- Ludwig-Maximilians-Universität München -
Institut für Statistik

Geostatistik mit GRASS und R

Bachelorarbeit
zur Erlangung des akademischen Grades
Bachelor of Science (B. Sc.)

von

Hannah Otterbach

Gutachter: Prof. Dr. Volker J. Schmid

München, den 16. Juni 2014

Inhaltsverzeichnis

1 Einleitung

Die vorliegende Arbeit beschäftigt sich mit der geostatistischen Analyse von Daten aus dem Geoinformationssystem *Geographical Resources Analysis Support System* GRASS mit der Statistik-Software R. Von Interesse ist die Anbindung zwischen beiden Programmen über das R-Paket `spgrass6` von Bivand (2013) und die Theorie geostatistischer Verfahren, explorative Datenanalyse, Variogramme und Kriging, sowie die programmiertechnische Umsetzung dieser in R mithilfe des R-Paketes `gstat` von Pebesma (2004).

Geostatistische Daten sind allgemein Daten mit räumlichem Bezug. Sie stammen dabei aus den verschiedensten Bereichen. So werden Methoden der Geostatistik nach Navratil (2006) beispielsweise sowohl zur Exploration von Erzlagerstätten, als auch zur Analyse von Bodenverunreinigungen auf Basis von Bodenproben oder Vorhersage von Temperaturwerten an bestimmten Punkten eingesetzt.
Die geostatistischen Daten bestehen nach Wackernagel (1995) üblicherweise aus Messungen $y(s_i)$ an einer begrenzten diskreten Anzahl von Lokationen s_i innerhalb eines zusammenhängenden Gebietes $D \subseteq \mathbb{R},\ d \in \mathbb{N}$. Ziel der hier vorgestellten geostatistischen Methoden ist es dann auf Basis dieser Messungen Schlüsse auf einen zugrunde liegenden stetigen stochastischen Prozess $Y(s)$ zu ziehen. Mithilfe von diesem lassen sich Werte für unbeobachtete Lokationen s_0 vorhersagen. Diese Vorhersage beruht auf der Annahme, dass eine räumliche Abhängigkeitsstruktur besteht, der Wert an einer unbeobachteten Lokation s_0 also von den umliegenden Lokationen abhängig ist. Speziell besteht die Annahme, dass sich nahe Lokationen ähnlicher sind, als weiter entfernte.

Bevor konkrete geostatistische Verfahren und ihre Anwendung in der Arbeit vorgestellt werden, wird in Kapitel 2 auf die beiden verwendeten Programme GRASS und R und die Anbindung zwischen diesen über das `spgrass6`-Paket eingegangen. Außerdem enthält das Kapitel einen kurzen Überblick über den in der Arbeit beispielhaft zur Veranschaulichung der geostatistischen Methoden genutzten Datensatz *Spearfish Sample Database* aus GRASS, sowie eine Einführung in den Umgang mit räumlichen Daten in R.
Das Kapitel 3 beinhaltet anschließend notwendige mathematische Grundlagen, um das Verständnis der vorgestellten geostatistischen Methoden zu erleichtern. Außerdem werden verschiedene Annahmen, die für den zugrunde liegenden stochastischen Prozess $Y(s)$ getroffen werden können und häufig für die Verwendung der geostatistischen Methoden vorausgesetzt werden, vorgestellt.
Die folgenden Kapitel 4, 5 und 6 enthalten schließlich die in der Arbeit vorgestellten geostatistischen Methoden. Begonnen wird in Kapitel 4 mit explorativen Visualisierungsmöglichkeiten. Dabei werden zum einen allgemeine Darstellungsmöglichkeiten und zum anderen spezielle Möglichkeiten, die den räumlichen Bezug der Daten berücksichtigen, vorgestellt. In Kapitel 5 wird gezeigt, wie die räumliche Abhängigkeitsstruktur mithilfe einer Funktion, die als *Variogramm* bezeichnet wird, modelliert werden kann. Vario-

gramme bieten dann die Basis für die hier vorgestellte statistische Prädiktions- und Interpolationsmethode *Kriging*. Diese berücksichtigt bei der Prädiktion von Werten an unbeobachteten Lokationen die räumliche Abhängigkeitsstruktur, indem die Werte der umliegenden Lokationen je nach Entfernung unterschiedlich gewichtet werden.
Zum Abschluss gibt das Kapitel 7 eine Zusammenfassung über die Arbeit und einen Ausblick über weitere nicht genannte Spezialfälle der vorgestellten geostatistischen Methoden und zusätzliche Möglichkeiten, geostatistische Daten in R zu analysieren.

Aufgebaut ist die Arbeit so, dass für die geostatistischen Verfahren in jedem Abschnitt jeweils vorweg der theoretische Hintergrund erklärt wird und dann die Anwendung mit R folgt. Verwendet wurden die zum Stand der Arbeit aktuellsten Versionen 3.1.0 beta für R und 6.4.3 für GRASS. Die in R verwendeten Funktionen für die geostatistischen Methoden stammen, soweit nicht anders gekennzeichnet, aus dem R-Paket `gstat` und alle weiteren angewandten Funktionen aus den standardmäßig in R implementierten Paketen.

Notation

Vorweg sind in der Tabelle 1 zur Übersicht und Vergleichbarkeit, da in der Literatur unterschiedliche Notationen existieren, einige in dieser Arbeit verwendeten Symbole für relevante Größen in der Geostatistik und ihre Bedeutung dargestellt.

Tabelle 1: Verwendete Symbolik

Symbol	Bedeutung
s	räumliche Lokation, Position im Raum
$Y(s)$	räumlicher stochastischer Prozess
D	durchgehendes Untersuchungsgebiet ($\mathbb{R}^d$)
d	Dimension des Untersuchungsgebietes D
h	Abstand zwischen zwei Punkten
$\gamma(h)$	Semivariogramm
$2\gamma(h)$	Variogramm

Zusätzlich zur Notation ist zu beachten, dass die in der Arbeit genannten Definitionen und Gleichungen nummeriert sind. Wird sich auf diese bezogen, werden die Nummern von Gleichungen mit $(\cdot)$ angegeben und die von Definitionen ohne Klammer. Weiterhin wird in der Arbeit häufig der angewendete R-Code für die geostatistischen Methoden angegeben. Dieser wird abgesetzt wie

```
> x <-
+ 1:10
> y <- 11:20
> x+y
[1] 12 14 16 18 20 22 24 26 28 30
```

dargestellt. Mit `>` vorweg werden die eingegebenen R-Befehle gekennzeichnet und ohne ist die resultierende Ausgabe angegeben. `+` stellt die Fortsetzung eines R-Befehls dar.

2 GRASS und R

Dieses Kapitel gibt Hintergrundinformationen zu den verwendeten Programmen. Abschnitt 2.1 beschreibt das Geoinformationssystem GRASS, sowie den beispielhaft für die später vorgestellten geostatistischen Methoden verwendeten Datensatz *Spearfish Sample Database*. Der folgende Abschnitt 2.2 geht auf die für die statistische Auswertungen genutzte Software R und den Umgang mit räumlichen Daten in dieser ein. Basierend auf den beiden Programmen ist zuletzt die Anbindung zwischen diesen durch das R-Paket `spgrass6` von Interesse, welche in Abschnitt 2.3 vorgestellt wird.

2.1 Geographical Resources Analysis Support System (GRASS)

Geographical Resources Analysis Support System (GRASS) ist ein Open Source Geoinformationssystem (GIS), welches unter `http://grass.osgeo.org/download/software/` verfügbar ist. Entwickelt wurde das Programm nach Neteler (2003) 1983 vom *U.S. Army Corps of Engineers/CERL (Construction Engineering Research Lab)* für militärische Planungszwecke. Ende der 80er Jahre wurde es dann der Öffentlichkeit zur Verfügung gestellt. Heute wird die Software größtenteils durch das *GRASS Development Team*, aber auch einige freie Programmierer weiterentwickelt. GRASS ist nach Neteler (2003) ein kombiniertes Raster-/Vektor-GIS, dass über 400 Programme zur Datenaggregierung, -analyse und -synthese von Raster-, Vektor- und Punktdaten enthält. Insbesondere enthält GRASS einige Funktionalitäten zum Umgang mit räumlichen und damit auch geostatistischen Daten. Bedienen lässt sich das Programm zum einen über eine integrierte Benutzeroberfläche und zum anderen über die Kommandozeile. Einen Überblick über die Software liefern zahlreiche Bücher, Manuals, Tutorials und Internetseiten, von denen einige auf der Homepage zu finden sind. Weiterhin existieren ein eigenes GRASS-Wiki unter `http://grasswiki.osgeo.org/wiki/`, sowie verschiedene Mailinglisten und Foren. Für einen ausführlichen aktuellen Überblick über das Programm ist Neteler und Mitasova (2008) zu empfehlen.

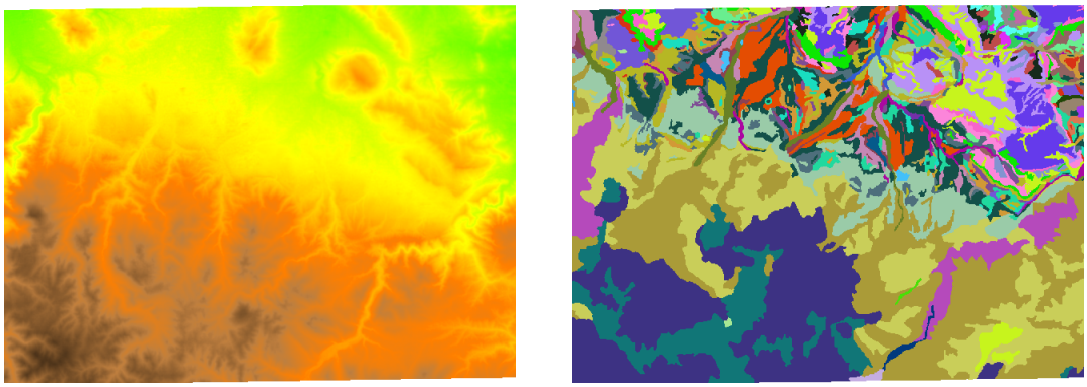
Quelle: GRASS Development Team (2012)

Abb. 2.1: GRASS-Logo

Spearfish Sample Database

An dem Beispieldatensatz *Spearfish Sample Database* werden die in der Arbeit vorgestellten geostatistischen Methoden veranschaulicht. Der Datensatz kann, analog zu der Software GRASS selbst, auf der Internetseite `http://http://grass.osgeo.org/download/`

`sample-data/` heruntergeladen werden. Beim Start von GRASS kann er dann als Lokation gesetzt werden. Auf der Internetseite befindet sich auch eine Dokumentation (GRASS Development Team; 1993) über den Datensatz. Er wird vom GRASS Development Team bereitgestellt, um den Nutzern von GRASS eine solide und vielfältige Datengrundlage zu bieten, mit der nahezu alle Funktionen und Möglichkeiten des Programms erlernt und getestet werden können. Nach Neteler (2003) enthält der Datensatz zwei topographische Karten im Maßstab 1:24.000 mit den Namen *Spearfish* und *Deadwood North* aus dem Westen von South Dakota in den USA. Die Karten beinhalten nach GRASS Development Team (1993) die Umgebung der Stadt Spearfish, sowie einen Großteil des *Black Hills National Forest*, wie unter anderem *Mount Rushmore*. Als Projektion liegt das Koordinatensystem *Universal Transverse Mercator* (UTM) zugrunde (Neteler; 2003).
Insgesamt lassen sich in GRASS aus dem *Spearfish Sample Database* 35 Rasterkarten und 40 Vektorkarten einlesen, wobei einige Karten sich aus anderen Karten ableiten oder als Kombination anderer Karten ergeben (GRASS Development Team; 1993). Vorhandene Rasterkarten enthalten beispielsweise Informationen über die Höhe, die Bodenbeschaffenheit, die Vegetationsbedeckung oder Walddichte und vorhandene Vektorkarten über die Straßen, Flüsse oder Gebietsgrenzen. Eine genaue Übersicht über die verfügbaren Karten befindet sich in der Dokumentation, wobei sich die Anzahl und Namen der dort genannten Karten nicht komplett mit denen deckt, die tatsächlich eingelesen werden können. In dieser Arbeit werden zur Veranschaulichung der geostatistischen Methoden



(a) Höhendaten `elevation.dem`

(b) Bodendaten `soils`

Abb. 2.2: Mit GRASS visualisierte Rasterkarten

die zwei Rasterkarten `elevation.dem` und `soils` des Datensatzes verwendet (siehe Abb. 2.2). Letztere enthält Informationen über 54 verschiedene Bodenarten in dem Gebiet und die erstgenannte ein Höhenmodell mit Werten zwischen 1066m und 1840m.
Prinzipiell werden für die hier vorgestellten geostatistischen Methoden stetige Daten vor-

ausgesetzt. Die Bodendaten sind eigentlich kategorial, werden aber aufgrund mangelnder stetiger Alternativen in der *Spearfish Sample Database* notwendigerweise als quasistetig angenommen.

2.2 R

R ist nach R Core Team (2014) eine Programmiersprache und -umgebung, in der eine große Bandbreite von statistischen Methoden und graphischen Darstellungsmöglichkeiten implementiert ist. Die Software ist daher für statistische Anwendungen weit verbreitet und wird auch in dieser Arbeit zur statistischen Analyse verwendet. R hat den Vorteil, dass es wie GRASS eine freie Software und erweiterbar ist. Das Programm kann, wie auch zahlreiche Pakete, unter `http://www.R-project.org/` heruntergeladen werden. Die Anwendung von R und den Paketen ist sehr gut dokumentiert, so können Manuals und auch einige Einführungen auf der Homepage gefunden werden. Außerdem existieren mit `http://de.wikibooks.org/wiki/GNU_R` und `http://rwiki.sciviews.org/doku.php` zwei R Wikis. Für die geostatistischen Methoden wurde das R-Paket `gstat` verwendet, da dies nach Bivand et al. (2013) im Vergleich zu anderen R-Paketen für geostatistische Daten den größten Umfang an Funktionen für die geostatistische Analyse bereitstellt. Auch die Autoren von Neteler und Mitasova (2008) bevorzugen `gstat` für ihre Auswertungen in R. Einen Überblick über die enthaltenen Funktionen und Möglichkeiten findet sich in Pebesma (2004). Allgemein müssen R-Pakete vor der Verwendung aufgerufen und einmalig installiert werden. Dies funktioniert beispielhaft für `gstat` und schließlich analog für andere R-Pakete nach dem folgenden Schema.

Quelle: R Core Team (2014)

Abb. 2.3: R-Logo

```
> install.packages("gstat", dependencies=TRUE)
> library(gstat)
```

Das `dependencies`-Argument ist dabei relevant, um Pakete, auf denen das zu installierende Paket basiert, mit zu installieren. Hilfe zu Paketen oder auch Funktionen kann in R über die beiden Funktionen

```
> ?gstat
> help(gstat)
```

erhalten werden. Dies ist in der Regel sehr nützlich, um einen Überblick über die Funktionsweise, sowie notwendige Argumente und ihre Eigenschaften von R-Funktionen zu bekommen oder aber eine Zusammenfassung über beinhaltete Funktionen eines R-Paketes zu erhalten.

Räumliche Daten in R

Allgemein werden Daten in R nach Ligges (2007) je nach ihrer Datenstruktur in unterschiedlichen Klassen abgespeichert. Unterschieden werden kann nach Ligges (2007) allgemein zwischen alten oder `S3`- und neuen oder `S4`-Klassen. Zu `S3`-Klassen zählen beispielsweise Vektoren, Matrizen, Arrays, Datensätze und Listen, welche jeweils beliebige weitere Objekte enthalten können. Der Unterschied dieser Klassen zu den `S4`-Klassen besteht nach Ligges (2007) darin, dass die `S4`-Klassen aus sogenannten *slots* bestehen, welche ebenfalls weitere Objekte enthalten, die nun üblicherweise wieder `S3`-Klassen zugeordnet sind. Ein weiterer Unterschied zwischen beiden Klassen ist nach Ligges (2007), dass auf Objekte in `S3`-Klassen mit `$` und auf Objekte von `S4`-Klassen mit `@` zugegriffen wird. Vorteil der `S4`-Klassen ist nun, dass Klassen, Methoden und Funktionen formal definiert sind und genau spezifiziert ist, welche Struktur ein Objekt einer Klasse hat, während in `S3`-Klassen auch Objekte existieren können, die keiner oder mehreren Klassen zugeordnet sind. (Ligges; 2007). Für genauere Informationen über Datenstrukturen in R sei auf Ligges (2007) verwiesen.

Klassen für räumliche Daten und damit auch Geodaten gehören nach Bivand et al. (2013) nun zu den `S4`-Klassen. Konkret sind die räumlichen Klassen in dem R-Paket `sp` implementiert, welches ausführlich in Bivand et al. (2013) beschrieben ist. Basierend darauf gibt dieser Abschnitt einen kurzen Überblick über die existierenden Klassen für räumliche Daten.
Zugrunde liegt allen die Klasse `Spatial`, welche aus den zwei *slots* `bbox` und `proj4string` besteht. Die `bbox` ist eine Matrix, die einen Bereich, bestehend aus Koordinaten in zwei Spalten (`min`, `max`) und zwei Zeilen mit den Hoch- und Rechtswerten (y- und x-Achse), darstellt. `proj4string` ist ein `CRS`-Objekt (*Coordinate Reference System*), in dem Informationen über das Bezugssystem der Koordinaten enthalten sind. Einen Überblick über die Klasse `Spatial` gibt die Funktion `getClass()` aus dem R-Paket `sp`.

```
> getClass("Spatial")
Class "Spatial" [package "sp"]

Slots:

Name:         bbox proj4string
Class:      matrix         CRS

Known Subclasses:
Class "SpatialPoints", directly
Class "SpatialGrid", directly
Class "SpatialLines", directly
Class "SpatialPolygons", directly
```

```
Class "SpatialPointsDataFrame", by class "SpatialPoints", distance 2
Class "SpatialPixels", by class "SpatialPoints", distance 2
Class "SpatialGridDataFrame", by class "SpatialGrid", distance 2
Class "SpatialLinesDataFrame", by class "SpatialLines", distance 2
Class "SpatialPixelsDataFrame", by class "SpatialPoints", distance 3
Class "SpatialPolygonsDataFrame", by class "SpatialPolygons", distance 2
```

Zu sehen ist, dass die Klasse `Spatial` wie zu erwarten aus dem R-Paket `sp` stammt und aus den beiden *slots* `bbox` und `proj4string` besteht. Außerdem sind die Subklassen aufgelistet. Die Subklassen enthalten ebenfalls die beiden genannten *slots*, wobei sie noch zusätzliche beinhalten können.
Geodaten, die nach Bivand et al. (2013) üblicherweise als Punkte, Linien, Polygone oder Raster vorliegen, werden nun, je nachdem wie sie bestehen, in einer der genannten Subklassen gespeichert. Punkte werden als `SpatialPoints`-Objekt abgespeichert, wo zusätzlich ein `coords`-*slot* enthalten ist, in dem nach Bivand (2005) die Koordinaten der Punkte gespeichert sind. Linien und Polygone, oder allgemein Vektordaten, werden dementsprechend als `SpatialLines`- oder `SpatialPolygons`-Objekt abgespeichert und Rasterdaten als `SpatialGrid`- oder `SpatialPixels`-Objekt. Der Unterschied zwischen den letzten beiden besteht nach Bivand (2005) darin, dass für die Klasse `SpatialPixels` kein vollständiges Raster vorliegen muss.

2.3 GRASS/R-Interface über `spgrass6`

Zwischen GRASS und R existiert eine Anbindung über das R-Paket `spgrass6` (Bivand; 2013), welches für komplexe statistische Analysen relevant ist, die in GRASS alleine nicht durchgeführt werden können. Vorgestellt wird das Paket beispielsweise in Bivand (2005), Bivand (2007) und Neteler und Mitasova (2008). Ein kompletter Überblick über das Paket findet sich unter Bivand (2013). Auf dieser genannten Literatur basiert auch die in diesem Kapitel vorgestellte Einführung in das Paket.

Dieser Abschnitt soll vor allen Dingen zeigen, wie GRASS-Daten in R eingelesen und dort mit Funktionen aus anderen R-Paketen bearbeitet werden können. Neben dieser Möglichkeit mit dem `spgrass6`-Paket existieren in dem Paket noch weitaus mehr Funktionen, mit denen beispielsweise von R aus Raster- und Vektorobjekte nach GRASS exportiert werden können (`writeRAST6()` und `writeVECT6()`) oder auch GRASS-Befehle von R aus gestartet werden können (`execGRASS()`). Da Funktionen dieser Art in der Arbeit nicht verwendet werden, sei für genauere Informationen auf die Dokumentation des Paketes in Bivand (2013) verwiesen.

Damit die Anbindung zwischen GRASS und R funktioniert, muss nach Bivand (2013) sicher gestellt sein, dass eine R-Version ≥ 2.12 und eine GRASS-Version ≥ 6.3 installiert ist. In R sollten außerdem die Pakete `XML` sowie `sp` vorhanden sein, da die in der

Anbindung genutzten Objektklassen die räumlichen Klassen aus sp sind. Arbeiten lässt sich mit beiden Programmen, indem R von GRASS aus über den Aufruf R in der Kommandozeile oder GRASS von R aus mit der Funktion initGRASS() gestartet wird. Für letztere Variante muss vor dem Start von GRASS spgrass6 installiert sein und aufgerufen werden, da die Funktion bereits aus dem Paket stammt, während dies bei der ersten Variante erst nach dem Aufruf von R möglich ist.

```
> library(spgrass6)
GRASS GIS interface loaded with GRASS version: GRASS 6.4.3 (2013)
and location: spearfish60
```

Mit dem Aufruf von spgrass6 werden nach Neteler und Mitasova (2008) automatisch die Metadaten über die in GRASS geöffnete Lokation in R übertragen. Hier zeigt sich, dass der bereits vorgestellte Datensatz *Spearfish Sample Database* zugrunde liegt. Ein Überblick über die Metadaten kann dann mit der Funktion gmeta6() gegeben werden.

```
> gmeta6()
gisdbase    /home/grassdata
location    spearfish60
mapset      PERMANENT
rows        466
columns     633
north       4928000
south       4914020
west        590010
east        609000
nsres       30
ewres       30
projection  +proj=utm +zone=13 +a=6378206.4 +rf=294.9786982 +no_defs
+nadgrids=/usr/lib/grass64/etc/nad/conus +to_meter=1.0
```

Zu sehen ist der Koordinatenbereich, den die Lokation umfasst, sowie das Bezugssystem des *Spearfish Sample Database*, UTM. Um nun Analysen der Daten aus GRASS durchführen zu können, müssen noch die jeweils interessierenden Karten aus dem Datensatz importiert werden. Dies funktioniert für Rasterdaten mit readRAST6() und für Vektordaten mit readVECT6(). Wie bereits unter Abschnitt 2.1 erwähnt, werden in der Arbeit beispielhaft die Höhendaten elevation.dem und die Bodendaten soils verwendet. Beide Datensätze sind Rasterkarten und werden daher folgendermaßen eingelesen.

```
# Einlesen der Höhendaten elevation.dem in R
> elev <- readRAST6("elevation.dem")

# Einlesen der Bodendaten soils in R
> soils <- readRAST6("soils")
```

Sowohl für `readRAST6()` als auch für `readVECT6()` existieren einige Argumente, die beim Einlesen je nach Gebrauch angewendet werden können. So kann beispielsweise mit `cat` angegeben werden, ob die in GRASS vergebenen Namen für Kategorien einer Variable importiert werden sollen. Werden keine zusätzlichen Argumente angegeben, werden nach Bivand et al. (2013) weder Namen für Kategorien noch Farben o. ä. importiert. Je nach Anwendung sind unterschiedliche Verwendungen der Argumente sinnvoll. Für diese Arbeit reicht das Einlesen ohne zusätzliche Argumente aus. Prinzipiell genügen diese genannten Schritte bereits, um mit den Daten aus GRASS in R arbeiten zu können.
R-Funktionen können nun nach dem Einlesen der GRASS-Daten wie üblich auf die Datensätze angewendet werden. Ein Überblick über die eingelesenen Daten kann beispielsweise analog zu Abschnitt 2.2 mit der Funktion `class()` erhalten werden. Neben der Information, in welcher Klasse die Daten abgespeichert wurden, ist jedoch in der Praxis von Interesse, wie die Daten aufgebaut sind und was sie enthalten. Um diese Informationen gemeinsam abzurufen, bietet sich die Funktion `summary()` an.

```
# Überblick über die Höhendaten elevation.dem
> summary(elev)
Object of class SpatialGridDataFrame
Coordinates:
         min     max
[1,]  590010  609000
[2,] 4914020 4928000
Is projected: TRUE
proj4string :
[+proj=utm +zone=13 +a=6378206.4 +rf=294.9786982
+no_defs +nadgrids=/usr/lib/grass64/etc/nad/conus
+to_meter=1.0]
Grid attributes:
  cellcentre.offset cellsize cells.dim
1            590025       30       633
2           4914035       30       466
Data attributes:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   1066    1200    1316    1354    1488    1840    2661
```

Beispielhaft ist hier die Zusammenfassung für die Höhendaten `elevation.dem` gegeben. Für die Bodendaten `soils` ergibt sie sich analog. Zu sehen ist, dass die Daten als `SpatialGridDataFrame`-Objekt abgespeichert wurden. Außerdem sind die beiden *slots* `bbox` und `proj4string` angegeben. Zusätzlich beinhaltet die Klasse `SpatialGridDataFrame` nach (Bivand et al.; 2013) *slots* mit Informationen über die Zellen der Rasterdaten, welche ebenfalls in der Ausgabe enthalten sind. Zuletzt ist eine univariate Statistik mit angegeben. Die Koordinaten der Höhendaten befinden sich zwischen 590010 und

609000 für die Rechtswerte und 4914020 und 4928000 für die Hochwerte. Die Höhen liegen, wie bereits in Abschnitt 2.1 erwähnt, zwischen 1066 und 1840 und als Bezugssystem ergibt sich analog zu dem Gesamtdatensatz *Spearfish Sample Database* UTM. Der Mittelwert der Höhen liegt bei 1354 und fehlende Werte sind 2661 vorhanden. Diese müssen allerdings auf die Gesamtzahl an Beobachtungen bezogen werden, welche sich mit

```
> length(elev)
[1] 294978
```

zu fast 300.000 ergibt.
Weitere Anwendungen der Daten, speziell geostatistische Analysen, werden in den folgenden Kapitel 4, 5 und 6 dargestellt.

