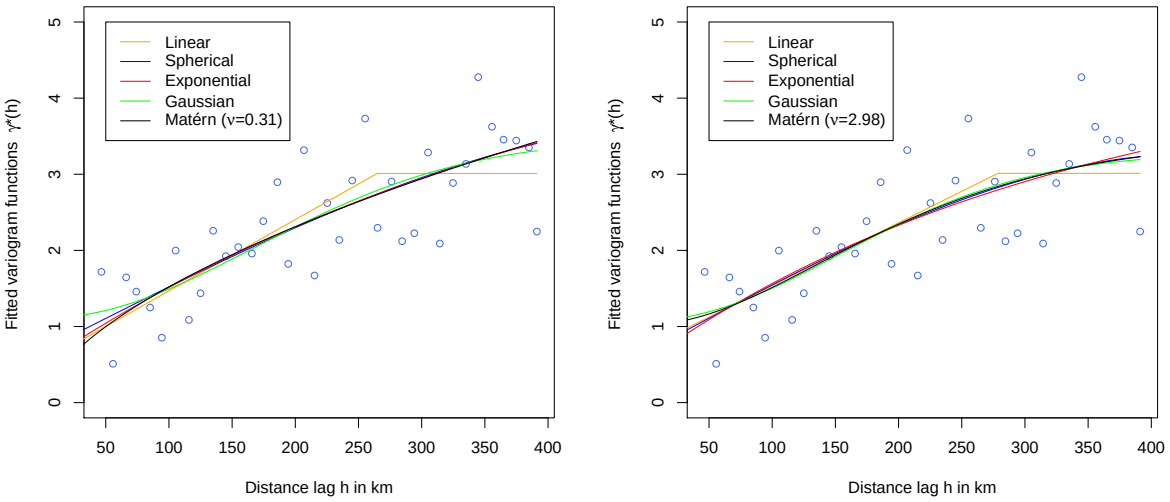
Figure 4.7: Matérn variogram functions with lowest sum of squares fitted to the empirical variogram



(a) Fitted variogram functions, where the parameters are estimated with weighted least squares (fit.method=1)

(b) Fitted variogram functions, where the parameters are estimated with nonweighted least squares (fit.method=6)

Figure 4.8: Fitted variogram models of 2010/11/28



(a) Fitted variogram functions, where the parameters are estimated with weighted least squares (fit.method=1)

(b) Fitted variogram functions, where the parameters are estimated with nonweighted least squares (fit.method=6)

Figure 4.9: Fitted variogram models of 2012/06/09

5 Kriging the Mean

The aim of this section about our first method of kriging prediction is to predict the value of the mean $\mu \in \mathbb{R}$ of the underlying random function $Z(\mathbf{x})$ from the sample points $\mathbf{x}_i,\ i = 1, \ldots, n$, where $\mathbf{x}$ is in the spatial domain $D \subseteq \mathbb{R}^d$ for $d \in \mathbb{N}$. The corresponding observed values $z(\mathbf{x}_i)$ are modeled again as realizations of the random variables $Z(\mathbf{x}_i)$ (Wackernagel 2003, p. 28).

This mean value can be computed with a weighted average of the observations, as for instance, Wackernagel (2003) explained a first intuitive approach using the arithmetic mean, where the weight of each point is the same and they all sum up to one. This approach can be used in the setting of uncorrelated samples. But since the data points are irregularly located in D, more general weights allow to take the knowledge of the spatial correlation of the samples into account, as usually closer points are more correlated than those, which are more distant (Wackernagel 2003, p. 28).

For this reason we want to introduce *kriging the mean*, which was originally stated by Georges Matheron. In the whole section we will follow closely the book by Wackernagel (2003, Chapter 4). The main results can also be found in Matheron (1971, pp. 118–119, 125–126) and Webster and Oliver (2007, pp. 181–183).

5.1 Model for Kriging the Mean

Kriging the mean as a geostatistical method for predicting the global mean μ of a geographical region D relies on the following model assumptions (see Wackernagel 2003, pp. 28–30):

Assumption 5.1 (Model for Kriging the Mean)

(i) The unknown mean $\mu \in \mathbb{R}$ exists at all points of the spatial domain D and is constant, i.e. $\mathbb{E}[Z(\mathbf{x})] = \mu\ \forall \mathbf{x} \in D$.

(ii) Additionally, the underlying random function $Z(\mathbf{x})$ is second-order stationary (cf. Definition 4.3, p. 14) with known covariance function $C(\mathbf{h}) := Cov\left(Z(\mathbf{x}), Z(\mathbf{x}+\mathbf{h})\right) = \mathbb{E}\left[Z(\mathbf{x})Z(\mathbf{x}+\mathbf{h})\right] - \mu^2$.

Under these assumptions, Matheron (1971, p. 126) and Wackernagel (2003, p. 28) defined the linear predictor for kriging the mean as follows:

Definition 5.2 (Predictor for Kriging the Mean)
The *predictor of kriging the mean* $M^*_{\boldsymbol{\omega}}$ of μ is the linear combination of the random function $Z(\mathbf{x})$ evaluated at each sample point $\mathbf{x}_i$

$$M^*_{\boldsymbol{\omega}} := \sum_{i=1}^{n} \omega_i Z(\mathbf{x}_i) = \boldsymbol{\omega}^T \mathbf{Z}$$

with unknown weights $\omega_i \in \mathbb{R}$ for $i = 1, \ldots, n$ and $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)^T \in \mathbb{R}^n$ the vector containing all weights.

5.2 Unbiasedness condition

In the first step, we want to avoid systematic bias in the prediction, i.e. the prediction error $M_{\boldsymbol{\omega}}^* - \mu$ has to be zero on average. Hence, according to Wackernagel (2003, p. 29), we have to force the weights ω_i to sum to one, i.e.

$$\sum_{i=1}^{n} \omega_i = 1 \;\Leftrightarrow\; \boldsymbol{\omega}^T \mathbb{1} = 1. \tag{5.1}$$

In fact, this yields the unbiasedness

$$\mathbb{E}\left[M_{\boldsymbol{\omega}}^* - \mu\right] = \mathbb{E}\left[\sum_{i=1}^{n} \omega_i Z(\mathbf{x}_i) - \mu\right] = \sum_{i=1}^{n} \omega_i \underbrace{\mathbb{E}[Z(\mathbf{x}_i)]}_{=\mu} - \mu = \mu\left(\underbrace{\sum_{i=1}^{n} \omega_i}_{=1} - 1\right) = 0.$$

5.3 Variance of the prediction error

Further, we want to compute the variance of the prediction error σ_E^2, since it provides information about the accuracy of our linear predictor $M_{\boldsymbol{\omega}}^*$. In accordance with Assumption 5.1 and following Wackernagel (2003, p. 30), we obtain for σ_E^2

$$\sigma_E^2 := Var(M_{\boldsymbol{\omega}}^* - \mu) = \underbrace{\mathbb{E}\left[(M_{\boldsymbol{\omega}}^* - \mu)^2\right]}_{=mse(M_{\boldsymbol{\omega}}^*)} - \left(\underbrace{\mathbb{E}[M_{\boldsymbol{\omega}}^* - \mu]}_{bias=0}\right)^2$$

$$\overset{(5.1)}{=} \mathbb{E}\left[(M_{\boldsymbol{\omega}}^*)^2\right] - 2\mu \underbrace{\mathbb{E}[M_{\boldsymbol{\omega}}^*]}_{=\mu} + \mu^2 = \mathbb{E}\left[\left(\sum_{i=1}^{n} \omega_i Z(\mathbf{x}_i)\right)^2\right] - \mu^2$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} \omega_i \omega_j \mathbb{E}\left[Z(\mathbf{x}_i) Z(\mathbf{x}_j)\right] - \mu^2 \underbrace{\sum_{i=1}^{n} \omega_i}_{=1} \underbrace{\sum_{j=1}^{n} \omega_j}_{=1}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} \omega_i \omega_j \underbrace{\left(\mathbb{E}[Z(\mathbf{x}_i) Z(\mathbf{x}_j)] - \mu^2\right)}_{=C(\mathbf{x}_i - \mathbf{x}_j)} = \sum_{i=1}^{n}\sum_{j=1}^{n} \omega_i \omega_j C(\mathbf{x}_i - \mathbf{x}_j).$$

Hence, we get the prediction variance

$$\sigma_E^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} \omega_i \omega_j C(\mathbf{x}_i - \mathbf{x}_j)$$

$$= \boldsymbol{\omega}^T \Sigma \boldsymbol{\omega} \geq 0,$$

which is nonnegative, since Σ is positive definite as the covariance matrix of the random vector $\mathbf{Z}$.

5.4 Minimal prediction variance

Our next aim is to derive the "optimal" weights $\omega_i,\ i = 1, \ldots, n$, which minimize the prediction variance σ_E^2 under the unbiasedness condition (5.1) of the predictor $M_{\boldsymbol{\omega}}^*$ (Wackernagel 2003, pp. 30–31). This leads us to the minimization problem

$$\text{minimum of } \boldsymbol{\omega}^T \Sigma \boldsymbol{\omega} \text{ subject to } \boldsymbol{\omega}^T \mathbb{1} = 1,$$

which is identical with minimizing a quadratic form subject to a linear constraint. We obtain the "best" weights $\boldsymbol{\omega}_{KM} = \left(\omega_1^{KM}, \ldots, \omega_n^{KM}\right)^T$ and thus achieve minimal prediction variance $\sigma_{KM}^2 := Var(M_{\boldsymbol{\omega}_{KM}}^* - \mu)$ for kriging the mean, the so-called *kriging variance*, by the next theorem:

Theorem 5.3 (Solution for Kriging the Mean)
The kriging weights $\boldsymbol{\omega}_{KM}$ and the minimal kriging variance σ_{KM}^2 are given by

$$\boldsymbol{\omega}_{KM} = \frac{\Sigma^{-1}\mathbb{1}}{\mathbb{1}^T\Sigma^{-1}\mathbb{1}},$$

$$\sigma_{KM}^2 = \frac{1}{\mathbb{1}^T\Sigma^{-1}\mathbb{1}}.$$

Proof:
Since we assumed Σ to be positive definite (see Assumption 2.9, p. 5), we can apply Theorem 2.15 (p. 8). Its setting coincides with our current setting for $A = \Sigma$, $X = \boldsymbol{\omega}$, $B = \mathbb{1}$ and $U = 1$, where a generalized inverse S^- of $B^T A^{-1} B = \mathbb{1}^T\Sigma^{-1}\mathbb{1} > 0$ is simply given by the reciprocal $\frac{1}{\mathbb{1}^T\Sigma^{-1}\mathbb{1}} > 0$, which is well-defined by virtue of Proposition 2.10 (p. 6), since Σ^{-1} is again positive definite.

Hence, we derive the minimal kriging variance

$$\sigma_{KM}^2 := \inf_{\mathbb{1}^T\boldsymbol{\omega}=1} \boldsymbol{\omega}^T\Sigma\boldsymbol{\omega} = \inf_{B^TX=U} X^TAX = U^TS^-U = S^- = \frac{1}{\mathbb{1}^T\Sigma^{-1}\mathbb{1}},$$

which is attained at

$$\boldsymbol{\omega}_{KM} := A^{-1}BS^-U = \Sigma^{-1}\mathbb{1}S^- = \frac{\Sigma^{-1}\mathbb{1}}{\mathbb{1}^T\Sigma^{-1}\mathbb{1}}.$$

□

Wackernagel (2003) also stated these results about the "optimal" weights ω_i^{KM} and the minimal kriging variance σ_{KM}^2, but with solving the above optimization problem with the method of Lagrange multipliers. We prefer to present this solution using Theorem 2.15, since we can conclude that the resulting variance is automatically minimal.

Remark 5.4
However, in applications inverting an $n \times n$ matrix with n large should be avoided and numerical more stable methods as e.g. a QR decomposition should be used for determining the kriging weights and the kriging variance. Hereby the target matrix $\Sigma \in \mathbb{R}^{n\times n}$ is

decomposed into the product of an orthogonal matrix $Q \in \mathbb{R}^{n\times n}$, i.e. $Q^T = Q^{-1}$, and an upper triangular matrix $R \in \mathbb{R}^{n\times n}$, such that $\Sigma = QR$.

In *R*, we can derive the QR decomposition of Σ using the commands *qr()*, *qr.Q()* and *qr.R()* in the standard package *base*, i.e.

```
> QRdecomposition<-qr(Sigma)
> Q<-qr.Q(QRdecomposition)
> R<-qr.R(QRdecomposition)

> all(Q%*%R==Sigma)

[1] TRUE
```

The inverse of Σ can also be directly computed by an QR decomposition with the command *qr.solve()*, which we will use in our prediction later:

```
> all(qr.solve(Sigma)%*%Sigma==Id)

[1] TRUE
```

5.5 Prediction for Kriging the Mean

In summary, after these calculations, we can write the *best linear unbiased predictor (BLUP)* $M^*_{\boldsymbol{\omega}_{KM}}$ of the mean μ for kriging the mean by applying Theorem 5.3, i.e. by inserting $\boldsymbol{\omega}_{KM}$, such that

$$M^*_{\boldsymbol{\omega}_{KM}} = \sum_{i=1}^{n} \omega_i^{KM} Z(\mathbf{x}_i) = \boldsymbol{\omega}_{KM}^T \mathbf{Z} = \left(\frac{\Sigma^{-1}\mathbb{1}}{\mathbb{1}^T\Sigma^{-1}\mathbb{1}} \right)^T \mathbf{Z} = \frac{\mathbb{1}^T\Sigma^{-1}\mathbf{Z}}{\mathbb{1}^T\Sigma^{-1}\mathbb{1}},$$

with kriging variance given in Theorem 5.3 and estimate $m^*_{\boldsymbol{\omega}_{KM}}$ of μ

$$m^*_{\boldsymbol{\omega}_{KM}} = \sum_{i=1}^{n} \omega_i^{KM} z(\mathbf{x}_i) = \boldsymbol{\omega}_{KM}^T \mathbf{z} = \left(\frac{\Sigma^{-1}\mathbb{1}}{\mathbb{1}^T\Sigma^{-1}\mathbb{1}} \right)^T \mathbf{z} = \frac{\mathbb{1}^T\Sigma^{-1}\mathbf{z}}{\mathbb{1}^T\Sigma^{-1}\mathbb{1}}.$$

Remark 5.5
Kriging the mean relies on the existence of a known covariance function $C(\mathbf{h})$ of $Z(\mathbf{x})$. Hence, for practical application one can either estimate the covariance function $C(\mathbf{h})$ itself (see Cressie 1993, Chapter 2), or the variogram function $\gamma(\mathbf{h})$ of $Z(\mathbf{x})$, which is often done in practice. In the case that the estimated variogram is bounded, one can obtain the corresponding covariance function according to Proposition 4.5 (Equivalence of variogram and covariance function, p. 15), but all equations and formulas in this section can also be transformed into terms of the variogram by replacing the covariance with the variogram terms.

5.6 Kriging the Mean in *R*

Finally, after the theoretical part about kriging the mean, we want to predict the mean value μ of our data set exemplarily for both dates 2010/11/28 and 2012/06/09 in *R*. Unfortunately, the procedure of kriging the mean is not directly implemented in the *R* package *gstat*. Hence, we first estimate the underlying variogram and fit it with a valid parametric variogram function by using *gstat* (see last section "The Variogram"). Afterwards, we apply Theorem 5.3 to determine the corresponding kriging weights, the kriging variance and finally the kriging estimate for μ.

Note that we have already fitted a variogram model function to the empirical variogram of our data in the last section. Thus we can use our previous results and can begin with our calculations in *R*.

```
> #Kriging the Mean:
>
> #1) Derive distance matrix from variogram cloud:
> #Gamma_dist(i,j) = |x_i-x_j|
> #Note that distances are always the same, only different values
> #for gamma (dissimilarities)
>
> Gamma_dist<-matrix(rep(0,l*l),nrow=l,ncol=l)
> k<-1
> for (i in 1:l){
+ for (j in 1:i){
+ if (i==j){
+ Gamma_dist[i,j]<-0
+ }
+ else{
+ Gamma_dist[i,j]<-vcloud1$dist[k]
+ Gamma_dist[j,i]<-vcloud1$dist[k]
+ k<-k+1
+ }
+ }
+ }

> #2.) Variogram matrix Gamma:
> #From last section "The Variogram": vfit1, vfit2
>
> Gamma1<-Gamma_dist
> Gamma2<-Gamma_dist
> for (i in 1:l){
+ for (j in 1:l){
+ Gamma1[i,j]<-matern(Gamma_dist[i,j],sum(vfit1$range),
+ sum(vfit1$psill),sum(vfit1$kappa))
```

```
+ Gamma2[i,j]<-matern(Gamma_dist[i,j],sum(vfit2$range),
+ sum(vfit2$psill),sum(vfit2$kappa))
+ }
+ }

> #3.) Covariance matrix Sigma:
>
> C01<-sum(vfit1$psill)
> C02<-sum(vfit2$psill)
> Sigma1<-C01*matrix(rep(1,length=l*l),nrow=l,ncol=l)-Gamma1
> Sigma2<-C02*matrix(rep(1,length=l*l),nrow=l,ncol=l)-Gamma2
> ones<-rep(1,l)

> #4.) Invert Sigma by a QR decomposition:
>
> Sigma1inv<-qr.solve(Sigma1)
> Sigma2inv<-qr.solve(Sigma2)

> #5.) Derive "optimal" weights:
>
> weights1<-(Sigma1inv%*%ones)/
+ ((ones%*%Sigma1inv%*%ones)[1,1])
> weights2<-(Sigma2inv%*%ones)/
+ ((ones%*%Sigma2inv%*%ones)[1,1])

> #6.) Derive minimal kriging variance:
>
> var1<-1/((ones%*%Sigma1inv%*%ones)[1,1])
> var2<-1/((ones%*%Sigma2inv%*%ones)[1,1])

> #7.) Predicted mean values:
>
> mean1<-t(weights1)%*%temp1
> mean2<-t(weights2)%*%temp2
```

Hence, we obtain our final results presented in Table 5.1, where the values for 2010/11/28 are printed in the first and for 2012/06/09 in the second row.

	Arithmetic mean	Predicted mean value	Kriging variance
2010/11/28	-2.78	-2.99	0.53
2012/06/09	14.97	14.51	0.66

Table 5.1: Results of prediction with kriging the mean applied to the temperature data in Germany

6 Simple Kriging

After predicting the mean value μ over a region (see last section "Kriging the Mean"), we want to predict the value of our underlying random function $Z(\mathbf{x})$ at any arbitrary point in our geographical region D. In other words, we want to predict the value of $Z(\mathbf{x})$ at some unsampled point $\mathbf{x}_0$, called the *prediction* or *estimation point* in literature (e.g. Wackernagel 2003).

Sometimes the mean μ of a random function $Z(\mathbf{x})$ for $\mathbf{x}$ in D is known or can be assumed from the underlying problem. In this case, the knowledge of the mean should be integrated in the model to improve the estimate of $Z(\mathbf{x})$ at $\mathbf{x}_0$ (Webster and Oliver 2007, p. 183). One way how we can achieve this, is to use *simple kriging*, which represents the simplest case of geostatistical prediction with kriging and which we will present in this section.

The most important references to which we will refer, are the books of Wackernagel (2003, pp. 24–26) and Cressie (1993, pp. 109–110, 359). But there exist a lot of other books as Webster and Oliver (2007, pp. 183–184) and Journel and Huijbregts (1978, pp. 561–562), just to name some. Notice that this kind of kriging was also originally stated by Georges Matheron (1962), see for instance Matheron (1971) in Chapter 3. But some similar versions have even appeared earlier, e.g. in Wold (1938).

6.1 Model for Simple Kriging

Wackernagel (2003, p. 24) and Cressie (1993, p. 359) summarized the model assumptions on the random function $Z(\mathbf{x})$ for simple kriging:

Assumption 6.1 (Model for Simple Kriging)

(i) The mean $\mu \in \mathbb{R}$ of $Z(\mathbf{x})$ for $\mathbf{x} \in D$ is known and constant, i.e. $\mu := \mathbb{E}[Z(\mathbf{x})]\ \forall \mathbf{x} \in D$.

(ii) Further, $Z(\mathbf{x})$ is supposed to be *second-order stationary* (cf. Definition 4.3, p. 14) with known covariance function $C(\mathbf{h}) := Cov(Z(\mathbf{x}), Z(\mathbf{x}+\mathbf{h})) = \mathbb{E}[Z(\mathbf{x})Z(\mathbf{x}+\mathbf{h})] - \mu^2$ for all $\mathbf{x}$, $\mathbf{x}+\mathbf{h}$ in the spatial domain D.

Due to the assumption of the knowledge of the constant mean μ, Wackernagel (2003, p. 25) also called simple kriging *kriging with known mean.*

The simple kriging predictor of $Z(\mathbf{x}_0)$ at the prediction point $\mathbf{x}_0$ uses the information at each sample $\mathbf{x}_i$, $i = 1, \ldots, n$, and the knowledge of the mean μ and the covariance $C(\mathbf{h})$. We define it as the following linear predictor in the sense of Wackernagel (2003, p. 25):

Definition 6.2 (Predictor for Simple Kriging)
The *simple kriging predictor* $Z^*_{\boldsymbol{\omega}}(\mathbf{x}_0)$ of $Z(\mathbf{x})$ at the prediction point $\mathbf{x}_0$ is defined as the sum of the mean μ and the weighted average of the differences of the random function

$Z(\mathbf{x})$ evaluated at each sample point $\mathbf{x}_i$ and the mean μ, i.e.

$$Z^*_{\boldsymbol{\omega}}(\mathbf{x}_0) := \mu + \sum_{i=1}^{n} \omega_i \left(Z(\mathbf{x}_i) - \mu\right) = \sum_{i=1}^{n} \omega_i Z(\mathbf{x}_i) + \mu \left(1 - \sum_{i=1}^{n} \omega_i\right)$$

$$= \mu + \boldsymbol{\omega}^T(\mathbf{Z} - \mu\mathbb{1}),$$

with $\omega_i \in \mathbb{R}$ being the weight of the corresponding residual $Z(\mathbf{x}_i)-\mu$ and $\boldsymbol{\omega} := (\omega_1, \ldots, \omega_n)^T \in \mathbb{R}^n$ the vector containing all weights of these residuals.

6.2 Unbiasedness condition

Fortunately, the unbiasedness of $Z^*_{\boldsymbol{\omega}}(\mathbf{x}_0)$ is automatically satisfied by its own definition, since

$$\mathbb{E}[Z^*_{\boldsymbol{\omega}}(\mathbf{x}_0) - Z(\mathbf{x}_0)] = \mu + \sum_{i=1}^{n} \omega_i \underbrace{\mathbb{E}[Z(\mathbf{x}_i) - \mu]}_{=0} - \underbrace{\mathbb{E}[Z(\mathbf{x}_0)]}_{=\mu} = \mu - \mu = 0.$$

Hence, we need no constraints on the weights $\omega_i,\ i = 1, \ldots, n$ (see Wackernagel 2003, p. 25). At first glance, this property seems to hold by chance, but this is just the way how the predictor is defined, with known μ integrated in the predictor (Webster and Oliver 2007, p. 184).

6.3 Variance of the prediction error

Furthermore, due to the unbiasedness of $Z^*_{\boldsymbol{\omega}}(\mathbf{x}_0)$, the prediction variance σ^2_E, which displays the quality of the linear predictor $Z^*_{\boldsymbol{\omega}}(\mathbf{x}_0)$, is given by its mean squared error $mse(Z^*_{\boldsymbol{\omega}}(\mathbf{x}_0))$. Consequently, it follows by Wackernagel (2003, p. 25):

$$\sigma^2_E := Var\left(Z^*_{\boldsymbol{\omega}}(\mathbf{x}_0) - Z(\mathbf{x}_0)\right) = \underbrace{\mathbb{E}\left[\left(Z^*_{\boldsymbol{\omega}}(\mathbf{x}_0) - Z(\mathbf{x}_0)\right)^2\right]}_{=mse(Z^*_{\boldsymbol{\omega}}(\mathbf{x}_0))} - \left(\underbrace{\mathbb{E}\left[Z^*_{\boldsymbol{\omega}}(\mathbf{x}_0) - Z(\mathbf{x}_0)\right]}_{bias=0}\right)^2$$

$$= \mathbb{E}\left[\left(\sum_{i=1}^{n} \omega_i (Z(\mathbf{x}_i) - \mu) + (\mu - Z(\mathbf{x}_0))\right)^2\right]$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} \omega_i\omega_j \mathbb{E}\left[(Z(\mathbf{x}_i) - \mu)(Z(\mathbf{x}_j) - \mu)\right] - 2\sum_{i=1}^{n} \omega_i \mathbb{E}[(Z(\mathbf{x}_i) - \mu)(Z(\mathbf{x}_0) - \mu)]$$

$$+\mathbb{E}[(Z(\mathbf{x}_0) - \mu)^2]$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} \omega_i\omega_j \underbrace{\left(\mathbb{E}[Z(\mathbf{x}_i)Z(\mathbf{x}_j)] - \mu^2\right)}_{=Cov(Z(\mathbf{x}_i),Z(\mathbf{x}_j))} - 2\sum_{i=1}^{n} \omega_i \underbrace{\left(\mathbb{E}[Z(\mathbf{x}_i)Z(\mathbf{x}_0)] - \mu^2\right)}_{=Cov(Z(\mathbf{x}_i),Z(\mathbf{x}_0))} + Var(Z(\mathbf{x}_0))$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} \omega_i\omega_j Cov(Z(\mathbf{x}_i), Z(\mathbf{x}_j)) - 2\sum_{i=1}^{n} \omega_i Cov(Z(\mathbf{x}_i), Z(\mathbf{x}_0)) + Cov(Z(\mathbf{x}_0), Z(\mathbf{x}_0)),$$

and hence we obtain

$$\sigma_E^2 = C(\mathbf{0}) + \sum_{i=1}^{n}\sum_{j=1}^{n}\omega_i\omega_j C(\mathbf{x}_i - \mathbf{x}_j) - 2\sum_{i=1}^{n}\omega_i C(\mathbf{x}_i - \mathbf{x}_0)$$

$$= C(\mathbf{0}) + \boldsymbol{\omega}^T\Sigma\boldsymbol{\omega} - 2\boldsymbol{\omega}^T\mathbf{c}_0 \geq 0.$$

The nonnegativity of the prediction variance follows from the representation

$$\sigma_E^2 = \mathbf{x}^T\Sigma_0\mathbf{x} \geq 0,$$

where $\mathbf{x} := \left(\boldsymbol{\omega}^T, -1\right)^T$ and Σ_0 denotes the covariance matrix of the random vector $(Z(\mathbf{x}_1), \ldots, Z(\mathbf{x}_n), Z(\mathbf{x}_0))^T$. For further details see the Appendix (p. 93), where the nonnegativity of the prediction variances of both following kriging methods is shown more precisely and analogously to our current setting.

6.4 Minimal prediction variance

In the next step, we want to minimize the prediction variance σ_E^2 of our linear predictor and hence want to maximize its accuracy. We observe that σ_E^2 is minimal, where its first derivative is zero, i.e. $\frac{\partial\sigma_E^2}{\partial\boldsymbol{\omega}} = 0$, and if the sufficient, but not necessary condition of positive definiteness of the Hessian matrix is fulfilled.