3 Mathematische Grundlagen

Um die in den nächsten Kapiteln dargestellten geostatistischen Methoden und Umformungen von Gleichungen leichter verstehen und nachvollziehen zu können, werden in diesem Kapitel einige mathematische Grundlagen erläutert. Dabei wird im Abschnitt 3.1 auf einige Grundbegriffe der Wahrscheinlichkeitstheorie eingegangen, der zweite Abschnitt 3.2 beinhaltet Grundlagen zur Parameterschätzung und ihrer Beurteilung und im dritten Abschnitt 3.3 werden mögliche Annahmen für stochastische Prozesse in der Geostatistik dargestellt. Angegeben sind die folgenden Definitionen und Begriffe für den eindimensionalen Fall. Im mehrdimensionalen Fall ergeben sie sich entsprechend.

3.1 Wahrscheinlichkeitstheoretische Grundbegriffe

Zugrunde liegen nahezu jeder statistischen Auswertung und damit auch geostatistischen Anwendungen *Zufallsvariablen*, die einen Zufallsvorgang beschreiben. Die aufgeführten Definitionen dieser und ihrer Eigenschaften finden sich, soweit nicht anders gekennzeichnet, in Fahrmeir et al. (2011).

Definition 3.1 (Zufallsvariable)
Eine Zufallsvariable Y ist eine Abbildung

$$Y : \Omega \longrightarrow \mathbb{R},\ \omega \mapsto Y(\omega) = y,$$

die bei einem zufälligen Vorgang für jedes Ereignis ω aus der Ereignismenge Ω eine reelle Zahl $Y(\omega) = y$ annimmt.

Der Wert $Y(\omega) = y$ wird dabei als *Realisation* bezeichnet. Von Interesse ist in der Regel die Verteilung der Zufallsvariablen Y, die mit der *Wahrscheinlichkeitsverteilung* angegeben wird.

Definition 3.2 (Wahrscheinlichkeitsverteilung)
Die Wahrscheinlichkeitsverteilung oder kurz Verteilung der Zufallsvariable Y ist die Zuordnung von Wahrscheinlichkeiten

$$P(Y \in B),$$

mit $P(Y \in B) \in [0, 1]$, wobei B eine Teilmenge von $\mathbb{R}$ darstellt.

Die Verteilung einer Zufallsvariablen kann durch ihre *Momente* charakterisiert werden. Die k-ten Momente einer Zufallsvariable Y ergeben sich nach Wackernagel (1995) mit

$$E(Y^k) = \int_{-\infty}^{\infty} y^k f(y) dy. \tag{3.1}$$

Die beiden in der Praxis am häufigsten verwendeten Momente sind das erste und das zweite Moment. Das erste Moment stellt den *Erwartungswert* dar und mithilfe von diesem und dem zweiten Moment lässt sich die *Varianz* berechnen.

Definition 3.3 (Erwartungswert)
Der Erwartungswert $E(Y)$ einer stetigen Zufallsvariablen Y mit Dichte $f(y)$ berechnet sich durch

$$\mu = E(Y) = \int_{-\infty}^{\infty} yf(y)dy.$$

Zwei wichtige Eigenschaften des Erwartungswertes, die für Umformungen in den folgenen Kapiteln relevant sind, sind die *Linearität* und *Additivität*.
Für die Linearkombination $X = aY + b$ folgt

$$\mathrm{E}(X) = \mathrm{E}(aY + b) = a\mathrm{E}(Y) + b \tag{3.2}$$

und für die Addition zweier (oder generell mehrerer) Zufallsvariablen $Y + X$

$$\mathrm{E}(Y + X) = \mathrm{E}(Y) + \mathrm{E}(X). \tag{3.3}$$

Definition 3.4 (Varianz)
Die Varianz $Var(Y)$ einer stetigen Zufallsvariablen Y mit Dichte $f(y)$ ist gegeben mit

$$\sigma^2 = Var(Y) = \int_{-\infty}^{\infty} (y - \mu)^2 f(y)dy.$$

Diese Darstellung ergibt sich aus der Differenz des zweiten und dem quadrierten ersten Moment:

$$\mathrm{Var}(Y) = \mathrm{E}(Y^2) - \mathrm{E}(Y)^2 = \mathrm{E}(Y^2) - \mu^2. \tag{3.4}$$

Neben den ersten beiden Momenten zur Beschreibung des Erwartungswertes und der Varianz ist das Verhalten von Zufallsvariablen zueinander in der Geostatistik von Bedeutung, da diese hier üblicherweise nicht unabhängig voneinander sind. Beschrieben werden kann dieses Verhalten durch die *Kovarianz* und *Korrelation*.

Definition 3.5 (Kovarianz)
Die Kovarianz von zwei Zufallsvariablen Y und X ist bestimmt durch

$$\begin{aligned} Cov(Y, X) &= E((Y - E(Y))(X - E(X))) \\ &= E(YX) - E(Y)E(X). \end{aligned} \tag{3.5}$$

Im Allgemeinen erfüllt die Kovarianz die Eigenschaft der *Symmetrie*

$$\mathrm{Cov}(Y, X) = \mathrm{Cov}(X, Y) \tag{3.6}$$

und ergibt im Fall derselben Zufallsvariable die Varianz, da

$$\mathrm{Cov}(Y, Y) = \mathrm{E}(Y^2) - \mathrm{E}(Y)^2$$

$$\overset{(3.4)}{=} \mathrm{Var}(Y). \tag{3.7}$$

Als Interpretation ergibt sich dann für positive Kovarianzen, dass die Zufallsvariablen Y und X einen gleichsinnigen Zusammenhang aufweisen, und negative Kovarianzen deuten auf einen gegensinnigen Zusammenhang hin. Die Kovarianz ist abhängig vom Maßstab der Zufallsvariablen, was die Interpretation erschwert. Daher wird sie häufig mithilfe des *Korrelationskoeffizienten* normiert.

Definition 3.6 (Korrelationskoeffizient)
Der Korrelationskoeffizient ist bestimmt durch

$$\rho(Y,X) = \frac{Cov(Y,X)}{\sqrt{Var(Y)}\sqrt{Var(X)}} = \frac{Cov(Y,X)}{\sigma_Y \sigma_X}.$$

Der Wertebereich von diesem ist aufgrund der Normierung mit

$$-1 \leq \rho(Y,X) \leq 1 \tag{3.8}$$

begrenzt. Gilt $\rho(Y,X) = 0$ kann von Unabhängigkeit und damit Unkorreliertheit der beiden Zufallsvariablen ausgegangen werden. Die Interpretation für den Korrelationskoeffizienten ergibt sich analog zu der Kovarianz, wobei für beide beachtet werden muss, dass nur lineare Zusammenhänge erkannt werden.

3.2 Parameterschätzung

Ziel der Geostatistik ist es, aufgrund von beobachteten Lokationen Schätzungen für Messwerte an unbekannten Lokationen zu machen. Die folgenden Grundlagen zur Parameterschätzung sind wieder Fahrmeir et al. (2011) entnommen. Grundsätzlich kann zwischen *Punktschätzung* und *Intervallschätzung* unterschieden werden, wobei in dieser Arbeit nur die Punktschätzung betrachtet wird, da die hier vorgestellten geostatistischen Schätzungen unter diese fallen.

Definition 3.7 (Schätzstatistik)
Eine Schätzstatistik für den Grundgesamtheitsparameter θ ist eine Funktion

$$T = g(Y_1, \ldots, Y_n)$$

der Stichprobenvariablen $Y_1, \ldots, Y_n$. Der konkrete Schätzwert resultiert dann aus den Realisationen $y_i, \ldots, y_n$.

Zwei Gleichungen aus Navratil (2006) geben Beispiele für Schätzstatistiken. Zum einen für den Erwartungswert μ durch

$$\hat{\mu} = g(Y_1, \ldots, Y_n) = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{3.9}$$

und zum anderen für die Varianz σ^2 mit

$$s^2 = g(Y_1, \ldots, Y_n) = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \hat{\mu})^2. \tag{3.10}$$

Optimalerweise weichen die geschätzten Werte nicht weit vom wahren Wert ab und die Abweichungen ergeben sich im Mittel zu Null. Diese Eigenschaft nennt sich *Erwartungstreue.*

Definition 3.8 (Erwartungstreue)
Eine Schätzstatistik $T = g(Y_1, \ldots, Y_n)$ heißt erwartungstreu für θ, wenn gilt

$$E_\theta(T) = \theta.$$

Neben der Erwartungstreue kann auch durch den *Bias* gezeigt werden, ob ein Schätzer verzerrt ist, also sich die Abweichungen im Mittel nicht zu Null ergeben.

Definition 3.9 (Bias)
Der Bias einer Schätzstatistik $T = g(Y_1, \ldots, Y_n)$ ist bestimmt durch

$$Bias_\theta(T) = E_\theta(T) - \theta.$$

Für die beiden Schätzstatistiken aus den Gleichungen (3.9) und (3.10) ließe sich beispielsweise zeigen, dass diese erwartungstreu und unverzerrt sind. Häufiger als der Bias wird zur Beurteilung der Schätzung die *mittlere quadratische Abweichung* (engl.: *mean squared error*) (MSE) betrachtet.

Definition 3.10 (Mittlere quadratische Abweichung (MSE))
Die erwartete mittlere quadratische Abweichung berechnet sich durch

$$\begin{aligned} MSE &= E\left((T-\theta)^2\right) \\ &= Var(T) + Bias(T)^2. \end{aligned}$$

Die Schwierigkeit besteht nun darin, einen Schätzer zu finden, der einen möglichst kleinen MSE aufweist. Dieser verringert sich im Falle eines erwartungstreuen Schätzers mit

$$\begin{aligned} \mathrm{MSE} &= \mathrm{Var}(T) + \underbrace{\mathrm{Bias}(T)^2}_{0} \\ &= \mathrm{Var}(T) \end{aligned} \tag{3.11}$$

zur Varianz. Methoden zur Konstruktion solcher Schätzer sind unter anderem das *Maximum-Likelihood-Prinzip* und die *Kleinste-Quadrate-Schätzung.*

Definition 3.11 (Maximum-Likelihood-Prinzip)
Das Maximum Likelihood Prinzip besteht darin, dass zu Realisationen $y_1, \ldots, y_n$ als Parameterschätzer der Parameter $\hat{\theta}$ gewählt wird, für den die Likelihood maximal ist, d.h.

$$L(\hat{\theta}) = \max_{\theta} L(\theta)$$

mit

$$L(\theta) = f(y_1, \ldots, y_n|\theta) = f(y_1|\theta) \cdots f(y_n|\theta).$$

Zur Maximierung der Likelihood wird der Einfachheit halber normalerweise die *Log-Likelihood* $\ln L(\theta)$ mit

$$\ln L(\theta) = \sum_{i=1}^{n} \ln f(y_i|\theta) \tag{3.12}$$

betrachtet. Den geschätzten Wert $\hat{\theta}$ erhält man dann durch Ableiten und Nullsetzen der Log-Likelihood. Analog können mit der Methode der kleinsten Quadrate Parameterschätzer konstruiert werden. Dabei wird versucht mithilfe einer Funktion den Zusammenhang zwischen zwei Variablen Y und X zu modellieren. Für ein einfaches lineares Regressionsmodell wird angenommen, dass Y durch eine Gerade

$$Y = \alpha + \beta X + \epsilon \tag{3.13}$$

beschrieben werden kann. Zugelassen wird bei dieser Modellierung ein Fehlerterm ϵ. Bestimmen lassen sich die Parameter α und β mithilfe der Kleinsten-Quadrate-Schätzung folgendermaßen.

Definition 3.12 (Kleinste-Quadrate-Schätzung)
Die Kleinste-Quadrate-Schätzung minimiert die Funktion

$$Q(\alpha, \beta) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - (\alpha + \beta))^2$$

bezüglich α und β. $Q(\alpha, \beta)$ beschreibt die quadrierten Differenzen zwischen den beobachteten und prognostizierten Werten y_i und $\hat{y}_i$. Die Kleinste-Quadrate-Schätzer für α und β ergeben sich dann zu

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$
$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

und der Fehlerterm ϵ berechnet sich für einzelne Realisationen mit

$$\hat{\epsilon}_i = y_i - \hat{y}_i = y_i - \hat{\alpha} + \hat{\beta} x_i.$$

3.3 Stochastische Prozesse

Geostatistische Anwendungen werden in der Regel durch *stochastische Prozesse* beschrieben. Zugrunde liegt nach Wackernagel (1995) die Annahme, dass der beobachtete Wert an einer Lokation $y(s_i)$ in dem Untersuchungsgebiet $D \subseteq \mathbb{R}^d, d \in \mathbb{N}$ als Realisation einer Zufallsvariablen $Y(s_i)$ für die Lokation s_i zustande kommt. Dabei existiert für jede mögliche Lokation $s, s \in D$ eine eigene Zufallsvariable. Die unendliche Menge all dieser Zufallsvariablen ergibt dann den stochastischen Prozess $Y(s)$, welcher demnach stetig auf dem Gebiet D definiert ist. Bezeichnet wird er auch als *Zufallsprozess* oder *Zufallsfunktion* (Webster und Oliver; 2007).
Üblicherweise besitzt ein stochastischer Prozess nach Navratil (2006) eine Abhängigkeitsstruktur nach dem Parameter s, welche sich hier als räumliche Abhängigkeitsstruktur ergibt, da mit s die Lage im Gebiet D angegeben ist. Beschrieben werden kann diese Abhängigkeitsstruktur im Allgemeinen durch die Kovarianz und speziell für geostatistische Anwendungen durch das *Variogramm* (siehe Kapitel 5) (Navratil; 2006). Problematisch ist nun nach Webster und Oliver (2007) bei dem Beschreiben der Abhängigkeitsstruktur, dass die Erwartungswerte der einzelnen Zufallsvariablen $Y(s_i), i = 1, \ldots, n$ nicht bekannt sind, da in praktischen Anwendungen nur jeweils eine Realisation pro Lokation s_i erfasst wird, sodass keine Erwartungswerte berechnet werden können. Diese sind aber, wenn man die Definition 3.5 für die Kovarianz betrachtet, notwendig, um die Abhängigkeitsstruktur zu beschreiben.
Um dieses Problem zu umgehen, wird für den stochastischen Prozess $Y(s)$ in praktischen Anwendungen die Annahme der *Stationarität* getroffen, die nach Webster und Oliver (2007) besagt, dass die Verteilung des stochastischen Prozesses $Y(s)$ Eigenschaften besitzt, die sich für alle Lokationen s nicht unterscheiden. Unterteilen lässt sich die Stationarität in *streng stationär*, *Stationarität zweiter Ordnung* und *intrinsisch stationär*, welche sich nach Wackernagel (1995) folgendermaßen ergeben.

Definition 3.13 (Strenge Stationarität)
Ein stochastischer Prozess $Y(s)$ wird als streng stationär bezeichnet, wenn seine Verteilung verschiebungsinvariant ist.

Definition 3.14 (Stationarität zweiter Ordnung)
Stationarität zweiter Ordnung bedeutet für einen stochastischen Prozess, dass der Erwartungswert und die Kovarianz verschiebungsinvariant sind. Konkret ist für einen Abstand $h = s - s'$

(i) der Erwartungswert $E(Y(s)) = \mu$ konstant $\forall s \in D$, sodass

$$E(Y(s)) = E(Y(s+h)),$$

(ii) und die Kovarianzfunktion $\forall s \in D$ nur von h abhängig, sodass

$$c(h) = Cov(Y(s), Y(s+h))$$

$$\overset{(3.5)}{=} E(Y(s)Y(s+h)) - \mu^2.$$

Definition 3.15 (Intrinsische Stationarität)
Ein stochastischer Prozess $Y(s)$ heißt intrinsisch stationär, wenn für das Inkrement $Y(s) - Y(s+h)$ gilt, dass

(i) der Erwartungswert $\mu(h)$ verschiebungsinvariant ist $\forall s \in D$ mit

$$E(Y(s) - Y(s+h)) = \mu(h) = 0$$

(ii) und die Varianz endlich ist und nur noch von dem Abstand h abhängt $\forall s \in D$ mit

$$Var(Y(s) - Y(s+h)) = E((Y(s) - Y(s+h))^2) < \infty.$$

Beachtet werden sollte bei der intrinsischen Stationarität nach Wackernagel (1995), dass die Existenz der ersten beiden Momente des stochastischen Prozesses $Y(s)$ bei Annahme dieser noch nicht gegeben ist.

Bei Annahme der Stationarität zweiter Ordnung ist die Kovarianzfunktion, wie bereits erwähnt, nur noch von dem Abstand h abhängig. Nach Wackernagel (1995) ergibt sich außerdem, dass sie begrenzt und kleiner als die Varianz ist, da

$$|c(h)| \leq c(0) \overset{(3.7)}{=} \text{Var}(Y(s)). \tag{3.14}$$

Für die Korrelationsfunktion eines stochastischen Prozesses mit Stationarität zweiter Ordnung folgt dann dementsprechend, dass diese nur von dem Abstand h abhängig ist, sodass

$$\rho(h) = \frac{c(h)}{c(0)}. \tag{3.15}$$

Neben der Stationarität wird häufig *Isotropie* vorausgesetzt. Nach Schmid und Feilke (2012) besagt diese, dass die Kovarianz zwischen zwei Lokationen nur noch auf der Distanz zwischen diesen und nicht mehr auf der räumlichen Position oder Richtung beruht.

Definition 3.16 (Isotropie und Anisotropie)
Ein stationärer stochastischer Prozess $Y = \{Y(s), s \in D \subseteq \mathbb{R}^d\}$ heißt isotropisch, wenn $c(h) = c(\|h\|)$, wobei $\|\cdot\|$ den Euklidischen Abstand bezeichnet. Ist dies nicht der Fall heißt er anisotrop.

Die Kovarianz eines stationären isotropischen stochastischen Prozesses würde damit das gleiche Ergebnis für unterschiedliche Punktpaare, die aber den selben Abstand haben, liefern.

4 Explorative Datenanalyse

Nach Fahrmeir et al. (2011) dient die explorative Datenanalyse dazu, sich einen Überblick über Strukturen und Besonderheiten in den Daten zu verschaffen, also Aussagen zur Verteilung der Daten zu treffen oder Ausreißer zu erkennen. Speziell bei räumlichen Daten ist zusätzlich die Betrachtung räumlicher Strukturen relevant. So wird nach Navratil (2006) nach Daten gesucht, die nicht zu den räumlichen Nachbarn passen. Für die explorative Datenanalyse existieren unterschiedliche Verfahren je nachdem, ob die räumliche Struktur beachtet wird oder nicht. Verfahren ohne Beachtung der räumlichen Struktur sind nach Fahrmeir et al. (2011) beispielsweise Boxplots und Schätzungen von Dichtekurven. Nach Navratil (2006) gehören auch Stamm-Blatt-Diagramme und Histogramme dazu. Berücksichtigt werden kann die räumliche Struktur nach Bivand et al. (2013) schließlich am besten mithilfe von Karten, da diese die Koordinaten beinhalten können oder auf ihnen basieren. Die unterschiedlichen Werte einer interessierenden Variablen können dann durch Farben oder Symbolgrößen dargestellt werden (Bivand et al.; 2013).

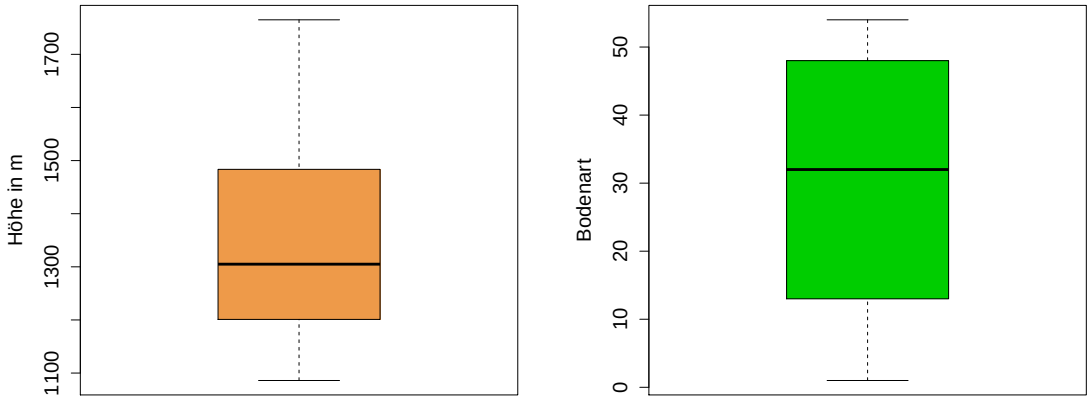
Für die folgenden geostatistischen Analysen, sowohl die explorative Datenanalyse als auch die in Kapitel 5 und 6 vorgestellten geostatistischen Methoden, Variogramme und Kriging, wird aufgrund der großen Anzahl an Beobachtungen jeweils eine Stichprobe von 500 aus den verwendeten Datensätzen `elevation.dem` und `soils` aus der *Spearfish Sample Database* gezogen. Dazu werden die Datensätze in die Klasse `data.frame` umgewandelt und nach der Stichprobenziehung durch Zuweisung der Koordinaten zurück in ein `Spatial`-Objekt transformiert, wobei sich ein `SpatialPointsDataFrame` ergibt, da nur noch einzelne Beobachtungen enthalten sind.

```
# Stichprobenziehung für die Höhendaten elevation.dem
> elev2 <- as.data.frame(elev)
> set.seed(123)
> elevsample=elev2[sample(1:nrow(elev2),500,replace=FALSE),]
> coordinates(elevsample) <- c("s1","s2")

# Stichprobenziehung für die Bodendaten soils
> soils2 <- as.data.frame(soils)
> set.seed(123)
> soilssample <- soils2[sample(1:nrow(soils2),500,replace=FALSE),]
> coordinates(soilssample) <- c("s1","s2")
```

Bevor nun die räumliche Struktur mit in die Analyse einbezogen wird, sind die genannten Möglichkeiten der explorativen Datenanalyse ohne Beachtung der räumlichen Struktur für die Höhen- und Bodendaten dargestellt. Die R-Funktionen zur Visualisierung stammen aus dem R-Paket `graphics`, welches standardmäßig in R enthalten ist.

Abb. 4.1 zeigt Boxplots der Daten. Die Abbildung der Höhendaten weist dabei auf eine linkssteile Verteilung mit einem Median bei etwa 1300 hin, während sich die Verteilung der Bodendaten dem Boxplot nach relativ symmetrisch mit einem Median bei 32 ergibt.

```
# Boxplot für die Höhendaten
> boxplot(elevsample$elevation.dem, col="tan2", ylab="Höhe in m")

# Boxplot für die Bodendaten
> boxplot(soilssample$soils, col="green3", ylab="Bodenart")
```



(a) Höhendaten (b) Bodendaten

Abb. 4.1: Einfache Boxplots

Variablen können auch in Abhängigkeit voneinander als Boxplot dargestellt werden.

```
# Boxplots der Höhendaten gruppiert nach Bodenarten
> boxplot(elevsample$elevation.dem ~ soilssample$soils,
+         col=terrain.colors(60), ylab="Höhe in m", xlab="Bodenart")
```

Dies zeigt Abb. 4.2 exemplarisch für die Höhendaten gruppiert nach Bodenarten. Diese Darstellung ist hier allerdings nur möglich, da die Bodendaten eigentlich kategorial sind. Für die restlichen Auswertungen werden sie wieder als quasistetig angenommen. Zu sehen ist, dass sich die Verteilung der Höhendaten für die vielen Bodenarten sehr unterschiedlich ergibt. Für einige Bodenkategorien muss dabei beachtet werden, dass teilweise weniger als fünf Beobachtungen in diesen vorliegen, weshalb eine Darstellung als Boxplot nicht sinnvoll ist, da dieser auf genau fünf Punkten (Median, 25-Prozent und 75-Prozent-Quantil, obere und untere Zaungrenze) basiert (Fahrmeir et al.; 2011).
Die Histogramme und die Schätzungen der Dichtekurven sind für die Höhen- und Bodendaten jeweils gemeinsam in Abb. 4.3 abgebildet. Bei Betrachtung dieser bestätigt

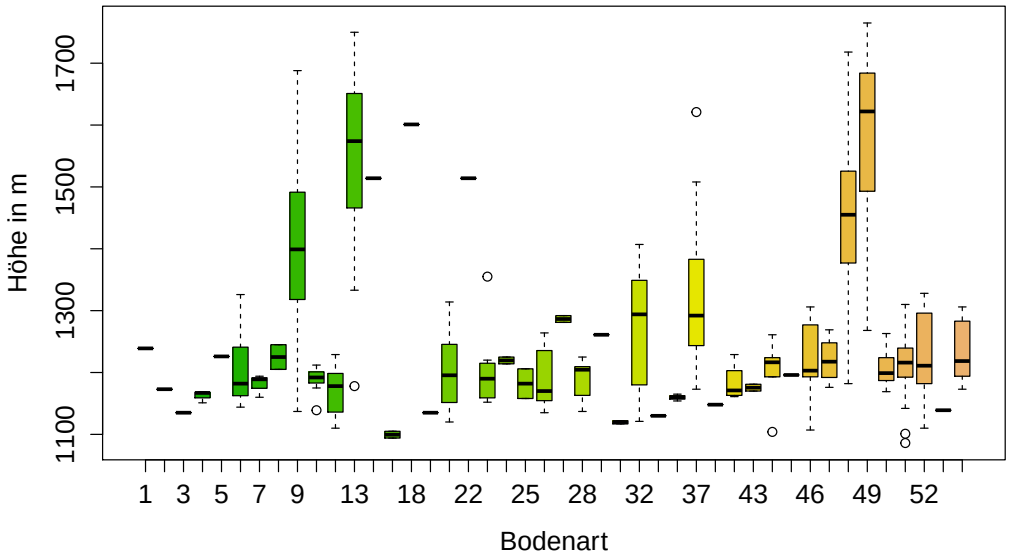
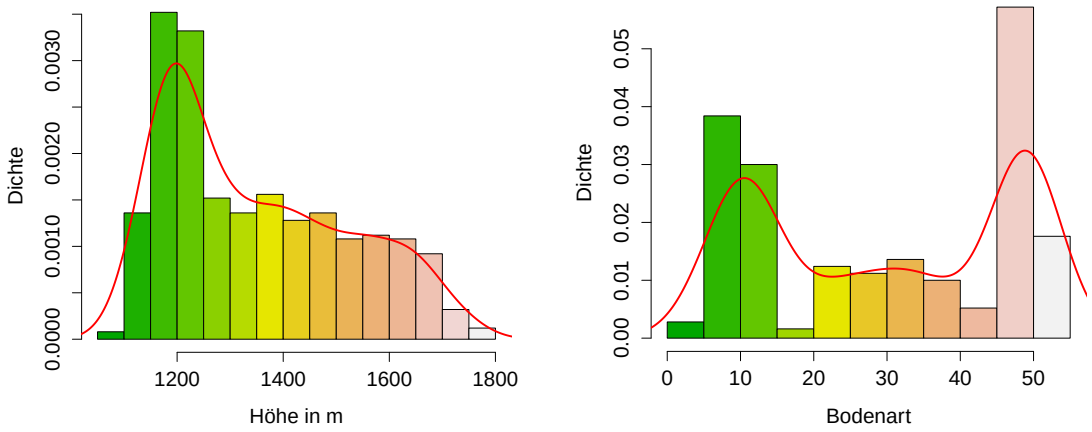