From the computation of σ_E^2 in 6.3 and since Σ is symmetric, we conclude in accordance with Wackernagel (2003, pp. 25–26):

$$\frac{\partial\sigma_E^2}{\partial\boldsymbol{\omega}} = \frac{\partial}{\partial\boldsymbol{\omega}}\left(C(\mathbf{0}) + \boldsymbol{\omega}^T\Sigma\boldsymbol{\omega} - 2\boldsymbol{\omega}^T\mathbf{c}_0\right) = \left(\Sigma + \Sigma^T\right)\boldsymbol{\omega} - 2\mathbf{c}_0 = 2\Sigma\boldsymbol{\omega} - 2\mathbf{c}_0 \stackrel{!}{=} 0$$

$$\Leftrightarrow \Sigma\boldsymbol{\omega} = \mathbf{c}_0.$$

The second derivative of σ_E^2 with respect to $\boldsymbol{\omega}$ is

$$\frac{\partial^2\sigma_E^2}{\partial\boldsymbol{\omega}^2} = \frac{\partial}{\partial\boldsymbol{\omega}}\left(2\Sigma\boldsymbol{\omega} - 2\mathbf{c}_0\right) = 2\Sigma,$$

which is actually positive definite due to Assumption 2.9 (p. 5).

Since this sufficient condition on the Hessian matrix hold, the computations above yield that we indeed achieve minimal prediction variance σ_E^2 if $\Sigma\boldsymbol{\omega} = \mathbf{c}_0$.

6.5 Equations for Simple Kriging

The last fact gives the conditions on the weights ω_i for $Z^*_{\boldsymbol{\omega}}(\mathbf{x}_0)$ being the *best linear unbiased predictor (BLUP)* of the value at $\mathbf{x}_0$. Hence, we can express the *equation system for simple kriging* (Wackernagel 2003, p. 26):

$$\sum_{j=1}^{n}\omega_j^{SK}C(\mathbf{x}_i - \mathbf{x}_j) = C(\mathbf{x}_i - \mathbf{x}_0),\ i = 1, \ldots, n$$

with "optimal" simple kriging weights $\omega_i^{SK} \in \mathbb{R}$ or analogously

$$\Sigma \boldsymbol{\omega}_{SK} = \mathbf{c}_0, \tag{6.1}$$

where $\boldsymbol{\omega}_{SK} := (\omega_1^{SK}, \ldots, \omega_n^{SK})^T \in \mathbb{R}^n$ denotes the vector providing the simple kriging weights.

With these observations, we can compute the weights ω_i^{SK} either by solving the linear system (6.1) of n equations step by step e.g. by the Gaussian elimination algorithm, or by Theorem 6.3 given by Cressie (1993, p. 110), which just inverts the covariance matrix Σ:

Theorem 6.3 (Solution of the simple kriging system)
The unique solution for $\boldsymbol{\omega}_{SK}$ of the simple kriging equation system (6.1) is

$$\boldsymbol{\omega}_{SK} = \begin{pmatrix} \omega_1^{SK} \\ \omega_2^{SK} \\ \vdots \\ \omega_n^{SK} \end{pmatrix} = \Sigma^{-1}\mathbf{c}_0. \tag{6.2}$$

Proof:
This follows immediately from the simple kriging system (6.1) and by the invertibility of Σ in Assumption 2.9 and Proposition 2.10 (pp. 5, 6) (Cressie 1993, p. 110).

□

6.6 Simple Kriging Variance

Furthermore, from the kriging system (6.1) and Theorem 6.3, we can directly obtain for the minimal *simple kriging variance* σ_{SK}^2 , which is defined as the variance of the prediction error $Z^*_{\boldsymbol{\omega}_{SK}}(\mathbf{x}_0) - Z(\mathbf{x}_0)$, according to Cressie (1993, pp. 110, 359):

$$\begin{aligned} \sigma_{SK}^2 &:= Var(Z^*_{\boldsymbol{\omega}_{SK}}(\mathbf{x}_0) - Z(\mathbf{x}_0)) = C(\mathbf{0}) + \boldsymbol{\omega}_{SK}^T \Sigma \boldsymbol{\omega}_{SK} - 2\boldsymbol{\omega}_{SK}^T \mathbf{c}_0 \\ &\overset{(6.1)}{=} C(\mathbf{0}) + \boldsymbol{\omega}_{SK}^T \mathbf{c}_0 - 2\boldsymbol{\omega}_{SK}^T \mathbf{c}_0 = C(\mathbf{0}) - \boldsymbol{\omega}_{SK}^T \mathbf{c}_0 \\ &\overset{(6.2)}{=} C(\mathbf{0}) - \mathbf{c}_0^T \Sigma^{-1} \mathbf{c}_0 = C(\mathbf{0}) - \sum_{i=1}^{n} \omega_i^{SK} C(\mathbf{x}_i - \mathbf{x}_0), \end{aligned} \tag{6.3}$$

where $C(\mathbf{0})$ equals the variance of the process $Z(\mathbf{x})$, since $C(\mathbf{0}) = C(\mathbf{x} - \mathbf{x}) = Var(Z(\mathbf{x}))$ holds for each point $\mathbf{x}$ in D.

6.7 Simple Kriging Prediction

Overall after these considerations, we can specify the simple kriging predictor, which is given by Cressie (1990, 1993), as

$$Z^*_{\boldsymbol{\omega}_{SK}}(\mathbf{x}_0) = \mu + \sum_{i=1}^{n} \omega_i^{SK}(Z(\mathbf{x}_i) - \mu) = \mu + \boldsymbol{\omega}_{SK}^T(\mathbf{Z} - \mu\mathbb{1}) = \mu + \mathbf{c}_0^T \Sigma^{-1}(\mathbf{Z} - \mu\mathbb{1})$$

$$= \mathbf{c}_0^T \Sigma^{-1} \mathbf{Z} + \mu(1 - \mathbf{c}_0^T \Sigma^{-1} \mathbb{1}),$$

with kriging variance in (6.3) and kriging estimate $z^*_{\boldsymbol{\omega}_{SK}}(\mathbf{x}_0)$ at the prediction point $\mathbf{x}_0$

$$z^*_{\boldsymbol{\omega}_{SK}}(\mathbf{x}_0) = \mu + \sum_{i=1}^{n} \omega_i^{SK}(z(\mathbf{x}_i) - \mu) = \mu + \boldsymbol{\omega}_{SK}^T(\mathbf{z} - \mu\mathbb{1}) = \mu + \mathbf{c}_0^T \Sigma^{-1}(\mathbf{z} - \mu\mathbb{1})$$

$$= \mathbf{c}_0^T \Sigma^{-1} \mathbf{z} + \mu(1 - \mathbf{c}_0^T \Sigma^{-1} \mathbb{1}).$$

Remark 6.4 (Exact interpolator)
Finally, consistent with Cressie (1993, p. 359), the simple kriging predictor $Z^*_{\boldsymbol{\omega}_{SK}}(\mathbf{x}_0)$ is called an *exact interpolator* because in the case that the prediction point $\mathbf{x}_0 = \mathbf{x}_i$ is identical with one of the data location points for $i \in \{1, \ldots, n\}$, then $Z^*_{\boldsymbol{\omega}_{SK}}(\mathbf{x}_0) = Z(\mathbf{x}_i)$, i.e. $\omega_i^{SK} = 1$ and $\omega_j^{SK} = 0$ for $j \in \{1, \ldots, n\},\ j \neq i$.

This holds, since the vector $(\omega_1^{SK}, \ldots, \omega_i^{SK}, \ldots, \omega_n^{SK})^T = (0, \ldots, 0, 1, 0, \ldots, 0)^T$ is a solution of

$$\underbrace{\begin{pmatrix} C(\mathbf{x}_1 - \mathbf{x}_1) & \cdots & C(\mathbf{x}_1 - \mathbf{x}_i) & \cdots & C(\mathbf{x}_1 - \mathbf{x}_n) \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ C(\mathbf{x}_i - \mathbf{x}_1) & \cdots & C(\mathbf{x}_i - \mathbf{x}_i) & \cdots & C(\mathbf{x}_i - \mathbf{x}_n) \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ C(\mathbf{x}_n - \mathbf{x}_1) & \cdots & C(\mathbf{x}_n - \mathbf{x}_i) & \cdots & C(\mathbf{x}_n - \mathbf{x}_n) \end{pmatrix}}_{=\Sigma} \underbrace{\begin{pmatrix} \omega_1^{SK} \\ \vdots \\ \omega_i^{SK} \\ \vdots \\ \omega_n^{SK} \end{pmatrix}}_{=\boldsymbol{\omega}_{SK}} = \underbrace{\begin{pmatrix} C(\mathbf{x}_1 - \mathbf{x}_i) \\ \vdots \\ C(\mathbf{x}_i - \mathbf{x}_i) \\ \vdots \\ C(\mathbf{x}_n - \mathbf{x}_i) \end{pmatrix}}_{=\mathbf{c}_0},$$

and $\mathbf{c}_0$ equals the *ith* column of the left-hand side matrix Σ. Additionally, by virtue of Theorem 6.3, the vector $(\omega_1^{SK}, \ldots, \omega_i^{SK}, \ldots, \omega_n^{SK})^T = (0, \ldots, 0, 1, 0, \ldots, 0)^T$ represents the unique solution of the simple kriging equations for $\mathbf{x}_0 = \mathbf{x}_i$ with $i \in \{1, \ldots, n\}$.

Hence, $Z^*_{\boldsymbol{\omega}_{SK}}(\mathbf{x}_0) = Z(\mathbf{x}_i)$ and $\sigma^2_{SK} \overset{(6.3)}{=} C(\mathbf{0}) - C(\mathbf{x}_i - \mathbf{x}_i) = 0$.

Remark 6.5
Similar to kriging the mean, simple kriging requires the existence of a known covariance function $C(\mathbf{h})$ of $Z(\mathbf{x})$. But since it is not known in practice, we have to estimate again either the covariance itself, or the corresponding underlying variogram. We can do that because the existence of a covariance function $C(\mathbf{h})$ of a second-order stationary process implies the existence of an equivalent (bounded) variogram function according to Proposition 4.5 (p. 15) in theory. But Webster and Oliver (2007, p. 183–184) commented that the estimated final variogram function has to be bounded. Afterwards, one has the choice if one would like to deduce the covariogram $C(\mathbf{h})$ from the estimated variogram or to rewrite all results in terms of the variogram function.

At this point, also notice that the property of the simple kriging predictor to be an exact interpolator only holds as long as we assume a theoretical variogram function $\gamma(\mathbf{h})$. This means that in practice, we will loose this property if we use a fitted variogram function, where a nugget component is added. This feature also holds for both following kriging methods.

6.8 Simple Kriging in *R*

At the end of this section about the theory of simple kriging, we want to present a way how simple kriging could be done in practical application in *gstat*. Therefore, as in the previous sections about the variogram and kriging the mean, we take the mean temperature data set in Germany of the two dates 2010/11/28 and 2012/06/09. At the end, we achieve a map of Germany including on the one hand the predicted temperature values of simple kriging and on the other hand their corresponding kriging variances.

We assume the arithmetic mean as the known mean value, since we do not have the "real" mean at hand. Fortunately, we can use the fitted Matérn variogram models *vfit1* and *vfit2* from section "The Variogram" and just have to update the gstat objects including the mean value. For a better understanding, the important and crucial *R* codes are printed below.

In the following sections we will not go into such detail as at this point, since the different kriging methods only vary in a few arguments.

```
> #Simple Kriging:
>
> #Update gstat object:
> #With known mean beta=arithmetic mean!
> g_sk1<-gstat(g1,model=vfit1,id="temp1",formula=temp1~1,
+ locations=~longkm+latkm,data=data1,beta=mean(temp1))

data:
temp1 : formula = temp1~1 ; data dim = 54 x 1 beta = -2.781481
variograms:
         model     psill    range kappa
temp1[1]   Nug 0.6464684   0.0000  0.00
temp1[2]   Mat 2.0341866 102.6184  1.49
~longkm + latkm

> g_sk2<-gstat(g2,model=vfit2,id="temp2",formula=temp2~1,
+ locations=~longkm+latkm,data=data2,beta=mean(temp2))

data:
temp2 : formula = temp2~1 ; data dim = 54 x 1 beta = 14.97407
variograms:
         model    psill    range kappa
temp2[1]   Nug 1.025047  0.00000  0.00
temp2[2]   Mat 2.442959 72.23161  2.98
~longkm + latkm

> #Additional 24 stations have latnewkm as latitude
> #and longnewkm as longitude
> newdat<-data.frame(longnewkm,latnewkm)
> coordinates(newdat)<-~longnewkm+latnewkm
```

```
> #Simple Kriging Prediction for additional 24 weather stations:
> p_sk1<-predict(g_sk1,newdata=newdat)

[using simple kriging]

> p_sk2<-predict(g_sk2,newdata=newdat)

[using simple kriging]
```

We print the first lines obtained from simple kriging prediction. The first column displays the coordinates of the prediction points, the second column the simple kriging estimates and the last column the kriging variances.

```
> p_sk1[1:5,] #2010/11/28

         coordinates temp1.pred temp1.var
1    (8.979, 48.216)  -1.052472 0.9595025
2  (9.8044, 51.9672)  -4.404861 0.9077604
3 (13.4367, 54.6817)  -1.695010 1.6049809
4    (7.979, 51.465)  -3.116559 0.8861442
5 (10.9431, 48.4261)  -2.235687 0.8815832

> p_sk2[1:5,] #2012/06/09

         coordinates temp2.pred temp2.var
1    (8.979, 48.216)   15.94579  1.332583
2  (9.8044, 51.9672)   14.58337  1.270921
3 (13.4367, 54.6817)   15.69993  2.052531
4    (7.979, 51.465)   13.20399  1.256293
5 (10.9431, 48.4261)   14.60842  1.267006
```

We want to comment that there exists another alternative function in *gstat* for univariate kriging prediction, implemented through the command *krige()*.

```
> krige_sk1<-krige(formula=temp1~1, locations=~longkm+latkm,
+ data=data1, newdata=newdat, model=vfit1, beta=mean(temp1))

[using simple kriging]

> #Kriging prediction estimates and variances coincide:
> all(p_sk1$temp1.pred==krige_sk1$var1.pred)

[1] TRUE

> all(p_sk1$temp1.var==krige_sk1$var1.var)

[1] TRUE
```

Furthermore, we generate a grid of longitude and latitude for our plot:

```
> grid<-expand.grid(x=seq(from=long.range[1],
+ to=long.range[2], by=step), y=seq(from=lat.range[1],
+ to=lat.range[2], by=step))

> coordinates(grid)<-~x+y
> gridded(grid)<-TRUE
```

Next, we apply simple kriging prediction to the coordinates of our grid and then plot the resulting estimates and variances:

```
> #Simple Kriging Prediction for grid:
> prediction_sk1<-predict(object=g_sk1, newdata=grid)

[using simple kriging]

> prediction_sk2<-predict(object=g_sk2, newdata=grid)

[using simple kriging]

> #Plot Simple Kriging Estimates:
> #prediction_sk1_plot same values as prediction_sk1,
> #but right coordinates for plotting, no conversion into km as unit
>
> image.plot(prediction_sk1_plot, 1, legend.only=FALSE,
+ zlim=temp.lim_sk1, breaks=breakstemp, col=coltemp, nlevel=20,
+ legend.width=3, asp=asp)
> contour(prediction_sk1_plot, 1, drawlabels=TRUE, col="brown",
+ add=TRUE)

> #Plot Simple Kriging Variances:
>
> image.plot(prediction_sk1_plot, 2, legend.only=FALSE,
+ zlim=var.lim_sk1, breaks=breaksvar, col=colvar, nlevel=20,
+ legend.width=3, asp=asp)
> contour(prediction_sk1_plot, 2, drawlabels=TRUE, col="brown",
+ add=TRUE)
```

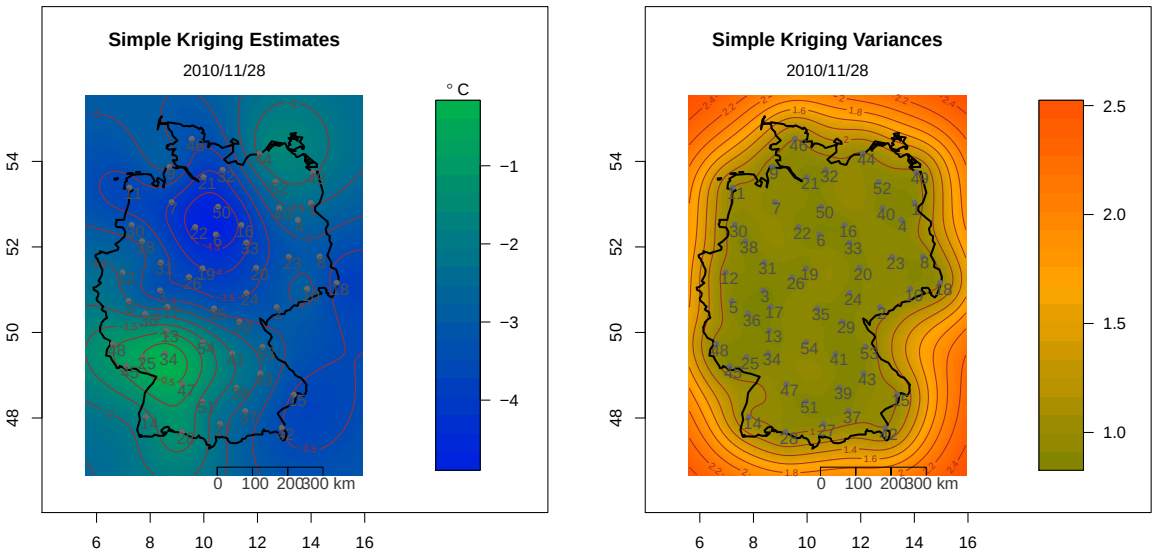
We put our results together and obtain the residuals of our additional 24 weather stations of 2010/11/28 and 2012/06/09 in Table 6.1. We print the differences of the observed, measured values and the predicted temperatures. In the last row we calculate the sum of squares, i.e. the sum of the squared residuals, to indicate whether the prediction is "good" or not. We observe that the estimates for the date 2010/11/28 seem to be closer to the observed data as it is for 2012/06/09, since their residual sum of squares are obviously lower. But one has to take care, since the "outlier" with a residual of -7.05 for 2012/06/09 has a strong impact on the amount of the sum of squares.

Longitude	Latitude	Residuals 2010/11/28	Residuals 2012/06/09
8.98	48.22	-3.05	-2.25
9.8	51.97	-1.10	0.32
13.44	54.68	1.50	-1.90
7.98	51.47	0.32	0.10
10.94	48.43	-0.46	0.09
7.31	50.04	-1.50	-2.37
6.7	53.6	1.68	-0.74
9.14	53.45	-1.13	-0.30
9.32	49.52	-0.74	-1.57
14.73	52.02	-0.08	-0.32
10.5	49.85	-0.33	-0.94
10.13	48.99	-1.23	-1.02
12.73	48.48	-0.29	-1.00
10.68	53.58	-0.13	0.03
13.14	49.11	-3.15	-7.05
13.94	53.32	-0.51	-0.34
9.22	50.51	-2.77	-4.03
11.14	52.97	0.73	-0.17
11.27	47.48	-0.86	-3.13
7.64	47.81	-0.06	0.36
11.14	50.5	-2.51	-4.33
10.88	51.67	-3.81	-3.13
8.57	52.45	0.48	0.27
12.46	52.12	-0.53	-1.07
Sum of Squares		62.17	126.34

Table 6.1: Residuals from simple kriging prediction of the additional 24 weather stations in Germany, where each residual equals the difference of the observed value and the prediction estimate obtained from simple kriging; sum of squares is the sum of all squared residuals

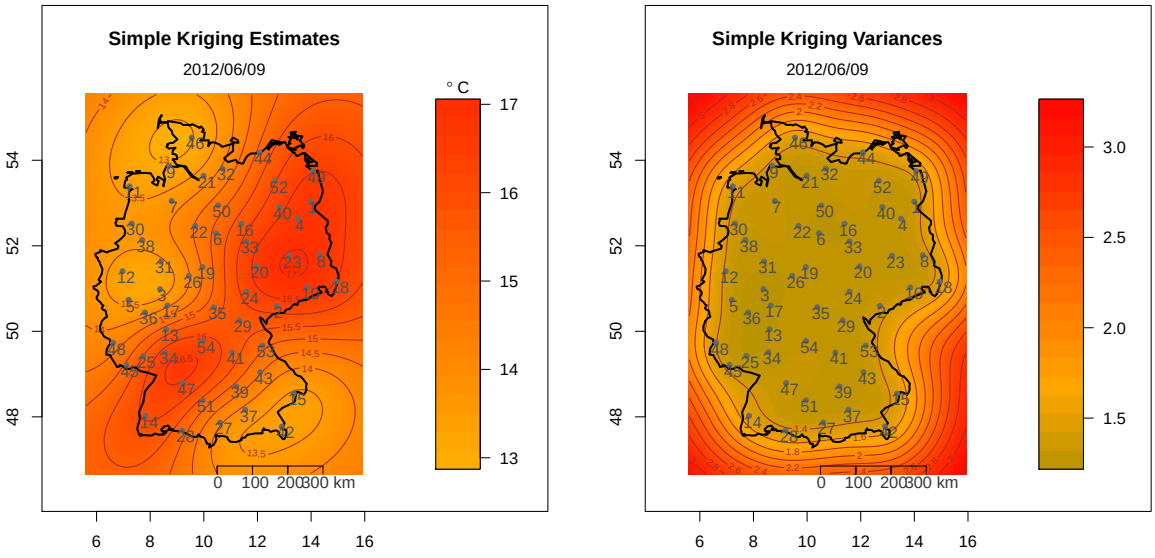
Finally, we obtain a plot, i.e. the map of Germany, where the predicted temperature values from simple kriging are printed on the left, and their corresponding kriging variances on the right, see Figure 6.1 for the date 2010/11/28 and Figure 6.2 for 2012/06/09.

Note that the prediction variances of 2012/06/09 within the map, and especially at the sample points, are higher than those variances of 2010/11/28. This results from the higher nugget component in the corresponding fitted model, 1.03 instead of 0.65.



(a) Simple Kriging Estimates (b) Simple Kriging Variances

Figure 6.1: Simple Kriging applied to the temperature data of 2010/11/28 in Germany



(a) Simple Kriging Estimates (b) Simple Kriging Variances

Figure 6.2: Simple Kriging applied to the temperature data of 2012/06/09 in Germany

7 Ordinary Kriging

In practice, in most cases the mean μ and the covariance function $C(\mathbf{h})$ of the underlying random function $Z(\mathbf{x})$ are unknown. Thus, unfortunately, simple kriging prediction is not really applicable, since it requires information about μ and $C(\mathbf{h})$ (see Cressie 1993, p. 359). Therefore we want to introduce prediction with the geostatistical method *ordinary kriging*, which - unlike to simple kriging - does not assume the knowledge of the mean and the covariance. For this reason, ordinary kriging represents the most common kriging method in practice and its aim is to predict the value of the random variable $Z(\mathbf{x})$ at an unsampled point $\mathbf{x}_0$ of a geographical region as well (Webster and Oliver 2007, p. 155).

In literature, there are a lot of books covering the subject ordinary kriging. In this section we will especially focus on Wackernagel (2003) in Chapter 11 and Cressie (1993, pp. 119–123, 360–361), for other references see Webster and Oliver (2007, pp. 155–159), Journel and Huijbregts (1978, pp. 304–313, 563–564) and Kitanidis (1997, pp. 65–74). Their theory relies on the work of Georges Matheron (1962), for instance see Matheron (1971) in Chapter 3, where kriging, and in particular ordinary kriging, was presented.

7.1 Model for Ordinary Kriging

Ordinary kriging refers to spatial prediction under the following two assumptions, specified by Cressie (1993, pp. 120–121, 360) and Wackernagel (2003, p. 80). It requires only weaker assumptions compared to simple kriging:

Assumption 7.1 (Model for Ordinary Kriging)

(i) The global, constant mean $\mu \in \mathbb{R}$ of the random function $Z(\mathbf{x})$ is unknown.

(ii) The data come from an intrinsically stationary random function $Z(\mathbf{x})$ with known variogram function $\gamma(\mathbf{h})$ (see Definition 4.1, p. 13), i.e.

$$\gamma(\mathbf{h}) = \frac{1}{2} Var(Z(\mathbf{x}+\mathbf{h}) - Z(\mathbf{x})) = \frac{1}{2} \mathbb{E}\left[(Z(\mathbf{x}+\mathbf{h}) - Z(\mathbf{x}))^2\right].$$

On the basis of these model assumptions, we define the predictor for ordinary kriging in consistence with Wackernagel (2003, p. 79) as follows:

Definition 7.2 (Predictor for Ordinary Kriging)

The *ordinary kriging predictor* $Z^*_{\boldsymbol{\omega}}(\mathbf{x}_0)$ of the value at $\mathbf{x}_0$ is the linear combination of $Z(\mathbf{x})$ evaluated at each sample $\mathbf{x}_i,\ i = 1, \ldots, n$

$$Z^*_{\boldsymbol{\omega}}(\mathbf{x}_0) := \sum_{i=1}^{n} \omega_i Z(\mathbf{x}_i) = \boldsymbol{\omega}^T \mathbf{Z},$$

where $\boldsymbol{\omega} := (\omega_1, \ldots, \omega_n)^T \in \mathbb{R}^n$ provides the unknown weights $\omega_i \in \mathbb{R}$ corresponding with the influence of the variable $Z(\mathbf{x}_i)$ in the computation of $Z^*_{\boldsymbol{\omega}}(\mathbf{x}_0)$.

7.2 Unbiasedness condition

To ensure the unbiasedness of the linear predictor $Z^*_{\boldsymbol{\omega}}(\mathbf{x}_0)$, Wackernagel (2003, p. 80) set the sum of the weights to one, i.e.

$$\sum_{i=1}^{n} \omega_i = 1 \;\Leftrightarrow\; \boldsymbol{\omega}^T \mathbb{1} = 1. \tag{7.1}$$

Thus, the expected error vanishes

$$\mathbb{E}[Z^*_{\boldsymbol{\omega}}(\mathbf{x}_0) - Z(\mathbf{x}_0)] = \mathbb{E}\left[\sum_{i=1}^{n} \omega_i Z(\mathbf{x}_i) - Z(\mathbf{x}_0) \underbrace{\sum_{i=1}^{n} \omega_i}_{=1}\right] = \sum_{i=1}^{n} \omega_i \underbrace{\mathbb{E}\left[Z(\mathbf{x}_i) - Z(\mathbf{x}_0)\right]}_{=0} = 0,$$

since the expected value of the increments is zero due to Assumption 7.1 (ii), where we supposed $Z(\mathbf{x})$ to be intrinsically stationary.