Abb. 4.2: Boxplots der Höhendaten gruppiert nach Bodenarten



(a) Höhendaten (b) Bodendaten

Abb. 4.3: Histogramme und Schätzungen der Dichtekurven

sich die Vermutung der linkssteilen Verteilung für die Höhendaten, die bereits in dem Boxplot in Abb. 4.1(a) zu erkennen war. Die Verteilung der Bodendaten erweist sich dagegen als bimodal, was in dem Boxplot in Abb. 4.1(b) nicht zu sehen war.

```
# Histogramm und Schätzung der Dichtekurve für die Höhendaten
> hist(elevsample$elevation.dem, col=terrain.colors(15), prob=TRUE,
+      xlab="Höhe in m", ylab="Dichte", main=NULL)
> lines(density(elevsample$elevation.dem), col="red")

# Histogramm und Schätzung der Dichtekurve für die Bodendaten
> hist(soilssample$soils, col=terrain.colors(11), prob=TRUE,
+      xlab="Bodenart", ylab="Dichte", main=NULL)
> lines(density(soilssample$soils), col="red")
```

Die gleiche Struktur der Verteilungen, die in der Darstellung der Histogramme mit den geschätzten Dichtekurven deutlich wird, zeigen auch die folgenden Stamm-Blatt-Diagramme. Ausreißer in den Daten sind weder in diesen noch in den Histogrammen oder Boxplots zu sehen.

```
# Stamm-Blatt-Diagramm für die Höhendaten
> stem(elevsample$elevation.dem)

The decimal point is 2 digit(s) to the right of the |

10 | 99
11 | 00111122222333344444444444444
11 | 55555555556666666666666666777777777777777777888888888888888888889+6
12 | 00000000000000000001111111111111111111111111222222222222222223333+11
12 | 55555555556666666777777888888999999999
13 | 00000000001111111111222222233344444444
13 | 55555566666667777777778888888888889999
14 | 00000000111111122222222233333344444
14 | 55555566666667777777778888889999999
15 | 01111112222223333444444444444
15 | 55566666667777778888888899999
16 | 01112222222233334444444
16 | 55555666666666888888999
17 | 0001222333
17 | 5667
```

```
# Stamm-Blatt-Diagramm für die Bodendaten
> stem(soilssample$soils)

The decimal point is 1 digit(s) to the right of the |

0 | 123444
0 | 566666666666666677788999999999999999999999999999999999999999999+11
1 | 000000000111111113333333333333333333333333333333333333333333333+6
1 | 778
2 | 011111111111111112333333333344
2 | 556666666666778888888888889
3 | 002222222222222222222222222224
3 | 555777777777777777777777778
4 | 111133444444
4 | 566666666777788888888888888888888888888888888888888888888888888+46
5 | 000000000000000000001111111111111111111111111112222222223444444
```

Die räumliche Struktur kann nun wie bereits erwähnt mithilfe von Karten und unterschiedlichen Farben oder Symbolgrößen berücksichtigt werden. Grafisch resultieren die beiden letztgenannten Möglichkeiten als Streudiagramme. Für die Darstellung durch unterschiedliche Farben existiert für räumliche Objekte beispielsweise speziell die Funktion `spplot()` aus dem R-Paket `sp` (Bivand et al.; 2013). Die Funktion vergibt verschiedene Farben proportional zu den Werten der dargestellten Variable. Dargestellt sind die Grafiken für die Höhen- und Bodendaten in Abb. 4.4 und Abb. 4.5.

```
# Unterschiedliche Farben für die Höhendaten
> spplot(obj=elevsample, zcol="elevation.dem",
+        col.regions=terrain.colors(100), cex=1.25,
+        colorkey=TRUE, par.settings=list(fontsize=list(text=15)),
+        scales=list(draw=TRUE))

# Unterschiedliche Farben für die Bodendaten
> spplot(obj=soilssample, zcol="soils",
+        col.regions=terrain.colors(100), cex=1.25,
+        colorkey=TRUE, par.settings=list(fontsize=list(text=15)),
+        scales=list(draw=TRUE))
```

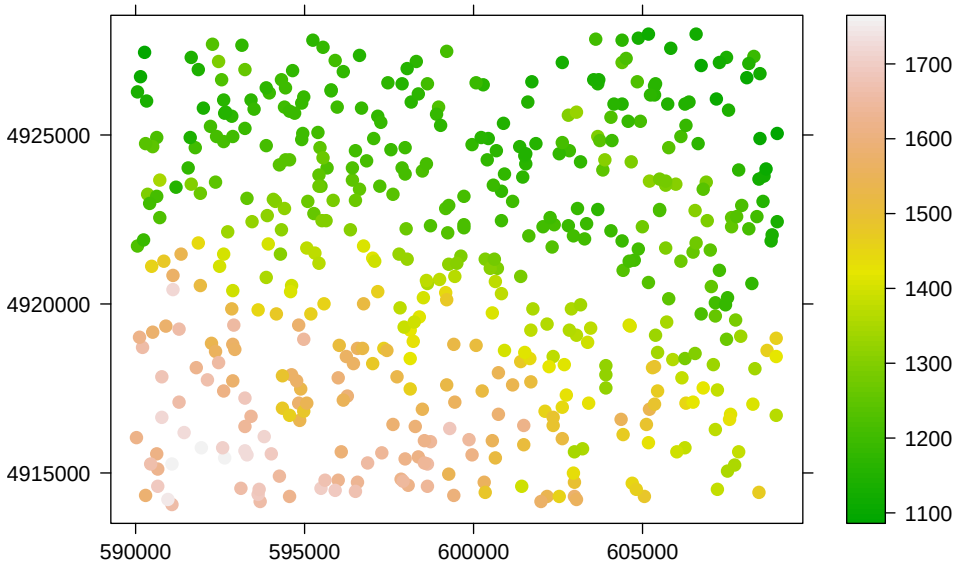


Abb. 4.4: Darstellung der räumlichen Struktur der Höhendaten durch unterschiedliche Farben

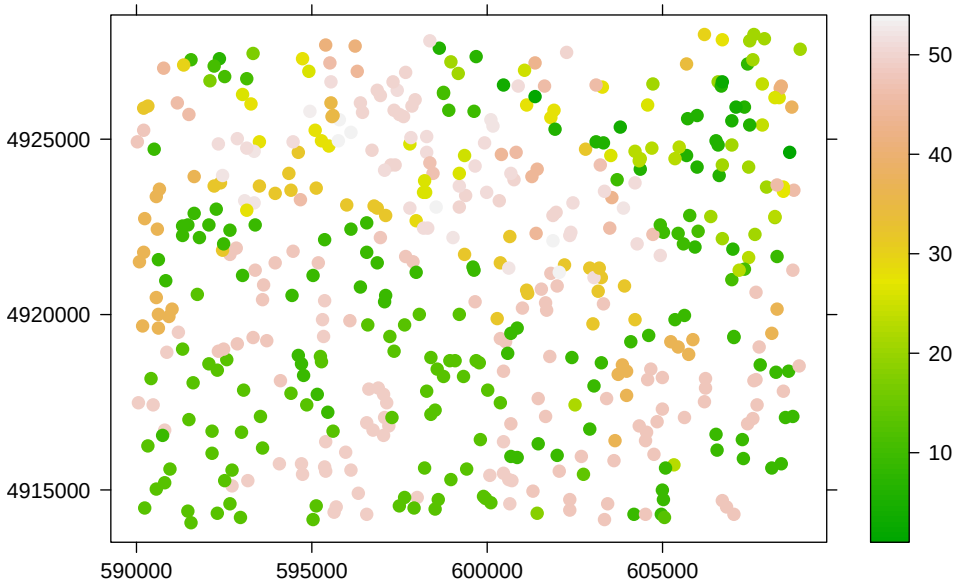


Abb. 4.5: Darstellung der räumlichen Struktur der Bodendaten durch unterschiedliche Farben

Eine Möglichkeit der Visualisierung mithilfe unterschiedlicher Symbolgrößen ist mit der Funktion `bubble()`, die ebenfalls aus dem `sp`-Paket stammt, gegeben. Dabei werden die Lokationen der Stichproben dargestellt und die Symbolgröße entspricht analog zu den Farben proportional den Werten für die Höhen- und Bodendaten. Die maximale Symbolgröße kann mit dem Argument `maxsize` beschränkt werden und die Einträge in der Legende mit `key.entries` angegeben werden. Abb. 4.6 und Abb. 4.7 zeigen beispielhafte Darstellungen für die Höhen- und Bodendaten.

```
# Unterschiedliche Symbolgrößen für die Höhendaten
> bubble(elevsample, zcol="elevation.dem", key.entries=1000+200*(0:4),
+        maxsize=2, fill=FALSE, col="black", do.sqrt=FALSE, main="",
+        par.settings=list(fontsize=list(text=15)),
+        scales=list(draw=TRUE))

# Unterschiedliche Symbolgrößen für die Bodendaten
> bubble(soilssample, zcol="soils", key.entries=6*(1:9),
+        maxsize=2, fill=FALSE, col="black", do.sqrt=FALSE, main="",
+        par.settings=list(fontsize=list(text=15)),
+        scales=list(draw=TRUE))
```

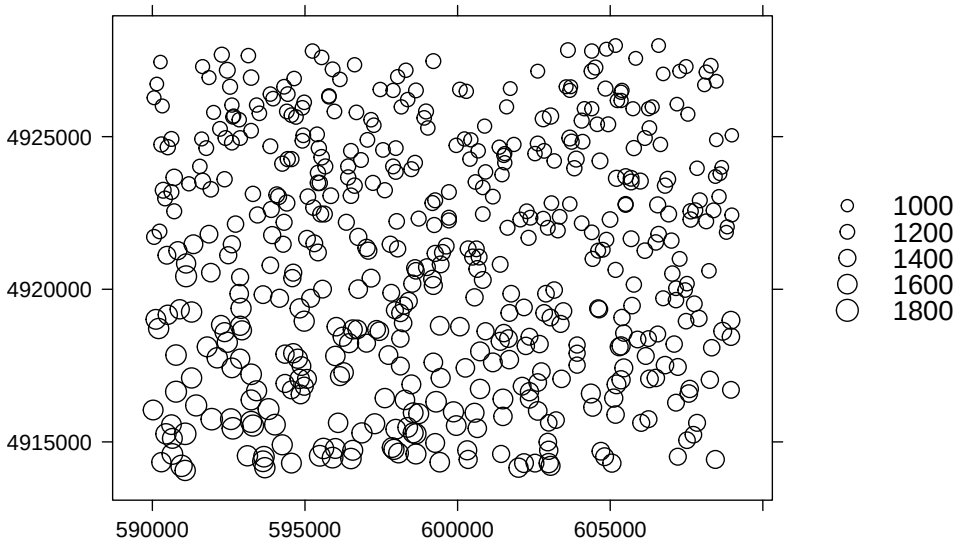


Abb. 4.6: Darstellung der räumlichen Struktur der Höhendaten durch unterschiedliche Symbolgrößen

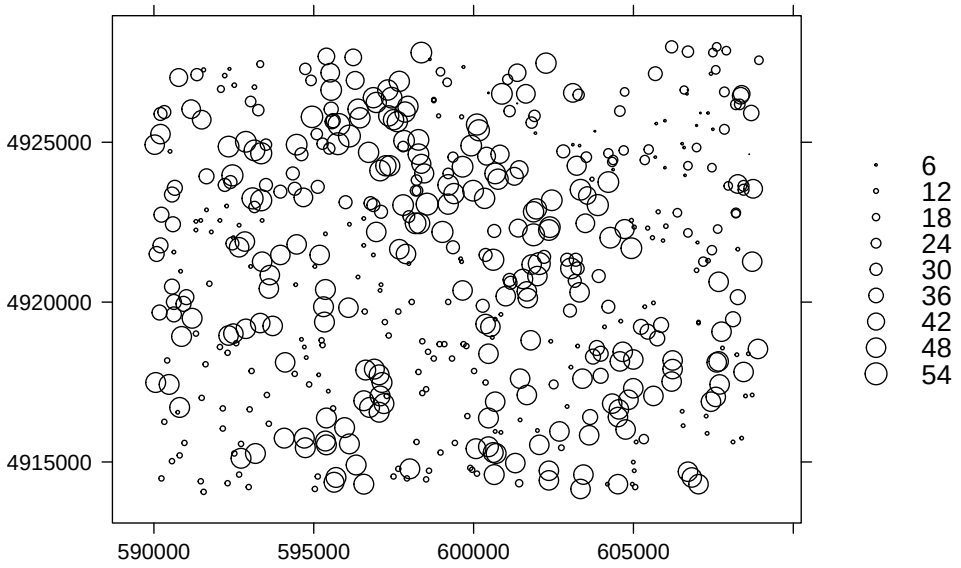


Abb. 4.7: Darstellung der räumlichen Struktur der Bodendaten durch unterschiedliche Symbolgrößen

In beide Varianten, unterschiedlichen Farben und unterschiedlichen Symbolgrößen, zeigen sich für die Höhen- und Bodendaten räumliche Strukturen. Speziell bei den Höhendaten lässt sich erkennen, dass die Werte für kleinere Hoch- und Rechtswerte größer sind und für größere Hoch- und Rechtswerte kleiner.
Liegen sehr viele Daten vor, existiert neben diesen Darstellungen außerdem die Möglichkeit der Visualisierung mithilfe der Funktion `image()`, die in dem `graphics`-Paket enthalten ist. Diese bildet alle Werte, auch nach Farben getrennt, ab. Grafisch ergibt sich dann eine Darstellung analog zu Abb. 2.2 in Abschnitt 2.1.

Eine weitere Möglichkeit, räumliche Daten explorativ darzustellen, ist nach Diggle und Ribeiro (2007) das Variogramm. In dieser Arbeit, wie auch häufig in der Literatur, wird es jedoch nicht direkt zur explorativen Analyse, sondern zur Modellierung des räumlichen Zusammenhangs zwischen den Lokationen und als Basis für die geostatistische Interpolationsmethode Kriging verwendet, weshalb Variogramme gesondert im folgenden Kapitel 5 vorgestellt werden.

5 Variogramme

Grundlage für die geostatistischen Auswertungen ist die Annahme, dass die betrachteten Beobachtungen eines räumlichen stochastischen Prozesses voneinander nicht unabhängig sind Wackernagel (1995). Hintergrund ist, dass gemessene Werte an benachbarten Lokationen ähnliche Ergebnisse liefern, im Gegensatz dazu, sich gemessene Werte von weit entfernten Punkten unterscheiden. Beispielhaft ist die Differenz der Niederschlagsmengen an zwei Punkten, die eine Distanz von 1 km aufweisen, wahrscheinlich geringer, als die zweier Punkte, die eine Distanz von 10 km aufweisen.
Beschreiben lässt sich diese räumliche Variation nach Webster und Oliver (2007) mithilfe von *Variogrammen*. Ziel von Variogrammen ist es, eine Funktion an die Daten anzupassen, die die räumliche Abhängigkeitsstruktur des zugrunde liegenden stochastischen Prozesses $Y(s)$ so gut wie möglich beschreibt.

Dieses Kapitel geht nun auf die Vorgehensweise ein, wie solch eine Funktion an die vorhandenen Daten angepasst werden kann. Zuerst wird mithilfe einer *Variogrammwolke* in Abschnitt 5.1 ein Überblick über den Zusammenhang zwischen allen Lokationen aus den Daten bezogen auf ihren Abstand h gegeben. Anschließend wird in Abschnitt 5.2 gezeigt, wie sich aus den Daten Schätzer für den Zusammenhang zwischen Lokationen mit Abstand h in Form eines *empirischen Variogramms* berechnen lassen. In dem folgenden Abschnitt 5.3 wird dann auf Eigenschaften der Variogrammfunktion eingegangen und in Abschnitt 5.4 das Kovariogramm und Korrelogramm vorgestellt. Abschnitt 5.5 beinhaltet schließlich die vollständige Schätzung der räumlichen Abhängigkeitsstruktur des zugrunde liegenden stochastischen Prozesses $Y(s)$ in Form eines theoretischen Variogramms, für das verschiedene Modelle zur Schätzung vorgestellt werden. Anschließend gehen die Abschnitte 5.6 und 5.7 auf Anwendungen ein, in denen spezielle Daten vorliegen. Im ersten Abschnitt werden anisotrope Variogramme vorgestellt, die im Falle von nicht isotropischen Daten berechnet werden und im zweiten Abschnitt werden Drift, Hole-Effekt und unbeschränkte Variogramme eingeführt.

Die allgemeine Variogrammfunktion ist indirekt bereits in Definition 3.15 als Varianz des Inkrements gegeben. Mit der zusätzlichen Annahme der Isotropie ist die Funktion für das Variogramm nach Webster und Oliver (2007) noch einmal aufgeführt.

Definition 5.1
Das Variogramm eines intrinsisch stationären isotropischen stochastischen Prozesses ergibt sich durch

$$\begin{aligned} 2\gamma(h) &= Var(Y(s) - Y(s+h)) \\ &= E\Big((Y(s) - Y(s+h))^2\Big) \end{aligned} \tag{5.1}$$

mit $h = \|s - s'\|$ als Abstand zwischen zwei Lokationen.

Aufgrund der Annahme der Isotropie ergibt sich das Variogramm als unabhängig von den Positionen zweier Lokationen und der Richtung, wie sie zueinander stehen. Große Werte für $2\gamma(h)$ bedeuten, dass die Korrelation zwischen den beiden Realisationen gering ist. Analog ist die räumliche Abhängigkeit zwischen zwei Realisationen hoch, wenn sich kleine Werte für $2\gamma(h)$ ergeben. In der Literatur wird das Variogramm häufig auch mit

$$\gamma(h) = \frac{1}{2}\mathrm{Var}(Y(s) - Y(s+h)) \tag{5.2}$$

angegeben, was jedoch genau genommen das *Semivariogramm*, nach Webster und Oliver (2007) auch bezeichnet als *Semivarianz* für die Distanz h, darstellt. In dieser Arbeit wird daher für das Variogramm die Funktion 2γ verwendet und das Semivariogramm ergibt sich mit γ.

Variogramm in R

In R lassen sich Variogramme nach Bivand et al. (2013) mithilfe der Funktion `variogram()` mit

```
> variogram(object, locations=coordinates(data), data, ...)
```

berechnen. Dabei werden die Semivariogrammwerte bezogen auf die Distanz zwischen den Lokationen dargestellt. Je nach Anwendung können der Funktion verschiedene weitere Argumente übergeben werden. Auf einige, in dieser Arbeit verwendeten Argumente, wird im Folgenden genauer eingegangen. Zur leichteren Handhabung bei der Berechnung der Variogramme und später dem Kriging in Kapitel 6 werden die Stichproben der Höhendaten `elevsample` und der Bodendaten `soilssample` in `gstat`-Objekte umgewandelt.

```
# gstat-Objekt für die Höhendaten
> g.elev <- gstat(id="elev", formula=elevation.dem~1,
+                 data=elevsample)
> g.elev
data:
elev : formula = elevation.dem‘~‘1 ; data dim = 500 x 1

# gstat-Objekt für die Bodendaten
> g.soils <- gstat(id="soils", formula=soils~1,
                   data=soilssample)
> g.soils
data:
soils : formula = soils‘~‘1 ; data dim = 500 x 1
```

Dies hat, da in der Arbeit für die geostatistischen Verfahren ausschließlich das R-Paket `gstat` verwendet wird, den Vorteil, dass bei der Berechnung die Angabe von einigen Argumenten entfällt. So muss für die Berechnung von Variogrammen keine Information mehr bezüglich `locations` und `data` übergeben werden, da diese bereits im `gstat`-Objekt gespeichert sind. Die Angabe der Konstante `~1` resultiert bei der Umwandlung in `gstat`-Objekte nach Bivand et al. (2013) aus der Annahme der intrinsischen Stationarität (siehe Definition 3.15), dass ein konstanter Erwartungswert zugrunde liegt.

5.1 Variogrammwolke

Die Variogrammfunktion ist in Anwendungen üblicherweise nicht bekannt, weshalb sie geschätzt werden muss. Dazu kann mithilfe einer *Variogrammwolke* ein erster Überblick über die räumliche Struktur der vorliegenden empirischen Daten gegeben werden. Außerdem lassen sich durch die Variogrammwolke nach Bivand et al. (2013) leicht Ausreißer identifizieren. Sie besteht nach Diggle und Ribeiro (2007) aus den empirischen Datenpunkten

$$(h_{ij}, \gamma^*_{ij}) : j > i, \tag{5.3}$$

wobei

$$2\gamma^*_{ij} = (y(s_i) - y(s_j))^2 \tag{5.4}$$

$$h_{ij} = \|s_i - s_j\|. \tag{5.5}$$

Dargestellt wird also die Varianz jedes Lokationspaares bezogen auf ihren euklidischen Abstand. Grafisch ergibt sich die Variogrammwolke dann als Scatterplot.

Die Punkte der Variogrammwolke sind nach Diggle und Ribeiro (2007) untereinander korreliert, da sich aus den nur n Realisationen der räumlichen Zufallsvariable mithilfe der *Gaußschen Summenformel*

$$\frac{n(n-1)}{2} \tag{5.6}$$

Punkte für die Variogrammwolke ergeben. Allein für die Stichproben der Höhen- und Bodendaten mit 500 Beobachtungen resultieren noch 124.750 Punkte. Wären die Stichproben (siehe Kapitel 4) nicht gezogen worden, läge die Zahl jeweils bei über 43 Milliarden Punkten.

Variogrammwolke in R

In R lässt sich eine Variogrammwolke durch Angabe von `cloud=TRUE` mit

```
> variogram(object, cloud=TRUE ...)
```

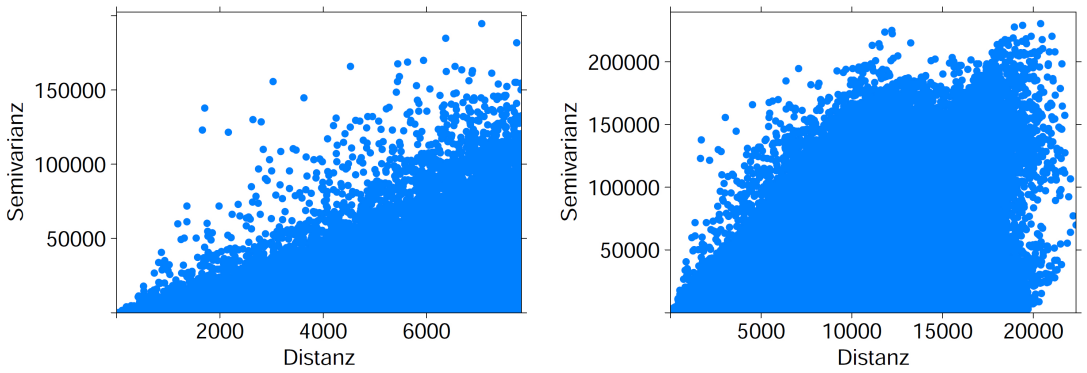
berechnen. Für die Höhen- und Bodendaten ist sie damit durch

```
# Variogrammwolke für die Höhendaten
> vcloud.elev <- variogram(object=g.elev, cloud=TRUE)

# Variogrammwolke für die Bodendaten
> vcloud.soils <- variogram(object=g.soils, cloud=TRUE)
```

bestimmt. Mit dem Aufruf von `vcloud.elev` oder `vcloud.soils` ließen sich nun alle resultierenden Werte der Variogrammwolken anzeigen. Aufgrund der vielen Punkte ist dies nicht sinnvoll, weshalb ein exemplarischer Überblick mit der Funktion `head()` gegeben wird.

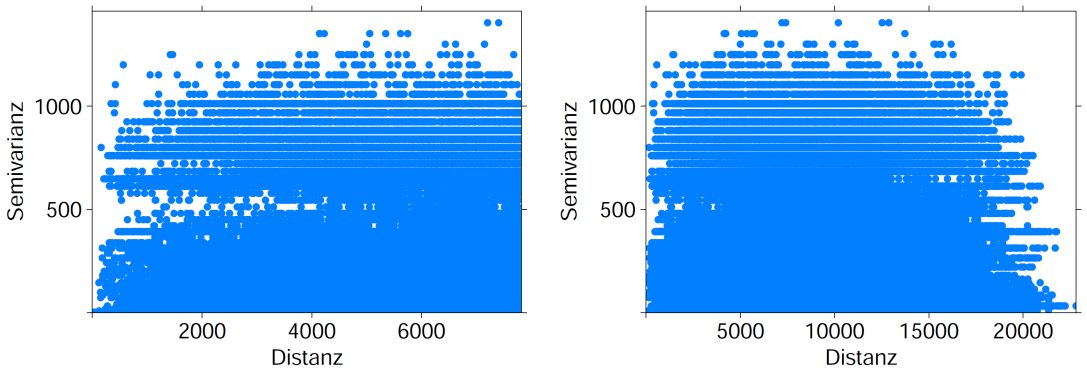
```
# Überblick über die Werte der Variogrammwolke für die Höhendaten
> head(vcloud.elev)
      dist  gamma dir.hor dir.ver   id left right
1 2047.266  364.5       0       0 elev    3     1
2 6185.095 9522.0       0       0 elev    3     2
3 5209.990 1568.0       0       0 elev    4     2
4 6804.249 3362.0       0       0 elev    4     3
5 5592.003 1624.5       0       0 elev    6     1
6 6024.251 3528.0       0       0 elev    6     3

# Überblick über die Werte der Variogrammwolke für die Bodendaten
> head(vcloud.soils)
      dist gamma dir.hor dir.ver    id left right
1 1965.401 112.5       0       0 soils    3     1
2 6964.266   2.0       0       0 soils    4     3
3 5266.773 648.0       0       0 soils    5     2
4 3491.619 882.0       0       0 soils    7     1
5 1650.273 364.5       0       0 soils    7     3
6 5448.119 312.5       0       0 soils    7     4
```

Von Interesse sind die ersten beiden Spalten. Die jeweils erste Spalte stellt die Distanz zwischen den Lokationen dar und die zweite Spalte den zugehörigen Variogrammwert. Grafisch sind die beiden sich hier ergebenden Variogrammwolken in Abb. 5.1(a) und 5.2(a) dargestellt.
In der Variogrammwolke für die Höhendaten steigen die Variogrammwerte mit zunehmender Distanz. Die Variogrammwolke der Bodendaten beinhaltet dagegen sowohl für kleine, als auch für große Distanzen hohe Variogrammwerte. Außerdem ist hier erkennbar, dass es sich eigentlich um eine kategoriale Variable handelt.