7.3 Variance of the prediction error

Further, in order to achieve an indicator of our estimate accuracy, we want to calculate the error variance σ^2_E of $Z^*_{\boldsymbol{\omega}}(\mathbf{x}_0)$. We can express this quantity by inserting the variogram function $\gamma(\mathbf{h})$ of $Z(\mathbf{x})$ according to the identity $\gamma(\mathbf{h}) = \frac{1}{2}\mathbb{E}\left[(Z(\mathbf{x}+\mathbf{h}) - Z(\mathbf{x}))^2\right]$. This yields in accordance with Cressie (1993, p. 121):

$$\sigma^2_E := Var\left(Z^*_{\boldsymbol{\omega}}(\mathbf{x}_0) - Z(\mathbf{x}_0)\right) \overset{(7.1)}{=} \mathbb{E}\left[(Z^*_{\boldsymbol{\omega}}(\mathbf{x}_0) - Z(\mathbf{x}_0))^2\right] = \mathbb{E}\left[\left(\sum_{i=1}^{n} \omega_i Z(\mathbf{x}_i) - Z(\mathbf{x}_0)\right)^2\right]$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} \omega_i\omega_j \mathbb{E}[Z(\mathbf{x}_i)Z(\mathbf{x}_j)] - 2\sum_{i=1}^{n} \omega_i \mathbb{E}[Z(\mathbf{x}_i)Z(\mathbf{x}_0)] + \mathbb{E}\left[(Z(\mathbf{x}_0))^2\right]$$

$$\overset{(*)}{=} -\sum_{i=1}^{n}\sum_{j=1}^{n} \omega_i\omega_j \underbrace{\frac{\mathbb{E}\left[(Z(\mathbf{x}_i) - Z(\mathbf{x}_j))^2\right]}{2}}_{=\gamma(\mathbf{x}_i - \mathbf{x}_j)} + 2\sum_{i=1}^{n} \omega_i \underbrace{\frac{\mathbb{E}\left[(Z(\mathbf{x}_i) - Z(\mathbf{x}_0))^2\right]}{2}}_{=\gamma(\mathbf{x}_i - \mathbf{x}_0)}.$$

Hence, we obtain for the prediction variance

$$\sigma^2_E = -\sum_{i=1}^{n}\sum_{j=1}^{n} \omega_i\omega_j \gamma(\mathbf{x}_i - \mathbf{x}_j) + 2\sum_{i=1}^{n} \omega_i \gamma(\mathbf{x}_i - \mathbf{x}_0)$$

$$= -\boldsymbol{\omega}^T \Gamma \boldsymbol{\omega} + 2\boldsymbol{\omega}^T \boldsymbol{\gamma}_0 = \boldsymbol{\omega}^T (2\boldsymbol{\gamma}_0 - \Gamma\boldsymbol{\omega}) \geq 0,$$

with semivariances $\gamma(\mathbf{x}_i - \mathbf{x}_j)$ of $Z(\mathbf{x})$ between the data points, $\gamma(\mathbf{x}_i - \mathbf{x}_0)$ between each data point and the unsampled point $\mathbf{x}_0$, symmetric variogram matrix $\Gamma_{i,j} := \gamma(\mathbf{x}_i - \mathbf{x}_j)$, $i, j = 1, \ldots, n$ and $\boldsymbol{\gamma}_0 := (\gamma(\mathbf{x}_1 - \mathbf{x}_0), \ldots, \gamma(\mathbf{x}_n - \mathbf{x}_0))^T \in \mathbb{R}^n$ (Webster and Oliver 2007, p. 156).

The nonnegativity of the prediction variance follows from the representation

$$\sigma_E^2 = -\mathbf{x}^T \Gamma_0 \mathbf{x} \geq 0,$$

where $\mathbf{x} := \begin{pmatrix} \boldsymbol{\omega} \\ -1 \end{pmatrix}$ and $\Gamma_0 := \left(\begin{array}{c|c} \Gamma & \boldsymbol{\gamma}_0 \\ \hline \boldsymbol{\gamma}_0^T & 0 \end{array}\right)$ denotes the variogram matrix of $(Z(\mathbf{x}_1), \ldots, Z(\mathbf{x}_n), Z(\mathbf{x}_0))^T$. This representation can easily be seen by

$$-\mathbf{x}^T \Gamma_0 \mathbf{x} = -\left(\boldsymbol{\omega}^T, -1\right) \Gamma_0 \begin{pmatrix} \boldsymbol{\omega} \\ -1 \end{pmatrix} = -\left(\boldsymbol{\omega}^T, -1\right) \begin{pmatrix} \Gamma\boldsymbol{\omega} - \boldsymbol{\gamma}_0 \\ \boldsymbol{\gamma}_0^T \boldsymbol{\omega} \end{pmatrix} = -\boldsymbol{\omega}^T \Gamma \boldsymbol{\omega} + 2\boldsymbol{\omega}^T \boldsymbol{\gamma}_0 = \sigma_E^2.$$

Since $\mathbf{x}^T \mathbb{1}_{n+1} = \boldsymbol{\omega}^T \mathbb{1} - 1 = 0$, the prediction variance is nonnegative due to the conditionally negative semidefiniteness of the variogram.

Furthermore, notice that Equation $(*)$ only holds, since $\sum_{j=1}^n \omega_j = 1$, which is identical with (7.1). Following Cressie (1993, p. 121), we can verify this fact:

$$\begin{aligned}
&-\sum_{i=1}^n \sum_{j=1}^n \omega_i \omega_j \frac{(Z(\mathbf{x}_i) - Z(\mathbf{x}_j))^2}{2} + 2\sum_{i=1}^n \omega_i \frac{(Z(\mathbf{x}_i) - Z(\mathbf{x}_0))^2}{2} \\
&= -\sum_{i=1}^n \omega_i \frac{(Z(\mathbf{x}_i))^2}{2} \underbrace{\sum_{j=1}^n \omega_j}_{=1} + \sum_{i=1}^n \sum_{j=1}^n \omega_i \omega_j Z(\mathbf{x}_i) Z(\mathbf{x}_j) - \sum_{j=1}^n \omega_j \frac{(Z(\mathbf{x}_j))^2}{2} \underbrace{\sum_{i=1}^n \omega_i}_{=1} \\
&\quad + \sum_{i=1}^n \omega_i \left(Z(\mathbf{x}_i)\right)^2 - 2\sum_{i=1}^n \omega_i Z(\mathbf{x}_i) Z(\mathbf{x}_0) + \left(Z(\mathbf{x}_0)\right)^2 \underbrace{\sum_{i=1}^n \omega_i}_{=1} \\
&= \sum_{i=1}^n \sum_{j=1}^n \omega_i \omega_j Z(\mathbf{x}_i) Z(\mathbf{x}_j) - 2\sum_{i=1}^n \omega_i Z(\mathbf{x}_i) Z(\mathbf{x}_0) + \left(Z(\mathbf{x}_0)\right)^2.
\end{aligned}$$

Equation $(*)$ simply follows by taking expectations.

7.4 Minimal prediction variance

Similar to kriging the mean and simple kriging, our next goal is to achieve minimal prediction variance σ_E^2 under the unbiasedness condition (7.1) and further to find the "optimal" weights ω_i. Therefore, we look for a solution of the minimization problem

$$\text{minimum of } \boldsymbol{\omega}^T \left(2\boldsymbol{\gamma}_0 - \Gamma\boldsymbol{\omega}\right) \text{ subject to } \boldsymbol{\omega}^T \mathbb{1} = 1.$$

For this purpose, we define the function φ to solve the problem with the method of Lagrange multipliers, such that

$$\varphi : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$$

$$(\boldsymbol{\omega}, \lambda) \mapsto \varphi(\boldsymbol{\omega}, \lambda) := -\boldsymbol{\omega}^T \Gamma \boldsymbol{\omega} + 2\boldsymbol{\omega}^T \boldsymbol{\gamma}_0 - 2\lambda(\boldsymbol{\omega}^T \mathbb{1} - 1).$$

Here, the Lagrange multiplier $\lambda \in \mathbb{R}$ is involved to guarantee the unbiasedness condition (7.1) (Cressie 1993, p. 121).

Therefore, for solving this optimization problem, we have to set the first order partial derivatives of φ with respect to $\boldsymbol{\omega}$ and λ to zero. This give the necessary conditions on $\boldsymbol{\omega}$ and λ.

First, we obtain the derivative of φ with respect to the weight vector $\boldsymbol{\omega}$

$$\frac{\partial \varphi(\boldsymbol{\omega}, \lambda)}{\partial \boldsymbol{\omega}} = -2\Gamma\boldsymbol{\omega} + 2\boldsymbol{\gamma}_0 - 2\lambda\mathbb{1} \stackrel{!}{=} 0,$$

which yields

$$\Gamma\boldsymbol{\omega} + \lambda\mathbb{1} = \boldsymbol{\gamma}_0.$$

Secondly, by differentiating φ with respect to the Lagrange parameter λ, we derive

$$\frac{\partial \varphi(\boldsymbol{\omega}, \lambda)}{\partial \lambda} = -2(\boldsymbol{\omega}^T \mathbb{1} - 1) \stackrel{!}{=} 0$$

if and only if the unbiasedness equation (7.1) is fulfilled (see Cressie 1993, pp. 121–122).

This way we indeed achieve the minimum of the prediction variance σ_E^2 subject to the constraints on the weights. This means that the necessary conditions are also sufficient. For the complete proof of this minimality, we refer to the Appendix (p. 93). There, we give another representation of the variance and apply Theorem 2.15 by Rao (1973), which yields the minimality and the nonnegativity of the variance. Note that we follow the Lagrange approach at this point, since it will give a nicer and more efficient solution for the parameters.

7.5 Equations for Ordinary Kriging

The last formulas show the way how we can achieve minimal error variance of our prediction, namely by applying weights ω_i satisfying these conditions (Wackernagel 2003, p. 81). Hence, we can write the *ordinary kriging system*, which was presented by Matheron for the first time. For instance see Matheron (1971, pp. 123–130), where he called it *kriging equations for stationary random function with unknown expectation*:

$$\sum_{j=1}^{n} \omega_j^{OK} \gamma(\mathbf{x}_i - \mathbf{x}_j) + \lambda_{OK} = \gamma(\mathbf{x}_i - \mathbf{x}_0) \text{ for } i = 1, \ldots, n$$

$$\sum_{j=1}^{n} \omega_j^{OK} = 1$$

We can also express this system in matrix formulation (Webster and Oliver 2007, pp. 158–159), such that

$$\Gamma \boldsymbol{\omega}_{OK} + \lambda_{OK} \mathbb{1} = \boldsymbol{\gamma}_0 \tag{7.2}$$

$$\boldsymbol{\omega}_{OK}^T \mathbb{1} = 1$$

$$\Leftrightarrow$$

$$\underbrace{\begin{pmatrix} \gamma(\mathbf{x}_1 - \mathbf{x}_1) & \gamma(\mathbf{x}_1 - \mathbf{x}_2) & \cdots & \gamma(\mathbf{x}_1 - \mathbf{x}_n) & 1 \\ \gamma(\mathbf{x}_2 - \mathbf{x}_1) & \gamma(\mathbf{x}_2 - \mathbf{x}_2) & \cdots & \gamma(\mathbf{x}_2 - \mathbf{x}_n) & 1 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ \gamma(\mathbf{x}_n - \mathbf{x}_1) & \gamma(\mathbf{x}_n - \mathbf{x}_2) & \cdots & \gamma(\mathbf{x}_n - \mathbf{x}_n) & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{pmatrix}}_{=:\tilde{\Gamma}} \begin{pmatrix} \omega_1^{OK} \\ \omega_2^{OK} \\ \vdots \\ \omega_n^{OK} \\ \lambda_{OK} \end{pmatrix} = \underbrace{\begin{pmatrix} \gamma(\mathbf{x}_1 - \mathbf{x}_0) \\ \gamma(\mathbf{x}_2 - \mathbf{x}_0) \\ \vdots \\ \gamma(\mathbf{x}_n - \mathbf{x}_0) \\ 1 \end{pmatrix}}_{=:\tilde{\gamma}_0},$$

where $\boldsymbol{\omega}_{OK} := (\omega_1^{OK}, \ldots, \omega_n^{OK})^T \in \mathbb{R}^n$ denotes the vector providing the optimal weights $\omega_i^{OK} \in \mathbb{R}$, $\lambda_{OK} \in \mathbb{R}$ the Lagrange multiplier of ordinary kriging, $\tilde{\boldsymbol{\gamma}}_0 := (\boldsymbol{\gamma}_0^T, 1)^T \in \mathbb{R}^{n+1}$ and block matrix $\tilde{\Gamma} := \left(\begin{array}{ccc|c} & \Gamma & & \begin{matrix}1\\ \vdots \\ 1\end{matrix} \\ \hline 1 & \cdots & 1 & 0 \end{array}\right) \in \mathbb{R}^{(n+1)\times(n+1)}$.

Afterwards, we can compute the ordinary kriging weights ω_i^{OK} again either by solving the linear system of $n+1$ equations in (7.1) and (7.2) by the Gaussian elimination algorithm. Or in the case that the matrix $\tilde{\Gamma}$ is nonsingular, we can derive the ordinary kriging weights ω_i^{OK} and the Lagrange multiplier λ_{OK} according to Webster and Oliver (2007, p. 159) as

$$\begin{pmatrix} \boldsymbol{\omega}_{OK} \\ \lambda_{OK} \end{pmatrix} = \begin{pmatrix} \omega_1^{OK} \\ \omega_2^{OK} \\ \vdots \\ \omega_n^{OK} \\ \lambda_{OK} \end{pmatrix} = \tilde{\Gamma}^{-1} \tilde{\boldsymbol{\gamma}}_0.$$

Another alternative way for solving the ordinary kriging equations is provided in the next theorem, which can be found in Cressie (1993, p. 122):

Theorem 7.3 (Solution of the ordinary kriging system)
Under the assumption of invertibility of the variogram matrix Γ, the unique solution for $\boldsymbol{\omega}_{OK}$ and λ_{OK} of the ordinary kriging system (7.1) and (7.2) can be specified as

$$\boldsymbol{\omega}_{OK} = \Gamma^{-1} \left[\boldsymbol{\gamma}_0 - \mathbb{1} \left(\frac{\mathbb{1}^T \Gamma^{-1} \boldsymbol{\gamma}_0 - 1}{\mathbb{1}^T \Gamma^{-1} \mathbb{1}} \right) \right], \tag{7.3}$$

$$\lambda_{OK} = \frac{\mathbb{1}^T \Gamma^{-1} \boldsymbol{\gamma}_0 - 1}{\mathbb{1}^T \Gamma^{-1} \mathbb{1}}. \tag{7.4}$$

Proof:
First notice that $\boldsymbol{\omega}_{OK}$ and λ_{OK} in Theorem 7.3 are well-defined, since Γ^{-1} exists. Further, we want to show that the theorem actually provides a solution of the ordinary kriging equations (7.1) and (7.2). In fact, this is really the case, since it holds

(i) $\Gamma\boldsymbol{\omega}_{OK} + \lambda_{OK}\mathbb{1} = \boldsymbol{\gamma}_0 - \mathbb{1}\left(\frac{\mathbb{1}^T\Gamma^{-1}\boldsymbol{\gamma}_0 - 1}{\mathbb{1}^T\Gamma^{-1}\mathbb{1}}\right) + \mathbb{1}\left(\frac{\mathbb{1}^T\Gamma^{-1}\boldsymbol{\gamma}_0 - 1}{\mathbb{1}^T\Gamma^{-1}\mathbb{1}}\right) = \boldsymbol{\gamma}_0$ and

(ii) $\boldsymbol{\omega}_{OK}^T\mathbb{1} = \left[\boldsymbol{\gamma}_0 - \mathbb{1}\left(\frac{\mathbb{1}^T\Gamma^{-1}\boldsymbol{\gamma}_0 - 1}{\mathbb{1}^T\Gamma^{-1}\mathbb{1}}\right)\right]^T\Gamma^{-1}\mathbb{1} = \boldsymbol{\gamma}_0^T\Gamma^{-1}\mathbb{1} - \mathbb{1}^T\Gamma^{-1}\mathbb{1}\left(\frac{\mathbb{1}^T\Gamma^{-1}\boldsymbol{\gamma}_0 - 1}{\mathbb{1}^T\Gamma^{-1}\mathbb{1}}\right)$
$= \boldsymbol{\gamma}_0^T\Gamma^{-1}\mathbb{1} - \left(\mathbb{1}^T\Gamma^{-1}\boldsymbol{\gamma}_0 - 1\right) = 1,$

due to the symmetry of the matrices Γ and Γ^{-1}. And finally, the given solution of the ordinary kriging system is in fact unique due to the assumption of nonsingularity of Γ.

□

7.6 Ordinary Kriging Variance

Now at this point, we are able to specify the minimized prediction variance, the *ordinary kriging variance*, which is defined as $\sigma_{OK}^2 := Var(Z^*_{\boldsymbol{\omega}_{OK}}(\mathbf{x}_0) - Z(\mathbf{x}_0))$. In accordance with Cressie (1993, p. 122), we can derive it from σ_E^2 (see 7.3) by inserting the ordinary kriging equations (7.1) and (7.2) and the last Theorem 7.3:

$$\sigma_{OK}^2 := Var(Z^*_{\boldsymbol{\omega}_{OK}}(\mathbf{x}_0) - Z(\mathbf{x}_0)) = \boldsymbol{\omega}_{OK}^T(2\boldsymbol{\gamma}_0 - \Gamma\boldsymbol{\omega}_{OK}) = -\boldsymbol{\omega}_{OK}^T(\underbrace{\Gamma\boldsymbol{\omega}_{OK} - \boldsymbol{\gamma}_0}) + \boldsymbol{\omega}_{OK}^T\boldsymbol{\gamma}_0$$

$$\overset{(7.2)}{=} \lambda_{OK}\underbrace{\boldsymbol{\omega}_{OK}^T\mathbb{1}} + \boldsymbol{\omega}_{OK}^T\boldsymbol{\gamma}_0 \overset{(7.1)}{=} \lambda_{OK} + \boldsymbol{\omega}_{OK}^T\boldsymbol{\gamma}_0 = \left(\boldsymbol{\omega}_{OK}^T, \lambda_{OK}\right)\begin{pmatrix}\boldsymbol{\gamma}_0 \\ 1\end{pmatrix}$$

$$\overset{(7.3),(7.4)}{=} \frac{\mathbb{1}^T\Gamma^{-1}\boldsymbol{\gamma}_0 - 1}{\mathbb{1}^T\Gamma^{-1}\mathbb{1}} + \left[\Gamma^{-1}\left(\boldsymbol{\gamma}_0 - \mathbb{1}\frac{\mathbb{1}^T\Gamma^{-1}\boldsymbol{\gamma}_0 - 1}{\mathbb{1}^T\Gamma^{-1}\mathbb{1}}\right)\right]^T\boldsymbol{\gamma}_0$$

$$\overset{\left(\Gamma^{-1}\right)^T=\Gamma^{-1}}{=} \frac{\mathbb{1}^T\Gamma^{-1}\boldsymbol{\gamma}_0 - 1}{\mathbb{1}^T\Gamma^{-1}\mathbb{1}} + \boldsymbol{\gamma}_0^T\Gamma^{-1}\boldsymbol{\gamma}_0 - \left(\frac{\mathbb{1}^T\Gamma^{-1}\boldsymbol{\gamma}_0 - 1}{\mathbb{1}^T\Gamma^{-1}\mathbb{1}}\right)\mathbb{1}^T\Gamma^{-1}\boldsymbol{\gamma}_0$$

$$= \boldsymbol{\gamma}_0^T\Gamma^{-1}\boldsymbol{\gamma}_0 + \frac{\mathbb{1}^T\Gamma^{-1}\boldsymbol{\gamma}_0 - 1}{\mathbb{1}^T\Gamma^{-1}\mathbb{1}}\left(1 - \mathbb{1}^T\Gamma^{-1}\boldsymbol{\gamma}_0\right) = \boldsymbol{\gamma}_0^T\Gamma^{-1}\boldsymbol{\gamma}_0 - \frac{\left(\mathbb{1}^T\Gamma^{-1}\boldsymbol{\gamma}_0 - 1\right)^2}{\mathbb{1}^T\Gamma^{-1}\mathbb{1}}.$$

Hence, we obtain for the minimal kriging variance

$$\sigma_{OK}^2 = \lambda_{OK} + \boldsymbol{\omega}_{OK}^T\boldsymbol{\gamma}_0 = \boldsymbol{\gamma}_0^T\Gamma^{-1}\boldsymbol{\gamma}_0 - \frac{\left(\mathbb{1}^T\Gamma^{-1}\boldsymbol{\gamma}_0 - 1\right)^2}{\mathbb{1}^T\Gamma^{-1}\mathbb{1}}$$

$$= -\sum_{i=1}^n\sum_{j=1}^n \omega_i^{OK}\omega_j^{OK}\gamma(\mathbf{x}_i - \mathbf{x}_j) + 2\sum_{i=1}^n \omega_i^{OK}\gamma(\mathbf{x}_i - \mathbf{x}_0)$$

$$= \lambda_{OK} + \sum_{i=1}^n \omega_i^{OK}\gamma(\mathbf{x}_i - \mathbf{x}_0). \tag{7.5}$$

7.7 Ordinary Kriging Prediction

Overall, we can write the ordinary kriging predictor $Z^*_{\boldsymbol{\omega}_{OK}}(\mathbf{x}_0)$ at $\mathbf{x}_0$ by applying Theorem 7.3 (Cressie 1990, 1993), such that

$$Z^*_{\boldsymbol{\omega}_{OK}}(\mathbf{x}_0) = \sum_{i=1}^{n} \omega_i^{OK} Z(\mathbf{x}_i) = \boldsymbol{\omega}_{OK}^T \mathbf{Z} = \left(\boldsymbol{\gamma}_0 - \mathbb{1}\frac{\mathbb{1}^T\Gamma^{-1}\boldsymbol{\gamma} - 1}{\mathbb{1}^T\Gamma^{-1}\mathbb{1}}\right)^T \Gamma^{-1}\mathbf{Z}$$

with corresponding kriging variance in (7.5) and kriging estimate $z^*_{\boldsymbol{\omega}_{OK}}(\mathbf{x}_0)$ at $\mathbf{x}_0$

$$z^*_{\boldsymbol{\omega}_{OK}}(\mathbf{x}_0) = \sum_{i=1}^{n} \omega_i^{OK} z(\mathbf{x}_i) = \boldsymbol{\omega}_{OK}^T \mathbf{z} = \left(\boldsymbol{\gamma}_0 - \mathbb{1}\frac{\mathbb{1}^T\Gamma^{-1}\boldsymbol{\gamma}_0 - 1}{\mathbb{1}^T\Gamma^{-1}\mathbb{1}}\right)^T \Gamma^{-1}\mathbf{z}.$$

Remark 7.4 (Exact interpolator)
Wackernagel (2003, p. 81) concluded that, in absence of a nugget component, the ordinary kriging predictor $Z^*_{\boldsymbol{\omega}_{OK}}(\mathbf{x}_0)$ is also an exact interpolator similar to simple kriging. This is meant in the sense that if the prediction point $\mathbf{x}_0$ equals a sample point $\mathbf{x}_i$ for $i \in \{1, \ldots, n\}$, then the predicted value coincides with the data value at that point, i.e. $Z^*_{\boldsymbol{\omega}_{OK}}(\mathbf{x}_0) = Z(\mathbf{x}_i)$ if $\mathbf{x}_0 = \mathbf{x}_i$, as well as $z^*_{\boldsymbol{\omega}_{OK}}(\mathbf{x}_0) = z(\mathbf{x}_i)$.

This holds, since $\tilde{\boldsymbol{\gamma}}_0 = (\boldsymbol{\gamma}_0^T, 1)^T = (\gamma(\mathbf{x}_1 - \mathbf{x}_i), \ldots, \gamma(\mathbf{x}_n - \mathbf{x}_i), 1)^T$ is equal to the ith column of the matrix $\tilde{\Gamma}$. Consequently, it follows that the vector $(\omega_1^{OK}, \ldots, \omega_i^{OK}, \ldots, \omega_n^{OK}, \lambda_{OK})^T = (0, \ldots, 1, \ldots, 0, 0)^T$ with $\omega_i^{OK} = 1$ and $\omega_j^{OK} = \lambda_{OK} = 0$ for $j \in \{1, \ldots, n\},\ j \neq i$, is a solution of the ordinary kriging system and, by virtue of Theorem 7.3, also unique.

Hence, we observe $Z^*_{\boldsymbol{\omega}_{OK}}(\mathbf{x}_0) = \sum_{j=1}^{n} \omega_j^{OK} Z(\mathbf{x}_j) = Z(\mathbf{x}_i)$ with ordinary kriging variance $\sigma^2_{OK} \overset{(7.5)}{=} \lambda_{OK} + \sum_{j=1}^{n} \omega_j^{OK} \gamma(\mathbf{x}_j - \mathbf{x}_i) = \gamma(\mathbf{x}_i - \mathbf{x}_i) = 0$.

7.8 Ordinary Kriging in terms of a known covariance

Our next aim is to present all important results of ordinary kriging prediction in terms of the covariance function $C(\mathbf{h})$ of $Z(\mathbf{x})$ for completeness, in particular following Cressie (1993, p. 123).

We will go on rather quickly without giving detailed explanations, since all computations are very similar to the calculations in ordinary kriging in the variogram case before. More details are presented in the books by Cressie (1993, p. 123) and Journel and Huijbregts (1978, pp. 304–307), which rely on Matheron (1971). Consistent with Cressie (1993), we strengthen the stationarity assumption on $Z(\mathbf{x})$ in Assumption 7.1 from instrinsic to second-order stationarity to ensure the existence of the covariance function. This means that we additionally suppose our underlying random function $Z(\mathbf{x})$ to be second-order stationary with known covariance $C(\mathbf{h}) := Cov(Z(\mathbf{x}), Z(\mathbf{x}+\mathbf{h})) = \mathbb{E}[Z(\mathbf{x})Z(\mathbf{x}+\mathbf{h})] - \mu^2\ \forall \mathbf{x}, \mathbf{x}+\mathbf{h} \in D$.

With the same definition of the predictor $Z^*_{\boldsymbol{\omega}}(\mathbf{x}_0)$ in Definition 7.2 and the same unbiasedness condition on the weights (7.1), we can rewrite the ordinary kriging equation

system and the other results. They simply follow by replacing the variogram terms in the computations according to the relation $\gamma(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h})$, given in Proposition 4.5 (p. 15). Hence, Cressie (1993, p. 123) concluded:

$$\sum_{j=1}^{n} \omega_j^{OK} C(\mathbf{x}_i - \mathbf{x}_j) - \lambda_{OK} = C(\mathbf{x}_i - \mathbf{x}_0) \text{ for } i = 1, \ldots, n$$

$$\sum_{j=1}^{n} \omega_j^{OK} = 1,$$

or equally in matrix notation

$$\begin{gathered} \Sigma \boldsymbol{\omega}_{OK} - \lambda_{OK} \mathbb{1} = \mathbf{c}_0 \\ \boldsymbol{\omega}_{OK}^T \mathbb{1} = 1 \\ \Leftrightarrow \\ \underbrace{\left(\begin{array}{ccc|c} & \Sigma & & 1 \\ & & & \vdots \\ & & & 1 \\ \hline 1 & \cdots & 1 & 0 \end{array}\right)}_{=:\tilde{\Sigma} \in \mathbb{R}^{(n+1)\times(n+1)}} \begin{pmatrix} \boldsymbol{\omega}_{OK} \\ -\lambda_{OK} \end{pmatrix} = \underbrace{\begin{pmatrix} \mathbf{c}_0 \\ 1 \end{pmatrix}}_{=:\tilde{\mathbf{c}}_0 \in \mathbb{R}^{n+1}} . \end{gathered} \tag{7.6}$$

This kriging system is identical with the kriging system in the variogram case after replacing Γ by Σ, $\boldsymbol{\gamma}_0$ by $\mathbf{c}_0$ and by multiplying λ_{OK} with -1.

Notice that as long as Σ is assumed to be nonsingular (cf. Assumption 2.9, p. 5), the matrix $\tilde{\Sigma}$ is also nonsingular. This follows by virtue of the Propositions 2.10 and 2.11 (p. 6):

$$\det(\tilde{\Sigma}) = \det(\Sigma) \det\left(0 - \mathbb{1}^T \Sigma^{-1} \mathbb{1}\right) = \underbrace{\det(\Sigma)}_{>0} \underbrace{(-\mathbb{1}^T \Sigma^{-1} \mathbb{1})}_{<0} < 0.$$

Hence, $\begin{pmatrix} \boldsymbol{\omega}_{OK} \\ \lambda_{OK} \end{pmatrix} = \tilde{\Sigma}^{-1} \tilde{\mathbf{c}}_0$ represents a solution of (7.1) and (7.6).

Additionally, similar to ordinary kriging with a variogram, (Cressie, 1993, p. 123) stated the theoretical solution for the "best" weights and the corresponding Lagrange multiplier as follows:

Theorem 7.5 (Solution for ordinary kriging with a known covariance)
The unique solution for $\boldsymbol{\omega}_{OK}$ and λ_{OK} of the ordinary kriging system (7.1) and (7.6) is

$$\boldsymbol{\omega}_{OK} = \Sigma^{-1} \left[\mathbf{c}_0 + \mathbb{1} \left(\frac{1 - \mathbb{1}^T \Sigma^{-1} \mathbf{c}_0}{\mathbb{1}^T \Sigma^{-1} \mathbb{1}} \right) \right],$$

$$\lambda_{OK} = \frac{1 - \mathbb{1}^T \Sigma^{-1} \mathbf{c}_0}{\mathbb{1}^T \Sigma^{-1} \mathbb{1}}.$$

Proof:
At this point we omit the proof, since it follows strictly the line of the proof of Theorem 7.3.

□

Afterwards we can compute the ordinary kriging variance σ^2_{OK}, which is again defined as the variance of the prediction error, by inserting $\boldsymbol{\omega}_{OK}$ and λ_{OK} from Theorem 7.5 in the prediction variance. In accordance with Cressie (1993, p. 123), we observe

$$\sigma^2_{OK} := Var(Z^*_{\boldsymbol{\omega}_{OK}}(\mathbf{x}_0) - Z(\mathbf{x}_0)) = C(\mathbf{0}) + \lambda_{OK} - \boldsymbol{\omega}^T_{OK}\mathbf{c}_0$$

$$= C(\mathbf{0}) - \left(\boldsymbol{\omega}^T_{OK}, -\lambda_{OK}\right)\begin{pmatrix}\mathbf{c}_0\\ 1\end{pmatrix} = C(\mathbf{0}) - \mathbf{c}_0^T\Sigma^{-1}\mathbf{c}_0 + \frac{\left(1 - \mathbb{1}^T\Sigma^{-1}\mathbf{c}_0\right)^2}{\mathbb{1}^T\Sigma^{-1}\mathbb{1}}, \tag{7.7}$$

as well as

$$\sigma^2_{OK} = C(\mathbf{0}) + \sum_{i=1}^{n}\sum_{j=1}^{n}\omega_i^{OK}\omega_j^{OK}C(\mathbf{x}_i - \mathbf{x}_j) - 2\sum_{i=1}^{n}\omega_i^{OK}C(\mathbf{x}_i - \mathbf{x}_0)$$

$$= C(\mathbf{0}) + \lambda_{OK} - \sum_{i=1}^{n}\omega_i^{OK}C(\mathbf{x}_i - \mathbf{x}_0). \tag{7.8}$$

After all, in summary, we can state the ordinary kriging predictor $Z^*_{\boldsymbol{\omega}_{OK}}(\mathbf{x}_0)$ (see Cressie 1990) as

$$Z^*_{\boldsymbol{\omega}_{OK}}(\mathbf{x}_0) = \sum_{i=1}^{n}\omega_i^{OK}Z(\mathbf{x}_i) = \boldsymbol{\omega}^T_{OK}\mathbf{Z} = \left[\mathbf{c}_0 + \mathbb{1}\left(\frac{1 - \mathbb{1}^T\Sigma^{-1}\mathbf{c}_0}{\mathbb{1}^T\Sigma^{-1}\mathbb{1}}\right)\right]^T\Sigma^{-1}\mathbf{Z},$$

with kriging variance in (7.7) and (7.8) and predicted value

$$z^*_{\boldsymbol{\omega}_{OK}}(\mathbf{x}_0) = \sum_{i=1}^{n}\omega_i^{OK}z(\mathbf{x}_i) = \boldsymbol{\omega}^T_{OK}\mathbf{z} = \left[\mathbf{c}_0 + \mathbb{1}\left(\frac{1 - \mathbb{1}^T\Sigma^{-1}\mathbf{c}_0}{\mathbb{1}^T\Sigma^{-1}\mathbb{1}}\right)\right]^T\Sigma^{-1}\mathbf{z}.$$

7.9 Ordinary Kriging in *R*

Finally, we want to show how ordinary kriging can be performed in *R*. Therefore we proceed very similar to simple kriging in the last section. But the main difference is that we do not have to know the "real" mean value μ of the mean temperature in Germany at the prespecified dates 2010/11/28 and 2012/06/09. Hence, we update the gstat objects without the argument *beta* first, and then we go forward exactly as in simple kriging. Thus, we omit most of the *R* code, since it is the same as in simple kriging except for updating the gstat objects.

At the end we obtain again the map of Germany colored once according to the prediction estimates of ordinary kriging, and once according to the minimal kriging variances.

```
> #Ordinary Kriging:
>
> #Update gstat objects without beta:
> g_ok1<-gstat(g1,model=vfit1,id="temp1",formula=temp1~1,
+ locations=~longkm+latkm,data=data1)
> g_ok2<-gstat(g2,model=vfit2,id="temp2",formula=temp2~1,
+ locations=~longkm+latkm,data=data2)

> #Ordinary Kriging Prediction for additional 24 weather stations:
>
> p_ok1<-predict(g_ok1,newdata=newdat)

[using ordinary kriging]

> p_ok2<-predict(g_ok2,newdata=newdat)

[using ordinary kriging]

> #First lines of prediction:
> p_ok1[1:5,] #2010/11/28

         coordinates temp1.pred temp1.var
1    (8.979, 48.216)  -1.052415 0.9595225
2  (9.8044, 51.9672)  -4.404775 0.9078066
3 (13.4367, 54.6817)  -1.692527 1.6435510
4    (7.979, 51.465)  -3.116538 0.8861469
5 (10.9431, 48.4261)  -2.235702 0.8815847

> p_ok2[1:5,] #2012/06/09

         coordinates temp2.pred temp2.var
1    (8.979, 48.216)   15.93937  1.332763
2  (9.8044, 51.9672)   14.57881  1.271012
3 (13.4367, 54.6817)   15.59252  2.103155
4    (7.979, 51.465)   13.20192  1.256312
5 (10.9431, 48.4261)   14.60961  1.267012
```

Note that in this case, the prediction estimates of ordinary kriging as well as the variances are very similar to those of simple kriging. This can be seen in their small differences, especially for 2010/11/28. For their minimal and maximal deviations, see Table 7.1.
Further, we present all residuals of our prediction estimates of the 24 additional stations. We print the differences between the measured temperatures and the prediction estimates for both dates 2010/11/28 and 2012/06/09 in Table 7.2, where the last line provides again the computed sum of the squared residuals, which are quite similar to those obtained from simple kriging.

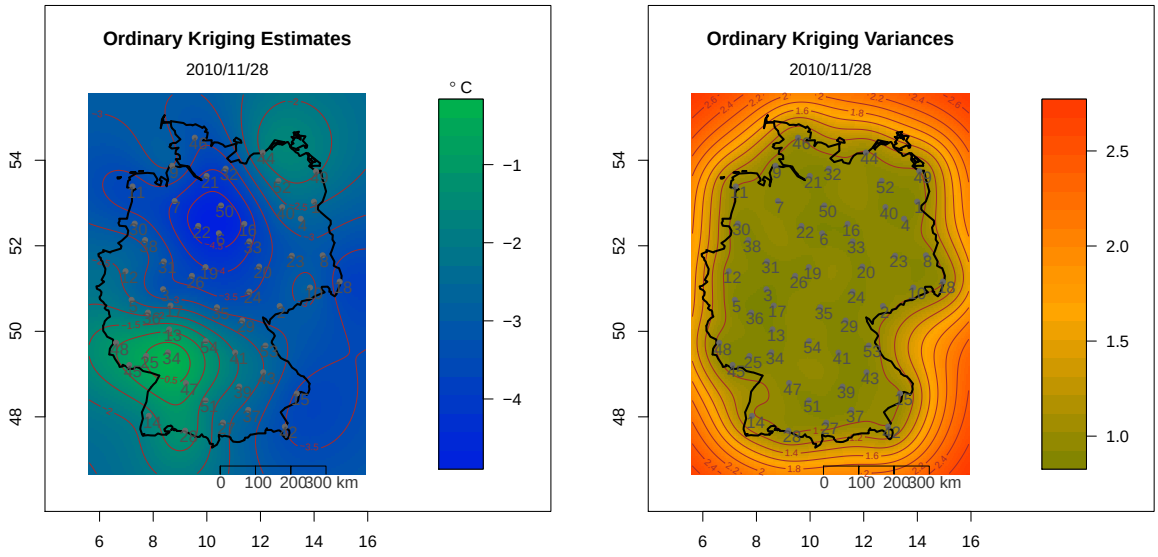
	2010/11/28	2012/06/09
Maximal difference estimates	0.00	0.11
Minimal difference estimates	0.00	0.00
Maximal difference variances	0.04	0.05
Minimal difference variances	0.00	0.00

Table 7.1: Absolute differences of the ordinary kriging and corresponding simple kriging estimates and variances of the additional 24 weather stations of 2010/11/28 and 2012/06/09 in Germany

Longitude	Latitude	Residuals 2010/11/28	Residuals 2012/06/09
8.98	48.22	-3.05	-2.24
9.8	51.97	-1.10	0.32
13.44	54.68	1.49	-1.79
7.98	51.47	0.32	0.10
10.94	48.43	-0.46	0.09
7.31	50.04	-1.50	-2.37
6.7	53.6	1.68	-0.66
9.14	53.45	-1.13	-0.30
9.32	49.52	-0.74	-1.57
14.73	52.02	-0.08	-0.28
10.5	49.85	-0.33	-0.94
10.13	48.99	-1.23	-1.02
12.73	48.48	-0.29	-0.99
10.68	53.58	-0.13	0.03
13.14	49.11	-3.15	-7.03
13.94	53.32	-0.51	-0.32
9.22	50.51	-2.77	-4.03
11.14	52.97	0.73	-0.17
11.27	47.48	-0.86	-3.08
7.64	47.81	-0.06	0.43
11.14	50.5	-2.51	-4.32
10.88	51.67	-3.81	-3.12
8.57	52.45	0.48	0.26
12.46	52.12	-0.53	-1.07
Sum of Squares		62.16	124.95

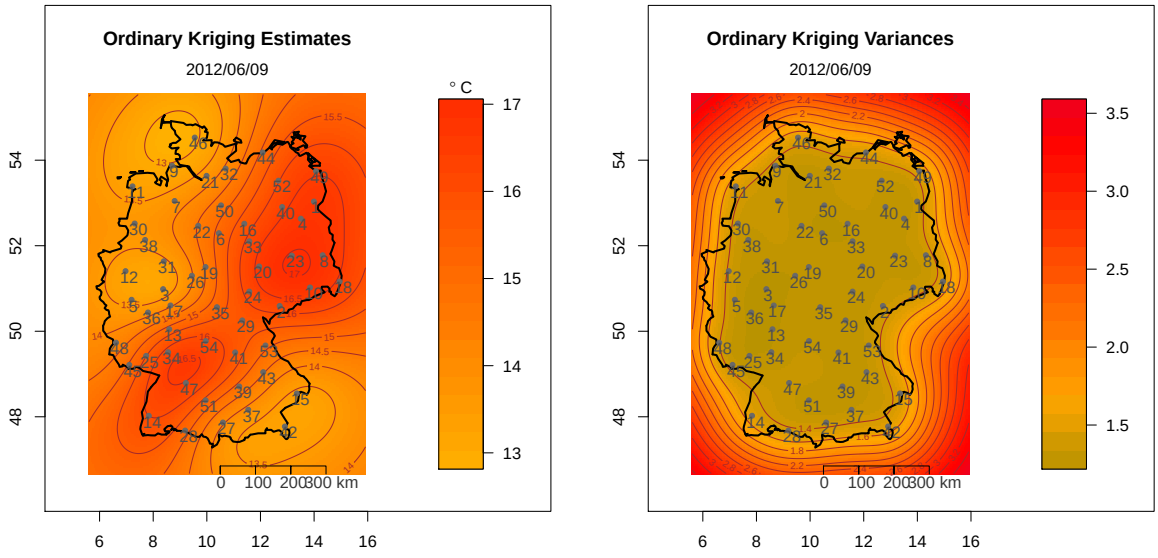
Table 7.2: Residuals from ordinary kriging prediction of the additional 24 weather stations in Germany, where each residual equals the difference of the observed value and the prediction estimate obtained from ordinary kriging; sum of squares is the sum of all squared residuals

And lastly, by using the same grid as in the section about simple kriging, we obtain our final result in form of a plot into the map of Germany of both dates, see Figure 7.1 for 2010/11/28 and Figure 7.2 for 2012/06/09, where the predicted temperature values are printed on the left and the corresponding prediction variances on the right. Notice that the plots of ordinary and simple kriging seem to be nearly the same, which is caused by very similar prediction estimates and variances for the coordinates of our grid.



(a) Ordinary Kriging Estimates (b) Ordinary Kriging Variances

Figure 7.1: Ordinary Kriging applied to the temperature data of 2010/11/28 in Germany



(a) Ordinary Kriging Estimates (b) Ordinary Kriging Variances

Figure 7.2: Ordinary Kriging applied to the temperature data of 2012/06/09 in Germany

8 Universal Kriging

Simple and ordinary kriging assume a stationary, i.e. constant mean μ of the underlying real-valued random function $Z(\mathbf{x})$. But in reality, the mean value of some spatial data cannot be assumed constant in general, it varies, since it also depends on the absolute location of the sample. For instance, the average temperature in Germany and Spain will probably differ.

For this sake, we want to introduce the spatial prediction method *universal kriging*, whose aim is to predict $Z(\mathbf{x})$ at unsampled places as well. It splits the random function into a linear combination of deterministic functions, the smoothly varying and nonstationary trend, or also called drift $\mu(\mathbf{x}) \in \mathbb{R}$, and a random component $Y(\mathbf{x}) := Z(\mathbf{x}) - \mu(\mathbf{x})$ representing the residual random function (Wackernagel 2003, p. 300). Thus, using universal kriging, we can take a trend in the mean over the spatial region D in our prediction into account, such that the variation in $Z(\mathbf{x})$ contains a systematic component in addition to the random one. For this reason universal kriging is also called *kriging in the presence of a drift* or even *kriging with a trend* and represents a generalization of ordinary kriging (Cressie 1993, p. 151).

We introduce this kriging method with a trend model in the upcoming section, following closely the books by Cressie (1993, pp. 151–156, 361–362) and Webster and Oliver (2007, pp. 195–200). The key results can also be found for instance in Wackernagel (2003, Chapter 38), Journel and Huijbregts (1978, pp. 313–320, 564–565) and Kitanidis (1997, Chapter 6). We also recommend Matheron (1971) in Chapter 4, where he even called it universal kriging.

8.1 Model for Universal Kriging

As mentioned above, there are a few model assumptions on the underlying random function $Z(\mathbf{x})$ for universal kriging varying from the assumptions of simple and ordinary kriging. These are presented in Cressie (1993, pp. 151, 361) and Webster and Oliver (2007, pp. 195–196) and we summarize them in the following assumption:

Assumption 8.1 (Model for Universal Kriging)

(i) Assume that $Z(\mathbf{x})$ can be decomposed into a deterministic, i.e. nonrandom trend function $\mu(\mathbf{x})$, and a real-valued residual random function $Y(\mathbf{x})$, such that

$$Z(\mathbf{x}) = \mu(\mathbf{x}) + Y(\mathbf{x}).$$

(ii) $Y(\mathbf{x})$ is supposed to be intrinsically stationary with zero mean and variogram function $\gamma_Y(\mathbf{h})$, called *residual variogram function* of $Z(\mathbf{x})$, i.e. $\forall \mathbf{x},\ \mathbf{x}+\mathbf{h} \in D$:

$$I\!E[Z(\mathbf{x})] = I\!E[\mu(\mathbf{x})] + \underbrace{I\!E[Y(\mathbf{x})]}_{=0} = \mu(\mathbf{x}) \text{ and}$$

$$\gamma_Y(\mathbf{h}) = \frac{1}{2} Var(Y(\mathbf{x}+\mathbf{h}) - Y(\mathbf{h})) = \frac{1}{2} I\!E[(Y(\mathbf{x}+\mathbf{h}) - Y(\mathbf{x}))^2].$$

(iii) Finally let $f_0, f_1, \ldots, f_L$ be deterministic functions of the geographical coordinates $\mathbf{x} \in D$ with $L \in \mathbb{N}$ the number of known and selectable basic functions $f_l : D \to \mathbb{R},\ l = 0, \ldots, L$. We assume $\mu(\mathbf{x})$ to be a linear combination of these functions evaluated at $\mathbf{x}$

$$\mu(\mathbf{x}) = \sum_{l=0}^{L} a_l f_l(\mathbf{x})$$

with unknown coefficients $a_l \in \mathbb{R}\backslash\{0\}$ for all $l = 0, \ldots, L$ and suppose $f_0(\mathbf{x}) = 1$ constant by convention.

Denote the drift coefficient vector by $\mathbf{a} := (a_0, a_1, \ldots, a_L)^T \in \mathbb{R}^{L+1}$ and let

$$F := \begin{pmatrix} 1 & f_1(\mathbf{x}_1) & \cdots & f_L(\mathbf{x}_1) \\ 1 & f_1(\mathbf{x}_2) & \cdots & f_L(\mathbf{x}_2) \\ \vdots & \vdots & \cdots & \vdots \\ 1 & f_1(\mathbf{x}_n) & \cdots & f_L(\mathbf{x}_n) \end{pmatrix} \in \mathbb{R}^{n \times (L+1)}, \text{ i.e. } F_{i,l+1} = f_l(\mathbf{x}_i),\ i = 1, \ldots, n \text{ and } l = 0, 1, \ldots, L.$$

In accordance with the assumptions above, Cressie (1993, p. 151) observed

$$Z(\mathbf{x}_i) = \mu(\mathbf{x}_i) + Y(\mathbf{x}_i) = \sum_{l=0}^{L} a_l f_l(\mathbf{x}_i) + Y(\mathbf{x}_i) = (F\mathbf{a} + \mathbf{Y})_i,\ i = 1, \ldots, n.$$

Hence, we obtain for the random vector $\mathbf{Z}$:

$$\mathbf{Z} = \begin{pmatrix} Z(\mathbf{x}_1) \\ \vdots \\ Z(\mathbf{x}_n) \end{pmatrix} = \underbrace{\begin{pmatrix} f_0(\mathbf{x}_1) & \cdots & f_L(\mathbf{x}_1) \\ \vdots & \ldots & \vdots \\ f_0(\mathbf{x}_n) & \cdots & f_L(\mathbf{x}_n) \end{pmatrix}}_{=F} \underbrace{\begin{pmatrix} a_0 \\ \vdots \\ a_L \end{pmatrix}}_{=\mathbf{a}} + \underbrace{\begin{pmatrix} Y(\mathbf{x}_1) \\ \vdots \\ Y(\mathbf{x}_n) \end{pmatrix}}_{=\mathbf{Y}} = F\mathbf{a} + \mathbf{Y}.$$

Remark 8.2 (Special case)
If $L = 0$, the drift component $\mu(\mathbf{x})$ reduces to the single term a_0 and consequently

$$Z(\mathbf{x}) = \mu(\mathbf{x}) + Y(\mathbf{x}) = a_0 + Y(\mathbf{x})$$

with unknown but constant mean a_0. Note that this is identical with the setting of ordinary kriging (Webster and Oliver 2007, p. 196).

Furthermore, we define the linear predictor for universal kriging as it is done in Cressie (1993, p. 152) and the same way as in ordinary kriging:

Definition 8.3 (Predictor for Universal Kriging)
The *universal kriging predictor* $Z^*_{\boldsymbol{\omega}}(\mathbf{x}_0)$ of the value of $Z(\mathbf{x})$ at the target point $\mathbf{x}_0$ is the linear sum

$$Z^*_{\boldsymbol{\omega}}(\mathbf{x}_0) := \sum_{i=1}^{n} \omega_i Z(\mathbf{x}_i) = \boldsymbol{\omega}^T \mathbf{Z}$$

with weights $\omega_i \in \mathbb{R},\ i = 1, \ldots, n$, corresponding to each evaluation of the random function $Z(\mathbf{x})$ at the sample point $\mathbf{x}_i$ and $\boldsymbol{\omega} := (\omega_1, \ldots, \omega_n)^T \in \mathbb{R}^n$.

Consistent with this definition and $\mathbf{Z} = F\mathbf{a} + \mathbf{Y}$, we obtain

$$Z^*_{\boldsymbol{\omega}}(\mathbf{x}_0) = \sum_{i=1}^{n} \omega_i \left(\underbrace{\sum_{l=0}^{L} a_l f_l(\mathbf{x}_i)}_{=\mu(\mathbf{x}_i)} + Y(\mathbf{x}_i) \right) = \boldsymbol{\omega}^T (F\mathbf{a} + \mathbf{Y}).$$

8.2 Unbiasedness condition

In the next step, we have to consider several conditions on the weights ω_i to ensure uniform unbiasedness of our linear predictor $Z^*_{\boldsymbol{\omega}}(\mathbf{x}_0)$. This means we want to avoid systematic bias in any situation. It follows by Cressie (1993, p. 152) and Wackernagel (2003, p. 301):

$$\mathbb{E}[Z^*_{\boldsymbol{\omega}}(\mathbf{x}_0) - Z(\mathbf{x}_0)] = \sum_{i=1}^{n} \omega_i \left(\mathbb{E}[\mu(\mathbf{x}_i)] + \underbrace{\mathbb{E}[Y(\mathbf{x}_i)]}_{=0} \right) - \left(\mathbb{E}[\mu(\mathbf{x}_0)] + \underbrace{\mathbb{E}[Y(\mathbf{x}_0)]}_{=0} \right)$$

$$\overset{\mu(\mathbf{x}) \text{ deterministic}}{=} \sum_{i=1}^{n} \omega_i \mu(\mathbf{x}_i) - \mu(\mathbf{x}_0) \overset{!}{=} 0$$

$$\Leftrightarrow \sum_{l=0}^{L} a_l \left(\sum_{i=1}^{n} \omega_i f_l(\mathbf{x}_i) - f_l(\mathbf{x}_0) \right) = 0.$$

Together with $a_l \neq 0$ in Assumption 8.1 (iii) and $\mathbf{f}_0 := (1, f_1(\mathbf{x}_0), \ldots, f_L(\mathbf{x}_0))^T \in \mathbb{R}^{L+1}$, the general unbiasedness of $Z^*_{\boldsymbol{\omega}}(\mathbf{x}_0)$ is satisfied if and only if (Kitanidis 1997, pp. 125–126)

$$\sum_{i=1}^{n} \omega_i f_l(\mathbf{x}_i) = f_l(\mathbf{x}_0) \text{ for } l = 0, \ldots, L \;\Leftrightarrow\; F^T \boldsymbol{\omega} = \mathbf{f}_0, \tag{8.1}$$

which Matheron (1971) named *universality conditions*.

The above equivalence simply follows by Cressie (1993, p. 152), since $\mathbb{E}[Z(\mathbf{x}_0)] = \mu(\mathbf{x}_0) = \sum_{l=0}^{L} a_l f_l(\mathbf{x}_0) = \mathbf{f}_0^T \mathbf{a}$ and $\mathbb{E}[Z^*_{\boldsymbol{\omega}}(\mathbf{x}_0)] = \mathbb{E}[\boldsymbol{\omega}^T \mathbf{Z}] = \boldsymbol{\omega}^T \mathbb{E}[\mathbf{Z}] = \boldsymbol{\omega}^T F\mathbf{a}$, where the expectation of $\mathbf{Z}$ is taken componentwise. Hence, if $\mathbb{E}[Z^*_{\boldsymbol{\omega}}(\mathbf{x}_0)] = \mathbb{E}[Z(\mathbf{x}_0)]$ should hold in general, it follows

$$\boldsymbol{\omega}^T F\mathbf{a} = \mathbf{f}_0^T \mathbf{a} \;\; \forall \mathbf{a} \in \mathbb{R}^{L+1}, \; a_l \neq 0 \; \forall l = 0, 1, \ldots, L$$

$$\Leftrightarrow \;\; \boldsymbol{\omega}^T F = \mathbf{f}_0^T.$$

Remark:
For the constant function $f_0(\mathbf{x}) = 1$ for $l = 0$ in (8.1), we observe the usual, previous condition on the weights, $\sum_{i=1}^{n} \omega_i = 1$, as in ordinary kriging or even in kriging the mean (Wackernagel 2003, p. 301).