(a) `default cutoff` (b) `cutoff=Inf`

Abb. 5.1: Variogrammwolken der Höhendaten



(a) `default cutoff` (b) `cutoff=Inf`

Abb. 5.2: Variogrammwolken der Bodendaten

In R ist nach Bivand et al. (2013) für die oben genannten Funktionen voreingestellt, dass nur Lokationen mit einer maximalen Distanz von einem Drittel der größten Differenz der Diagonalen der `bbox` in die Berechnung der Variogrammwolke einbezogen werden. Die maximale mit einbezogene Distanz liegt bei diesen Variogrammwolken etwa bei 7800. Mithilfe der Angabe des `cutoff`-Arguments lassen sich auch größere, oder allgemein individuelle, Grenzen für die maximal einbezogene Distanz setzen. Zum Vergleich ist der Wert für `cutoff` in den Grafiken 5.1(b) und 5.2(b) auf `Inf` gesetzt, sodass auch für den maximalen Abstand zwischen zwei Lokationen in der Stichprobe ein Variogrammwert berechnet wird.

```
# Variogrammwolke für die Höhendaten mit cutoff=Inf
> vcloud.elev2 <- variogram(object=g.elev, cutoff=Inf, cloud=TRUE)

# Variogrammwolke für die Bodendaten mit cutoff=Inf
> vcloud.soils2 <- variogram(object=g.soils, cutoff=Inf, cloud=TRUE)
```

In den zugehörigen Grafiken lässt sich erkennen, dass bis zu dreimal so große Distanzen einbezogen wurden, im Gegensatz zu den Grafiken für die Variogrammwolken, in denen sich der Wert standardgemäß ergeben hat. Nach Bivand et al. (2013) ist es allerdings selten sinnvoll, Variogrammwerte für große Distanzen zu betrachten, da aufgrund von weniger vorhandenen Beobachtungen im Bereich der großen Distanzen die Strukturen stark schwanken und spätere Schätzungen auf Basis dieser unsicherer werden. Daher wird für die folgenden Auswertungen der voreingestellte `cutoff` verwendet.

5.2 Empirisches Variogramm

Das eigentliche Ziel der Variogrammschätzung ist es, den Zusammenhang zwischen den Lokationen durch eine Funktion zu beschreiben. Um dieser nach dem Überblick über die Daten durch die Variogrammwolke ein Stück näher zu kommen werden für explizite Abstände h aus den empirischen Daten nach Navratil (2006) mit

$$2\hat{\gamma}(h) = \frac{1}{N(h)} \sum_{i=1}^{N(h)} (y(s_i) - y(s_i + h))^2, \tag{5.7}$$

Schätzer berechnet, die gemeinsam das *empirische Variogramm* darstellen. Mit $N(h)$ ist dabei die Anzahl der Punktpaare mit Abstand h angegeben. Berechnet wird die Summe der quadrierten Differenzen aller Messwerte mit dem Abstand h, welche durch die Anzahl dieser Lokationspaare geteilt wird. Damit ergibt sich optimalerweise für jeden möglichen Abstand h ein Variogrammwert. In der Praxis existiert für viele Abstände h allerdings gar kein Wert, da nicht jeder mögliche Abstand erfasst wurde. Außerdem sollten nach Hattermann (2014) mindestens 30 Punktpaare mit dem gleichen Abstand vorhanden sein, um das Variogramm zu berechnen, was ebenfalls in der Praxis problematisch ist, da die Lokationen nicht wie Gitterdaten angeordnet, sondern irregulär im Raum verteilt sind. Dadurch ergeben sich zwar einige Abstände in dem gleichen Wertebereich, jedoch nicht genau äquivalent. Aufgrund dieser Problematik werden die Abstände zur Berechnung des empirischen Variogramms nach Navratil (2006) häufig in äquidistante Klassen unterteilt. Die zugehörigen Werte $2\hat{\gamma}$ ergeben sich dann als Mittelwert pro Klasse für die Messwerte, deren Abstand h

$$(k-1)u < h < ku \tag{5.8}$$

mit $k \in \mathbb{N}$ erfüllt, wobei u die Breite der Klassen angibt (Diggle und Ribeiro; 2007). Je nach Wahl von u kann sich dadurch die geschätzte Variogrammfunktion sehr unterschiedlich ergeben. Schätzen lässt sie sich nach Navratil (2006) mit

$$2\hat{\gamma}(h_k) = \frac{1}{N(h)} \sum_{N(h)} (y(s_i) - y(s_j))^2, \tag{5.9}$$

wobei $N(h)$ die Anzahl der Messwerte pro Klasse, $y(s_i)$ und $y(s_j)$ die Messwerte, deren Entfernungen in der Klasse h liegen und h_k den Mittelwert der Abstände pro Klasse angibt. Dieser Mittelwert h_k lässt sich im zweidimensionalen Fall laut Navratil (2006) durch

$$h_k = \frac{1}{N(h)} \sum_{i,j=1}^{N(h)} \sqrt{(s_i - s_j)^2 + (y_i - y_j)^2}. \tag{5.10}$$

berechnen. Mithilfe der empirischen Variogrammfunktion kann dann nach Diggle und Ribeiro (2007) das theoretische Variogramm geschätzt werden (siehe Abschnitt 5.5). In der Literatur existieren für das empirische Variogramm auch die Bezeichnungen *experimentelles Variogramm* oder *Stichprobenvariogramm*.

Empirisches Variogramm in R

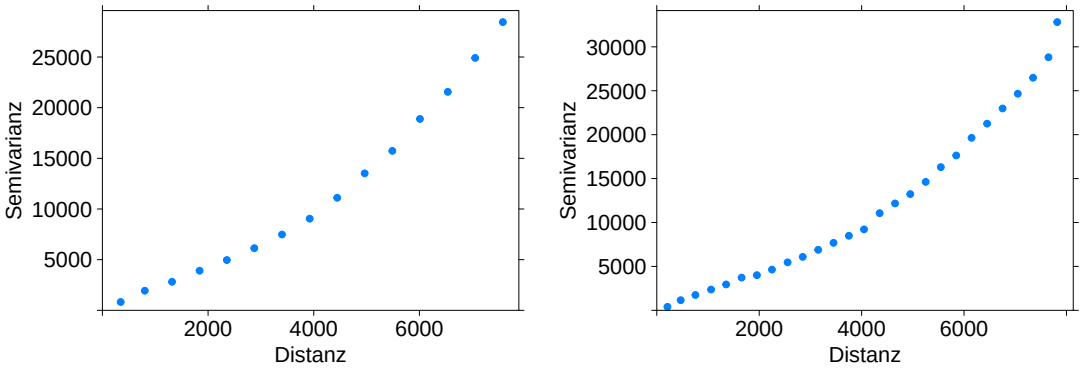
Die empirischen Variogramme können ohne zusätzliche Angaben berechnet werden.

```
# Empirisches Variogramm für die Höhendaten
> vemp.elev <- variogram(object=g.elev)

# Empirisches Variogramm für die Bodendaten
> vemp.soils <- variogram(object=g.soils)
```

Grafisch dargestellt sind sie in Abb. 5.3(a) und 5.4(a). R verwendet dabei als Klassenbreite standardgemäß ein Fünfzehntel des `cutoff`-Wertes, womit sich für die Höhen- und Bodendaten eine Breite von etwa $u = 520$ und entsprechend 15 Klassen für das empirische Variogramm ergeben.

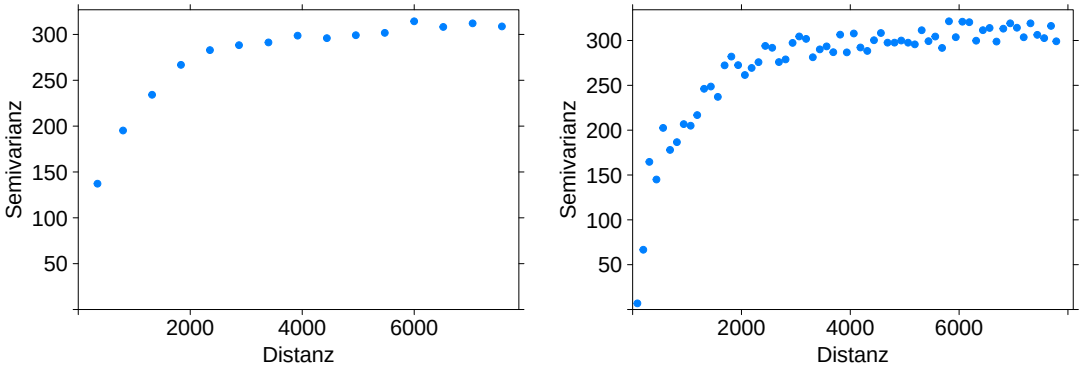
$$\begin{aligned} K_1 &= \{(1-1) \cdot 520 \leq |h| < 1 \cdot 520\} \\ &= \{0 \leq |h| < 520\} \\ &\vdots \\ K_{15} &= \{(15-1) \cdot 520 \leq |h| < 15 \cdot 520\} \\ &= \{7280 \leq |h| < 7800\}. \end{aligned}$$

Ein Punktpaar, dessen Abstand 5340 beträgt, wird demnach, gemeinsam mit allen Punktpaaren, deren Abstand zwischen 5200 und 5720 liegen, beispielsweise in die elfte Klasse mit einberechnet.



(a) `default width` (b) `width=300`

Abb. 5.3: Empirische Variogramme der Höhendaten



(a) `default width` (b) `width=125`

Abb. 5.4: Empirische Variogramme der Bodendaten

Bei Betrachtung der empirischen Variogramme lässt sich für beide Datensätze ein abnehmender Zusammenhang mit zunehmender Distanz erkennen. Zusätzlich zeigt sich bei den Bodendaten eine schnell einsetzende Sättigung, sodass ab einer bestimmten Distanz der Zusammenhang nicht weiter abnimmt. Sichtbar wird dieser Zusammenhang auch bei Betrachtung der resultierenden Variogrammwerte. Diese können vollständig gezeigt

werden, da nur noch 15 Werte, einer pro Klasse, vorhanden sind.

```
# Überblick über die empirischen Variogrammwerte für die Höhendaten
> vemp.elev
     np      dist      gamma dir.hor dir.ver        id
1   360  346.5181   821.9417       0       0 elevation
2  1115  802.4614  1934.7058       0       0 elevation
3  1818 1318.2088  2815.3553       0       0 elevation
4  2475 1839.7219  3908.2386       0       0 elevation
5  3048 2355.6455  4961.4685       0       0 elevation
6  3654 2875.1311  6132.1045       0       0 elevation
7  4015 3399.6812  7481.3624       0       0 elevation
8  4390 3923.6316  9046.2503       0       0 elevation
9  4771 4444.2381 11106.9258       0       0 elevation
10 5043 4964.0425 13518.9488       0       0 elevation
11 5189 5487.8707 15735.6969       0       0 elevation
12 5320 6010.3592 18888.8028       0       0 elevation
13 5241 6537.3497 21561.5840       0       0 elevation
14 5559 7054.9452 24910.7296       0       0 elevation
15 5549 7579.1120 28445.9440       0       0 elevation

# Überblick über die empirischen Variogrammwerte für die Bodendaten
> vemp.soils
     np      dist    gamma dir.hor dir.ver    id
1   360  343.7490 137.2222       0       0 soils
2  1120  799.8131 195.2098       0       0 soils
3  1824 1318.5961 234.1965       0       0 soils
4  2412 1834.5285 266.8194       0       0 soils
5  3012 2352.0633 282.8933       0       0 soils
6  3641 2871.2013 288.2870       0       0 soils
7  3972 3396.9237 291.2706       0       0 soils
8  4371 3917.8242 298.6789       0       0 soils
9  4761 4440.0357 295.9456       0       0 soils
10 4956 4957.5317 299.1600       0       0 soils
11 5225 5474.7260 301.7151       0       0 soils
12 5247 5999.1216 314.3262       0       0 soils
13 5396 6518.7700 308.1648       0       0 soils
14 5612 7043.6475 312.0930       0       0 soils
15 5621 7566.3977 308.7360       0       0 soils
```

Die Variogrammwerte sind nach den Klassen geordnet. Analog zu den Variogrammwerten und Distanzen steigen für die Höhen- und Bodendaten auch die Anzahl der Punktpaare,

die pro Klasse einberechnet wurden, welche in der ersten Spalte `np` angegeben sind. Lediglich in der 15. Klasse der Höhendaten sind wieder etwas weniger Punkte als in der vorherigen vorhanden.
Ist man in R an einer kleineren Klassenbreite und damit mehr Klassen und empirischen Variogrammwerten interessiert, lässt sich dies mit Angabe von `width` bewerkstelligen.

```
# Empirisches Variogramm mit kleinerer Klassenbreite für die Höhendaten
> vemp.elev2 <- variogram(object=g.elev, width=300)

# Empirisches Variogramm mit kleinerer Klassenbreite für die Bodendaten
> vemp.soils2 <- variogram(object=g.soils, width=125)
```

Die Grafiken für beide Variogramme mit einer kleineren Klassenbreite und damit mehr Schätzern, für die Höhendaten $u = 300$ und für die Bodendaten $u = 125$, sind jeweils in den Abbildungen 5.3(b) und 5.4(b) veranschaulicht.
Neben anderen Klassenbreiten können auch nicht äquivalente Klassengrößen mithilfe von `boundaries` übergeben werden.

```
> variogram(object, boundaries=c(0,50,100, seq(250,5250, 250)))
```

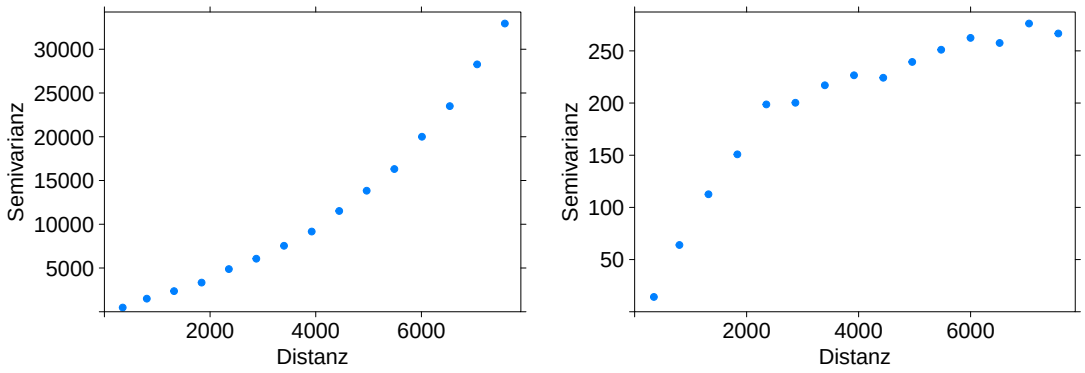
Eine kleinere Klassenbreite und damit mehr Klassen und Mittelwerte h_k, die in den Klassen berechnet werden, führen nach Bivand et al. (2013) dazu, dass zwar mehr Schätzer für die Variogrammfunktion vorliegen, diese Schätzer aber eine größere Verzerrung aufweisen, da weniger Beobachtungen pro Klasse vorhanden sind.

Wie bereits in Abschnitt 5.1 erwähnt, lassen sich mithilfe der Variogrammwolke leicht Ausreißer identifizieren. Werden diese nicht aus den Daten entfernt, existiert nach Bivand et al. (2013) die Möglichkeit, robuste Messungen für das empirische Variogramm zu berechnen, indem das Argument `cressie=TRUE` gesetzt wird. Damit soll sicher gestellt werden, dass die Ausreißer die Schätzung des Variogramms nicht verzerren. Einen Eindruck, wie sich die empirischen Variogramme daraufhin für die Höhen- und Bodendaten verändern, gibt Abb. 5.5.

```
# Empirisches Variogramm mit cressie=TRUE für die Höhendaten
> vemp.elev3 <- variogram(object=g.elev, cressie=TRUE)

# Empirisches Variogramm mit cressie=TRUE für die Bodendaten
> vemp.soils3 <- variogram(object=g.soils, cressie=TRUE)
```

5.3 Eigenschaften und Parameter des Variogramms

Variogramme erfüllen nach Wackernagel (1995) bestimmte Eigenschaften. Zum einen ist das Variogramm eine *monoton wachsende* Funktion, was daran liegt, dass die Varianz



(a) Höhendaten (b) Bodendaten

Abb. 5.5: Empirische Variogramme mit `cressie=TRUE`

des Inkrements mit zunehmendem Abstand h steigt.

$$\gamma(h) \geq 0 \tag{5.11}$$

Dies ist insofern sinnvoll, da das Variogramm den Zusammenhang zwischen zwei Lokationen beschreibt und dieser Zusammenhang mit zunehmendem Abstand h abnehmen sollte.
Weiterhin ist das Variogramm aufgrund der Annahme der Stationarität *symmetrisch*,

$$2\gamma(h) = 2\gamma(-h). \tag{5.12}$$

Dann gilt nach der Definition des Variogramms für Punkte mit dem Abstand $h = 0$ und damit im Ursprung

$$\gamma(0) = \mathrm{Var}(Y(s) - Y(s+0)) \tag{5.13}$$
$$= 0. \tag{5.14}$$

Zusätzlich gilt, dass das Variogramm langsamer als $|h|^2$ für $h \to \infty$ steigt, da sonst der Erwartungswert $\mu(h)$ aus Definition 3.15 nicht Null sein kann. Zuletzt muss das Variogramm eine *konditional negativ definite* Funktion sein. Ausführliche Darstellungen und Beweise zu den genannten Eigenschaften finden sich in Webster und Oliver (2007).

Darstellen lässt sich das Variogramm nach Diggle und Ribeiro (2007) auch durch die Kovarianzfunktion und Korrelationsfunktion mit

$$\begin{aligned} 2\gamma(h) &= Var(Y(s) - Y(s+h)) \\ &= Var(Y(s)) + Var(Y(s+h)) - 2Cov(S(h), S(s+h)) \end{aligned}$$

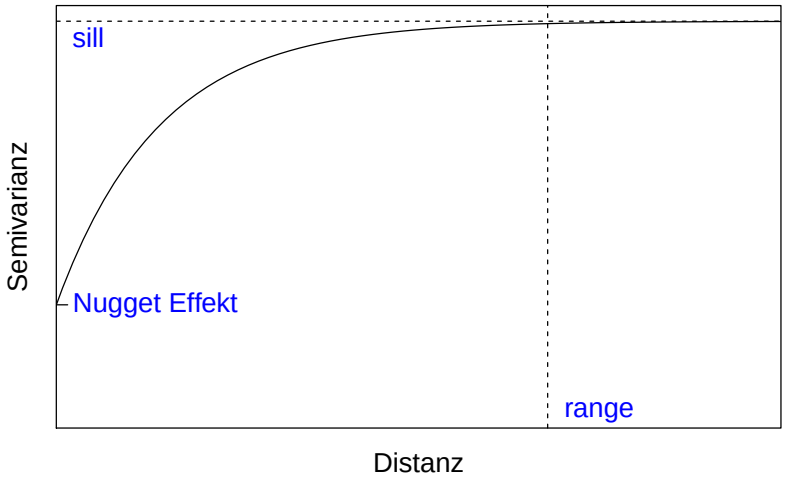
$$\begin{aligned} &\overset{3.14}{=} c(0) + c(0) - 2c(h) \\ &= 2(c(0) - c(h)) \qquad (5.15) \\ &= 2(\sigma^2 - c(h)) \\ &\overset{(3.15)}{=} 2(\sigma^2(1 - \rho(h))), \qquad (5.16) \end{aligned}$$

wobei nach Wackernagel (1995) beachtet werden sollte, dass dies nur unter Annahme der schwachen Stationarität gilt, da die Kovarianzfunktion auf Basis dieser definiert wurde.

Neben diesen genannten Eigenschaften existieren Variogrammparameter. So ist es in der Praxis nach Webster und Oliver (2007) selten der Fall, dass sich die Variogrammfunktion für $h = 0$ zu Null ergibt. Tatsächlich geht das Variogramm $\gamma(h)$ für $h \to 0$ nach Diggle und Ribeiro (2007) gegen einen Wert $\tau^2 \neq 0$, sodass

$$\begin{aligned} 2\gamma(h) &= \tau^2 + \sigma^2(1 - \rho(h)) \\ 2\gamma(0) &= \tau^2 \qquad (5.17) \end{aligned}$$

resultiert. Dieser Intercept τ^2, eine Unstetigkeit im Punkt $h = 0$, wird als *Nugget Effekt*, bezeichnet (Diggle und Ribeiro; 2007). Er entsteht dadurch, dass Mikrovariabilitäten vorhanden sein können (auch als *weißes Rauschen* bekannt), Messungen nur für Punkte mit einem sehr kleinen Abstand $h \neq 0$ möglich sind, sodass für $h \to 0$ extrapoliert werden muss oder Messungenauigkeiten vorliegen (Navratil; 2006). Der Nugget Effekt kann also als Varianz der Messabweichung interpretiert werden (Diggle und Ribeiro; 2007). Für



Quelle: Eigene Darstellung nach Hattermann (2014)

Abb. 5.6: Variogrammparameter (Nugget Effekt, *range*, *sill*)

Messwerte von Lokationen, die sehr weit voneinander entfernt liegen, für die also h sehr groß wird, ergibt sich $\rho(h) \overset{h\to\infty}{\longrightarrow} 0$. Demnach besteht keine Korrelation mehr zwischen diesen Werten, sondern sie sind unabhängig voneinander. Navratil (2006) definiert als *range* dann einen zugehörigen Abstand h, an dem $\sigma^2 - \gamma(h) < \epsilon$ gilt. Für $\gamma(h)$ resultiert in diesem Fall $\gamma(h) \overset{h\to\infty}{=} \tau^2 + \sigma^2$ als annähernd konstant. Der Grenzwert $\gamma(h) = \tau^2 + \sigma^2$ wird als *sill* bezeichnet (Diggle und Ribeiro; 2007). Zusätzlich bennen Diggle und Ribeiro (2007) den Begriff *practical range* für den Fall, dass $\rho(h)$ nicht 0 wird. Dann wird der *practical range* für den Abstand h_0 definiert, an dem $\rho(h_0) = 0.05$ gilt, die Korrelation zwischen zwei Werten also sehr gering ist. Dargestellt sind die Variogrammparameter in Abb. 5.3.

5.4 Kovariogramm und Korrelogramm

Neben dem Variogramm bieten das *Kovariogramm* $c(\cdot)$ und das *Korrelogramm* $\rho(\cdot)$ Möglichkeiten, den räumlichen Zusammenhang zwischen Realisationen von räumlichen Zufallsvariablen zu modellieren. Beide stehen, lediglich mit anderem Namen, für die bereits in Definition 3.14 und Gleichung (3.15) vorgestellte, stationäre Kovarianzfunktion bzw. stationäre Korrelationsfunktion. Das Kovariogramm beschreibt die Kovarianz und das Korrelogramm die Korrelation zweier Lokationen bezogen auf den Abstand h. Beide stehen über Gleichung (5.15) und (5.17) zu sehen ist, in direktem Zusammenhang mit dem Variogramm:

$$2c(h) = 2(\sigma^2 - \gamma(h)) \tag{5.18}$$

$$2\rho(h) = 2\left(1 - \frac{\gamma(h)}{\sigma^2}\right). \tag{5.19}$$

Speziell ergibt sich das Kovariogramm aus der Differenz des Variogrammwertes $\gamma(h)$ und der Varianz $c(0) = \sigma^2$ und spiegelt somit das Variogramm. Die beiden verhalten sich also entgegengesetzt, weshalb für das Kovariogramm die umgekehrten Eigenschaften des Variogramms gelten (siehe Abschnitt 5.3). Es ist monoton fallend, geht für große Abstände h gegen Null und im Ursprung gegen $\tau^2 + \sigma^2$. Diese Eigenschaften beschreiben daher gut die Annahme, dass Lokationen mit einem geringen Abstand h eine höhere Korrelation aufweisen und diese mit steigendem Abstand sinkt. Weiterhin müssen das Kovariogramm und Korrelogramm nach Webster und Oliver (2007) *positiv semidefinit* sein. Zusätzlich sind $\rho(\cdot)$ und $c(\cdot)$ analog zu $\gamma(\cdot)$, wie schon in Abschnitt 3.1 erwähnt, symmetrisch.

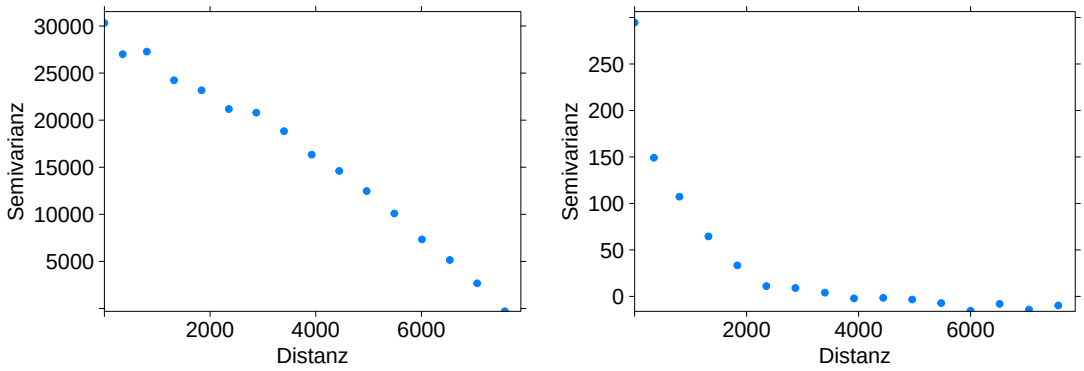
Kovariogramm in R

Ein Kovariogramm lässt sich in R mit der Angabe von `covariogram=TRUE` berechnen.

```
# Kovariogramm für die Höhendaten
```

```
> vko.elev <- variogram(object=g.elev, covariogram=TRUE)

# Kovariogramm für die Bodendaten
> vko.soils <- variogram(object=g.soils, covariogram=TRUE)
```

Die sich für die Höhen- und Bodendaten ergebenden Kovariogramme sind in Abb. 5.7 dargestellt. Vergleicht man sie mit den empirischen Variogrammen aus Abb. 5.3 und 5.4, zeigt sich die Spiegelung dieser durch die Kovariogrammfunktionen.



(a) Höhendaten (b) Bodendaten

Abb. 5.7: Kovariogramme

5.5 Theoretisches Variogramm

Grundsätzlich ist das theoretische Variogramm in Definition 5.1 angegeben. Da jedoch nur eine diskrete Anzahl an Lokationen beobachtet wird, der stochastische Prozess aber als stetig auf dem Untersuchungsgebiet definiert ist, ergibt sich daraus nach Navratil (2006) keine eindeutig bestimmte Funktion $\gamma(h)$. Notwendig ist daher die Anpassung einer Funktion an das empirische Variogramm, welches auf der diskreten Anzahl der empirischen Werte beruht. Die angepasste Funktion wird dann als *theoretisches Variogramm* bezeichnet (Diggle und Ribeiro; 2007). Dabei ist nach Webster und Oliver (2007) nicht jede Funktion eine mögliche Variogrammfunktion, sondern die Eigenschaften aus Abschnitt 5.3 müssen erfüllt sein. Da diese Eigenschaften nach Diggle und Ribeiro (2007) nicht immer leicht zu überprüfen sind, werden einige Standardfunktionen zur Modellierung von Variogrammen genutzt, von denen bekannt ist, dass sie diese Eigenschaften erfüllen.

Unterschieden wird bei der Variogrammmodellierung nach Webster und Oliver (2007)

zwischen *beschränkten* und *unbeschränkten* Modellen, je nachdem, ob die Varianz beschränkt ist. In dieser Arbeit wird sich auf eine Auswahl der beschränkten Modelle konzentriert, da diese nach Webster und Oliver (2007) häufiger verwendet werden. In der Literatur werden die Modelle häufig auch für die Kovarianzfunktion $c(\cdot)$ oder die Korrelationsfunktion $\rho(\cdot)$ angegeben. Da diese sich gegenseitig ergeben (siehe Abschnitt 5.4) ist allerdings irrelevant, für welche Funktion die Modelle definiert werden.

Die folgenden Variogrammmodelle und ihre Eigenschaften sind in Webster und Oliver (2007) zu finden. Mit a wird der *range* und mit c der *sill* bezeichnet. Mit r ist ein Distanzparameter gegeben, der das räumliche Ausmaß der Modelle angibt. r ersetzt dabei den *range* a, da in einigen Modellen nach Webster und Oliver (2007) nur asymptotisch ein *sill* erreicht wird und daher kein endlicher *range* vorhanden ist.

- Nugget-Effekt Modell

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ \sigma^2 & h > 0 \end{cases} \tag{5.20}$$

- Sphärisches Modell

$$\gamma(h) = \begin{cases} c\left(\frac{3h}{2a} - \frac{1}{2}\left(\frac{h}{a}\right)^3\right) & h \leq a \\ c & h > a \end{cases} \tag{5.21}$$

- Exponentialmodell

$$\gamma(h) = c\left(1 - \exp\left(-\frac{h}{r}\right)\right) \tag{5.22}$$

- Gauss-Modell

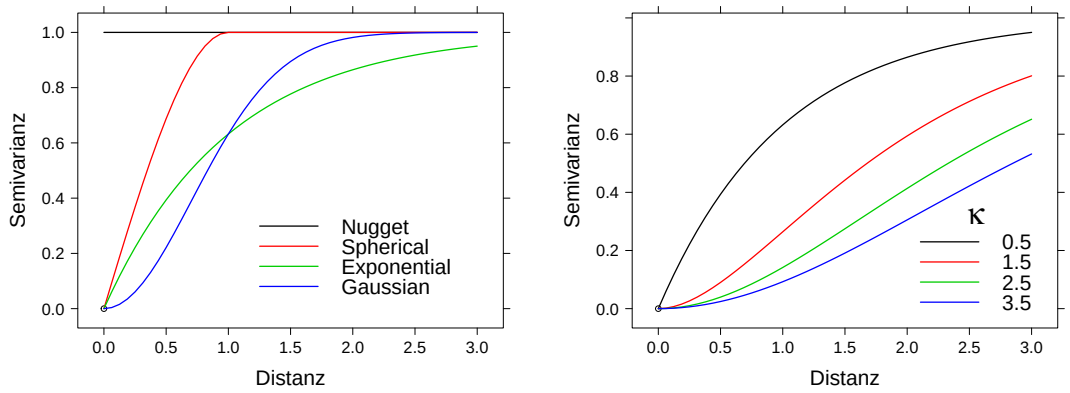
$$\gamma(h) = c\left(1 - \exp\left(-\left(\frac{h}{r}\right)\right)\right) \tag{5.23}$$

- Matérn-Modelle

$$\gamma(h) = c\left(1 - \frac{1}{2^{\kappa-1}\Gamma(\kappa)}\left(\frac{h}{r}\right)^{\kappa} K_{\kappa}\left(\frac{h}{r}\right)\right) \tag{5.24}$$

  $K_{\kappa}(\cdot)$ ist dabei die *Besselfunktion* der Ordnung κ, wobei κ einen Glättungsparameter mit $0 \leq \kappa < \infty$ angibt und $\Gamma(\cdot)$ steht für die *Gammafunktion*. Für festgelegte κ reduziert sich die Matérn-Funktion nach Diggle und Ribeiro (2007). So ergibt sie sich für $\kappa = 0.5$ beispielsweise als Exponentialmodell und für $\kappa \to \infty$ resultiert das Gauss-Modell.

Für die Modelle, die den *sill* nur asymptotisch erreichen, wird häufig der *practical range* verwendet. Näherungsweise ergibt er sich für das Exponentialmodell zu $3r$ und für das Gauss-Modell mit $\sqrt{3}r$. Neben den hier genannten Modellen sind in Webster und Oliver (2007) noch einige weitere, sowie auch die hier erwähnten, ausführlicher angegeben.



(a) Nugget-Effekt-, Sphärisches, Exponential- und Gaussmodell

(b) Matérn-Modelle

Abb. 5.8: Variogrammfunktionen mit *range* $a = 1$ und *sill* $c = 1$

Grafisch sind die genannten Modelle in Abb. 5.8 dargestellt, wobei in der rechten Abbildung Matérn-Funktionen für $\kappa \in \{0.5, 1.5, 2.5, 3.5\}$ und in der linken die anderen hier vorgestellten Funktionen gezeigt sind. Je nach vorliegenden Daten reicht die einfache Annahme der hier genannten Variogrammfunktionen nach Webster und Oliver (2007) nicht aus und es ist sinnvoll, die Modelle miteinander zu kombinieren, um die beste Anpassung zu erhalten. Häufig handelt es sich dabei beispielsweise um eine Kombination des Nugget-Effekt-Modells mit einem der anderen Modelle, da das Variogramm, wie bereits in Abschnitt 5.3 erwähnt, in der Regel nicht durch den Ursprung geht.
Neben der Annahme eines Modells müssen die Parameter *range* und *sill* und der Nugget-Effekt geschätzt werden. Dies kann nach Diggle und Ribeiro (2007) beispielsweise durch die in Abschnitt 3.2 vorgestellte Kleinste-Quadrate-Methode oder Maximum-Likelihood-Methode geschehen. Die Schätzungen ergeben sich jedoch als komplex, weshalb sie hier nicht weiter ausgeführt werden. Nähere Erklärungen und auch weitere Methoden zur Schätzung finden sich in Diggle und Ribeiro (2007).

Theoretisches Variogramm in R

In R wird das theoretische Variogramm berechnet, indem an das empirische Variogramm eine Funktion angepasst wird. Dies geschieht mithilfe der Funktion `fit.variogram()`. Dabei muss nach Bivand et al. (2013) vorweg überlegt werden, welches Modell zur Anpassung verwendet wird; sowie geeignete Werte für die Variogrammparameter *partial sill*, *range* und *Nugget* und eine passende Schätzmethode gewählt werden.

In dem `gstat`-Paket sind neben den im vorherigen Abschnitt genannten Variogrammmodellen zusätzlich weitere grundlegende Modelle implementiert. Eine Übersicht über den Verlauf des Großteils dieser kann mit `show.vgms()` erhalten werden, womit Abb. 5.5 resultiert. Wofür die Kürzel, wie beispielsweise `Per` stehen, zeigt der Aufruf von `vgm()`. Mit diesem Aufruf ist neben der Grafik ebenfalls ein Überblick über die implementierten Modelle gegeben, wobei hier nur ein Ausschnitt dargestellt ist.

```
> vgm()
   short                                      long
1    Nug                                Nug (nugget)
2    Exp                           Exp (exponential)
3    Sph                             Sph (spherical)
4    Gau                              Gau (gaussian)
5    Exc                Exclass (Exponential class)
6    Mat                                Mat (Matern)
7    Ste Mat (Matern, M. Stein's parameterization)
8    Cir                               Cir (circular)
9    Lin                                 Lin (linear)
10   Bes                                 Bes (bessel)
11   Pen                       Pen (pentaspherical)
12   Per                               Per (periodic)
13   Wav                                   Wav (wave)
14   Hol                                   Hol (hole)
15   Log                          Log (logarithmic)
16   Pow                                 Pow (power)
17   Spl                                 Spl (spline)
```

Ist man außerdem an dem Verlauf expliziter Modelle interessiert, kann beispielsweise

```
> show.vgms(model=c("Nug","Sph","Exp","Gau"), sill=1, range=1,
+           as.groups=TRUE)
> show.vgms(model="Mat", sill=1, range=1,
+           kappa.range=c(0.5,1.5,2.5,3.5), as.groups=TRUE)
```

angegeben werden, womit sich inhaltlich Abb. 5.8 ergibt.

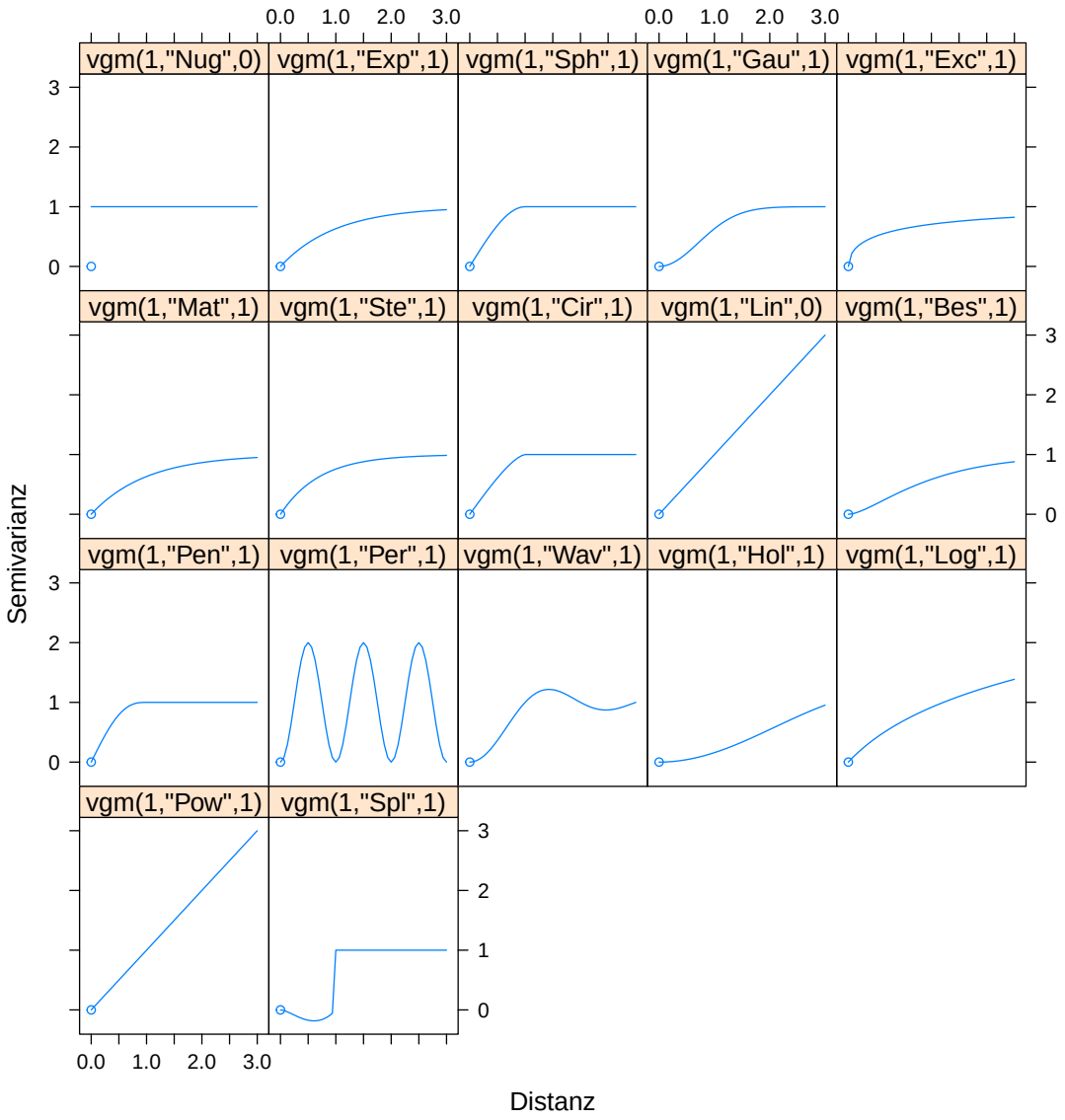


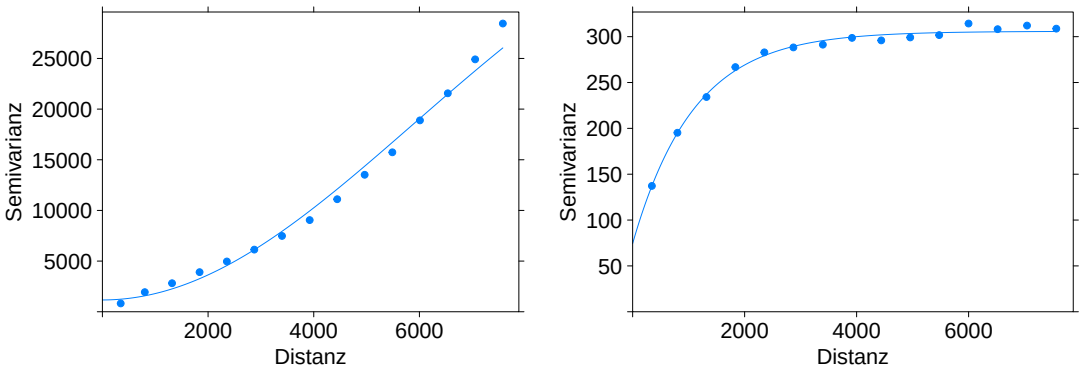
Abb. 5.9: Übersicht über den Verlauf implementierter Variogrammodelle in `gstat`

Die Anpassung des theoretischen Variogramms wird nach Bivand et al. (2013) üblicherweise auf Basis von Vorkenntnissen über das Verhalten der untersuchten Variablen und per Augenmaß vollführt, wobei an R das Modell, sowie Startwerte für die Variogrammparameter übergeben werden. Die optimalen Werte werden dann in R berechnet. Nach Bivand et al. (2013) werden in praktischen Anwendungen größtenteils das Sphärische (`Sph`), Exponential- (`Exp`), Gauss-(`Gau`), Matérn- (`Mat`) oder Potenzmodell (`Pow`) verwendet.
Entsprechend, nach den Abbildungen 5.3 und 5.4, ergeben sich das Gauss-Modell für die Höhendaten und das Exponentialmodell für die Bodendaten per Augenmaß. Anhand der Grafiken werden auch die Werte für die Variogrammparameter geschätzt und R übergeben. Bei der Ausführung der Anpassung zeigt sich dann, dass diese etwas höher als die in R berechneten Werte liegen. Abgebildet sind die beiden angepassten Funktionen in Abb. 5.10.

```
# Anpassung des theoretischen Variogramms für die Höhendaten
> vfit.elev <- fit.variogram(vemp.elev, vgm(60000,"Gau",10000,0))
> vfit.elev
  model     psill   range
1   Nug  1163.372    0.00
2   Gau 44243.377 8338.85

# Anpassung des theoretischen Variogramms für die Bodendaten
> vfit.soils <- fit.variogram(vemp.soils, vgm(280,"Exp",2000,0))
> vfit.soils
  model     psill    range
1   Nug  74.44808    0.000
2   Exp 231.41231 1087.191
```

Ist in der Praxis tatsächlich einer der Werte für *sill* oder *range* bekannt, kann dieser mithilfe von `fit.sills` oder `fit.ranges` festgehalten werden und wird nicht von R geschätzt.

Die Parameterschätzung erfolgt in R nach Bivand et al. (2013) über die Gewichtung der Abweichung zwischen den empirischen Variogrammwerten und der angepassten Variogrammfunktion. Mithilfe des Arguments `fit.method` können konkrete Optionen für die Gewichtung ausgewählt werden. Ein Teil dieser Optionen ist nach Bivand et al. (2013) in Tabelle 2 angegeben. `fit.method=6` (OLS) steht für *Ordinary Least Squares* und entspricht der Methode der kleinsten Quadrate (siehe Definition 3.12). Für die ML-Schätzung ist extra `fit.variogram.reml()` anzugeben, wobei REML für *Restricted Maximum Likelihood* steht, also eine Beschränkung der Maximum-Likelihood-Schätzung (siehe Definition 3.11).



(a) Höhendaten (b) Bodendaten

Abb. 5.10: Angepasste Variogrammfunktionen an die Höhendaten und Bodendaten

Tabelle 2: Optionen für das Argument `fit.method` der Funktion `fit.variogram`

`fit.method`	Gewichtung
1	N_h
2	$N_h/(\gamma(h))^2$
6	no weights (OLS)
7	N_h/h^2

Quelle: Darstellung nach Pebesma (2014a)

Einen exemplarischen Vergleich verschiedener Variogrammmodelle und Optionen für `fit.method` für die Bodendaten bietet Tabelle 3. Die Spalten für die Annahme des Exponentialmodells und des Matérnmodells ergeben sich interessanterweise gleich, was daran liegt, dass $\kappa = 0.5$ geschätzt wurde und das Matérnmodell dann, wie bereits erwähnt, als Exponentialmodell resultiert. Beide weisen für alle vier verschiedenen Anpassungsmethoden den kleinsten *SSErr* (*Sum of squared error*) auf und bieten damit die beste Anpassung an das empirische Variogrammmodell. Damit erweist sich die Wahl der Exponentialfunktion per Augenmaß als sinnvoll. Ebenso könnte natürlich in diesem Fall die Matérnfunktion verwendet werden.
Voreingestellt ist `fit.method=7`. Nach Bivand et al. (2013) liefert diese Option in den meisten Fällen gute Ergebnisse, was die für die Bodendaten ausgetesteten Varianten in Tabelle 3 bestätigen. Lediglich für Datensätze, in denen beispielsweise doppelte Realisationen vorhanden sind, kann es laut Bivand et al. (2013) passieren, dass die Variogrammwerte für Distanzen nahe Null unplausibel und unendlich groß werden.

Tabelle 3: Vergleich verschiedener Variogrammfunktionen und verschiedener Optionen für `fit.method` für die Bodendaten

		Modell			
		Sphärisch	Exponential	Gauss	Matérn
`fit.method=1`	*Nugget*	136.23	87.89	152.45	87.89
	partial sill	167. 34	220.08	150.76	220.08
	range	3365.20	1185.58	1555.24	1185.58
	κ	-	-	-	0.5
	SSErr	3135119	1160633	2988994	1160633
`fit.method=2`	*Nugget*	109.21	76.29	122.63	76.29
	partial sill	193.56	230.79	174.84	230.79
	range	2974.94	1114.23	1151.36	1114.23
	κ	-	-	-	0.5
	SSErr	49.81	19.68	163.96	19.68
`fit.method=6`	*Nugget*	111.66	77.00	132.94	77.00
	partial sill	190.00	230.16	167.70	230.16
	range	2954.66	1119.12	1344.55	1119.12
	κ		-	-	0.5
	SSErr	889.18	258.00	1019.11	258.00
`fit.method=7`	*Nugget*	102.33	74.45	125.53	74.45
	partial sill	193.65	231.41	166.70	231.41
	range	2642.27	1087.19	1179.46	1087.19
	κ		-	-	0.5
	SSErr	0.35	0.07	0.56	0.07

Für die weiteren Auswertungen werden die Annahmen des Gauss-Modells für die Höhendaten und des Exponentialmodells für die Bodendaten beibehalten, d. h. die beiden definierten Objekte `vfit.elev` und `vfit.soils` werden als Basis für das Kriging verwendet.

5.6 Anisotropes Variogramm

Bisher wurde das Variogramm für isotrope Daten vorgestellt, sodass $\gamma(h) = \gamma(\|h\|)$ gilt. Damit wurde unterstellt, dass nur der euklidische Abstand der Lokationen $s_i - s_j$ relevant ist, nicht aber ihre genaue Position und damit die Richtung, in der die Lokationen zueinander stehen. In vielen Anwendungen verhält sich der Zusammenhang zwischen Lokationen in unterschiedlichen Richtungen jedoch voneinander abweichend, womit es sich um anisotrope Daten handelt. Dies ist beispielsweise bei Messungen in Sedimentschichten der Fall, da sie in der Regel in vertikaler Richtung deutlich höhere Unterschiede aufweisen als in horizontaler Richtung. Auch Baumbewuchs kann, entlang eines Sees oder im Gebirge, stark von den physischen Gegebenheiten abhängig sein, sodass der Zusammenhang zwischen zwei Lokationen nicht mehr richtungsunabhängig ist. Um anisotrope Daten zu identifizieren, können die empirischen Variogramme für verschiedene Richtungen einzeln betrachtet werden (Wackernagel; 1995). Ergeben sich diese unterschiedlich, kann von Anisotropie ausgegangen werden. Unterschieden wird nach Hattermann (2014) zwischen zwei Fällen der Anisotropie:

Zonale Anisotropie liegt vor, wenn sich die Variation in einer konkreten Richtung stark unterscheidet, also ein unterschiedlicher *sill* resultiert.

Geometrische Anisotropie ist gegeben, wenn für verschiedene Richtungen der *sill* gleich ist und sich der *range* unterschiedlich ergibt.

In der Literatur werden unterschiedliche Verfahren genannt, um mit anisotropen Daten umzugehen. Allgemein wird bei Daten mit unterschiedlichen Richtungseffekten nicht mehr der euklidische Abstand $h = \|s_i - s_j\|$ betrachtet. Eine Lösung für die geometrische Anisotropie im zweidimensionalen Fall findet sich in Fahrmeir et al. (2009). Dazu wird der euklidische Abstand durch

$$\sqrt{(s_i - s_j)' R(\phi)' D(\delta) R(\phi)(s_i - s_j)} \tag{5.25}$$

ersetzt, wobei angenommen wird, dass die geometrisch anisotropen Daten mithilfe dieser Gleichung in isotrope Daten transformiert werden können. $R(\phi)$ stellt dabei eine Rotationsmatrix mit Anisotropiewinkel $\phi \in [0, 2\pi]$ dar,

$$\begin{pmatrix} \cos(\psi_A) & \sin(\psi_A) \\ -\sin(\psi_A) & \cos(\psi_A) \end{pmatrix} \tag{5.26}$$

und $D(\delta)$ eine Dehnungsmatrix mit Anisotropie-Verhältnis $\delta \geq 1$,

$$\begin{pmatrix} \delta^{-1} & 0 \\ 0 & 1 \end{pmatrix}. \tag{5.27}$$

Etwas allgemeiner sieht die Lösung für geometrische Anisotropie in Navratil (2006) mit

$$2\gamma(h) = 2\gamma^0(|A \cdot h|). \tag{5.28}$$

aus, wobei A eine $d \times d$-Matrix und γ^0 ein isotropes Semivariogramm ist, welches durch den Faktor $|A \cdot h|$ in Abhängigkeit von der Richtung angepasst wird. Die Transformation durch A wird hier nicht genauer spezifiziert. In Wackernagel (1995) ist eine Lösung für den Fall der zonalen Anisotropie gegeben. Dort wird im zweidimensionalen Raum eine Aufspaltung der Variogrammfunktion in zwei Terme

$$2\gamma(h) = 2(\gamma_1(h) + \gamma_2(h)) \tag{5.29}$$

vorgeschlagen, wobei $\gamma_1(h)$ ein isotropes Variogramm und $\gamma_2(h)$ ein geometrisch anisotropes Variogramm darstellt. Genauere Erläuterungen zum Umgang mit anisotropischen Daten finden sich in Wackernagel (1995).

Anisotropie in R

Für den Umgang mit anisotropen Daten existieren in R nach (Bivand et al.; 2013) zwei mögliche Varianten. Zum einen können Variogramme für verschiedene Richtungen einzeln betrachtet werden und zum anderen kann eine Variogrammkarte genutzt werden.

Variogramme für unterschiedliche Richtungen lassen sich in R durch Angabe von `alpha` berechnen. Mit `alpha` werden dabei die Richtungen in Grad im Uhrzeigersinn von der y-Achse weg übergeben. `alpha=0` gibt beispielsweise die nördliche Richtung an und `alpha=90` die östliche Richtung. Nach Pebesma (2014b) werden die Punktpaare aus dem Datensatz dann immer in das jeweilige Variogramm einberechnet, dessen Richtung am wenigsten weit entfernt ist. So beinhaltet das Variogramm in nördlicher Richtung alle Punktpaare zwischen -22.5 Grad und 22.5 Grad.