8.3 Variance of the prediction error

In the following we want to compute the variance of the prediction error $Z^*_{\boldsymbol{\omega}}(\mathbf{x}_0) - Z(\mathbf{x}_0)$, which "acts" again as a measure of the accuracy of our linear predictor. It can be calculated by inserting the residual variogram function $\gamma_Y(\mathbf{h})$. Following Cressie (1993, pp. 152–153), we derive:

$$\sigma_E^2 := Var(Z^*_{\boldsymbol{\omega}}(\mathbf{x}_0) - Z(\mathbf{x}_0)) \overset{(8.1)}{=} \mathbb{E}[(Z^*_{\boldsymbol{\omega}}(\mathbf{x}_0) - Z(\mathbf{x}_0))^2]$$

$$= \mathbb{E}\left[\left(\sum_{i=1}^n \omega_i Z(\mathbf{x}_i) - Z(\mathbf{x}_0)\right)^2\right] = \mathbb{E}\left[\left(\sum_{i=1}^n \omega_i Y(\mathbf{x}_i) - Y(\mathbf{x}_0) + \underbrace{\sum_{i=1}^n \omega_i \mu(\mathbf{x}_i) - \mu(\mathbf{x}_0)}_{=0}\right)^2\right]$$

$$= \sum_{i=1}^n \sum_{j=1}^n \omega_i \omega_j \mathbb{E}[Y(\mathbf{x}_i)Y(\mathbf{x}_j)] - 2\sum_{i=1}^n \omega_i \mathbb{E}[Y(\mathbf{x}_0)Y(\mathbf{x}_i)] + \mathbb{E}[(Y(\mathbf{x}_0))^2]$$

$$\overset{(**)}{=} -\sum_{i=1}^n \sum_{j=1}^n \omega_i \omega_j \underbrace{\frac{\mathbb{E}[(Y(\mathbf{x}_i) - Y(\mathbf{x}_j)^2]}{2}}_{=\gamma_Y(\mathbf{x}_i - \mathbf{x}_j)} + 2\sum_{i=1}^n \omega_i \underbrace{\frac{\mathbb{E}[(Y(\mathbf{x}_i) - Y(\mathbf{x}_0)^2]}{2}}_{=\gamma_Y(\mathbf{x}_i - \mathbf{x}_0)}$$

$$= -\sum_{i=1}^n \sum_{j=1}^n \omega_i \omega_j \gamma_Y(\mathbf{x}_i - \mathbf{x}_j) + 2\sum_{i=1}^n \omega_i \gamma_Y(\mathbf{x}_i - \mathbf{x}_0).$$

Hence, the prediction variance σ_E^2 is

$$\sigma_E^2 = -\sum_{i=1}^n \sum_{j=1}^n \omega_i \omega_j \gamma_Y(\mathbf{x}_i - \mathbf{x}_j) + 2\sum_{i=1}^n \omega_i \gamma_Y(\mathbf{x}_i - \mathbf{x}_0)$$

$$= -\boldsymbol{\omega}^T \Gamma_Y \boldsymbol{\omega} + 2\boldsymbol{\omega}^T \boldsymbol{\gamma}_{Y,0} \geq 0$$

with symmetric residual variogram matrix $\Gamma_Y \in \mathbb{R}^{n \times n}$, $(\Gamma_Y)_{i,j} := \gamma_Y(\mathbf{x}_i - \mathbf{x}_j)$, $i, j = 1, \ldots, n$ and $\boldsymbol{\gamma}_{Y,0} := (\gamma_Y(\mathbf{x}_1 - \mathbf{x}_0), \ldots, \gamma_Y(\mathbf{x}_n - \mathbf{x}_0))^T \in \mathbb{R}^n$.

The nonnegativity of the prediction variance is due to the similar representation of σ_E^2 as in ordinary kriging in the last section, simply by replacing the variogram terms $\gamma(\mathbf{h})$ with the terms of the residual variogram function $\gamma_Y(\mathbf{h})$.

Note that similar to ordinary kriging, identity $(**)$ only holds as long as $\sum_{i=1}^n \omega_i = 1$, which displays the unbiasedness condition for $l = 0$ in (8.1). Since $(**)$ is the same equation as $(*)$ applied to $Y(\mathbf{x})$ instead of $Z(\mathbf{x})$ (cf. ordinary kriging, p. 53), we have nothing to show. Remember that in $(*)$ the random function $Z(\mathbf{x})$ was supposed to be intrinsically stationary with constant mean μ and variogram function $\gamma(\mathbf{h})$. In our current setting, $Y(\mathbf{x})$ has constant mean 0 and is intrinsically stationary with variogram $\gamma_Y(\mathbf{h})$. This justifies the application of $(*)$ in this situation.

8.4 Minimal prediction variance

Afterwards, in order to minimize the prediction error variance σ_E^2, i.e. to maximize the precision of our linear predictor $Z_{\boldsymbol{\omega}}^*(\mathbf{x}_0)$, we have to solve the following constrained optimization problem given by Cressie (1993, pp. 152–153):

$$\text{minimum of } -\boldsymbol{\omega}^T\Gamma_Y\boldsymbol{\omega} + 2\boldsymbol{\omega}^T\boldsymbol{\gamma}_{Y,0} \text{ subject to } \boldsymbol{\omega}^T F = \mathbf{f}_0^T.$$

We can derive a solution of this kind of problem again by using the method of Lagrange multipliers similarly to the last sections. Therefore, we define the function φ as σ_E^2 plus an additional term involving the Lagrange parameters to guarantee the uniform unbiasedness, i.e.

$$\varphi : \mathbb{R}^n \times \mathbb{R}^{L+1} \to \mathbb{R}$$

$$(\boldsymbol{\omega}, \boldsymbol{\lambda}) \mapsto \varphi(\boldsymbol{\omega}, \boldsymbol{\lambda}) := -\boldsymbol{\omega}^T\Gamma_Y\boldsymbol{\omega} + 2\boldsymbol{\omega}^T\boldsymbol{\gamma}_{Y,0} - 2(\boldsymbol{\omega}^T F - \mathbf{f}_0^T)\boldsymbol{\lambda},$$

with Lagrange parameter vector $\boldsymbol{\lambda} := (\lambda_0, \lambda_1, \ldots, \lambda_L)^T \in \mathbb{R}^{L+1}$ providing the $L+1$ Lagrange multipliers for each single condition in (8.1) (see Cressie 1993, p. 152).

In the first step, we set the first order partial derivatives of φ with respect to $\boldsymbol{\omega}$ and $\boldsymbol{\lambda}$ to zero, which yield the "critical points", i.e. the necessary conditions on the parameters of our optimization problem to give a minimum.

First, we obtain by differentiating φ with respect to the weight vector $\boldsymbol{\omega}$

$$\frac{\partial\varphi}{\partial\boldsymbol{\omega}}(\boldsymbol{\omega}, \boldsymbol{\lambda}) = -2\Gamma_Y\boldsymbol{\omega} + 2\boldsymbol{\gamma}_{Y,0} - 2F\boldsymbol{\lambda} \stackrel{!}{=} 0$$

$$\Leftrightarrow \Gamma_Y\boldsymbol{\omega} + F\boldsymbol{\lambda} = \boldsymbol{\gamma}_{Y,0}.$$

And second, the partial derivative with respect to the Lagrange parameter $\boldsymbol{\lambda}$ yields

$$\frac{\partial\varphi}{\partial\boldsymbol{\lambda}}(\boldsymbol{\omega}, \boldsymbol{\lambda}) = -2(\boldsymbol{\omega}^T F - \mathbf{f}_0^T) \stackrel{!}{=} 0$$

$$\Leftrightarrow F^T\boldsymbol{\omega} = \mathbf{f}_0 \Leftrightarrow (8.1).$$

At this point, as in ordinary kriging in the last section, we omit the proof that these necessary conditions indeed yield the minimum. The interested reader may also find a proof in the Appendix (p. 94) in the case of the existence of the covariance. The proof will use again Theorem 2.15 by Rao (1973). But we prefer to follow this Lagrange approach for further computations. The reasons are the same as for ordinary kriging and are mentioned later in the Appendix.

Hence, we conclude that the prediction variance σ_E^2 is minimal if the necessary equations above hold, which are called *optimality conditions* by Matheron (1971).

8.5 Equations for Universal Kriging

After taken the constraints on the weights ω_i to ensure the uniform unbiasedness and to achieve minimal observation variance of our prediction into account, we can write the system for universal kriging as it is stated by Cressie (1993, pp. 153–154) and Webster and Oliver (2007, p. 197–198):

$$\sum_{j=1}^{n} \omega_j^{UK} \gamma(\mathbf{x}_i - \mathbf{x}_j) + \sum_{l=0}^{L} \lambda_l^{UK} f_l(\mathbf{x}_i) = \gamma(\mathbf{x}_i - \mathbf{x}_0) \text{ for } i = 1, \dots, n$$

$$\sum_{j=1}^{n} \omega_j^{UK} f_l(\mathbf{x}_j) = f_l(\mathbf{x}_0) \text{ for } l = 0, 1, \dots, L$$

with "optimal" kriging weights $\omega_i^{UK} \in \mathbb{R},\ i = 1, \dots, n$ and Lagrange parameters $\lambda_l^{UK} \in \mathbb{R},\ l = 0, 1, \dots, L$. This is as well as in matrix notation

$$\Gamma_Y \boldsymbol{\omega}_{UK} + F \boldsymbol{\lambda}_{UK} = \boldsymbol{\gamma}_{Y,0} \tag{8.2}$$

$$F^T \boldsymbol{\omega}_{UK} = \mathbf{f}_0$$

$$\Leftrightarrow$$

$$\begin{pmatrix} \Gamma_Y & F \\ F^T & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\omega}_{UK} \\ \boldsymbol{\lambda}_{UK} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\gamma}_{Y,0} \\ \mathbf{f}_0 \end{pmatrix}$$

with universal kriging weight vector $\boldsymbol{\omega}_{UK} := (\omega_1^{UK}, \dots, \omega_n^{UK})^T \in \mathbb{R}^n$, Lagrange parameter vector $\boldsymbol{\lambda}_{UK} := (\lambda_0^{UK}, \lambda_1^{UK}, \dots, \lambda_L^{UK})^T \in \mathbb{R}^{L+1}$, $0 \in \mathbb{R}^{(L+1)\times(L+1)}$ and symmetric block matrix

$$\begin{pmatrix} \Gamma_Y & F \\ F^T & 0 \end{pmatrix} = \left(\begin{array}{ccc|cccc} \gamma_Y(\mathbf{x}_1 - \mathbf{x}_1) & \cdots & \gamma_Y(\mathbf{x}_1 - \mathbf{x}_n) & 1 & f_1(\mathbf{x}_1) & \cdots & f_L(\mathbf{x}_1) \\ \gamma_Y(\mathbf{x}_2 - \mathbf{x}_1) & \cdots & \gamma_Y(\mathbf{x}_2 - \mathbf{x}_n) & 1 & f_1(\mathbf{x}_2) & \cdots & f_L(\mathbf{x}_2) \\ \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ \gamma_Y(\mathbf{x}_n - \mathbf{x}_1) & \cdots & \gamma_Y(\mathbf{x}_n - \mathbf{x}_n) & 1 & f_1(\mathbf{x}_n) & \cdots & f_L(\mathbf{x}_n) \\ \hline 1 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ f_1(\mathbf{x}_1) & \cdots & f_1(\mathbf{x}_n) & 0 & 0 & \cdots & 0 \\ \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ f_L(\mathbf{x}_1) & \cdots & f_L(\mathbf{x}_n) & 0 & 0 & \cdots & 0 \end{array}\right)$$

$\in \mathbb{R}^{(n+L+1)\times(n+L+1)}$.

It follows immediately that the matrix F needs a full column rank if the system should have an unique solution. This gives a condition on the choice of the deterministic functions f_l, as for instance it implies that f_0 is the single constant function among all f_l (see Wackernagel 2003, p. 301).

In the case where the whole block matrix $\begin{pmatrix} \Gamma & F \\ F^T & 0 \end{pmatrix}$ is invertible, which is only possible if the number of observations is larger than or equal to the number of basic functions, i.e.

$n \geq L+1$, we can compute the universal kriging weights and the Lagrange parameter according to Webster and Oliver (2007, pp. 197–198) by

$$\begin{pmatrix} \boldsymbol{\omega}_{UK} \\ \boldsymbol{\lambda}_{UK} \end{pmatrix} = \begin{pmatrix} \Gamma_Y & F \\ F^T & 0 \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{\gamma}_{Y,0} \\ \mathbf{f}_0 \end{pmatrix}.$$

Additionally, one can solve this linear system of $n+L+1$ equations with $n+L+1$ unknown variables ω_i^{UK} and λ_l^{UK} canceling out the variables step by step. But there exists another common way how this solution is stated in literature, too. We provide this solution in the next theorem below, whose results can be found for instance in Cressie (1993, p. 153):

Theorem 8.4 (Solution of the universal kriging system)
Under the assumption of invertibility of Γ_Y and $F^T\Gamma_Y^{-1}F$, the solution for $\boldsymbol{\omega}_{UK}$ and $\boldsymbol{\lambda}_{UK}$ of the universal kriging system (8.1) and (8.2) is uniquely determined by

$$\boldsymbol{\omega}_{UK} = \Gamma_Y^{-1}\left[\boldsymbol{\gamma}_{Y,0} - F\left(F^T\Gamma_Y^{-1}F\right)^{-1}\left(F^T\Gamma_Y^{-1}\boldsymbol{\gamma}_{Y,0} - \mathbf{f}_0\right)\right], \tag{8.3}$$

$$\boldsymbol{\lambda}_{UK} = \left(F^T\Gamma_Y^{-1}F\right)^{-1}\left(F^T\Gamma_Y^{-1}\boldsymbol{\gamma}_{Y,0} - \mathbf{f}_0\right). \tag{8.4}$$

Proof:
First of all note that $\boldsymbol{\omega}_{UK}$ and $\boldsymbol{\lambda}_{UK}$ in Theorem 8.4 are well-defined due to the assumption of nonsingularity of Γ_Y and $F^T\Gamma_Y^{-1}F$. Further, they actually represent a solution of the universal kriging system (8.1) and (8.2), which we can simply verify by

(i) $\Gamma_Y\boldsymbol{\omega}_{UK} + F\boldsymbol{\lambda}_{UK}$
$= \left[\boldsymbol{\gamma}_{Y,0} - F\left(F^T\Gamma_Y^{-1}F\right)^{-1}\left(F^T\Gamma_Y^{-1}\boldsymbol{\gamma}_{Y,0} - \mathbf{f}_0\right)\right] + F\left[\left(F^T\Gamma_Y^{-1}F\right)^{-1}\left(F^T\Gamma_Y^{-1}\boldsymbol{\gamma}_{Y,0} - \mathbf{f}_0\right)\right]$
$= \boldsymbol{\gamma}_{Y,0} - F\left(F^T\Gamma_Y^{-1}F\right)^{-1}\underbrace{\left(F^T\Gamma_Y^{-1}\boldsymbol{\gamma}_{Y,0} - \mathbf{f}_0 - F^T\Gamma_Y^{-1}\boldsymbol{\gamma}_{Y,0} + \mathbf{f}_0\right)}_{=0} = \boldsymbol{\gamma}_{Y,0}$ and

(ii) $F^T\boldsymbol{\omega}_{UK}$
$= F^T\Gamma_Y^{-1}\left[\boldsymbol{\gamma}_{Y,0} - F\left(F^T\Gamma_Y^{-1}F\right)^{-1}\left(F^T\Gamma_Y^{-1}\boldsymbol{\gamma}_{Y,0} - \mathbf{f}_0\right)\right]$
$= F^T\Gamma_Y^{-1}\boldsymbol{\gamma}_{Y,0} - \underbrace{F^T\Gamma_Y^{-1}F\left(F^T\Gamma_Y^{-1}F\right)^{-1}}_{=Id_{L+1}}\left(F^T\Gamma_Y^{-1}\boldsymbol{\gamma}_{Y,0} - \mathbf{f}_0\right) = \mathbf{f}_0.$

Since Γ_Y and $F^T\Gamma_Y^{-1}F$ are assumed to be invertible, the solution for $\boldsymbol{\omega}_{UK}$ and $\boldsymbol{\lambda}_{UK}$ is unique.

□

8.6 Universal Kriging Variance

After these steps, we are able to specify the minimal prediction variance, the *universal kriging variance* σ_{UK}^2. It is defined as the variance of the "optimal" linear predictor minus the random variable to be predicted, $Z^*_{\boldsymbol{\omega}_{UK}}(\mathbf{x}_0) - Z(\mathbf{x}_0)$. We achieve σ_{UK}^2 by inserting the kriging equations (8.1), (8.2) and Theorem 8.4 in the terms of σ_E^2 (see 8.3). Hence, we conclude following Cressie (1993, p. 154):

$$\sigma^2_{UK} := Var(Z^*_{\boldsymbol{\omega}_{UK}}(\mathbf{x}_0) - Z(\mathbf{x}_0)) = -\boldsymbol{\omega}^T_{UK}\underbrace{\Gamma_Y\boldsymbol{\omega}_{UK}}+2\boldsymbol{\omega}^T_{UK}\boldsymbol{\gamma}_{Y,0}$$

$$\overset{(8.2)}{=} \boldsymbol{\omega}^T_{UK}(F\boldsymbol{\lambda}_{UK} - \boldsymbol{\gamma}_{Y,0}) + 2\boldsymbol{\omega}^T_{UK}\boldsymbol{\gamma}_{Y,0} = \underbrace{\boldsymbol{\omega}^T_{UK}F}\,\boldsymbol{\lambda}_{UK} + \boldsymbol{\omega}^T_{UK}\boldsymbol{\gamma}_{Y,0}$$

$$\overset{(8.1)}{=} \mathbf{f}_0^T\boldsymbol{\lambda}_{UK} + \boldsymbol{\omega}^T_{UK}\boldsymbol{\gamma}_{Y,0} = \left(\boldsymbol{\omega}^T_{UK}, \boldsymbol{\lambda}^T_{UK}\right)\begin{pmatrix}\boldsymbol{\gamma}_{Y,0}\\ \mathbf{f}_0\end{pmatrix}$$

$$\overset{(8.3),(8.4)}{=} \left[(F^T\Gamma_Y^{-1}F)^{-1}(F^T\Gamma_Y^{-1}\boldsymbol{\gamma}_{Y,0} - \mathbf{f}_0)\right]^T\mathbf{f}_0$$

$$+ \left[\boldsymbol{\gamma}_{Y,0} - F(F^T\Gamma_Y^{-1}F)^{-1}(F^T\Gamma_Y^{-1}\boldsymbol{\gamma}_{Y,0} - \mathbf{f}_0)\right]^T\Gamma_Y^{-1}\boldsymbol{\gamma}_{Y,0}$$

$$= \left(F^T\Gamma_Y^{-1}\boldsymbol{\gamma}_{Y,0} - \mathbf{f}_0\right)^T\left(F^T\Gamma_Y^{-1}F\right)^{-1}\mathbf{f}_0 + \boldsymbol{\gamma}^T_{Y,0}\Gamma_Y^{-1}\boldsymbol{\gamma}_{Y,0}$$

$$- \left(F^T\Gamma_Y^{-1}\boldsymbol{\gamma}_{Y,0} - \mathbf{f}_0\right)^T\left(F^T\Gamma_Y^{-1}F\right)^{-1}F^T\Gamma_Y^{-1}\boldsymbol{\gamma}_{Y,0}.$$

This holds, since Γ_Y and $F^T\Gamma_Y^{-1}F$, and thus their inverse matrices, are symmetric. Hence, we obtain for the minimized kriging variance according to Cressie (1993, p. 154):

$$\sigma^2_{UK} = \left(\boldsymbol{\omega}^T_{UK}, \boldsymbol{\lambda}^T_{UK}\right)\begin{pmatrix}\boldsymbol{\gamma}_{Y,0}\\ \mathbf{f}_0\end{pmatrix}$$

$$= \boldsymbol{\gamma}^T_{Y,0}\Gamma_Y^{-1}\boldsymbol{\gamma}_{Y,0} - (F^T\Gamma_Y^{-1}\boldsymbol{\gamma}_{Y,0} - \mathbf{f}_0)^T(F^T\Gamma_Y^{-1}F)^{-1}(F^T\Gamma_Y^{-1}\boldsymbol{\gamma}_{Y,0} - \mathbf{f}_0)$$

$$= -\sum_{i=1}^n\sum_{j=1}^n\omega_i^{UK}\omega_j^{UK}\gamma_Y(\mathbf{x}_i - \mathbf{x}_j) + 2\sum_{i=1}^n\omega_i^{UK}\gamma_Y(\mathbf{x}_i - \mathbf{x}_0)$$

$$= \sum_{i=1}^n\omega_i^{UK}\gamma_Y(\mathbf{x}_i - \mathbf{x}_0) + \sum_{l=0}^L\lambda_l f_l(\mathbf{x}_0). \tag{8.5}$$

8.7 Universal Kriging Prediction

In total, we observe the "optimal", that is to say the *best linear unbiased predictor (BLUP)* $Z^*_{\boldsymbol{\omega}_{UK}}(\mathbf{x}_0)$ of universal kriging by inserting the above results of Cressie (1993):

$$Z^*_{\boldsymbol{\omega}_{UK}}(\mathbf{x}_0) = \sum_{i=1}^n\omega_i^{UK}Z(\mathbf{x}_i) = \boldsymbol{\omega}^T_{UK}\mathbf{Z} = \left[\boldsymbol{\gamma}_{Y,0} - F(F^T\Gamma_Y^{-1}F)^{-1}(F^T\Gamma_Y^{-1}\boldsymbol{\gamma}_{Y,0} - \mathbf{f}_0)\right]^T\Gamma_Y^{-1}\mathbf{Z}$$

with corresponding kriging variance computed in (8.5) and finally with kriging estimate $z^*_{\boldsymbol{\omega}_{UK}}(\mathbf{x}_0)$ at the location of interest $\mathbf{x}_0$

$$z^*_{\boldsymbol{\omega}_{UK}}(\mathbf{x}_0) = \sum_{i=1}^n\omega_i^{UK}z(\mathbf{x}_i) = \boldsymbol{\omega}^T_{UK}\mathbf{z} = \left[\boldsymbol{\gamma}_{Y,0} - F(F^T\Gamma_Y^{-1}F)^{-1}(F^T\Gamma_Y^{-1}\boldsymbol{\gamma}_{Y,0} - \mathbf{f}_0)\right]^T\Gamma_Y^{-1}\mathbf{z}.$$

Remark 8.5 (Exact interpolator)
As in simple and ordinary kriging, the universal kriging predictor is also an exact interpolator in theory (see Cressie 1993, p. 360), as in the case that the prediction point $\mathbf{x}_0$ equals a data point $\mathbf{x}_i$ for $i \in \{1, \ldots, n\}$, it follows that $Z^*_{\boldsymbol{\omega}_{UK}}(\mathbf{x}_0) = Z(\mathbf{x}_i)$. This holds, since the vector $(\boldsymbol{\gamma}_0^T, \mathbf{f}_0^T)^T$ is identical with the *ith* column of the matrix $\begin{pmatrix} \Gamma_Y & F \\ F^T & 0 \end{pmatrix}$.

Hence, we obtain the unique solution of the universal kriging system defined by $\omega_i^{UK} = 1,\ \omega_j^{UK} = 0$ for $j \neq i$ and all Lagrange parameters $\lambda_l^{UK} = 0,\ l = 0, 1, \ldots, L$. It follows that $Z^*_{\boldsymbol{\omega}_{UK}}(\mathbf{x}_0) = \boldsymbol{\omega}_{UK}^T \mathbf{Z} = Z(\mathbf{x}_i)$ with kriging variance $\sigma^2_{UK} \overset{(8.5)}{=} \sum_{i=1}^n \omega_i^{UK} \gamma_Y(\mathbf{x}_i - \mathbf{x}_0) + \sum_{l=0}^L \lambda_l^{UK} f_l(\mathbf{x}_0) = \gamma_Y(\mathbf{x}_i - \mathbf{x}_i) = 0$ as long as there is no nugget component.

8.8 Universal Kriging in terms of a known covariance

For completeness, we want to express our results in terms of a known covariance function $C_Y(\mathbf{h})$ of the random function $Y(\mathbf{x})$ similar to the section about ordinary kriging. We proceed in particular according to Cressie (1993, p. 154–155) and Wackernagel (2003, pp. 300–307), but we also recommend Chauvet and Galli (1982) in Chapter 2.

Consistent with Cressie (1993) and in addition to Assumption 8.1, we have to strengthen the stationarity assumption on $Y(\mathbf{x})$ to second-order instead of intrinsic stationarity. This implies additionally the existence of a known covariance function $C_Y(\mathbf{h})$ of $Y(\mathbf{x})$, such that $C_Y(\mathbf{h}) := Cov(Y(\mathbf{x}), Y(\mathbf{x}+\mathbf{h})) = \mathbb{E}\left[Y(\mathbf{x})Y(\mathbf{x}+\mathbf{h})\right]$ for all $\mathbf{x}, \mathbf{x}+\mathbf{h} \in D$, since $Y(\mathbf{x})$ is of zero mean.

After these considerations, we can rewrite the universal kriging system and all other statements into terms of the covariance $C_Y(\mathbf{h})$ by replacing the corresponding variogram terms $\gamma_Y(\mathbf{h})$. This works due to the equivalence of the covariance function and a bounded variogram by virtue of Proposition 4.5 (p. 15), which is applicable since $Y(\mathbf{x})$ is assumed to be second-order stationary.

Following Wackernagel (2003, pp. 300–301), we assume further that the predictor $Z^*_{\boldsymbol{\omega}}(\mathbf{x}_0)$ is defined the same way as previously in Definition 8.3. This implies the same unbiasedness conditions (8.1) of the linear predictor and we obtain the universal kriging equations (see Cressie 1993, pp. 153–154)

$$\sum_{j=1}^n \omega_j^{UK} C_Y(\mathbf{x}_i - \mathbf{x}_j) - \sum_{l=0}^L \lambda_l^{UK} f_l(\mathbf{x}_i) = C_Y(\mathbf{x}_i - \mathbf{x}_0),\ i = 1, \ldots, n$$

$$\sum_{j=1}^n \omega_j^{UK} f_l(\mathbf{x}_j) = f_l(\mathbf{x}_0),\ l = 0, 1, \ldots, L,$$

as well as in matrix formulation

$$\begin{aligned} \Sigma_Y \boldsymbol{\omega}_{UK} - F \boldsymbol{\lambda}_{UK} &= \mathbf{c}_{Y,0} \\ F^T \boldsymbol{\omega}_{UK} &= \mathbf{f}_0 \end{aligned} \tag{8.6}$$

$$\Leftrightarrow$$

$$\begin{pmatrix} \Sigma_Y & F \\ F^T & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\omega}_{UK} \\ -\boldsymbol{\lambda}_{UK} \end{pmatrix} = \begin{pmatrix} \mathbf{c}_{Y,0} \\ \mathbf{f}_0 \end{pmatrix}$$

with symmetric covariance matrix $\Sigma_Y \in \mathbb{R}^{n \times n}$ of the random vector $\mathbf{Y} := (Y(\mathbf{x}_1), \ldots, Y(\mathbf{x}_n))^T$, i.e. $(\Sigma_Y)_{i,j} := C_Y(\mathbf{x}_i - \mathbf{x}_j)$ for $i, j = 1, \ldots, n$, $\mathbf{c}_{Y,0} := (C_Y(\mathbf{x}_1 - \mathbf{x}_0), \ldots, C_Y(\mathbf{x}_n - \mathbf{x}_0))^T$ and symmetric block matrix $\begin{pmatrix} \Sigma & F \\ F^T & 0 \end{pmatrix} \in \mathbb{R}^{(n+L+1)\times(n+L+1)}$.

Notice that the only thing which is different from the variogram case is that Γ_Y is replaced by Σ_Y, $\boldsymbol{\gamma}_{Y,0}$ by $\mathbf{c}_{Y,0}$ and the changing sign of the Lagrange parameter vector $\boldsymbol{\lambda}_{UK}$.

If the system above is to be solved uniquely, it is again necessary that the matrix F is of full column rank, i.e. that the $L + 1$ basic function $f_0, f_1, \ldots, f_L$ must be linearly independent on the samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ (cf. Wackernagel 2003, p. 301).