Für die Höhen- und Bodendaten sind in Abb. 5.11 und Abb. 5.12 Variogramme für vier verschiedene Richtungen von 0 bis 135 Grad dargestellt. Diese Betrachtung reicht aus, da sich die Variogramme zwischen 180 und 360 Grad aufgrund der Tatsache, dass das Variogramm symmetrisch definiert ist, gleich zu diesen ergeben (Pebesma (2014b)). Bei den Höhendaten ergeben sich die Variogramme für 0 und 45 Grad, sowie für 90 und 135 jeweils ähnlich, untereinander aber unterschiedlich, während sich die Variogramme für die unterschiedlichen Richtungen der Bodendaten insgesamt ähnlich ergeben. Um Funktionen an die einzelnen direktionalen Variogramme anzupassen, muss zusätzlich zu den Angaben im omnidirektionalen Fall das Argument `anis` angegeben werden.

```
# Direktionale Variogramme für die Höhendaten
> vdir.elev <- variogram(object=g.elev, alpha=c(0,45,90,135))
> vanis.elev <- vgm(60000, "Gau", 10000, 0, anis=c(90,0.7))
```

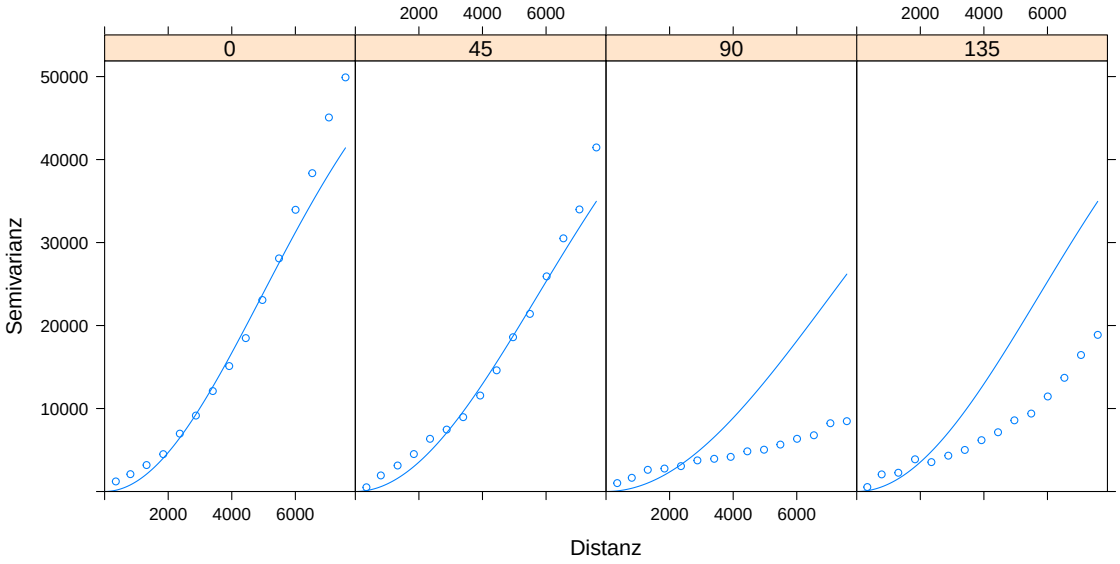


Abb. 5.11: Variogramme für vier unterschiedliche Richtungen für die Höhendaten

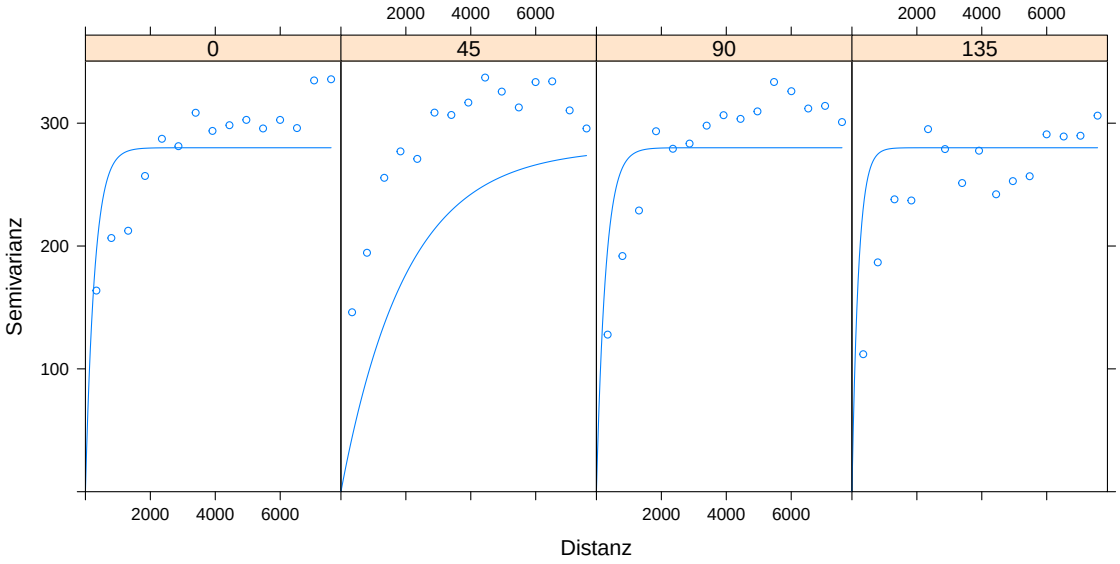


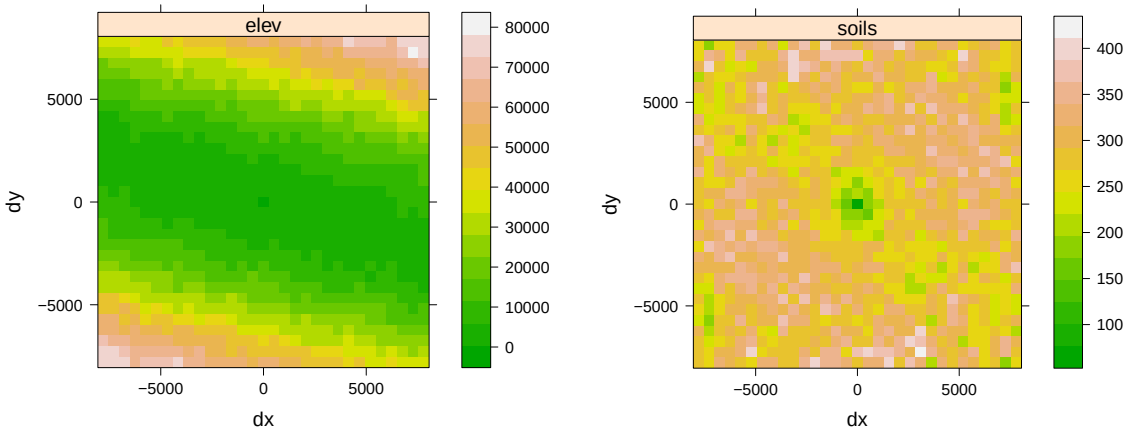
Abb. 5.12: Variogramme für vier unterschiedliche Richtungen für die Bodendaten

```
# Direktionale Variogramme für die Bodendaten
> vdir.soils <- variogram(object=g.soils, alpha=c(0,45,90,135))
> vanis.soils <- vgm(280,"Exp",2000,0, anis=c(45,0.1))
```

Bei der Berechnung einer Variogrammkarte werden nach Pebesma (2014b) im Gegensatz zu direktionalen Variogrammen nicht mehr die Richtung und der Abstand getrennt betrachtet, sondern gemeinsam. Dargestellt sind in der Variogrammkarte symmetrisch um den Abstand $(0, 0)$ in jeder Rasterzelle (x, y) der Mittelwert der Variogrammwerte der Lokationspaare, die einen Abstandsvektor $h = (x, y)$ aufweisen. x stellt hier wie gehabt den Rechtswert und y den Hochwert der Koordinaten dar.

Variogrammkarten, wie Abb. 5.13 für die Höhen- und Bodendaten, lassen sich in R nach Bivand et al. (2013) mit der Angabe von `map=TRUE` erzeugen. Notwendig ist außerdem die Angabe von dem Argument `cutoff`, sowie einer Rastergröße mit `width`. Zusätzlich kann das Argument `treshold` beim Erstellen der Grafik gesetzt werden, damit nicht zu wenige Punktpaare in einigen Rasterzellen resultieren.

```
# Variogrammkarte für die Höhendaten
> vmap.elev <- variogram(object=g.elev, map=TRUE, cutoff=7800, width=520)

# Variogrammkarte für die Bodendaten
> vmap.soils <- variogram(object=g.soils, map=TRUE, cutoff=7800,
                          width=520)
```



(a) Höhendaten (b) Bodendaten

Abb. 5.13: Variogrammkarten

Die Farben in der Variogrammkarte geben nun die Variogrammwerte an. Bei den Höhendaten ist so zu sehen, dass sich die Höhen an Lokationen in x-Richtung ähnlich sind und in y-Richtung unterscheiden. Bei den Bodendaten lässt sich kein direkter Zusammenhang in verschiedenen Richtungen erkennen. Diese Grafiken bestätigen also die erkannten Zusammenhänge bezogen auf die Richtung aus den direktionalen Variogrammen in Abb. 5.11 und Abb. 5.12.

5.7 Weitere Spezialfälle

Weitere Spezialfälle neben anisotropen Daten, die bei der geostatistischen Auswertung bedacht werden müssen, sind nach Webster und Oliver (2007) beispielsweise der *Drift*, der *Hole-Effekt* oder unbeschränkte Variogramme.

Wie bereits in Abschnitt 5.5 erwähnt, wird nach Webster und Oliver (2007) zwischen beschränkten und unbeschränkten Modellierungen für Variogrammfunktionen unterschieden. Unbeschränkte Variogramme sind nach Webster und Oliver (2007) Variogramme, bei denen die Werte keinen *sill* erreichen. Bei solchen Variogrammen steigen die Werte immer weiter mit zunehmender Distanz. Dies deutet nach Webster und Oliver (2007) darauf hin, dass keine schwache Stationarität angenommen werden kann, sondern maximal intrinsische Stationarität, wobei keine Kovarianzfunktion existiert.

Ähnlich ergibt sich der Fall eines Drifts, welcher nach Diggle und Ribeiro (2007) und Webster und Oliver (2007) auch als *Trend* bezeichnet wird. Nach Hattermann (2014) kann es vorkommen, dass der *sill* nur scheinbar erreicht wird und die Variogrammwer-

te danach wieder ansteigen. Die Variogrammwolke der Höhendaten mit dem Argument `cutoff=Inf` in Abb. 5.1(b) weist beispielsweise auf einen solchen Fall hin. Dies impliziert nach Webster und Oliver (2007), dass kein konstanter Erwartungswert des stochastischen Prozesses $Y(s)$ auf dem Untersuchungsgebiet D vorliegt und somit keine Stationarität gegeben ist.

Weiterhin kann ein *Hole-Effekt* in den Daten vorliegen. Auf ihn lässt sich nach Webster und Oliver (2007) schließen, wenn eine wellenförmige Struktur in den Daten vorliegt, sodass die Variogrammwerte nach dem Erreichen ihres Maximums kleiner werden und daraufhin wieder ansteigen. Nach Hattermann (2014) deutet dies darauf hin, dass Regelmäßigkeiten in den Daten vorliegen, sodass nach bestimmten Abständen h wieder ähnliche Werte für $Y(s)$ resultieren.

6 Kriging

In der Regel ist es nicht möglich, an jeder Lokation im untersuchten Raum die interessierenden Größen zu erfassen. Gründe dafür können nach Navratil (2006) finanzielle Aspekte sein, da sich die Erfassung einer höheren Zahl an Lokationen nicht rentiert oder räumliche Aspekte, da es nicht an jeder Lokation möglich sein muss, Messungen vorzunehmen. Ziel ist es daher, Vorhersagen für unbeobachtete Lokationen zu treffen.
Das *Kriging* bietet nun die Möglichkeit, durch die vorhandenen diskreten Daten Rückschlüsse auf den zugrunde liegenden stetigen stochastischen Prozess $Y(s)$ zu ziehen und damit sein Verhalten und Werte an Lokationen s_0, an denen keine Beobachtungen gemacht wurden, vorherzusagen (Bivand et al.; 2013). Die Bezeichnung Kriging geht nach Wackernagel (1995) auf den Bergbauingeneur D.G. Krige zurück, welcher erstmals geostatistische Methoden in den 1950er Jahren zur Bestimmung in Südafrika anwendete.
Ganz allgemein, werden bei Interpolationsverfahren die umliegenden Lokationen gewichtet. Häufig werden nahe Lokationen höher gewichtet und weiter entfernte Lokationen niedriger. Das Kriging schließt nun zusätzlich die Größe der Ähnlichkeit zwischen Lokationen in Form des Variogramms mit ein.
Die Kriging-Prädiktion kann nach Navratil (2006) auch als *Best Linear Unbiased Estimator* (BLUE) bezeichnet werden. Dem liegt nach Navratil (2006) zugrunde, dass die Varianzen der Schätzwerte minimal sein sollen (*Best*), sowie die Schätzung auf einer linearen Funktion basiert (*Linear*) und erwartungstreu ist (*Unbiased*).

Dieses Kapitel behandelt drei häufig verwendete Kriging-Methoden, das *Simple Kriging* in Abschnitt 6.1, wofür ein bekannter konstanter Erwartungswert des stochastischen Prozesses vorausgesetzt wird, das *Ordinary Kriging* in Abschnitt 6.2, wenn der Erwartungswert zwar als konstant angenommen wird, aber nicht bekannt ist, und das *Universal Kriging* in Abschnitt 6.3 für Fälle, in denen der Erwartungswert nicht als konstant angenommen werden kann, da ein Drift in den Daten vorliegt.

Die genannten Eigenschaften und Schätzungen für die drei Kriging-Methoden ergeben sich, soweit nicht anders gekennzeichnet, nach Wackernagel (1995).

Kriging in R

In dieser Arbeit wird zur Durchführung des Krigings in R analog zu Bivand et al. (2013) die Funktion `krige()` verwendet. Die Funktion `krige()` ist dabei nach Pebesma (2004) eine Anwendung der Funktion `predict()`, die dementsprechend zu den gleichen Ergebnissen führt.

```
> krige(formula, locations, data, newdata, model, ...)
```

Zusätzlich zu den Angaben, die bereits für die Berechnung der Variogramme übergeben wurden, kommt die Angabe einer angepassten Variogrammfunktion `model` und eines

neuen Datensatzes oder räumlichen Objektes `newdata`, welche Lokationen, für die die Prädiktion gemacht werden soll, enthalten, hinzu. Als angepasste Variogrammfunktionen fungieren die in Kapitel 5 berechneten Funktionen `vfit.elev` für die Höhendaten `elevation.dem` und `vfit.soils` für die Bodendaten `soils`. Die Lokationen, für die die Vorhersage gemacht werden soll, werden in Form von Koordinatennetzen `grid.elev` und `grid.soils` übergeben. Diese sind neu erzeugt worden und stellen den Koordinatenbereich dar, der jeweils von den Stichproben für die Höhendaten bzw. für die Bodendaten abgedeckt wird. Daher berechnen sich die Netze aus den jeweils größten und kleinsten enthaltenen Koordinaten. Diese entsprechen prinzipiell den Werten aus dem in Abschnitt 2.2 erwähnten Slot `bbox`. Für die hier berechneten Koordinatennetze wird eine Zellgröße von 100 übergeben.

```
# Berechnung des Koordinatennetzes für die Höhendaten
> s1.range.e <- as.integer(range(elevsample@coords[,1]))
> s2.range.e <- as.integer(range(elevsample@coords[,2]))

> grid.elev <- expand.grid(s1=seq(from=s1.range.e[1], to=s1.range.e[2],
                                  by=100),
                           s2=seq(from=s2.range.e[1], to=s2.range.e[2],
                                  by=100))
> coordinates(grid.elev) <- ~s1+s2
> gridded(grid.elev) <- TRUE

# Berechnung des Koordinatennetzes für die Bodendaten
> s1.range.s <- as.integer(range(soilssample@coords[,1]))
> s2.range.s <- as.integer(range(soilssample@coords[,2]))

> grid.soils <- expand.grid(s1=seq(from=s1.range.s[1], to=s1.range.s[2],
                                   by=100),
                            s2=seq(from=s2.range.s[1], to=s2.range.s[2],
                                   by=100))
> coordinates(grid.soils) <- ~s1+s2
> gridded(grid.soils) <- TRUE
```

Außerdem werden die Stichproben `elevsample` und `soilssample`, auf Basis derer die Vorhersage für die beiden erzeugten Koordinatennetze gemacht wird, als Listen abgespeichert, um diese später in die grafische Darstellung der Prädiktionen mit einzubinden.

```
# Punkte der Höhendaten
> p.elev <- list("sp.points", elevsample, pch = 4, col = "black",
+                cex=0.5)
```

```
# Punkte der Bodendaten
> p.soils <- list("sp.points", soilssample, pch = 4, col = "black",
+                 cex=0.5)
```

6.1 Simple Kriging

Die Methode des *Simple Kriging* wird angewendet, wenn der Erwartungswert bekannt ist, weshalb es auch als *kriging with known mean* bezeichnet wird. Neben dem bekannten Erwartungswert muss die Bedingung der schwachen Stationarität (siehe Definition 3.14) erfüllt sein. Dann ergibt sich der Erwartungswert als konstant mit

$$\mathrm{E}(Y(s)) = \mathrm{E}(Y(s+h)) = \mu \tag{6.1}$$

und die Kovarianzfunktion, ebenfalls als bekannt, mit

$$\begin{aligned} c(h) &= \mathrm{Cov}(Y(s), Y(s+h)) \\ &= \mathrm{E}(Y(s)Y(s+h)) - \mu^2. \end{aligned} \tag{6.2}$$

Der Krige-Schätzer für eine unbeobachtete Lokation s_0 ist mit

$$\hat{Y}(s_0) = \mu + \sum_{i=1}^{n} \lambda_i \left(Y(s_i) - \mu\right) \tag{6.3}$$

gegeben, wobei mit λ_i die Abweichungen der beobachteten Lokationen von dem bekannten Mittelwert gewichtet werden. Von Interesse für die Güte des Schätzers ist, wie groß die Abweichung zwischen dem geschätzten Wert und dem wahren Wert an der Lokation s_0

$$Y(s_0) - \hat{Y}(s_0) \tag{6.4}$$

ist. Im Mittel sollte idealerweise kein Unterschied zwischen beiden Werten bestehen, sodass $\hat{Y}(s_0)$ einen unverzerrten Schätzer darstellt. Diese Erwartungstreue ergibt sich mit Gleichung (6.3), da

$$\begin{aligned} \mathrm{E}(Y(s_0) - \hat{Y}(s_0)) &= \mathrm{E}(Y(s_0)) - \mathrm{E}(\hat{Y}(s_0)) \\ &= \mu - \mathrm{E}\left(\mu + \sum_{i=1}^{n} \lambda_i (Y(s_i) - \mu)\right) \\ &= \mu - \mu + \sum_{i=1}^{n} \lambda_i (\mathrm{E}(Y(s_i)) - \mu) \\ &= \sum_{i=1}^{n} \lambda_i (\mu - \mu) \end{aligned}$$

$$= 0. \tag{6.5}$$

Dabei zeigt sich, dass durch die Definiton des Krige-Schätzers für die Gewichte λ_i keine Einschränkungen getroffen werden müssen, um die Erwartungstreue zu gewährleisten. Da der Schätzer $\hat{Y}s_0$ keine Verzerrung aufweist, resultiert die Varianz der Schätzung als *mittlerer quadratische Fehler* (MSE) von $\hat{Y}s_0$, da

$$\begin{aligned}
\sigma_E^2 &= \mathrm{Var}(Y(s_0) - \hat{Y}(s_0)) \\
&= \underbrace{\mathrm{E}\left((Y(s_0) - \hat{Y}(s_0))^2\right)}_{MSE(\hat{Y}s_0)} - \underbrace{\left(\mathrm{E}(Y(s_0) - \hat{Y}(s_0))\right)^2}_{Bias(\hat{Y}s_0)=0} \\
&= \mathrm{E}\left(\left(Y(s_0) - \mu + \sum_{i=1}^{n} \lambda_i (Y(s_i) - \mu)\right)^2\right) \\
&= \mathrm{E}\left((Y(s_0) - \mu)^2\right) - 2\sum_{i=1}^{n} \lambda_i \mathrm{E}\left((Y(s_0) - \mu)(Y(s_i) - \mu)\right) \\
&\quad + \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \mathrm{E}\left((Y(s_i) - \mu)(Y(s_j) - \mu)\right) \\
&= \mathrm{Var}(Y(s_0)) - 2\sum_{i=1}^{n} \lambda_i \left(\mathrm{E}(Y(s_0)Y(s_i)) - \mu^2\right) \\
&\quad + \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \left(\mathrm{E}(Y(s_i)Y(s_j)) - \mu^2\right) \\
&= c(s_0 - s_0) - 2\sum_{i=1}^{n} \lambda_i c(s_0 - s_i) + \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j c(s_i - s_j).
\end{aligned} \tag{6.6}$$

Die Varianz der Schätzung lässt sich damit durch die Kovarianzfunktionen $c(\cdot)$ berechnen. Ziel ist es, diese bezüglich der Gewichte $\lambda_i, i = 1 \ldots, n$ mit

$$\frac{\partial \sigma_E^2}{\partial \lambda_i} = 0 \tag{6.7}$$

zu minimieren. Für die optimalen Gewichte λ_i ist dann das *Simple Kriging system* (SK) mit

$$\underbrace{\begin{pmatrix} c(s_1 - s_1) & \cdots & c(s_1 - s_n) \\ \vdots & \ddots & \vdots \\ c(s_n - s_1) & \cdots & c(s_n - s_n) \end{pmatrix}}_{C} \underbrace{\begin{pmatrix} \lambda_1^{SK} \\ \vdots \\ \lambda_n^{SK} \end{pmatrix}}_{\lambda^{SK}} = \underbrace{\begin{pmatrix} c(s_0 - s_1) \\ \vdots \\ c(s_0 - s_n) \end{pmatrix}}_{c_0} \tag{6.8}$$

zu lösen. Dies lässt sich ebenfalls durch

$$\sum_{j=1}^{n} \lambda_j^{SK} c(s_i - s_j) = c(s_i - s_0) \quad \text{für } i = 1, \ldots, n \tag{6.9}$$

und für λ^{SK} nach Navratil (2006) mit

$$\lambda^{SK} = C^{-1} c_0. \tag{6.10}$$

darstellen. Die Matrix C stellt die Kovarianzmatrix dar, die die jeweilige Kovarianz zweier Lokationen enthält. In dem Vektor λ^{SK} sind die zugehörigen Gewichte λ_i^{SK} angegeben und der Vektor c_0 enthält die Kovarianzen zwischen der beobachteten Lokation s_0 und den bekannten Lokationen $s_1, \ldots, s_n$. Als zugehörige Krige-Varianz für das *Simple Kriging* resultiert abschließend

$$\begin{aligned}
\sigma_{SK}^2 &= c(s_0 - s_0) - 2\sum_{i=1}^{n} \lambda_i c(s_0 - s_i) + \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j c(s_i - s_j) \\
&= c(s_0 - s_0) - 2\sum_{i=1}^{n} \lambda_i c(s_0 - s_i) + \sum_{i=1}^{n} \lambda_i c(s_0 - s_i) + c(s_0 - s_0) \\
&= c(s_0 - s_0) - \sum_{i=1}^{n} \lambda_i c(s_0 - s_i),
\end{aligned} \tag{6.11}$$

indem Gleichung (6.9) in Gleichung (6.6) eingesetzt wird. Das *Simple Kriging* liefert im Vergleich zu anderen Kriging-Methoden nach Navratil (2006) die kleinste Varianz σ_{SK}^2. Dennoch wird in der Praxis in der Regel das *Ordinary Kriging* (siehe Abschnitt 6.2), welches die am häufigsten genutzte Kriging-Methode darstellt, verwendet, da selten der Erwartungswert bekannt ist.

Simple Kriging in R

Damit die Funktion `krige()` das *Simple Kriging* berechnet, muss der Funktion ein bekannter Erwartungswert `beta` übergeben werden. Da für die Höhen- und Bodendaten kein Wissen über den Erwartungswert vorliegt, wird der Mittelwert der ursprünglich aus GRASS eingelesenen Daten `elev` und `soils`, die jeweils aus knapp 300.000 Beobachtungen bestehen, verwendet.

```
> # Simple Kriging für die Höhendaten
> sk.elev <- krige(elevation.dem~1, elevsample, newdata=grid.elev,
+                  model=vfit.elev,  id="sk.elev",
+                  beta=mean(elev$elevation.dem, na.rm=TRUE))
[using simple kriging]
```

```
# Simple Kriging für die Bodendaten
> sk.soils <- krige(soils~1, soilssample, newdata=grid.soils,
+                   model=vfit.soils, id="sk.soils",
+                   beta=mean(soils$soils, na.rm=TRUE))
[using simple kriging]
```

Die Vorhersagen für die Höhen- und Bodendaten, basierend auf dem *Simple Kriging* sind in Abb. 6.1 und Abb. 6.3 dargestellt. In den jeweiligen Abb. 6.2 und Abb. 6.4 sind die zugehörigen Varianzen angegeben.

```
# Grafische Darstellung der Simple Kriging Prädiktion und der zugehörigen
# Varianzen für die Höhendaten
> spplot(sk.elev, zcol="sk.elev.pred", sp.layout=list(p.elev),
+        col.regions=terrain.colors(20), contour=TRUE, col="brown",
+        labels=TRUE, pretty=TRUE, colorkey=list(labels=list(cex=1.25)),
+        scales=list(draw=TRUE, cex=1.25))
> spplot(sk.elev, zcol="sk.elev.var", sp.layout=list(p.elev),
+        col.regions=heat.colors(100), contour=TRUE, col="brown",
+        colorkey=list(labels=list(cex=1.25)),
+        scales=list(draw=TRUE, cex=1.25))

# Grafische Darstellung der Simple Kriging Prädiktion und der zugehörigen
# Varianzen für die Bodendaten
> spplot(sk.soils, zcol="sk.soils.pred", sp.layout=list(p.soils),
+        col.regions=terrain.colors(20), contour=TRUE, col="brown",
+        labels=TRUE, pretty=TRUE, colorkey=list(labels=list(cex=1.25)),
+        scales=list(draw=TRUE, cex=1.25))
> spplot(sk.soils, zcol="sk.soils.var", sp.layout=list(p.soils),
+        col.regions=heat.colors(20), contour=TRUE, col="brown",
+        colorkey=list(labels=list(cex=1.25)),
+        scales=list(draw=TRUE, cex=1.25))
```

Abb. 6.4 zeigt dabei besonders deutlich, dass die Varianzen für die durch das *Simple Kriging* geschätzten Werte um die Lokationen aus dem Stichprobendatensatz herum gering sind und für Bereiche, bei denen die nächsten vorhandenen Stichprobenwerte weit entfernt sind, hoch sind. Abb. 6.2 zeigt nahezu konstante Varianzen für das gesamte geschätzte Gebiet. Dies liegt daran, dass für die Höhendaten ein feineres Rasternetz vorliegen müsste, um kleinere Varianzen um die Stichprobenpunkte zu erreichen.