Furthermore, if the inverse of the whole block matrix exists, then we can simply obtain the universal kriging weights and the Lagrange multiplier from

$$\begin{pmatrix} \boldsymbol{\omega}_{UK} \\ -\boldsymbol{\lambda}_{UK} \end{pmatrix} = \begin{pmatrix} \Sigma_Y & F \\ F^T & 0 \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{c}_{Y,0} \\ \mathbf{f}_0 \end{pmatrix}.$$

Another theoretical representation of the solution is again additionally presented by Cressie (1993, p. 154) in the next theorem. Hereby the target parameters $\boldsymbol{\omega}_{UK}$ and $\boldsymbol{\lambda}_{UK}$ are computed directly from F, Σ_Y, $\mathbf{c}_{Y,0}$ and $\mathbf{f}_0$ without using the large block matrix above:

Theorem 8.6 (Solution for universal kriging with a known covariance)
Under the assumption of invertibility of $F^T \Sigma_Y F$, the unique solution for $\boldsymbol{\omega}_{UK}$ and $\boldsymbol{\lambda}_{UK}$ of the universal kriging system (8.1) and (8.6) is given by

$$\boldsymbol{\omega}_{UK} = \Sigma_Y^{-1} \left[\mathbf{c}_{Y,0} + F(F^T \Sigma_Y^{-1} F)^{-1} (\mathbf{f}_0 - F^T \Sigma_Y^{-1} \mathbf{c}_{Y,0}) \right],$$

$$\boldsymbol{\lambda}_{UK} = (F^T \Sigma_Y^{-1} F)^{-1} (\mathbf{f}_0 - F^T \Sigma_Y^{-1} \mathbf{c}_{Y,0}).$$

Proof:
We skip the proof since Theorem 8.6 could be proved by following the same steps as in the proof of Theorem 8.4.

□

Subsequently, we can express the minimized universal kriging variance σ^2_{UK}. It is again defined as the variance of the difference of the optimal linear predictor $Z^*_{\boldsymbol{\omega}_{UK}}(\mathbf{x}_0)$ and the to predicted variable $Z(\mathbf{x}_0)$. With help of Theorem 8.6 and the kriging equations (8.1) and (8.6), we obtain the kriging variance in accordance with Cressie (1993, p. 155) as follows:

$$\sigma^2_{UK} := Var(Z^*_{\boldsymbol{\omega}_{UK}}(\mathbf{x}_0) - Z(\mathbf{x}_0))$$

$$= C_Y(\mathbf{0}) + \boldsymbol{\omega}^T_{UK}\Sigma_Y\boldsymbol{\omega}_{UK} - 2\boldsymbol{\omega}^T_{UK}\mathbf{c}_{Y,0}$$

$$= C_Y(\mathbf{0}) - \begin{pmatrix}\boldsymbol{\omega}^T_{UK}, -\boldsymbol{\lambda}^T_{UK}\end{pmatrix}\begin{pmatrix}\mathbf{c}_{Y,0}\\ \mathbf{f}_0\end{pmatrix}$$

$$= C_Y(\mathbf{0}) - \mathbf{c}^T_{Y,0}\Sigma^{-1}_Y\mathbf{c}_{Y,0} + (\mathbf{f}_0 - F^T\Sigma^{-1}_Y\mathbf{c}_{Y,0})^T(F^T\Sigma^{-1}_Y F)^{-1}(\mathbf{f}_0 - F^T\Sigma^{-1}_Y\mathbf{c}_{Y,0}), \tag{8.7}$$

as well as

$$\sigma^2_{UK} = C_Y(\mathbf{0}) + \sum_{i=1}^{n}\sum_{j=1}^{n}\omega^{UK}_i\omega^{UK}_j C_Y(\mathbf{x}_i - \mathbf{x}_j) - 2\sum_{i=1}^{n}\omega^{UK}_i C_Y(\mathbf{x}_i - \mathbf{x}_0)$$

$$= C_Y(\mathbf{0}) + \sum_{l=0}^{L}\lambda^{UK}_l f_l(\mathbf{x}_0) - \sum_{i=1}^{n}\omega^{UK}_i C_Y(\mathbf{x}_i - \mathbf{x}_0). \tag{8.8}$$

In summary, by virtue of Theorem 8.6 and the last calculations, we obtain the "optimal" universal kriging predictor $Z^*_{\boldsymbol{\omega}_{UK}}(\mathbf{x}_0)$ such that

$$Z^*_{\boldsymbol{\omega}_{UK}}(\mathbf{x}_0) = \sum_{i=1}^{n}\omega^{UK}_i Z(\mathbf{x}_i) = \boldsymbol{\omega}^T_{UK}\mathbf{Z}$$

$$= \left[\mathbf{c}_{Y,0} + F(F^T\Sigma^{-1}_Y F)^{-1}(\mathbf{f}_0 - F^T\Sigma^{-1}_Y\mathbf{c}_{Y,0})\right]^T\Sigma^{-1}_Y\mathbf{Z}.$$

The kriging variance is observed in (8.7) and (8.8) and finally the corresponding kriging estimate $z^*_{\boldsymbol{\omega}_{UK}}(\mathbf{x}_0)$ at $\mathbf{x}_0$ can be written as

$$z^*_{\boldsymbol{\omega}_{UK}}(\mathbf{x}_0) = \sum_{i=1}^{n}\omega^{UK}_i z(\mathbf{x}_i) = \boldsymbol{\omega}^T_{UK}\mathbf{z}$$

$$= \left[\mathbf{c}_{Y,0} + F(F^T\Sigma^{-1}_Y F)^{-1}(\mathbf{f}_0 - F^T\Sigma^{-1}_Y\mathbf{c}_{Y,0})\right]^T\Sigma^{-1}_Y\mathbf{z}.$$

For a better understanding of universal kriging prediction, especially with the new subject "drift function", we want to present an easy example given by Kitanidis (1997, pp. 126–127). The differing signs of the Lagrange parameters compared with Kitanidis (1997) come from their different definition in the corresponding minimization problem.

Example 8.7 (Linear drift $\boldsymbol{\mu}(\mathbf{x})$)
Typically, the spatial trend $\mu(\mathbf{x})$ can be, and is modeled as a polynomial function in the spatial coordinates $\mathbf{x} \in D$ (Journel and Huijbregts 1978, pp. 314–315). For instance, in the simple case of a "linear drift", i.e. that a linear model is assumed over a two-dimensional spatial domain with coordinates of location $(x, y) \in D \subseteq \mathbb{R}^2$, one would set $L = 2$ and get $f_0(x, y) = 1$, $f_1(x, y) = x$ and $f_2(x, y) = y$ as drift functions.

Hence, the linear drift

$$\mu(x, y) = a_0 + a_1 x + a_2 y$$

and

$$Z(x,y) = a_0 + a_1 x + a_2 y + Y(x,y).$$

This implies that the universal kriging system with prediction point (x_0, y_0) includes three unbiasedness conditions, for $l = 0, 1$ and 2, such that it can be written as

$$\begin{cases} \sum_{j=1}^{n} \omega_j^{UK} C_Y((x_i,y_i)-(x_j,y_j)) - \lambda_0^{UK} - \lambda_1^{UK} x_i - \lambda_2^{UK} y_i \\ = C_Y\left((x_i,y_i)-(x_0,y_0)\right), \ i = 1,\ldots,n \\ \\ \sum_{j=1}^{n} \omega_j^{UK} = 1 \\ \sum_{j=1}^{n} \omega_j^{UK} x_j = x_0 \\ \sum_{j=1}^{n} \omega_j^{UK} y_j = y_0. \end{cases}$$

The universal kriging variance turns out to be

$$\sigma_{UK}^2 = C_Y(\mathbf{0}) + \lambda_0^{UK} + \lambda_1^{UK} x_0 + \lambda_2^{UK} y_0 - \sum_{i=1}^{n} \omega_i^{UK} C_Y\left((x_i,y_i)-(x_0,y_0)\right).$$

Remark 8.8 (Estimation of drift coefficients and residual variogram)
Finally, for the procedure of prediction with universal kriging in practice, we have to pay attention to two further issues:

(i) First of all, we have to estimate the drift function $\mu(\mathbf{x})$, since we do not have it at hand. For instance, this could be done using the "generalized-least-squares estimator" $\hat{\mathbf{a}} = \left(F^T \Sigma_Y^{-1} F\right)^{-1} F^T \Sigma_Y^{-1} \mathbf{Z}$ of the coefficient vector $\mathbf{a}$, which determines $\mu(\mathbf{x})$ completely (cf. Cressie 1993, p. 156). For further details, we refer the interested reader to the book of Wackernagel (2003, pp. 302–303) and to section 4.2 "Optimal Estimation of the Drift" in Matheron (1971).

(ii) Furthermore, in our theoretical considerations we assumed the underlying residual variogram function $\gamma_Y(\mathbf{h})$ of $Y(\mathbf{x})$ to be known. But in practice, we need to estimate it from our data. This turns out to be more problematic, since the nonstationary random function $Z(\mathbf{x})$ is decomposed into the unknown trend $\mu(\mathbf{x})$ and the unknown residual random function $Y(\mathbf{x})$. Hence, both components need to be estimated first and then, based on these estimates $\hat{\mu}(\mathbf{x})$ and $\hat{Y}(\mathbf{x})$, the variogram function $\gamma_Y(\mathbf{h})$ of $Y(\mathbf{x})$ can be estimated. But at this point the problem arises that the nonstationary drift $\mu(\mathbf{x})$ does not allow the direct estimation of the variogram (and covariogram $C_Y(\mathbf{h})$) neither from the empirical variogram, nor from the estimated residual (Webster and Oliver 2007, p. 195). Matheron (1971) called this the problem of "identification of the underlying variogram". More information and details can be found in the books by Wackernagel (2003, pp. 303–306) and Cressie (1993, pp. 165–170). We also recommend Chauvet and Galli (1982) in Chapter 4 and section 4.6 "Indeterminability of the Underlying Variogram" in Matheron (1971).

Note that in *R* the residual variogram is automatically estimated if *formula* in *variogram()* contains at least one argument different from 1, e.g. *formula = temp* $\sim$ $x+y+xy+x^2+y^2$ in the case of a quadratic trend and coordinates x and y.

8.9 Universal Kriging in R

As universal kriging prediction represents the last kriging method in this thesis, and also our most general model, we want to consider four different linear drifts. We suppose the mean value of the temperatures in Germany of our two dates as a linear trend of

(i) longitude,

(ii) latitude,

(iii) longitude and latitude and

(iv) longitude, latitude and elevation

evaluated at our given 54 sample points.

In *gstat*, universal kriging prediction can be performed similarly to simple and ordinary kriging. But in contrast, we have to insert our considered trend functions in the estimation of the variogram. Unfortunately, we cannot use our estimated variogram functions from the section "The Variogram", since we need the residual variogram functions (see Remark 8.8).

After estimating and fitting the residual variogram, we can proceed analogously to the last two sections. For this reason, we omit most of the *R* code and only print these codes varying from simple and ordinary kriging.

In the last case where we include the height of each weather station into our prediction, we will perform ordinary kriging for predicting the elevation values at the unsampled places of our grid in the first step. And afterwards we go on with universal kriging as in the other cases.

Hence, we begin with the preparation for kriging prediction and estimate the residual variogram functions of all four trend models, printing the *R* code exemplarily for the third trend function.

```
> #Universal Kriging:
>
> #1: linear trend in longitude
> #2: linear trend in latitude
> #3: linear trend in long- and latitude
> #4: linear trend in long-, latitude and elevation

> #1.) Gstat objects:
> g_uk1_3<-gstat(id="temp1", formula=temp1~longkm+latkm,
+ locations=~longkm+latkm,data=data1)
> g_uk2_3<-gstat(id="temp2", formula=temp2~longkm+latkm,
+ locations=~longkm+latkm,data=data2)
```

```
> #2.) Variogram cloud:
> vcloud_uk1_3<-variogram(object=g_uk1_3, formula=temp1~longkm+latkm,
+ cutoff=Inf,cloud=TRUE)
> vcloud_uk2_3<-variogram(object=g_uk2_3, formula=temp2~longkm+latkm,
+ cutoff=Inf,cloud=TRUE)

> #3.) Empirical variogram:
> vemp_uk1_3<-variogram(object=g_uk1_3, formula=temp1~longkm+latkm,
+ cutoff=max(vcloud_uk1_3$dist)/2,width=10)
> vemp_uk2_3<-variogram(object=g_uk2_3, formula=temp2~longkm+latkm,
+ cutoff=max(vcloud_uk2_3$dist)/2,width=10)

> #4.) Fitting the empirical variogram:
> #Same procedure as in "The Variogram", the final results are:
>
> #2010/11/28: sum of squares = 5.87
> vfituk1_3<-fit.variogram(object=vemp_uk1_3,
+ model=vgm(psill=1,model="Mat",range=100,nugget=1,kappa=1.61),
+ fit.sills=TRUE,fit.ranges=TRUE,fit.method=6)

> #2012/06/09: sum of squares = 10.55
> vfituk2_3<-fit.variogram(object=vemp_uk2_3,
+ model=vgm(psill=1,model="Mat",range=100,nugget=1,kappa=99.91),
+ fit.sills=TRUE,fit.ranges=TRUE,fit.method=6)

> #5.) Update gstat objects:
> g_uk1_3up<-gstat(g_uk1_3,model=vfituk1_3, id="temp1",
+ formula=temp1~longkm+latkm,locations=~longkm+latkm,data=data1)

data:
temp1 : formula = temp1~longkm + latkm ; data dim = 54 x 1
variograms:
         model     psill    range kappa
temp1[1]   Nug 0.5682006  0.00000  0.00
temp1[2]   Mat 1.5313091 72.44415  1.61
~longkm + latkm

> g_uk2_3up<-gstat(g_uk2_3,model=vfituk2_3, id="temp2",
+ formula=temp2~longkm+latkm,locations=~longkm+latkm,data=data2)

data:
temp2 : formula = temp2~longkm + latkm ; data dim = 54 x 1
variograms:
         model     psill    range kappa
temp2[1]   Nug 0.9909754 0.000000  0.00
temp2[2]   Mat 1.2947727 7.527346 99.91
~longkm + latkm
```

We finish with our preparation of the variogram and can begin with universal kriging prediction:

```
> #Universal Kriging Prediction for additional 24 weather stations:
>
> p_uk1_3<-predict(g_uk1_3up,newdata=newdat)

[using universal kriging]

> p_uk2_3<-predict(g_uk2_3up,newdata=newdat)

[using universal kriging]

> #First lines of prediction:
> p_uk1_3[1:5,] #2010/11/28

         coordinates temp1.pred temp1.var
1    (8.979, 48.216)  -1.058979 0.9264181
2  (9.8044, 51.9672)  -4.358222 0.8672984
3 (13.4367, 54.6817)  -2.231965 1.7439544
4    (7.979, 51.465)  -3.077680 0.8372743
5 (10.9431, 48.4261)  -2.305909 0.8259260

> p_uk2_3[1:5,] #2012/06/09

         coordinates temp2.pred temp2.var
1    (8.979, 48.216)   15.85785  1.319660
2  (9.8044, 51.9672)   14.67715  1.260069
3 (13.4367, 54.6817)   15.80469  2.130284
4    (7.979, 51.465)   13.16339  1.241139
5 (10.9431, 48.4261)   14.55123  1.247551
```

For the fourth trend function, we perform a grid of longitude, latitude and additionally of the elevation values, which we gained from ordinary kriging. This means that we inserted the kriging estimates for the elevation data in our grid.

```
> prediction_uk1_4<-predict(object=g_uk1_4up, newdata=griduk4)

[using universal kriging]

> prediction_uk2_4<-predict(object=g_uk2_4up, newdata=griduk4)

[using universal kriging]
```

In order to compare how close the resulting kriging estimates are, related to the measured values, we print the corresponding residuals of all four linear trend models in Table 8.1 for 2010/11/28 and Table 8.2 for 2012/06/09. We observe that the last linear trend in longitude, latitude and elevation seems to provide the best fit to our data in both cases, especially for 2012/06/09, where the amount of the sum of squares is only one-tenth compared with the other trend functions and compared with simple and ordinary kriging. This makes sense, since we included the most information in our prediction compared with the other ones, namely the longitude, latitude and elevation.

Longitude	Latitude	long (1)	lat (2)	long and lat (3)	long, lat and elev (4)
8.98	48.22	-3.04	-3.06	-3.04	-0.76
9.8	51.97	-1.10	-1.13	-1.14	-0.53
13.44	54.68	1.70	1.79	2.03	2.04
7.98	51.47	0.31	0.26	0.28	-0.49
10.94	48.43	-0.47	-0.39	-0.39	-0.35
7.31	50.04	-1.51	-1.51	-1.51	-0.16
6.7	53.6	1.48	1.85	1.60	2.70
9.14	53.45	-1.13	-1.25	-1.23	-1.56
9.32	49.52	-0.72	-0.86	-0.85	-0.22
14.73	52.02	0.00	-0.07	0.08	-0.15
10.5	49.85	-0.33	-0.33	-0.33	-0.29
10.13	48.99	-1.21	-1.18	-1.19	-0.09
12.73	48.48	-0.29	-0.29	-0.29	-0.21
10.68	53.58	-0.13	-0.03	-0.06	0.21
13.14	49.11	-3.10	-3.19	-3.14	3.17
13.94	53.32	-0.46	-0.47	-0.43	-0.38
9.22	50.51	-2.77	-2.78	-2.76	0.30
11.14	52.97	0.72	0.79	0.78	1.31
11.27	47.48	-0.86	-0.94	-0.92	1.26
7.64	47.81	-0.18	0.05	-0.08	0.84
11.14	50.5	-2.52	-2.47	-2.47	1.00
10.88	51.67	-3.83	-3.89	-3.88	-2.04
8.57	52.45	0.50	0.45	0.44	1.32
12.46	52.12	-0.54	-0.54	-0.54	0.14
Sum of Squares		61.95	65.06	64.35	36.60

Table 8.1: Residuals from universal kriging prediction of the additional 24 weather stations in Germany of 2010/11/28, where the last line provides the sum of the squared residuals and the columns are sorted by the different trend functions: linear trend in longitude (1), latitude (2), longitude and latitude (3), longitude, latitude and elevation (4)

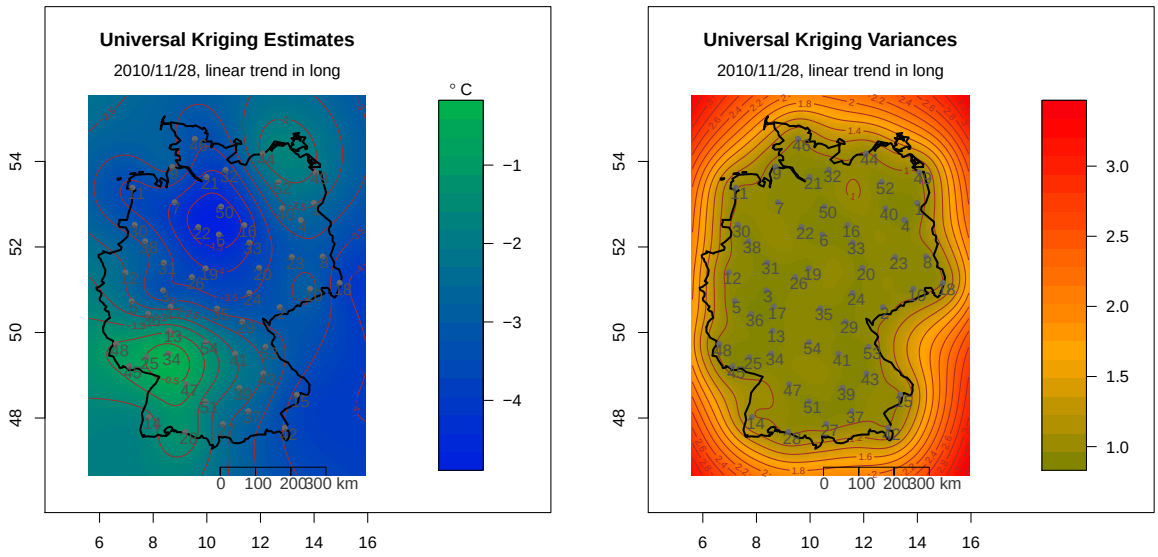
Longitude	Latitude	long (1)	lat (2)	long and lat (3)	long, lat and elev (4)
8.98	48.22	-1.80	-2.24	-2.16	1.14
9.8	51.97	0.27	0.32	0.22	0.46
13.44	54.68	-2.05	-1.78	-2.00	-1.48
7.98	51.47	0.11	0.10	0.14	-0.25
10.94	48.43	0.20	0.09	0.15	-0.13
7.31	50.04	-2.09	-2.37	-2.35	-0.39
6.7	53.6	-0.51	-0.65	-0.45	-0.02
9.14	53.45	-0.21	-0.30	-0.34	-0.31
9.32	49.52	-1.57	-1.56	-1.80	-0.86
14.73	52.02	-0.25	-0.28	-0.22	-0.86
10.5	49.85	-1.51	-0.94	-0.97	-0.69
10.13	48.99	-1.30	-1.01	-1.13	-0.15
12.73	48.48	-1.29	-0.99	-1.17	-0.72
10.68	53.58	0.12	0.03	0.07	0.08
13.14	49.11	-7.17	-7.03	-7.19	2.25
13.94	53.32	-0.44	-0.31	-0.35	-0.32
9.22	50.51	-3.95	-4.03	-3.87	0.19
11.14	52.97	0.01	-0.17	-0.08	-0.34
11.27	47.48	-3.12	-3.09	-3.34	0.34
7.64	47.81	0.63	0.42	0.72	0.11
11.14	50.5	-4.11	-4.33	-4.22	0.55
10.88	51.67	-3.22	-3.12	-3.17	0.04
8.57	52.45	0.08	0.27	0.04	0.03
12.46	52.12	-1.01	-1.07	-0.96	0.11
Sum of Squares		125.97	125.00	128.62	12.30

Table 8.2: Residuals from universal kriging prediction of the additional 24 weather stations in Germany of 2010/11/28, where the last line provides the sum of the squared residuals and the columns are sorted by the different trend functions: linear trend in longitude (1), latitude (2), longitude and latitude (3), longitude, latitude and elevation (4)

Finally, we obtain the plots of the universal kriging prediction estimates and variances for each of the first three linear trends once for 2010/11/28, see Figures 8.1-8.3, and once for 2012/06/09, see Figures 8.4-8.6.

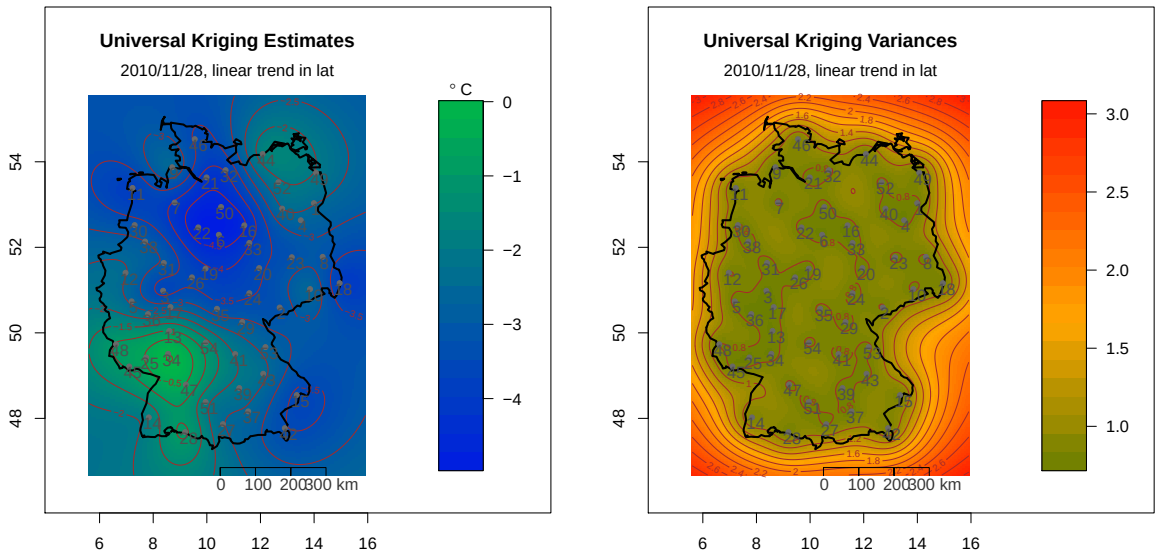
For the fourth kind of drift function which we consider, a linear trend in longitude, latitude and elevation, we obtain the ordinary kriging estimates and variances of the elevation data for the later universal kriging prediction first, see Figure 8.7. Afterwards we can use these predicted elevation values for universal kriging prediction and we get the corresponding plots shown below on Figure 8.8 for 2010/11/28 and on Figure 8.9 for 2012/06/09.

Note that the elevation estimates are the same for both dates and we do not have to estimate them twice.



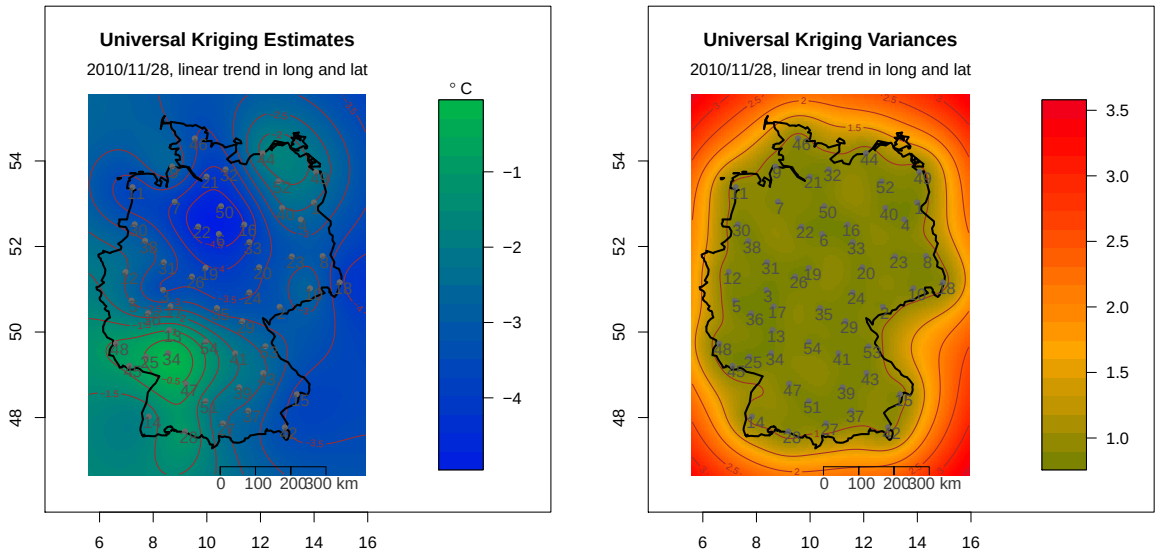
(a) Universal Kriging Estimates (b) Universal Kriging Variances

Figure 8.1: Universal Kriging with a linear trend in longitude applied to the temperature data of 2010/11/28 in Germany



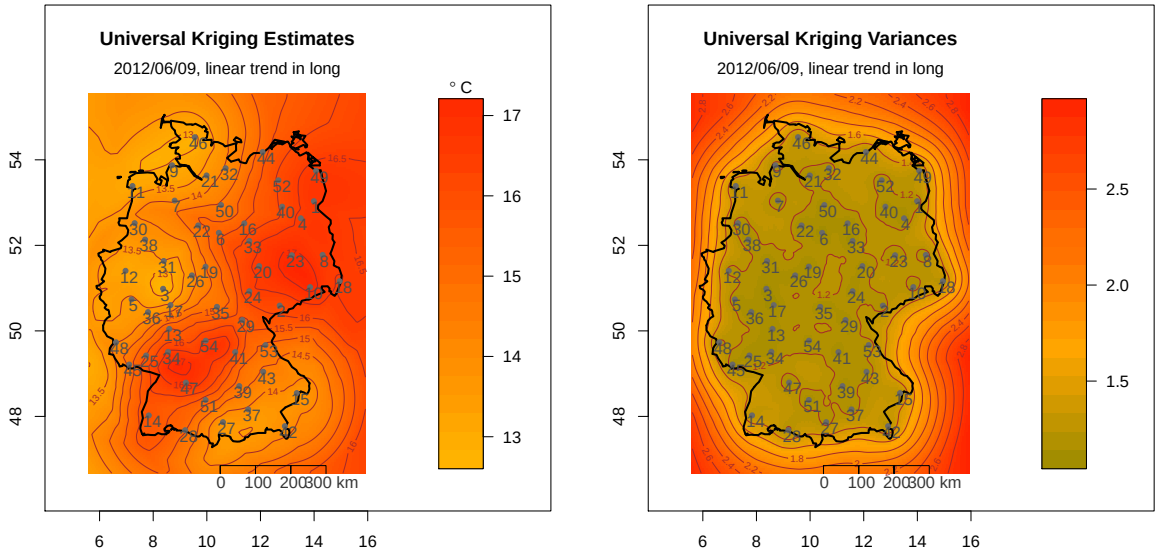
(a) Universal Kriging Estimates (b) Universal Kriging Variances

Figure 8.2: Universal Kriging with a linear trend in latitude applied to the temperature data of 2010/11/28 in Germany



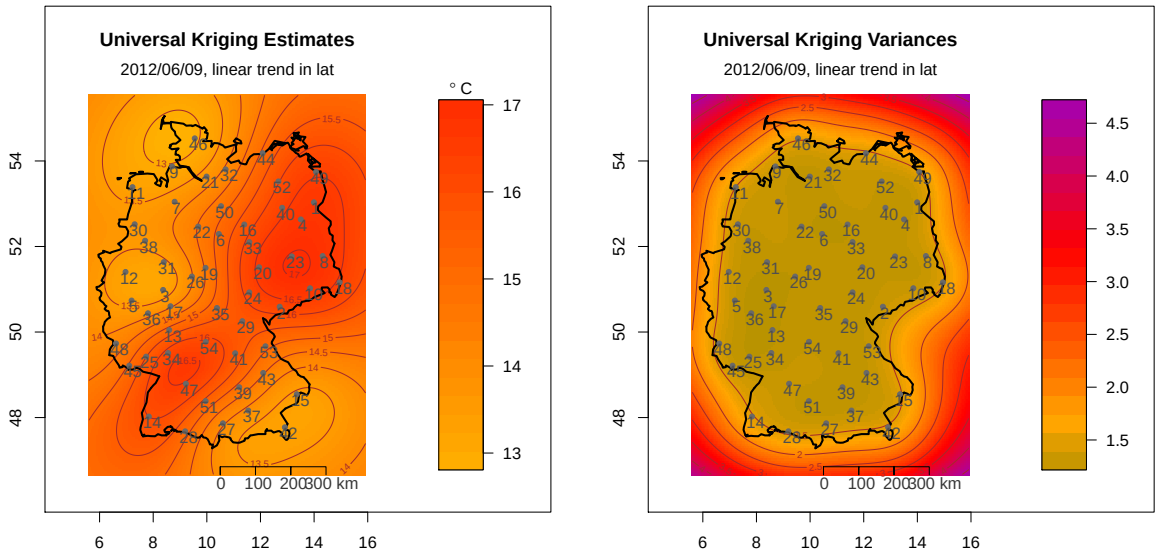
(a) Universal Kriging Estimates (b) Universal Kriging Variances

Figure 8.3: Universal Kriging with a linear trend in longitude and latitude applied to the temperature data of 2010/11/28 in Germany



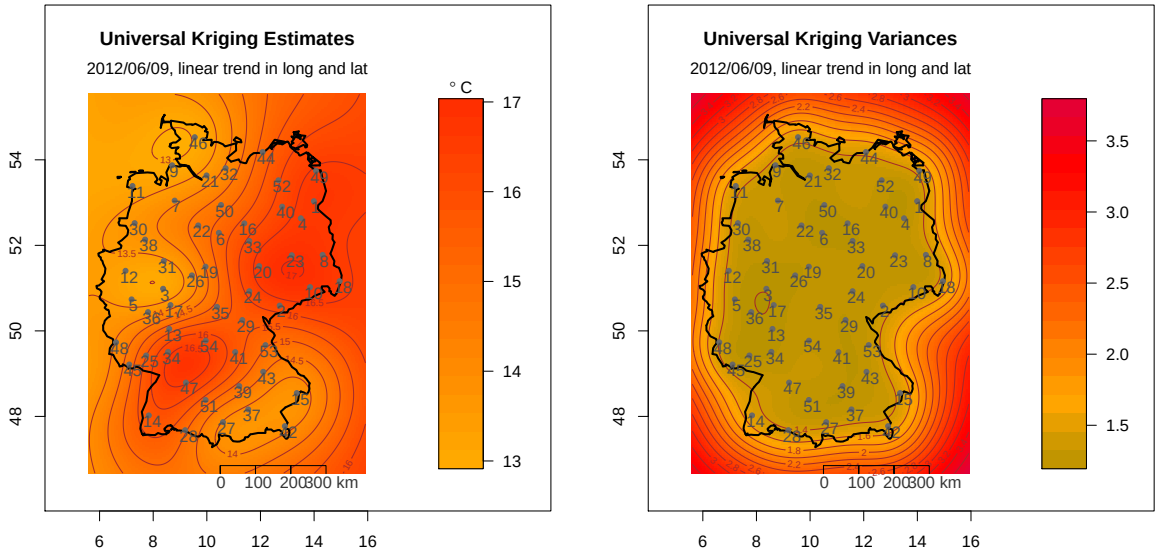
(a) Universal Kriging Estimates (b) Universal Kriging Variances

Figure 8.4: Universal Kriging with a linear trend in longitude applied to the temperature data of 2012/06/09 in Germany



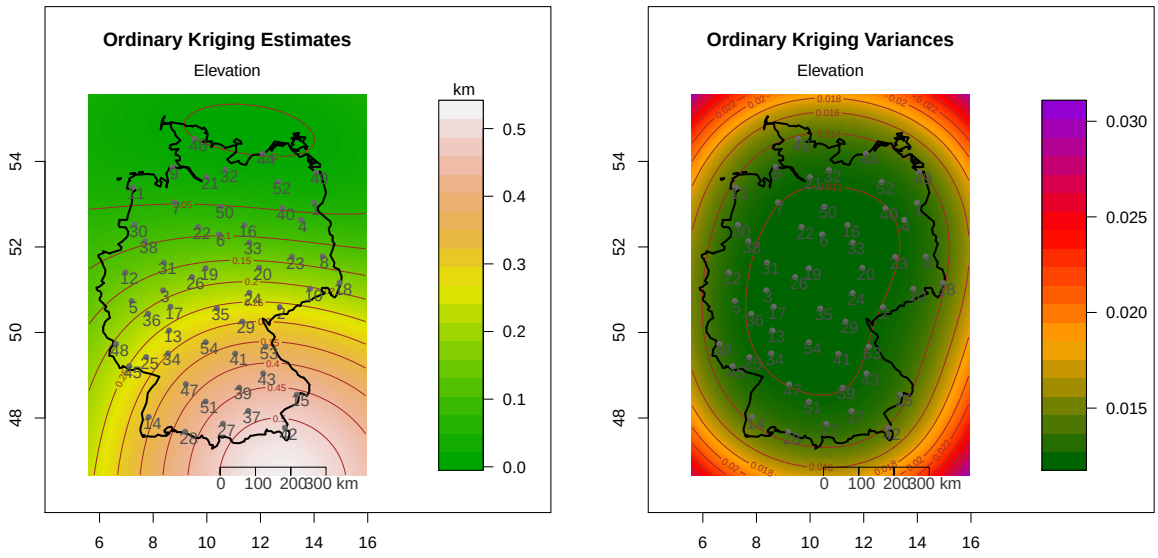
(a) Universal Kriging Estimates (b) Universal Kriging Variances

Figure 8.5: Universal Kriging with a linear trend in latitude applied to the temperature data of 2012/06/09 in Germany



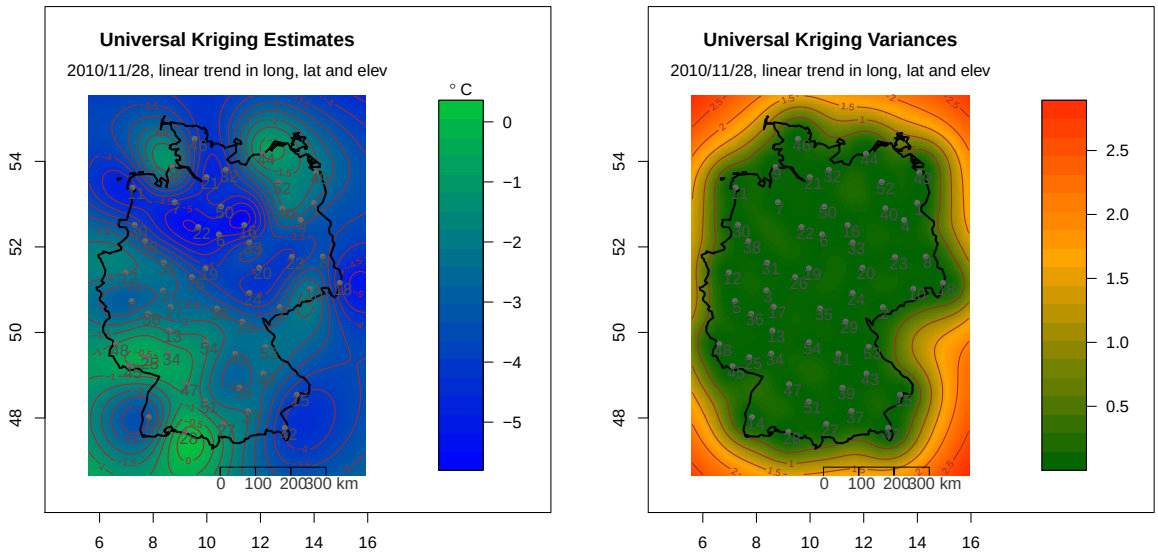
(a) Universal Kriging Estimates (b) Universal Kriging Variances

Figure 8.6: Universal Kriging with a linear trend in longitude and latitude applied to the temperature data of 2012/06/09 in Germany



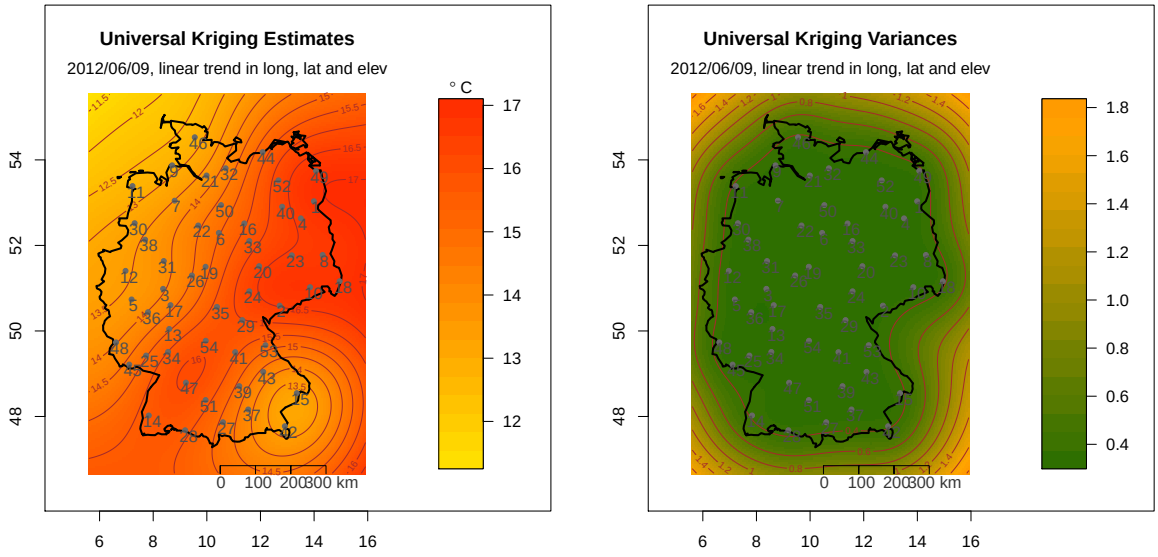
(a) Ordinary Kriging Estimates (b) Ordinary Kriging Variances

Figure 8.7: Ordinary Kriging applied to the elevation data of the data set of weather stations in Germany



(a) Universal Kriging Estimates (b) Universal Kriging Variances

Figure 8.8: Universal Kriging with a linear trend in longitude, latitude and elevation applied to the temperature data of 2010/11/28 in Germany



(a) Universal Kriging Estimates

(b) Universal Kriging Variances

Figure 8.9: Universal Kriging with a linear trend in longitude, latitude and elevation applied to the temperature data of 2012/06/09 in Germany

9 Summary and Outlook

We started this thesis by introducing the quantity variogram $\gamma(\mathbf{h})$. It relies on an intrinsic stationarity assumption on the underlying random function $Z(\mathbf{x})$ for $\mathbf{x}$ in a geographical space D. Its object of interest is to measure the spatial dependence of $Z(\mathbf{x})$ between two points in D. At the beginning we defined the variogram function theoretically and presented some important properties. This limits the choice of a valid variogram function in its estimation later. Afterwards we estimated a variogram function from our underlying samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and corresponding observations $z(\mathbf{x}_1), \ldots, z(\mathbf{x}_n)$.

For this reason we calculated and plotted the variogram cloud, which gives a first impression of the spatial structure. Then we grouped "similar" lags $\mathbf{h}$ and formed the empirical variogram, which provides a first estimate of the underlying variogram for all lags. Unfortunately, this obtained experimental variogram cannot be used for prediction, since it does not satisfy some conditions for validity in general (see Proposition 4.7, p. 16). Therefore, we fitted a valid parametric variogram model function to the empirical variogram. We did this fit using least squares, but there exist several other methods such as the restricted maximum likelihood (REML). At the end of our preparation, we obtained a fitted, valid variogram model function. We used this for our prediction, since kriging relies on the knowledge of some kind of spatial structure of $Z(\mathbf{x})$, i.e. variogram or covariance.

In the main part of this thesis, we introduced spatial prediction with the four geostatistical methods kriging the mean, simple, ordinary and universal kriging. Kriging the mean is used to predict the mean value of an underlying random function $Z(\mathbf{x})$ over a domain D, while the last three types serve to predict the value of $Z(\mathbf{x})$ at any arbitrary unsampled point of interest $\mathbf{x}_0$, called the prediction point. For this reason simple, ordinary and universal kriging are also called *punctual* kriging methods.

In most cases, the mean value of some object of interest is estimated by calculating the arithmetic mean of the observed data, since this is an intuitive approach. This makes sense if the data are distributed on an uniform grid or if $Z(\mathbf{x})$ and $Z(\mathbf{y})$ are uncorrelated $\forall \mathbf{x} \neq \mathbf{y} \in D$. But as in practice, the samples are irregularly placed in space (e.g. the 54 weather stations in our data set), this approach could be very missleading. Consider the extreme case, where most sample points are located really close to each other (e.g. if there were 30 of the 54 weather stations near Munich). Their observed values will probably be very similar and hence should get less weight in the prediction. Otherwise it may happen that we obtain a nonreasonable estimate for the mean value, for instance if the measured values close to Munich are very different from the rest (e.g. 6°C higher). Hence, we introduced kriging the mean, which takes the spatial dependence of the sample points into account and is therefore a good alternative for prediction with irregularly spaced data.

Finally, we presented three punctual kriging methods. All of them rely on spatial structure as well. That is to say that the weights of each random variable $Z(\mathbf{x}_i)$ in the linear predictor vary and depend on the underlying variogram or covariance. The main idea behind is that sample points $\mathbf{x}_i$ near to the prediction point $\mathbf{x}_0$ should get more weight in the calculation of the estimate, since they are to influence the value at $\mathbf{x}_0$ more than

those which are quite far away. Exemplary for our data set, Neuburg in Bavaria should get more weight than Berlin in the prediction with Munich as prediction point. Note that far away samples from $\mathbf{x}_0$ can even get negative weights in the computation.

The hierarchical presentation of the kriging methods is due to their nested construction, i.e. simple kriging is nested in ordinary kriging and ordinary kriging in the universal version. First of all, simple kriging assumes a known and constant mean μ. Since this is not really applicable, we weakened our assumptions on the mean by introducing ordinary kriging, i.e. we supposed μ to be still constant, but unknown. This is a more realistic model, since in most cases we will not know the mean a priori. Finally, since a stationary mean value over a very large region is often not reasonable, e.g. for a region like Germany or even larger, we presented the theory of universal kriging. Hereby $\mu(\mathbf{x})$ is no longer assumed to be constant, but an unknown linear combination of some known deterministic, i.e. nonrandom basic functions. Hence, we integrated a trend in the mean in our model to improve our estimates.

As a summary, we want to present a brief overview of the considered kriging methods. Here we can only compare simple, ordinary and universal kriging, since they have - unlike to kriging the mean - the same objective. Table 9.1 gives a short overview providing the most important assumptions and results of simple, ordinary and universal kriging, where Table 9.2 summarizes kriging the mean.

	Simple Kriging	Ordinary Kriging	Universal Kriging
Assumptions on $\mu(\mathbf{x})$:	known, constant μ	unknown, constant μ	unknown, $\mu(\mathbf{x}) = \sum_{l=0}^{L} a_l f_l(\mathbf{x})$
Assumptions on $Z(\mathbf{x})$:	Second-order stationary	Intrinsically stationary	$Z(\mathbf{x}) = \mu(\mathbf{x}) + Y(\mathbf{x})$, $Y(\mathbf{x})$ intrinsically stat., $\mathbb{E}[Y(\mathbf{x})] = 0$
Linear predictor:	$\mu + \boldsymbol{\omega}^T(\mathbf{Z} - \mu\mathbb{1})$	$\boldsymbol{\omega}^T\mathbf{Z}$	$\boldsymbol{\omega}^T\mathbf{Z}$
Kriging Equations:	$\Sigma\boldsymbol{\omega} = \mathbf{c}_0$	$\Gamma\boldsymbol{\omega} + \lambda\mathbb{1} = \boldsymbol{\gamma}_0$ $\boldsymbol{\omega}^T\mathbb{1} = 1$	$\Gamma_Y\boldsymbol{\omega} + F\boldsymbol{\lambda} = \boldsymbol{\gamma}_{Y,0}$ $F^T\boldsymbol{\omega} = \mathbf{f}_0$
Kriging Variance:	$C(\mathbf{0}) - \boldsymbol{\omega}^T\mathbf{c}_0$	$\boldsymbol{\omega}^T\boldsymbol{\gamma}_0 + \lambda_{OK}$	$\boldsymbol{\omega}^T\boldsymbol{\gamma}_{Y,0} + \mathbf{f}_0^T\boldsymbol{\lambda}$

Table 9.1: Overview punctual kriging methods simple, ordinary and universal kriging

	Kriging the Mean
Assumptions on $\mu(\mathbf{x})$:	unknown, constant mean μ
Assumptions on $Z(\mathbf{x})$:	Second-order stationary
Linear predictor $Z_{\boldsymbol{\omega}}(\mathbf{x}_0)$:	$\boldsymbol{\omega}^T\mathbf{Z}$
Kriging estimate for μ	$\frac{\mathbb{1}^T\Sigma^{-1}\mathbf{z}}{\mathbb{1}^T\Sigma^{-1}\mathbb{1}}$
Kriging Variance:	$\frac{1}{\mathbb{1}^T\Sigma^{-1}\mathbb{1}}$

Table 9.2: Summary of Kriging the Mean for predicting the mean value over a region

Afterwards, to get a good overview of the *gstat* functions in *R* for performing kriging prediction, we summarize the most important functions, their crucial arguments and give a short description, which can be seen on Table 9.3 on the next page.

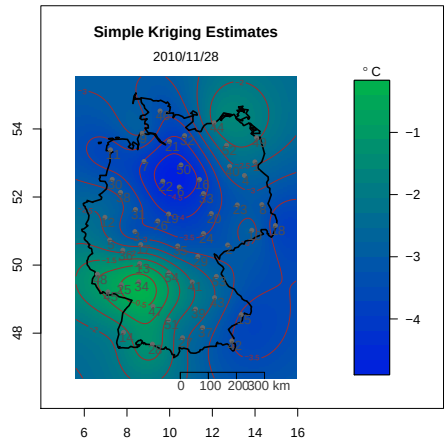
Additionally, we want to point out that there exist several other kinds of kriging besides our four introduced ones, for instance *Factorial*, *Indicator*, *Disjunctive* or *Bayesian Kriging*. We recommend Cressie (1993) and Webster and Oliver (2007) for further details.

Last but not least, since we only considered random processes $Z(\mathbf{x})$ for a fixed point in time, i.e. for a fixed date, we can also extend our model by taking the time into account. Then, the random function $Z(\mathbf{x}, t)$ is a function of both, the location $\mathbf{x}$ and the time t. Hence, the aim is to predict the value of $Z(\mathbf{x}, t)$ at any tuple $(\mathbf{x}_0, t_0)$, which requires spatio-temporal data and brings us into the context of spatio-temporal prediction. Cressie and Wikle (2011) present some spatio-temporal statistical models in Chapter 6, for instance spatio-temporal kriging in form of simple and ordinary kriging. Fortunately, the package *gstat* provides additionally some functions such as *variogramST()*, *vgmST()*, *fit.StVariogram()* or *krigeST()* for spatio-temporal kriging in *R*. For performing, the interested reader may have a look at the corresponding gstat manual "Spatio-temporal geostatistics using gstat" by Pebesma (2013).

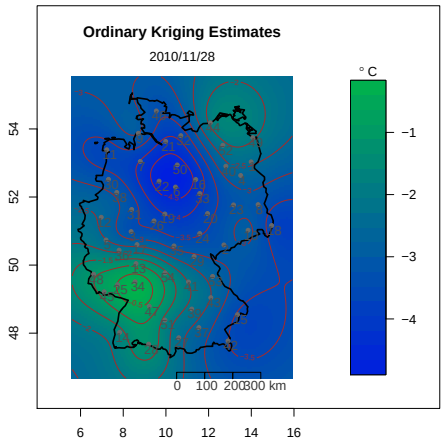
We finish this thesis by printing all resulting plots from the last sections right at the end, for comparison of the different methods and their effects on the estimates and variances of our data set of mean temperatures by eye. We begin with the date 2010/11/28 and plot all estimates first and then all variances, see Figures 9.1 and 9.2. The same applies to 2012/06/09 afterwards, see Figures 9.3 and 9.4. We conclude that in both cases, universal kriging with a linear trend in longitude, latitude and elevation seems to be the most suitable method, since its prediction variances are the lowest and the estimates of our additional 24 weather stations provides the closest fit to the measured data.

Function	Arguments	Description
gstat()		**Creates gstat object for data preparation** (compare to sections 4.2, 6.8 and 8.9)
	g	for creating (=NULL) or updating a gstat object
	id	identifier of new variable (temp)
	formula	temp∼trend
	locations	∼long+lat; set coordinates
	data	data frame containing all variables and coordinates
	model	variogram model for updating object, output of fit.variogram()
	beta	known constant mean; only in simple kriging
variogram()		**Creates variogram cloud and empirical variogram** (compare to sections 4.2, 4.3 and 8.9)
	object	gstat object (not updated)
	cutoff	maximum distance to take into account
	width	width of each vector classes (equidistant)
	cloud	=TRUE for deriving variogram cloud, default=FALSE
vgm()		**Provides theoretical parametric variogram models** (compare to section 4.5)
	psill	(start) value of partial sill
	model	parametric variogram model, e.g. *Lin* or *Gau*
	range	(start) value of range
	nugget	(start) value of nugget
	kappa	value of kappa (for Matérn model)
fit.variogram()		**Fits all parameters to the empirical variogram with least squares** (compare to sections 4.5 and 8.9)
	object	empirical variogram to be fitted, output of variogram()
	model	variogram model, output of vgm() (includes starting values for estimation)
	fit.method	set weights for least squares
predict()		**Derives the prediction of inserted new data locations** (compare to sections 6.8, 7.9 and 8.9)
	object	updated gstat object
	newdata	data frame with new prediction coordinates and variables (for trend)

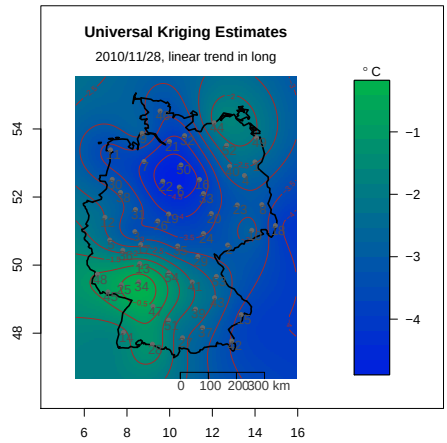
Table 9.3: Overview of the most important *R* functions of the package *gstat*



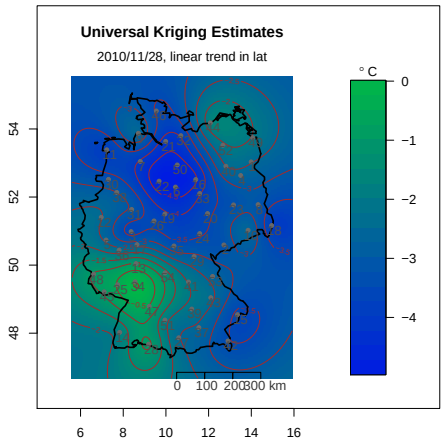
(a) Simple Kriging Estimates



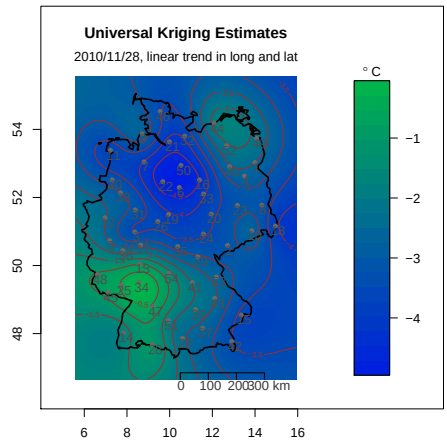
(b) Ordinary Kriging Estimates



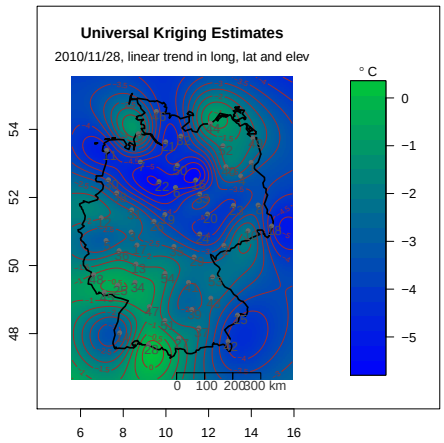
(c) Universal Kriging Estimates with a linear trend in longitude