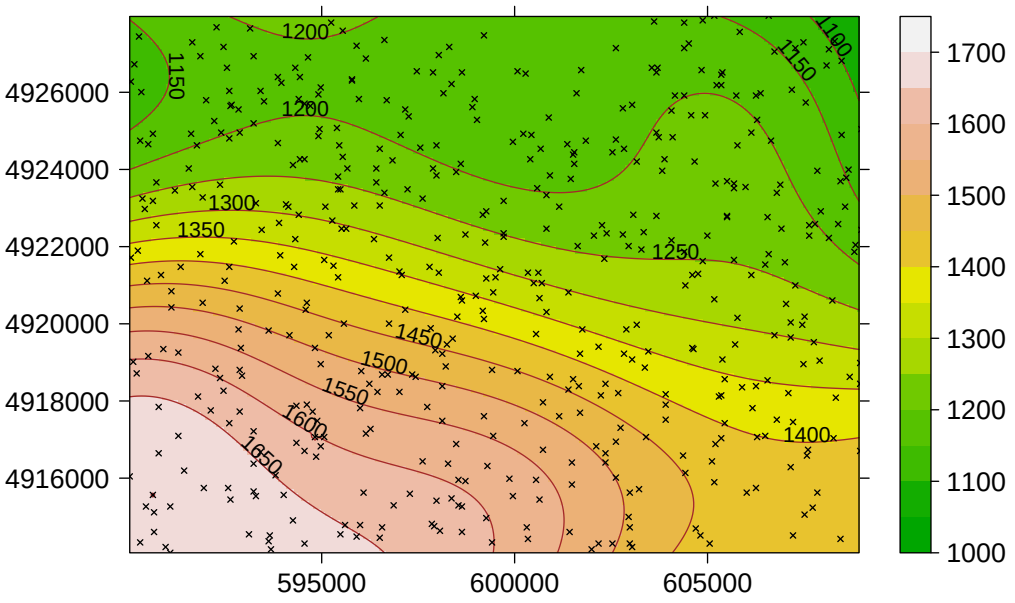


Abb. 6.1: *Simple Kriging* Prädiktion für die Höhendaten

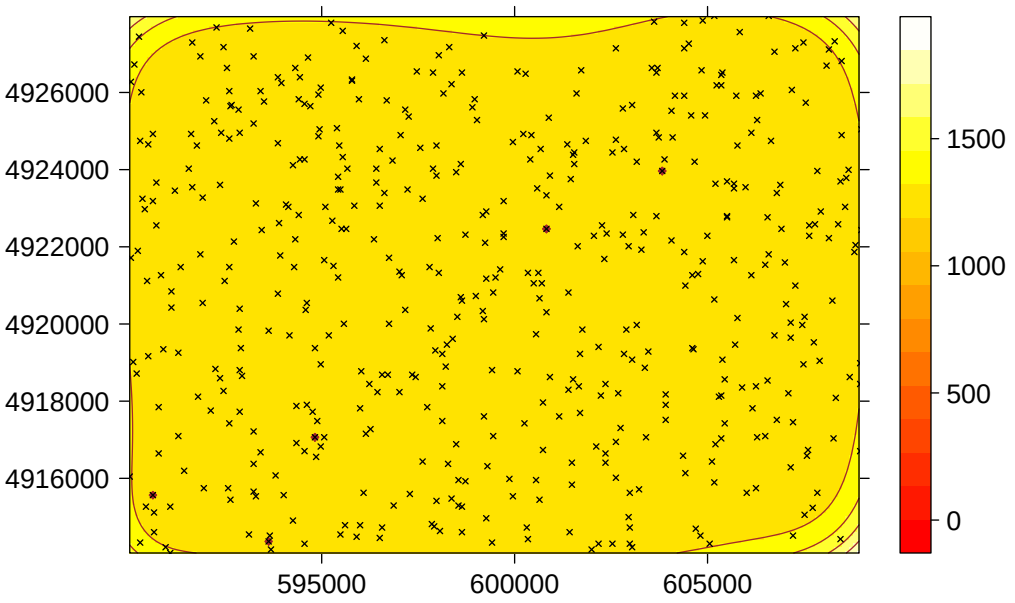


Abb. 6.2: *Simple Kriging* Varianz für die Höhendaten

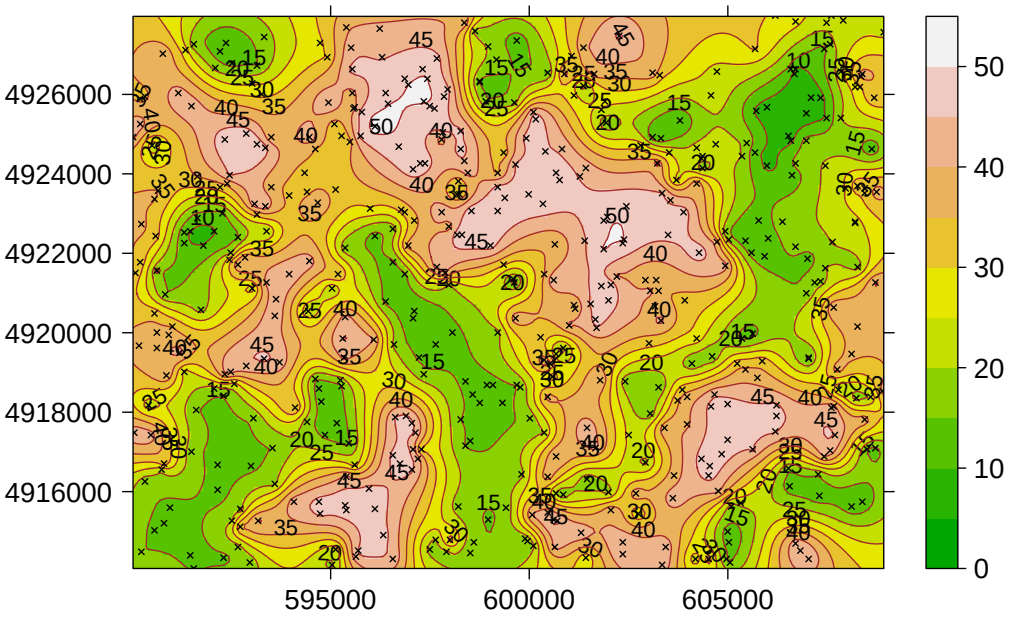


Abb. 6.3: *Simple Kriging* Prädiktion für die Bodendaten

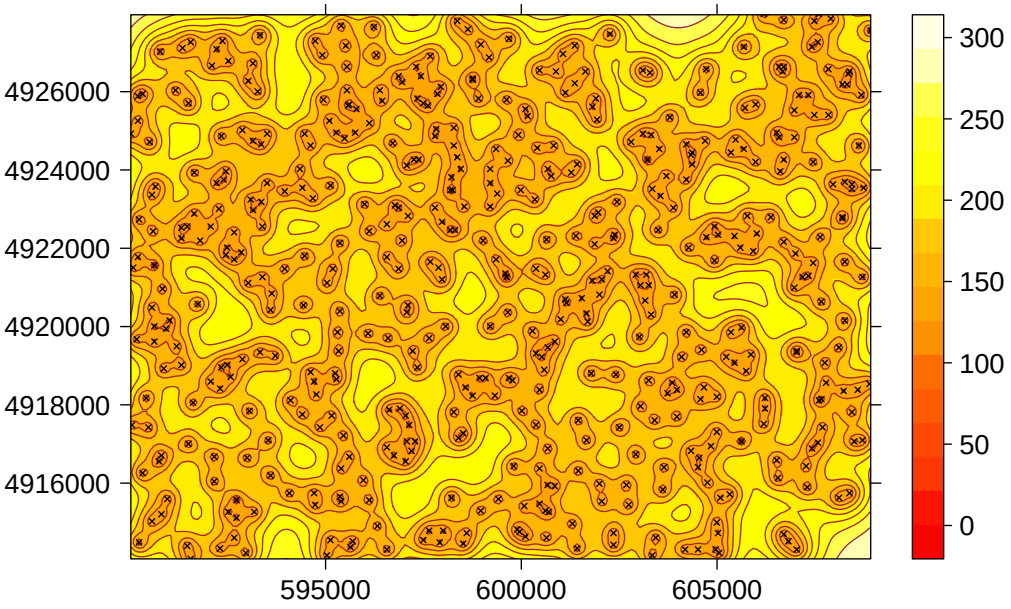


Abb. 6.4: *Simple Kriging* Varianz für die Bodendaten

6.2 Ordinary Kriging

Der Unterschied des *Ordinary Kriging* zum *Simple Kriging* liegt in dem Wissen über den Erwartungswert, welcher beim *Ordinary Kriging* als unbekannt, aber nach Navratil (2006) ebenfalls mit

$$\mathrm{E}(Y(s)) = \mu \tag{6.12}$$

als einheitlich und existent für alle Lokationen s angenommen wird. Der Krige-Schätzer für eine nicht beobachtete Lokation s_0 ergibt sich beim *Ordinary Kriging* durch eine Linearkombination von den gewichteten beobachteten Lokationen mit

$$\hat{Y}(s_0) = \sum_{i=1}^{n} \lambda_i Y(s). \tag{6.13}$$

Mit λ_i werden wieder die bisher unbekannten Gewichte der Lokationen dargestellt. Beim *Ordinary Kriging* müssen sie nach Navratil (2006) die Gleichung

$$\sum_{i=1}^{n} \lambda_i = 1 \tag{6.14}$$

erfüllen, damit der Schätzer $\hat{Y}(s_0)$ unverzerrt und damit erwartungstreu ist. Die Erwartungstreue kann mit

$$\begin{aligned} \mathrm{E}(Y(s_0) - \hat{Y}(s_0)) &= \mathrm{E}(Y(s_0)) - \mathrm{E}(\hat{Y}(s_0)) \qquad (6.15)\\ &= \mu - \mathrm{E}\left(\sum_{i=1}^{n} \lambda_i Y(s_i)\right) \\ &= \mu - \sum_{i=1}^{n} \lambda_i \mathrm{E}(Y(s_i)) \\ &= \mu - \sum_{i=1}^{n} \lambda_i \mu \\ &= 0. \qquad (6.16) \end{aligned}$$

gezeigt werden. Im Mittel besteht also auch hier kein Unterschied zwischen dem geschätzten Wert $\hat{Y}(s_0)$ und dem in den meisten Fällen unbekannten wahren Wert $Y(s_0)$. Die Varianz der Schätzung resultiert analog wie beim *Simple Kriging* als MSE, welcher sich mit

$$\begin{aligned} \sigma_E^2 &= \mathrm{E}\left((Y(s_0) - \hat{Y}(s_0))^2\right) \\ &= \mathrm{E}\left(\left(Y(s_0) - \sum_{i=1}^{n} \lambda_i Y(s_i)\right)^2\right) \end{aligned}$$

$$
\begin{aligned}
&= \mathrm{E}\left((Y(s_0))^2\right) - 2\sum_{i=1}^{n}\lambda_i \mathrm{E}(Y(s_0)Y(s_i)) + \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j \mathrm{E}(Y(s_i)Y(s_j)) \\
&= \mathrm{E}\left((Y(s_0))^2\right) - 2\sum_{i=1}^{n}\lambda_i \mathrm{E}(Y(s_0)Y(s_i)) + \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j \mathrm{E}(Y(s_i)Y(s_j)) \\
&\quad - \sum_{i=1}^{n}\lambda_i \mathrm{E}\left((Y(s_i))^2\right) + \sum_{i=1}^{n}\lambda_i \mathrm{E}\left((Y(s_i))^2\right) \\
&= \mathrm{E}\left((Y(s_0))^2\right) - 2\sum_{i=1}^{n}\lambda_i \mathrm{E}(Y(s_0)Y(s_i)) + \sum_{i=1}^{n}\lambda_i \mathrm{E}\left((Y(s_i))^2\right) \\
&\quad - \frac{1}{2}\sum_{i=1}^{n}\lambda_i \mathrm{E}\left((Y(s_i))^2\right) - \frac{1}{2}\sum_{i=1}^{n}\lambda_i \mathrm{E}\left((Y(s_i))^2\right) + \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j \mathrm{E}(Y(s_i)Y(s_j)) \\
&= \sum_{i=1}^{n}\lambda_i \mathrm{E}\left((Y(s_0)^2) - 2Y(s_0)Ys_i) + (Y(s_i))^2\right)^2 \\
&\quad - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j \mathrm{E}\left((Y(s_i)^2) - 2Y(s_i)Y(s_j) + (Y(s_j))^2\right) \\
&= \sum_{i=1}^{n}\lambda_i \mathrm{E}\left((Y(s_0) - Y(s_i))^2\right) - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j \mathrm{E}\left((Y(s_i) - Y(s_j))^2\right) \\
&= 2\sum_{i=1}^{n}\lambda_i\gamma(s_0 - s_i) - \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j\gamma(s_i - s_j) \qquad (6.17)
\end{aligned}
$$

als lediglich vom Variogramm $\gamma(\cdot)$ abhängig ergibt. Um die optimalen Gewichte λ_i für jede Lokation s_i bestimmen zu können, muss nun σ_E^2 bezüglich der λ_i minimiert werden. Nach Hattermann (2014) entsteht dabei das Problem eines überbestimmten Gleichungssystems, sodass mehr Gleichungen als zu bestimmende λ_i existieren. Lösen lässt sich dieses Problem nach Navratil (2006) mithilfe eines *Lagrange-Multiplikators* m, womit sich als resultierende zu minimierende Funktion

$$
2\sum_{i=1}^{n}\lambda_i\gamma(s_0 - s_i) - \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j\gamma(s_i - s_j) - 2m\left(\sum_{i=1}^{n}\lambda_i - 1\right) \qquad (6.18)
$$

ergibt. Differenziert man diese nach λ_i und m und setzt sie gleich Null, ist das Ergebnis das *Ordinary Kriging system* (OK)

$$
\underbrace{\begin{pmatrix} \gamma(s_1 - s_1) & \cdots & \gamma(s_1 - s_n) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \gamma(s_n - s_1) & \cdots & \gamma(s_n - s_n) & 1 \\ 1 & \cdots & 1 & 0 \end{pmatrix}}_{\Gamma}
\underbrace{\begin{pmatrix} \lambda_1^{OK} \\ \vdots \\ \lambda_n^{OK} \\ m^{OK} \end{pmatrix}}_{\lambda^{OK}}
= \underbrace{\begin{pmatrix} \gamma(s_0 - s_1) \\ \vdots \\ \gamma(s_0 - s_n) \\ 1 \end{pmatrix}}_{\gamma_0}. \qquad (6.19)
$$

Dieses lässt sich auch durch

$$\begin{cases} \sum_{j=1}^{n} \lambda_j^{OK} \gamma(s_i - s_j) + m^{OK} = \gamma(s_0 - s_i) & \text{für } i = 1, \ldots, n \\ \sum_{j=1}^{n} \lambda_j^{OK} = 1 \end{cases} \tag{6.20}$$

und für λ^{OK} nach Navratil (2006) mit

$$\lambda^{OK} = \Gamma^{-1} \gamma_0 \tag{6.21}$$

darstellen. Die symmetrische Matrix Γ enthält die jeweiligen Zusammenhänge zwischen zwei Lokationen, angegeben mit dem Semivariogrammwert $\gamma(\cdot)$, der Vektor λ^{OK} stellt die zugehörigen Gewichte sowie den Lagrange-Multiplikator m dar und der Vektor γ_0 enthält die resultierenden Zusammenhänge zwischen den beobachteten Lokationen $s_1, \ldots, s_n$ und der geschätzten Lokation s_0, wieder angegeben mit dem Semivariogrammwert $\gamma(\cdot)$. Die Varianz des Schätzers kann durch Einsetzen von Gleichung (6.20) in Gleichung (6.17) berechnet werden und wird auch als *Krige-Varianz* bezeichnet.

$$\begin{aligned} \sigma_{OK}^2 &= 2 \sum_{i=1}^{n} \lambda_i \gamma(s_0 - s_i) - \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j \gamma(s_i - s_j) \\ &= \sum_{i=1}^{n} \lambda_i \gamma(s_0 - s_i) - m \end{aligned} \tag{6.22}$$

Das *Ordinary Kriging* ist ein exakter Interpolator. Dies bedeutet, dass die neu berechneten geschätzten Werte für eine Lokation s_0 mit den wahren Werten an dieser Lokation übereinstimmen, wenn s_0 eine der beobachteten Lokationen s darstellt.

Ordinary Kriging in R

Die Berechnung des *Ordinary Kriging* kann in R ohne weitere Angaben mit der Funktion `krige()` erfolgen.

```
# Ordinary Kriging für die Höhendaten
> ok.elev <- krige(formula=elevation.dem~1, elevsample,
+                  model=vfit.elev, newdata=grid.elev,
+                  id="ok.elev")
[using ordinary kriging]

# Ordinary Kriging für die Bodendaten
> ok.soils <- krige(formula=soils~1, soilssample,
+                   newdata=grid.soils, model=vfit.soils,
+                   id="ok.soils")
[using ordinary kriging]
```

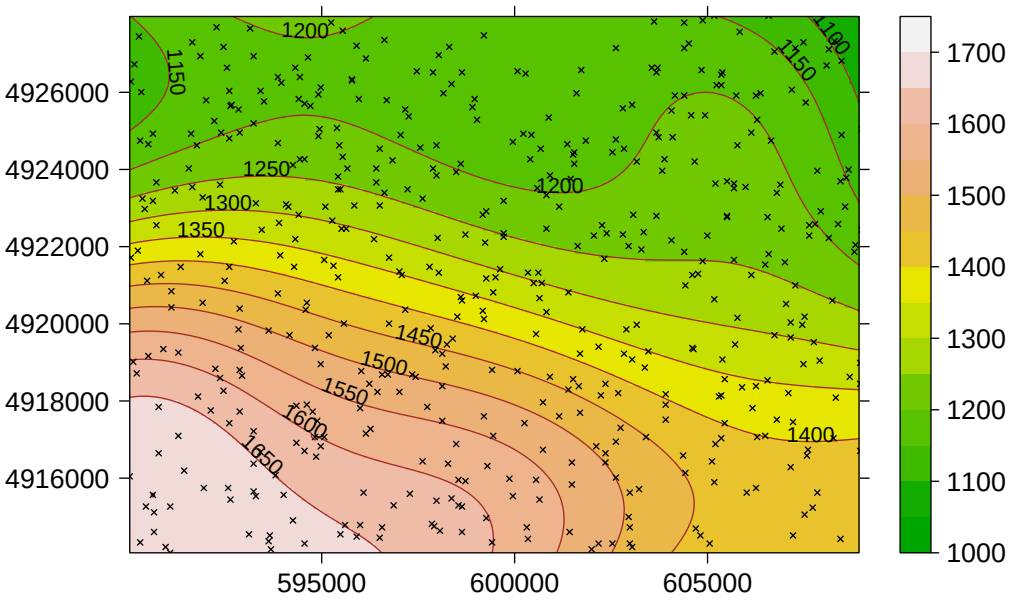


Abb. 6.5: *Ordinary Kriging* Prädiktion für die Höhendaten

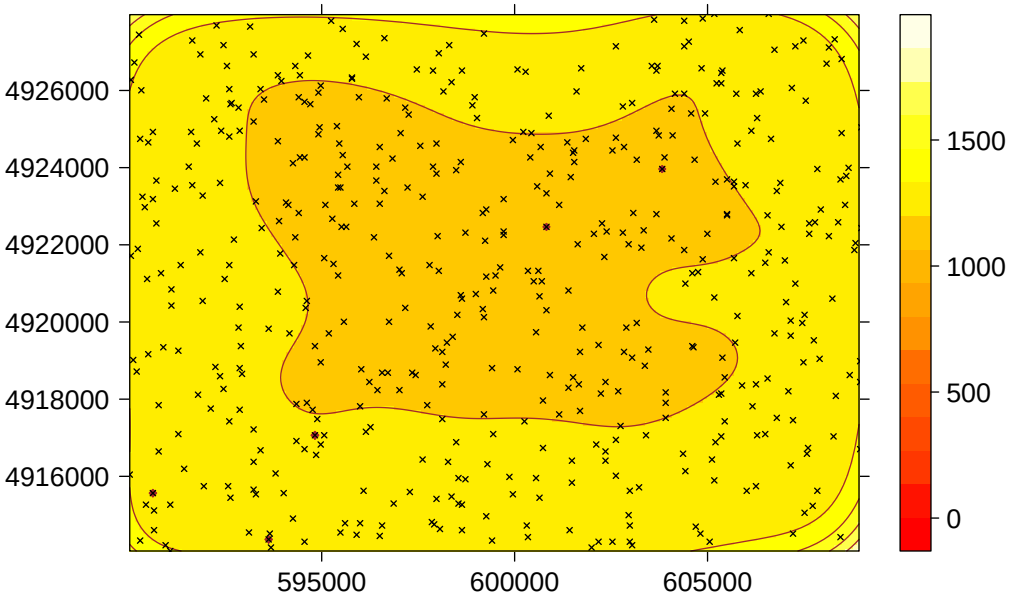


Abb. 6.6: *Ordinary Kriging* Varianz für die Höhendaten

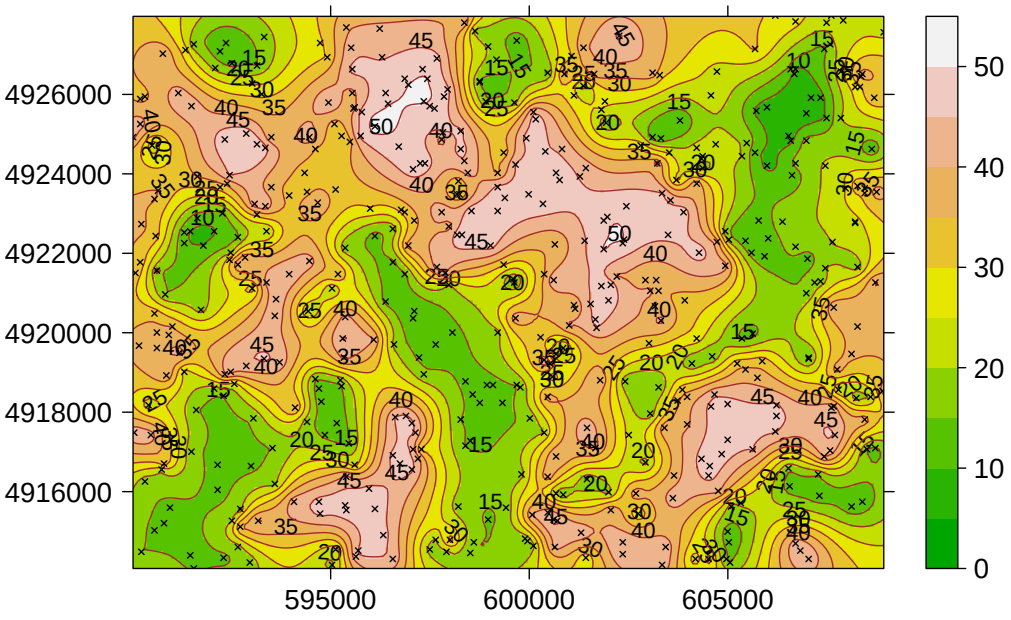


Abb. 6.7: *Ordinary Kriging* Prädiktion für die Bodendaten

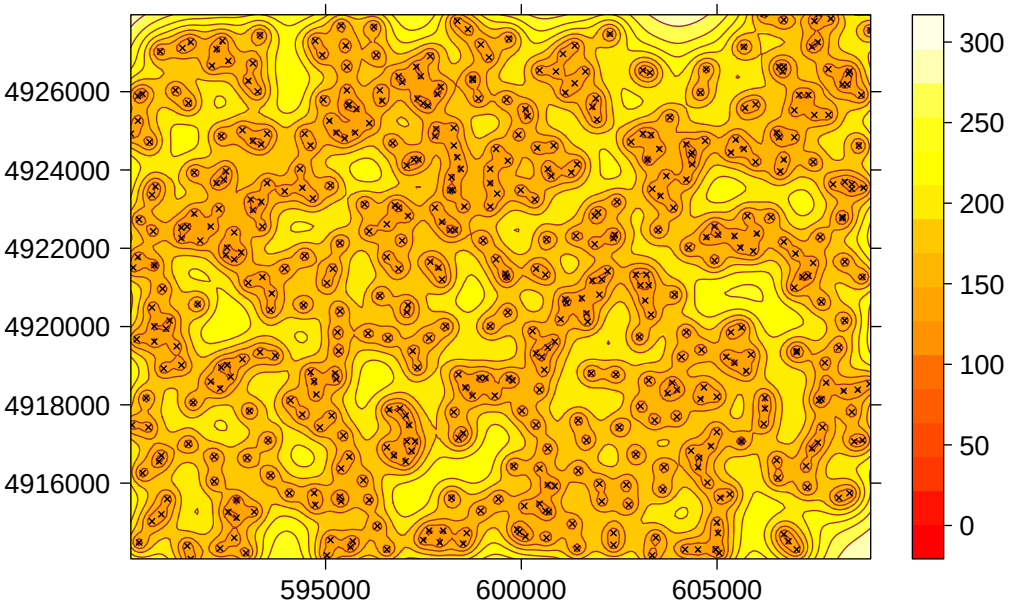


Abb. 6.8: *Ordinary Kriging* Varianz für die Bodendaten

Der R-Code für die grafische Darstellung der Prädiktion mit dem *Ordinary Kriging* ergibt sich analog zu dem für das *Simple Kriging* mit der Angabe von `ok.elev` und `ok.soils`, weshalb er im Folgenden nicht noch einmal aufgeführt ist. Grafisch dargestellt ist die Prädiktion per *Ordinary Kriging* für die Höhendaten in Abb. 6.5 und die zugehörigen Varianzen in Abb. 6.6 und für die Bodendaten in Abb. 6.7 und Abb. 6.8.