(d) Universal Kriging Estimates with a linear trend in latitude

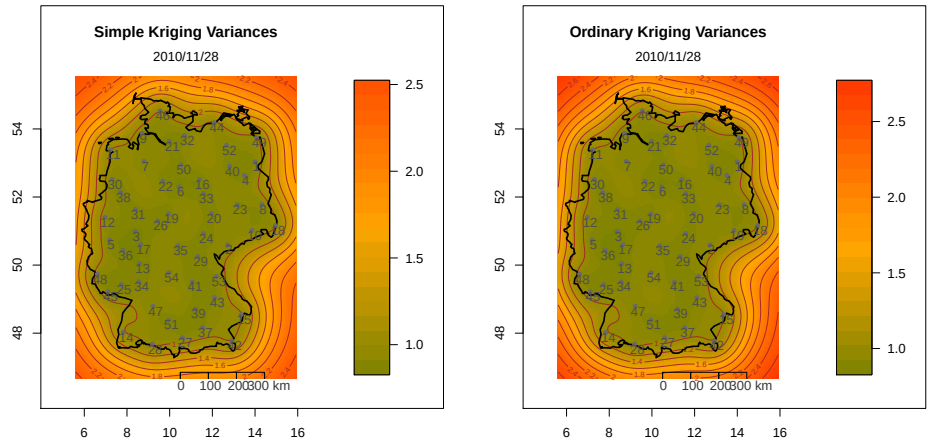


(e) Universal Kriging Estimates with a linear trend in longitude and latitude



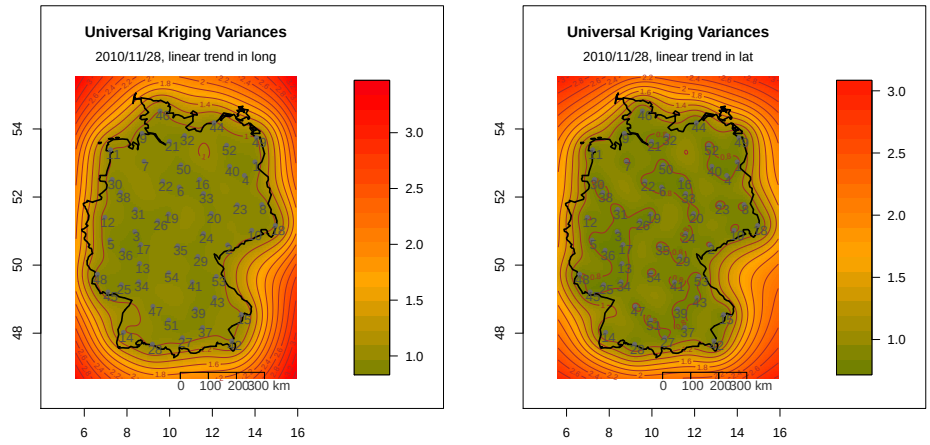
(f) Universal Kriging Estimates with a linear trend in longitude, latitude and elevation

Figure 9.1: Kriging Estimates of all considered kriging methods applied to the temperature data of 2010/11/28 in Germany



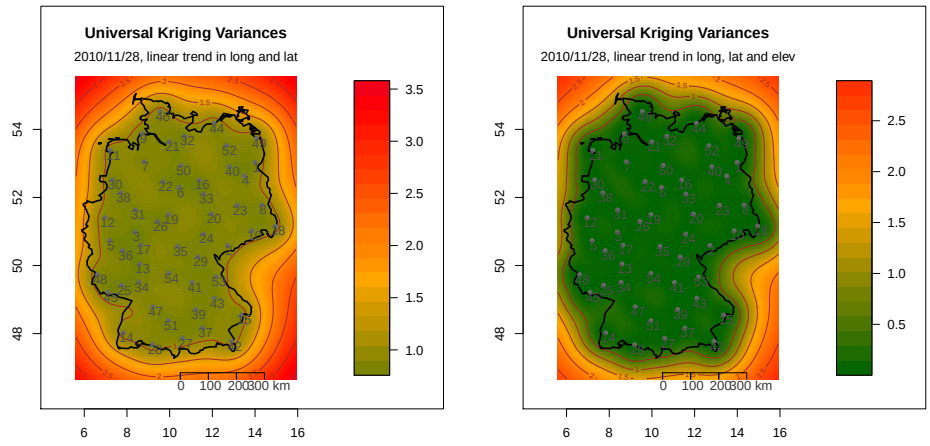
(a) Simple Kriging Variances

(b) Ordinary Kriging Variances



(c) Universal Kriging Variances with a linear trend in longitude

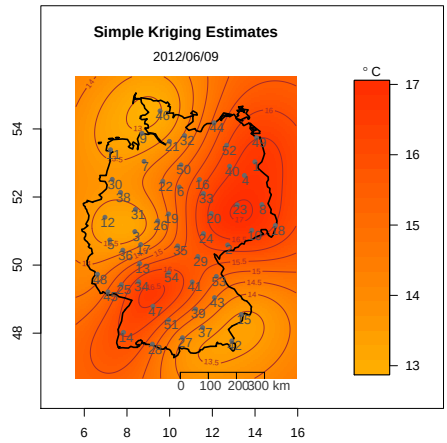
(d) Universal Kriging Variances with a linear trend in latitude



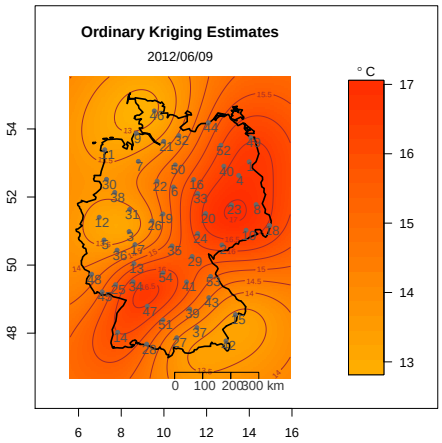
(e) Universal Kriging Variances with a linear trend in longitude and latitude

(f) Universal Kriging Variances with a linear trend in longitude, latitude and elevation

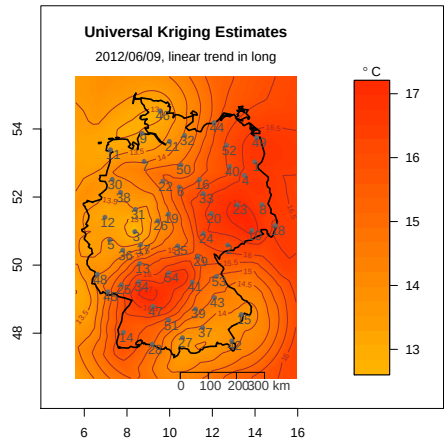
Figure 9.2: Kriging Variances of all considered kriging methods applied to the temperature data of 2010/11/28 in Germany



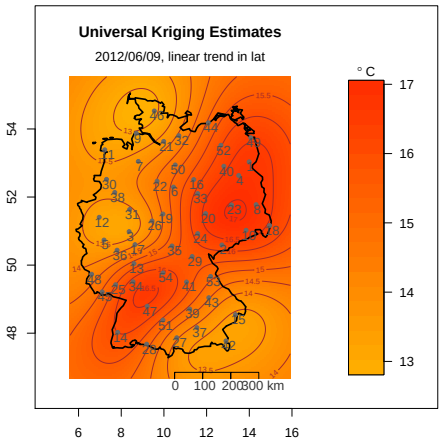
(a) Simple Kriging Estimates



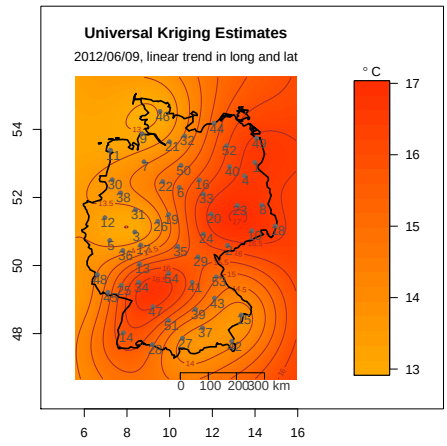
(b) Ordinary Kriging Estimates



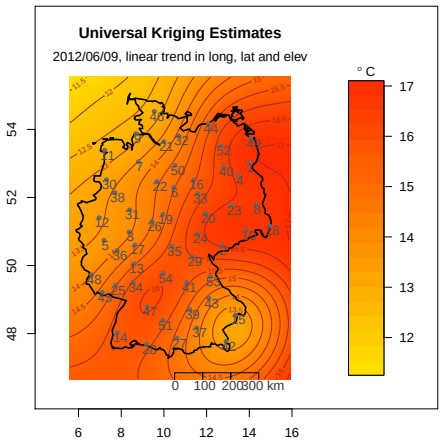
(c) Universal Kriging Estimates with a linear trend in longitude



(d) Universal Kriging Estimates with a linear trend in latitude

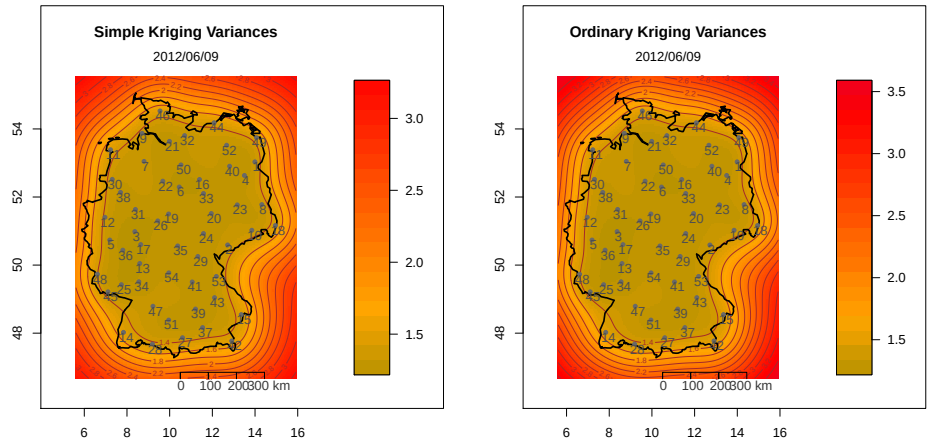


(e) Universal Kriging Estimates with a linear trend in longitude and latitude



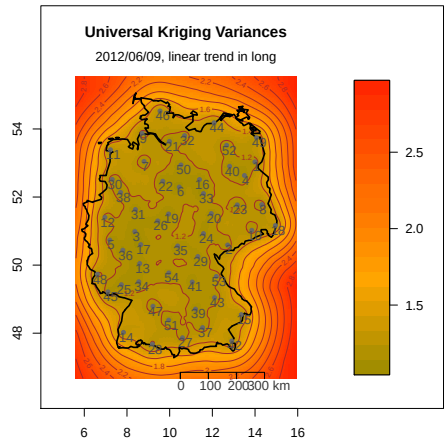
(f) Universal Kriging Estimates with a linear trend in longitude, latitude and elevation

Figure 9.3: Kriging Estimates of all considered kriging methods applied to the temperature data of 2012/06/09 in Germany

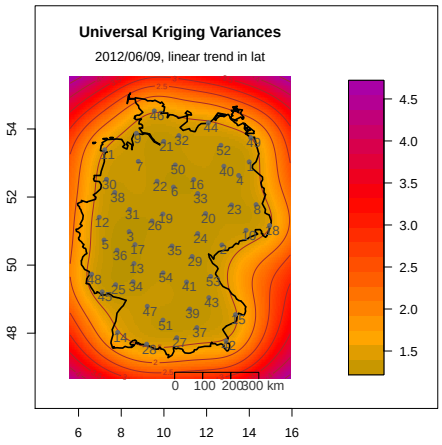


(a) Simple Kriging Variances

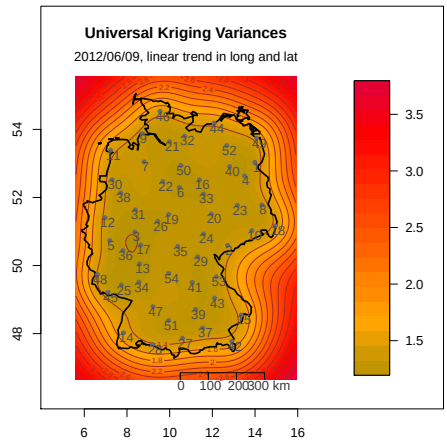
(b) Ordinary Kriging Variances



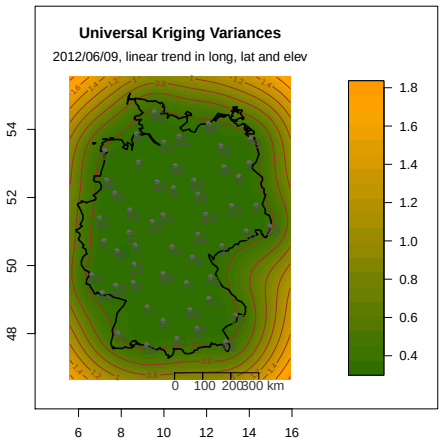
(c) Universal Kriging Variances with a linear trend in longitude



(d) Universal Kriging Variances with a linear trend in latitude



(e) Universal Kriging Variances with a linear trend in longitude and latitude



(f) Universal Kriging Variances with a linear trend in longitude, latitude and elevation

Figure 9.4: Kriging Variances of all considered kriging methods applied to the temperature data of 2012/06/09 in Germany

A Appendix

It remains to show that the kriging equations indeed yield the minimum for the prediction variances in ordinary and universal kriging, since we omitted both proofs ealier in this thesis. We prove these facts only for the covariance case, i.e. where $Z(\mathbf{x})$ or respectively $Y(\mathbf{x})$ are assumed to be second-order stationary with covariance function $C(\mathbf{h})$, respectively $C_Y(\mathbf{h})$. The corresponding prediction variances can be obtained by similar calculations compared with the variogram case, or easily by employing $\gamma(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h})$. Note that in ordinary and universal kriging, these variances only differ in the sense that we insert once the covariance matrix Σ of $\mathbf{Z}$ (ordinary kriging), and once the covariance matrix Σ_Y of the random vector $\mathbf{Y}$ of the residual process (universal kriging).

Our aim is to apply Theorem 2.15 (p. 8) given by Rao (1973) to show that our former obtained necessary conditions on the weights yield the minimum. We begin with the proof of the minimality of the prediction variance in ordinary kriging and finish with universal kriging.

A.1 Minimality of the prediction variance in ordinary kriging

Recall the minimization problem in section 7.4 (p. 53) translated into the covariance context:

$$\text{minimum of } \sigma_E^2 = C(\mathbf{0}) + \boldsymbol{\omega}^T \Sigma \boldsymbol{\omega} - 2\boldsymbol{\omega}^T \mathbf{c}_0 \text{ subject to } \boldsymbol{\omega}^T \mathbb{1} = 1$$

$$\text{Let } \Sigma_0 := \begin{pmatrix} C(\mathbf{x}_1 - \mathbf{x}_1) & \cdots & C(\mathbf{x}_n - \mathbf{x}_1) & C(\mathbf{x}_0 - \mathbf{x}_1) \\ & & & \\ C(\mathbf{x}_1 - \mathbf{x}_n) & \cdots & C(\mathbf{x}_n - \mathbf{x}_n) & C(\mathbf{x}_0 - \mathbf{x}_n) \\ C(\mathbf{x}_1 - \mathbf{x}_0) & \cdots & C(\mathbf{x}_n - \mathbf{x}_0) & C(\mathbf{x}_0 - \mathbf{x}_0) \end{pmatrix} = \left(\begin{array}{c|c} \sum & \mathbf{c}_0 \\ \hline \mathbf{c}_0^T & C(\mathbf{0}) \end{array}\right)$$

$\in \mathbb{R}^{(n+1)\times(n+1)}$, which is symmetric and positive definite, since it is the covariance matrix of the random vector $(Z(\mathbf{x}_1), \ldots, Z(\mathbf{x}_n), Z(\mathbf{x}_0))^T \in \mathbb{R}^{n+1}$.

Then consider the following, alternative representation of the prediction variance σ_E^2:

$$\sigma_E^2 = \left(\boldsymbol{\omega}^T, -1\right) \Sigma_0 \begin{pmatrix} \boldsymbol{\omega} \\ -1 \end{pmatrix} = \mathbf{x}^T \Sigma_0 \mathbf{x} \geq 0,$$

where $\mathbf{x} := \left(\boldsymbol{\omega}^T, -1\right)^T$ and the positive definiteness of Σ_0 ensures the nonnegativity of the prediction variance σ_E^2. This really holds, since

$$\left(\boldsymbol{\omega}^T, -1\right) \Sigma_0 \begin{pmatrix} \boldsymbol{\omega} \\ -1 \end{pmatrix} = \left(\boldsymbol{\omega}^T, -1\right) \begin{pmatrix} \Sigma\boldsymbol{\omega} - \mathbf{c}_0 \\ \mathbf{c}_0^T\boldsymbol{\omega} - C(\mathbf{0}) \end{pmatrix} = \boldsymbol{\omega}^T\Sigma\boldsymbol{\omega} - \boldsymbol{\omega}^T\mathbf{c}_0 - \mathbf{c}_0^T\boldsymbol{\omega} + C(\mathbf{0})$$

$$= C(\mathbf{0}) + \boldsymbol{\omega}^T\Sigma\boldsymbol{\omega} - 2\boldsymbol{\omega}^T\mathbf{c}_0 = \sigma_E^2.$$

Hence, we can rewrite the above minimization problem into the equivalent and more compact system

$$\text{minimum of } \mathbf{x}^T\Sigma_0\mathbf{x} \text{ subject to } B^T\mathbf{x} = U,$$

where $\mathbf{x} \in \mathbb{R}^{n+1}$, $B := (\mathbf{e}_{n+1}, \mathbb{1}_{n+1}) \in \mathbb{R}^{(n+1)\times 2}$ with unit vector $\mathbf{e}_{n+1} \in \mathbb{R}^{n+1}$, and $U := (-1, 0)^T \in \mathbb{R}^2$.

This way we wrote our optimization problem in the setting of Theorem 2.15 (p. 8), since Σ_0 is positive definite:

Set $A = \Sigma_0$ and let B and U be given as defined above. Further let S^- be a generalized inverse of $B^T A^{-1} B$, e.g. $S^- = \left(B^T A^{-1} B\right)^{-1}$, since $B^T A^{-1} B$ is invertible.

Hence, we minimize the equivalent system $\mathbf{x}^T A \mathbf{x}$ subject to $B^T \mathbf{x} = U$. Theorem 2.15 yields that we in fact achieve the minimum for the variance, which is given by

$$\inf_{B^T\mathbf{x}=U} \mathbf{x}^T \Sigma_0 \mathbf{x} = \inf_{B^T\mathbf{x}=U} \mathbf{x}^T A \mathbf{x} = U^T S^- U = U^T \left(B^T \Sigma_0^{-1} B\right)^{-1} U$$

and is attained at

$$\mathbf{x}_* = A^{-1} B S^- U = \Sigma_0^{-1} B \left(B^T \Sigma_0^{-1} B\right)^{-1} U.$$

Note that we did not follow this approach earlier in this thesis, since the resulting solution for the minimized variance and the "optimal" weights requires the inversion of the matrices Σ_0 and $B^T \Sigma_0 B$. This would be not really efficient and costs a lot of time. Since the matrix Σ_0 has to be updated every time for each new prediction point $\mathbf{x}_0$, we would have to invert or decompose both matrices for each single prediction point again. This would be inefficient and expensive, since we want to perform prediction for many different prediction points on a grid.

Hence we followed the Lagrange approach. Its final solution only requires the new set-up of $\mathbf{c}_0$ for a new prediction point. Thus, we have to invert or decompose Γ only once and can use its inverse or decomposition for each prediction point again. The necessary conditions, i.e. the first partial derivatives equated to zero, are thus sufficient, since we proved that we in fact achieve the uniquely determined minimum.

A.2 Minimality of the prediction variance in universal kriging

The same applies to the universal kriging case, where we have to deal with the residual covariances and some more constraints for uniform unbiasedness. Recall the minimization problem for the prediction variance in universal kriging as given in section 8.4 (p. 67). We translate it into the covariance setting:

$$\text{minimum of } \sigma_E^2 = C(\mathbf{0}) + \boldsymbol{\omega}^T \Sigma_Y \boldsymbol{\omega} - 2\boldsymbol{\omega}^T \mathbf{c}_{Y,0} \text{ subject to } F^T \boldsymbol{\omega} = \mathbf{f}_0$$

Fortunately, we can rewrite this problem into the following equivalent problem analogously to above. For this, let $\tilde{F} := \left(\begin{array}{ccc|c} & F & & \begin{matrix}0\\ \vdots \\ 0\end{matrix} \\ \hline 0 & \cdots & 0 & 1 \end{array}\right) \in \mathbb{R}^{(n+1)\times(n+1)}$ and $\tilde{\mathbf{f}}_0 := \left(\mathbf{f}_0^T, -1\right)^T$.

Further denote the symmetric and positive definite covariance matrix of the random vector $(Y(\mathbf{x}_1), \ldots, Y(\mathbf{x}_n), Y(\mathbf{x}_0))^T$ by $\Sigma_{Y,0} \in \mathbb{R}^{(n+1)\times(n+1)}$. We obtain

$$\text{minimum of } \sigma_E^2 = \mathbf{x}^T \Sigma_{Y,0} \mathbf{x} \geq 0 \text{ subject to } \tilde{F}^T \mathbf{x} = \tilde{\mathbf{f}}_0,$$

where $\mathbf{x} \in \mathbb{R}^{n+1}$.

Since our minimization problem can be written in this compact form, we infer that it coincides with the setting of Theorem 2.15. Hence, we can apply this theorem again, since $\Sigma_{Y,0}$ is positive definite and the constraints can be written in the special form above. Therefore, we infer that we achieve the minimum in our former computations. I.e. that the necessary conditions on the weights are sufficient for obtaining the unique minimum of the prediction variance subject to the unbiasedness conditions.

Finally, we conclude that in both cases, ordinary and universal kriging, the prediction variances are nonnegative and the necessary conditions for minimality of the underlying constrained minimization problems of the Lagrange approaches yield in fact the unique minimum and are hence sufficient. We prefered following the Lagrange approach, since it yields a nicer, more efficient solution for the kriging weights, variances and estimates.

References

Bohling, G. (2005). Introduction to Geostatistics and Variogram Analysis. Kansas Geological Survey. Available at: http://people.ku.edu/~gbohling/cpe940/Variograms.pdf.

Chauvet, P. and A. Galli (1982). *Universal Kriging*. Fontainebleau: Centre de Géostatistique et Morphologie Mathématique, École Nationale Supérieure des Mines de Paris.

Cressie, N. A. C. (1988). Spatial Prediction and Ordinary Kriging. *Mathematical Geology 20(4)*, 405–421.

Cressie, N. A. C. (1990). The Origins of Kriging. *Mathematical Geology 22(3)*, 239–252.

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

Cressie, N. A. C. and D. M. Hawkins (1980). Robust Estimation of the Variogram. *Journal of the International Association for Mathematical Geology 12(2)*, 115–125.

Cressie, N. A. C. and C. K. Wikle (2011). *Statistics for Spatio-Temporal Data*. Wiley Series in Probability and Statistics. Hoboken, New Jersey: John Wiley & Sons, Inc.

Durrett, R. (2010). *Probability: Theory and Examples* (4th ed.). Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press.

Fahrmeir, L., A. Hamerle, and G. Tutz (1996). *Multivariate statistische Verfahren* (2nd ed.). Berlin, New York: Walter de Gruyter.

Georgii, H.-O. (2013). *Stochastics: Introduction to Probability and Statistics* (2nd ed.). De Gruyter Textbook. Berlin, Boston: Walter de Gruyter.

Haskard, K. A. (2007). An anisotropic Matérn spatial covariance model: REML estimation and properties. The University of Adelaide. Available at: http://digital.library.adelaide.edu.au/dspace/handle/2440/47972.

Journel, A. G. and C. J. Huijbregts (1978). *Mining Geostatistics*. London, San Diego: Academic Press, Harcourt Brace & Company.

Kitanidis, P. K. (1997). *Introduction to Geostatistics: Applications in Hydrogeology*. Stanford-Cambridge Program. Cambridge: Cambridge University Press.

Kolmogorov, A. N. (1941a). Interpolation and extrapolation of stationary random sequences. *Izvestiia Akademii Nauk SSSR, Seriia Matematicheskiia 5*, 3–14. [Translation (1962), Memo RM-3090-PR, Rand Corp., Santa Monica, CA.].

Kolmogorov, A. N. (1941b). The local structure of turbulence in an incompressible fluid at very large Reynolds numbers. *Doklady Akademii Nauk SSSR 30*, 301–305. [Reprinted (1961), in Turbulence: Classic Papers on Statistical Theory, S.K. Friedlander and L. Topping, eds. Interscience Publishers, New York, 151–155].

Krige, D. G. (1951). A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of the Chemical, Metallurgical and Mining Society of South Africa 52*, 119–139.

Matérn, B. (1960). Spatial Variation. *Meddelanden fran Statens Skogsforskningsinstitut 49(5)*. [Second edition (1986), Lecture Notes in Statistics, No. 36, New York: Springer].

Matheron, G. (1962). *Traité de Géostatistique Appliquée*. Mémoires du Bureau de Recherches Géologiques et Minières, Tome I(14). Paris: Éditions Technip.

Matheron, G. (1963). Principles of Geostatistics. *Economic Geology 58*, 1246–1266.

Matheron, G. (1969). *Le Krigeage Universel*. Les Cahiers du Centre de Morphologie Mathématique de Fontainebleau, Fascicule 1. Fontainebleau: École Nationale Supérieure des Mines de Paris.

Matheron, G. (1971). *The Theory of Regionalized Variables and Its Applications*. Les Cahiers du Centre de Morphologie Mathématique de Fontainebleau, No. 5. Fontainebleau: École Nationale Supérieure des Mines de Paris.

Matheron, G. (1989). *Estimating and Choosing: An Essay on Probability in Practice*. Berlin, Heidelberg: Springer.

Nguyen, H. T. and G. S. Rogers (1989). *Fundamentals of Mathematical Statistics*. Springer Texts in Statistics. Volume I: Probability for Statistics. New York: Springer.

Pebesma, E. J. (2001). *Gstat user's manual*. Dept. of Physical Geography, Utrecht University. gstat 2.3.3. May 29, 2001. Available at: http://www.gstat.org/gstat.pdf.

Pebesma, E. J. (2013). *Spatio-temporal geostatistics using gstat*. Institute of Geoinformatics, University of Münster. February 19, 2013. Available at: http://cran.r-project.org/web/packages/gstat/vignettes/st.pdf.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications* (2nd ed.). Wiley Series in Probability and Mathematical Statistics. New York: John Wiley & Sons, Inc.

Rasmussen, C. E. and C. K. I. Williams (2006). *Gaussian Processes for Machine Learning*. Adaptative Computation and Machine Learning Series. Massachusetts Institute of Technology: The MIT Press. 79–104.

Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Series in Statistics. New York: Springer.

Wackernagel, H. (2003). *Multivariate Geostatistics: An Introduction with Applications* (3rd ed.). Berlin, Heidelberg: Springer.

Webster, R. and M. A. Oliver (2007). *Geostatistics for Environmental Scientists* (2nd ed.). Statistics in Practice. Chichester: John Wiley & Sons, Ltd.

Wold, H. (1938). *A Study in the Analysis of Stationary Time Series*. Uppsala: Almqvist and Wiksells.