6.3 Universal Kriging

Beide bisher vorgestellten Kriging-Methoden haben vorausgesetzt, dass die Bedingung der schwachen Stationarität erfüllt ist und der Erwartungswert als konstant gegeben ist. Ist dies nicht der Fall, kann die Methode des *Universal Kriging* verwendet werden. Solch ein Fall resultiert beispielsweise, wie bereits in Abschnitt 5.7 erklärt, wenn ein Drift im Variogramm zu erkennen ist, weshalb das *Universal Kriging* nach Webster und Oliver (2007) auch als *Kriging with drift* bezeichnet wird. Um diesen bei der Prädiktion zu berücksichtigen, wird angenommen, dass sich der stochastische Prozess $Y(s)$ aus

$$Y(s) = \mu(s) + \delta(s) \tag{6.23}$$

zusammensetzt. $\mu(s)$ beschreibt einen Drift und $\delta(s)$ einen stochastischen Fehlerprozess. $\delta(s)$ ist dabei schwach stationär und nimmt im Mittel den Wert Null an, sodass sich als Erwartungswert für $Y(s)$,

$$\mathrm{E}(Y(s)) = \mu(s), \tag{6.24}$$

der Drift ergibt. Für den Drift selber wird angenommen, dass er sich als Linearkombination von deterministischen Funktionen $f_0, f_1, \ldots, f_L$,

$$\mu(s) = \sum_{l=0}^{L} a_l f_l(s), \tag{6.25}$$

mit $a_l \neq 0$ ergibt. Außerdem ist $f_0(s)$ mit

$$f_0(s) = 1 \tag{6.26}$$

als konstant definiert. Der Krige-Schätzer für das *Universal Kriging* ergibt sich vorerst äquivalent zu dem des *Ordinary Kriging* mit den Gewichten λ_i zu

$$\hat{Y}(s_0) = \sum_{i=1}^{n} \lambda_i Y(s_i). \tag{6.27}$$

Damit auch für diesen Schätzer bei nicht konstantem Erwartungswert die Erwartungstreue erfüllt ist, müssen nun die beiden folgenden Bedingungen gelten:

$$\mu(s_0) - \sum_{i=1}^{n} \lambda_i \mu(s_i) = 0 \tag{6.28}$$

und

$$\sum_{l=0}^{L} a_l \left(f_l(s_0) - \sum_{i=1}^{n} \lambda_i f_l(s_i) \right) = 0. \tag{6.29}$$

Letztere Gleichung impliziert zusammen mit der Annahme $a_l \neq 0$, dass

$$\sum_{i=1}^{n} \lambda_i f_l(s_i) = f_l(s_0) \tag{6.30}$$

gelten muss. Die Erwartungstreue ist dann durch alle drei Gleichungen mit

$$\begin{aligned}
\mathrm{E}(Y(s_0) - \hat{Y}(s_0)) &= \mathrm{E}(Y(s_0)) - \sum_{i=1}^{n} \lambda_i \mathrm{E}(Y(s_i)) \\
&= \mathrm{E}(\mu(s_0)) + \mathrm{E}(\delta(s_0)) - \sum_{i=1}^{n} \lambda_i \left(\mathrm{E}(\mu(s_i)) + \mathrm{E}(\delta(s_i)) \right) \\
&= \mu(s_0) - \sum_{i=1}^{n} \lambda_i \mu(s_i) \\
&= \sum_{l=0}^{L} a_l f_l(s_0) - \sum_{i=1}^{n} \lambda_i \sum_{l=0}^{L} a_l f_l(s_i) \\
&= \sum_{l=0}^{L} a_l \left(f_l(s_0) - \sum_{i=1}^{n} \lambda_i f_l(s_i) \right) \\
&= 0
\end{aligned} \tag{6.31}$$

gewährleistet. Die Varianz der Schätzung resultiert nach Navratil (2006) analog zu dem *Ordinary Kriging* durch

$$\begin{aligned}
\sigma_E^2 &= \mathrm{E}\left(\left(Y(s_0) - \sum_{i=1}^{n} \lambda_i Y(s_i) \right)^2 \right) \\
&= -\sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j \gamma(s_i - s_j) + 2 \sum_{i=1}^{n} \lambda_i \gamma(s_0 - s_i)
\end{aligned} \tag{6.32}$$

und mit der Annahme von L Lagrange-Multiplikatoren m als

$$-\sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j \gamma(s_i - s_j) + 2 \sum_{i=1}^{n} \lambda_i \gamma(s_0 - s_i) - 2 \sum_{j=1}^{p+1} m_{j-1} \left(\sum_{i=1}^{n} \lambda_i f_{j-1}(s_i) - f_{j-1}(s_0) \right). \tag{6.33}$$

Das *Universal Kriging system* (UK) ergibt sich dann mit

$$\underbrace{\begin{pmatrix} \gamma(s_1-s_1) & \cdots & \gamma(s_1-s_n) & 1 & f_1(s_1) & \cdots & f_L(s_1) \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma(s_n-s_1) & \cdots & \gamma(s_n-s_n) & 1 & f_1(s_n) & \cdots & f_L(s_n) \\ 1 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ f_1(s_1) & \cdots & f_L(s_1) & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ f_1(s_n) & \cdots & f_L(s_n) & 0 & 0 & \cdots & 0 \end{pmatrix}}_{\Gamma} \underbrace{\begin{pmatrix} \lambda_1^{UK} \\ \vdots \\ \lambda_n^{UK} \\ m_0^{UK} \\ m_1^{UK} \\ \vdots \\ m_L^{UK} \end{pmatrix}}_{\lambda^{UK}} = \underbrace{\begin{pmatrix} \gamma(s_0-s_1) \\ \vdots \\ \gamma(s_0-s_n) \\ 1 \\ f_1(s_0) \\ \vdots \\ f_L(s_0) \end{pmatrix}}_{\gamma_0}, \tag{6.34}$$

welches sich auch durch

$$\begin{cases} \sum_{j=1}^{n} \lambda_j^{UK} \gamma(s_i - s_j) - \sum_{l=0}^{L} m_l^{UK} f_l(s_i) = \gamma(s_0 - s_i) & \text{für } i = 1, \ldots, n \\ \sum_{j=1}^{n} \lambda_j^{UK} f_l(s_j) = f_l(s_0) & \text{für } l = 0, 1, \ldots, L \end{cases} \tag{6.35}$$

und für λ^{UK} nach Navratil (2006) als

$$\lambda^{UK} = \Gamma^{-1} \gamma_0 \tag{6.36}$$

darstellen lässt. Die Matrix Γ ist wieder symmetrisch. Sie enthält neben den Semivariogrammwerten für jede Lokationskombination die Funktionen $f_l, l = 0, 1, \ldots, L$ für alle Lokationen, wobei die Werte für $f_0(s_i)$ aufgrund der Eigenschaft als Konstante mit eins angegeben sind. Der Vektor λ^{UK} besteht neben den einzelnen Gewichten zusätzlich aus den L Lagrange-Multiplikatoren. Auf der rechten Seite des Gleichheitszeichens resultieren dann im Vektor γ_0 neben den Semivariogrammwerten für s_0 und die jeweilige Lokation s_i die Funktionen f_l. Diese Funktionen müssen linear unabhängig sein, damit eine Lösung des *Universal Kriging system* existiert.
Die Varianz des Schätzers lässt sich abschließend beim *Universal Kriging* durch Einsetzen der Gleichung (6.35) in Gleichung (6.32) mit

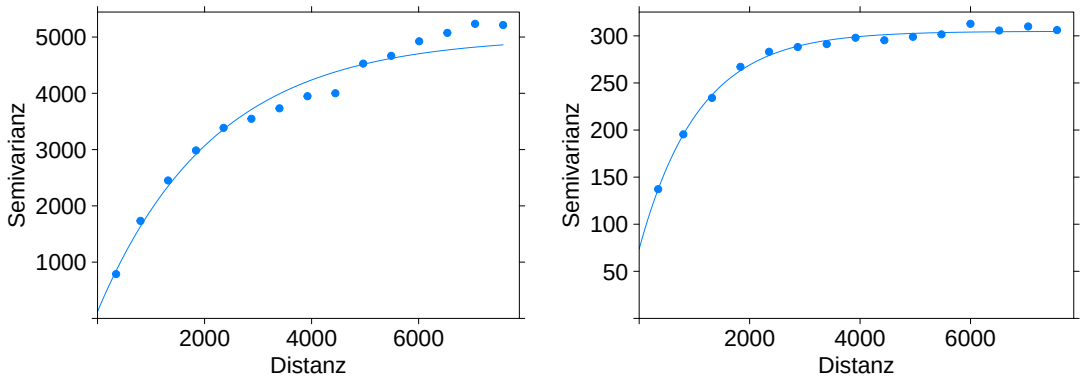
$$\begin{aligned} \sigma_{UK}^2 &= -\sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j \gamma(s_i - s_j) + 2 \sum_{i=1}^{n} \lambda_i \gamma(s_0 - s_i) \\ &= \sum_{i=1}^{n} \lambda_i \gamma(s_0 - s_i) + \sum_{l=0}^{L} m_l f_l(s_i) \end{aligned} \tag{6.37}$$

berechnen.

Universal Kriging in R

Für das *Universal Kriging* in R muss vorweg ein neues `gstat`-Objekt und für dieses eine angepasste Variogrammfunktion berechnet werden, da bei dem *Universal Kriging* der Erwartungswert nicht mehr als konstant angenommen wird, was bei den vorherigen `gstat`-Objekten durch das Übergeben von `~1` (siehe Abschnitt 5) impliziert wurde. Stattdessen wird ein linearer Zusammenhang mit anderen Einflussgrößen betrachtet. Beispielhaft wird für die Höhen- und Bodendaten der lineare Zusammenhang mit ihren Koordinaten untersucht, welcher in der bisherigen Variogrammberechnung nicht berücksichtigt wurde. Da dementsprechend auch neue Variogramme resultieren, werden auch neue Variogrammfunktionen angepasst, welche in Abb. 6.9 abgebildet sind. An dem angepassten Variogrammmodell für die Bodendaten hat sich dabei nichts verändert.

```
# Neues gstat-Objekt für die Höhendaten
> g.uk.elev <- gstat(id="uk.elev", formula=elevation.dem~s1+s2,
+                    data=elevsample)
> g.uk.elev
data: uk.elev : formula = elevation.dem‘~‘s1 + s2 ; data dim = 500 x 1

# Neue Anpassung einer Variogrammfunktion für die Höhendaten
> vemp.uk.elev <- variogram(object=g.uk.elev)
> vfit.uk.elev <- fit.variogram(vemp.uk.elev, vgm(5000,"Exp",6000,0))
> vfit.uk.elev
  model     psill    range
1   Nug  123.8418    0.000
2   Exp 4886.0038 2168.819

# Neues gstat-Objekt für die Bodendaten
> g.uk.soils <- gstat(id="uk.soils", formula=soils~s1+s2,
+                    data=soilssample)
> g.uk.soils
data: uk.soils : formula = soils‘~‘s1 + s2 ; data dim = 500 x 1

# Neue Anpassung einer Variogrammfunktion für die Bodendaten
> vemp.uk.soils <- variogram(object=g.uk.soils)
> vfit.uk.soils <- fit.variogram(vemp.uk.soils, vgm(280,"Exp",2000,0))
> vfit.uk.soils
  model     psill    range
1   Nug  73.67236    0.000
2   Exp 231.19847 1071.748
```



(a) Höhendaten (b) Bodendaten

Abb. 6.9: Angepasste Variogrammfunktionen für den linearen Zusammenhang mit den Koordinaten `s1+s2`

Auf Basis dieser neuen angepassten Funktionen kann dann das *Universal Kriging* berechnet werden, wobei dieselbe Formel, die zur Berechnung des empirischen Variogramms verwendet wird, genutzt werden muss.

```
# Universal Kriging für die Höhendaten
> uk.elev <- krige(formula=elevation.dem~s1+s2, elevsample,
+                  model=vfit.uk.elev, newdata=grid.elev,
+                  id="uk.elev")
[using universal kriging]

# Universal Kriging für die Bodendaten
> uk.soils <- krige(formula=soils~s1+s2, soilssample,
+                   model=vfit.uk.soils, newdata=grid.soils,
+                   id="uk.soils")
[using universal kriging]
```

Die resultierenden Prädiktionen und zugehörigen Varianzen für die Höhen- und Bodendaten sind in Abb. 6.10 und Abb. 6.11, sowie in Abb. 6.12 und Abb. 6.13 abgebildet. Die Grafiken ergeben sich in R wieder analog zu den Grafiken des *Simple Kriging* und *Ordinary Kriging* mit dem Unterschied, dass hier die *Universal Kriging* Schätzungen `uk.elev` und `uk.soils` verwendet werden.

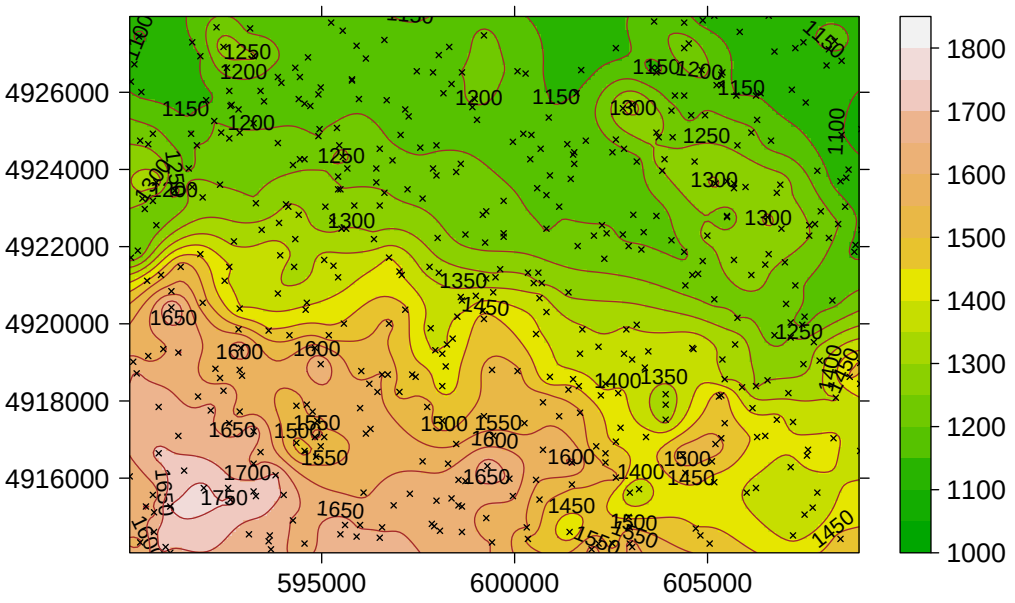


Abb. 6.10: *Universal Kriging* Prädiktion für die Höhendaten

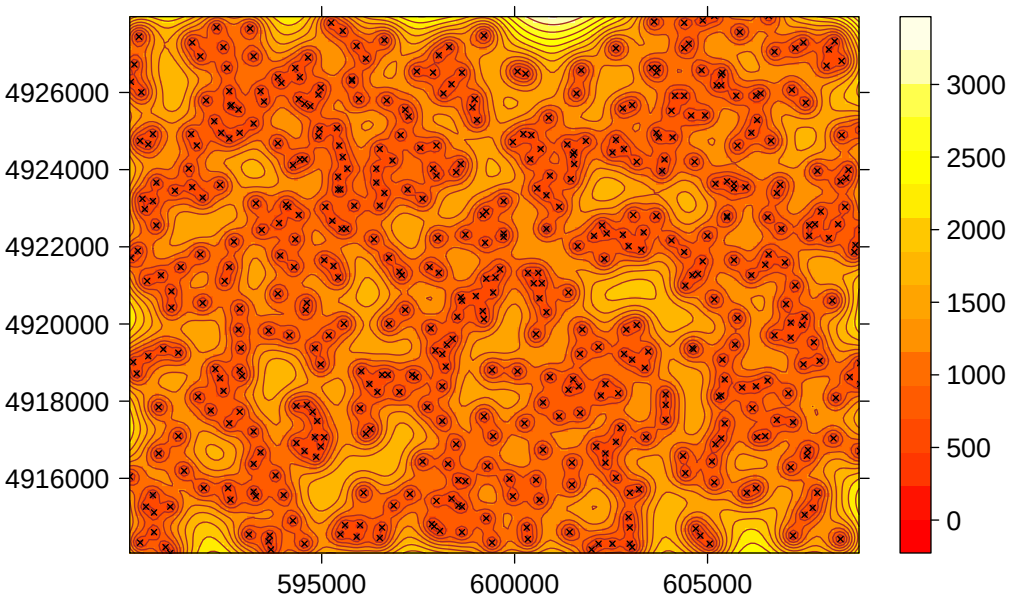


Abb. 6.11: *Universal Kriging* Varianz für die Höhendaten

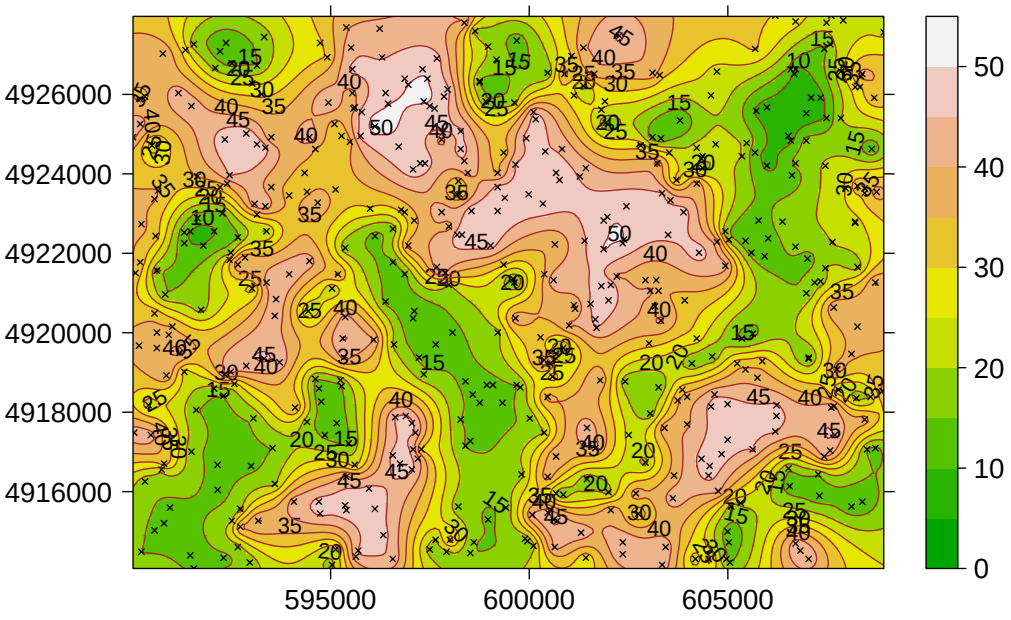


Abb. 6.12: *Universal Kriging* Prädiktion für die Bodendaten

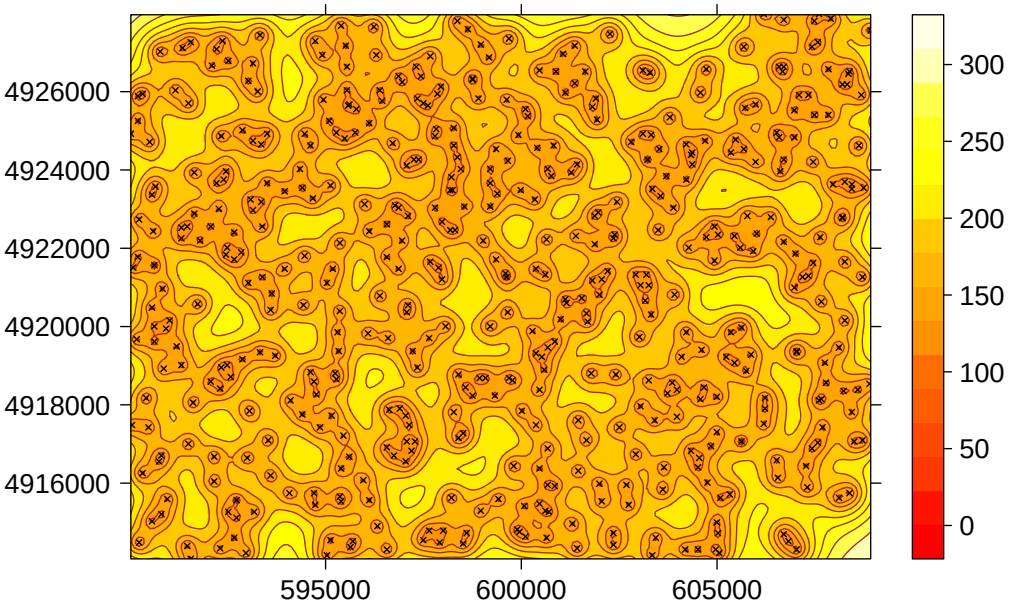


Abb. 6.13: *Universal Kriging* Varianz für die Bodendaten

7 Zusammenfassung und Ausblick

In der vorliegenden Arbeit wurde die Anbindung zwischen dem Geoinformationssystem *Geographical Analysis Support System* und der Statistik-Software R über das R-Paket `spgrass6` vorgestellt. Mithilfe von diesem lassen sich Daten aus GRASS in R einlesen. Das Paket ermöglicht sowohl das Bearbeiten der geostatistischen Daten durch das Anwenden von Funktionen aus R auf diese, als auch die Steuerung von GRASS-Funktionalitäten von R aus. Auf dieser Basis wurde die Anwendung geostatistischer Verfahren in R auf eingelesene Daten aus GRASS gezeigt. Beispielhaft wurden dazu die Höhendaten `elevation.dem` und die Bodendaten `soils` aus dem Datensatz *Spearfish Sample Database* verwendet und für das Umsetzen der geostatistischen Methoden das R-Paket `gstat` genutzt.

Basis der Anwendung geostatistischer Methoden ist die Annahme eines zugrunde liegenden stochastischen Prozesses $Y(s)$ für Lokationen s auf einem Untersuchungsgebiet D, dessen Eigenschaften von Interesse sind, um sein Verhalten für unbeobachtete Lokationen vorherzusagen.

Beschreiben lässt sich dieses Verhalten unter anderem mit explorativen Datenanalysemöglichkeiten, die einen ersten Überblick über die vorhandenen Daten und damit den stochastischen Prozess liefern. Für räumliche Daten bietet sich dabei die Darstellung als Streudiagramm der Koordinaten der Lokationen an, wobei die gemessenen Werte mit unterschiedlichen Farben oder Symbolgrößen hervorgehoben werden können. In R können für diese Modellierungen speziell die Funktionen `spplot()` und `bubble()` aus dem R-Paket `sp` verwendet werden.

Von Interesse ist in der Geostatistik in der Regel die räumliche Abhängigkeitsstruktur des stochastischen Prozesses $Y(s)$, welche mithilfe von Variogrammen beschrieben wird. Zur Darstellung des räumlichen Zusammenhangs können Variogrammwolken berechnet und empirische Variogramme erstellt werden, welche die Grundlage für die Anpassung einer Funktion für die räumliche Abhängigkeitsstruktur bieten. Zur Anpassung werden verschiedene Variogrammmodelle und Schätzmethoden für die Variogrammparameter zur Verfügung gestellt, wobei Spezialfälle, wie Anisotropie, ein Drift oder ein Hole-Effekt berücksichtigt werden müssen. Die Berechnung der Variogramme erfolgt in R mit der Funktion `variogram()`, wobei mit Angabe des Arguments `cloud=TRUE` die Variogrammwolke resultiert. Eine Funktion an das empirische Variogramm lässt sich dann mit `fit.variogram()` durch Übergabe eines Variogrammmodells und geschätzter Variogrammparameter in der Funktion `vgm()` anpassen, wobei mit dem Argument `fit.method` verschiedene Methoden der Parameterschätzung übergeben werden können.

Vorhersagen für unbeobachtete Lokationen lassen sich mithilfe von Kriging-Methoden treffen. Erläutert wurden die drei Kriging-Methoden *Simple Kriging*, *Ordinary Kriging* und *Universal Kriging*. Ziel dieser ist die Prädiktion auf Basis von Stichprobenwerten, die je nach räumlicher Abhängigkeitsstrukur von $Y(s)$ unterschiedlich gewichtet werden. Das *Simple Kriging* ergibt sich als Spezialfall des *Ordinary Kriging* und wird für Anwendungen, in denen der Erwartungswert des stochastischen Prozesses $Y(s)$ bekannt und konstant ist, genutzt. Das *Ordinary Kriging* kann dementsprechend in Anwendungen verwendet werden, in denen kein bekannter, aber ein konstanter Erwartungswert vorliegt. Diese Methode ergibt sich als Spezialfall des *Universal Kriging*, welches als Prädiktionsmethode genutzt wird, wenn ein Drift in den Daten vorliegt und damit kein konstanter Erwartungswert mehr gegeben ist. Für den Drift wird angenommen, dass dieser sich als Linearkombination von deterministischen Funktionen ergibt. Alle drei Methoden haben gemeinsam, dass sie zur Gruppe der BLUE-Schätzer (*Best Linear Unbiased Estimator*) gehören. In R lassen sich die Methoden mit der Funktion `krige()` berechnen. Bei Übergabe eines Erwartungswertes wird das *Simple Kriging* verwendet und bei Angabe von Einflussgrößen das *Universal Kriging*. Die Standardeinstellung ist damit das in Anwendungen am häufigsten verwendete *Ordinary Kriging*.

Ausblick

Aufgrund des begrenzten Umfangs beschränkt die Arbeit sich auf die Darstellung der notwendigen Funktionen aus dem R-Paket `spgrass6` zum Einlesen von GRASS-Daten in R und der grundlegenden Theorie geostatistischer Verfahren sowie die Anwendung der vorgestellten Methodik mit R.

Speziell wurde nicht auf die Anwendung von GRASS-Funktionalitäten mittels der Funktion `execGRASS()` aus `spgrass6` eingegangen. Allgemein könnte ein Vergleich der Berechnung von geostatistischen Methoden in R zu der Berechnung dieser mit GRASS von Interesse sein. Ebenfalls existieren nach Bivand et al. (2013) neben `gstat` beispielsweise mit den R-Paketen `spatial`, `RandomFields` und `geoR` weitere Möglichkeiten geostatistische Analysen in R durchzuführen. Alle Pakete können für die in dieser Arbeit vorgestellten Modellierungen ebenfalls verwendet werden. Einen Überblick und einige Vor- und Nachteile dieser R-Pakete findet sich in Bivand et al. (2013).

Bei der Darstellung der Variogramme wurde die Theorie der Parameterschätzung des theoretischen Variogrammes nur am Rande erwähnt. Für eine ausführliche Darstellung sei auf Diggle und Ribeiro (2007) verwiesen. Auch für die in Abschnitt 5.6 und 5.7 vorgestellten Eigenschaften, Anisotropie, Drift, Hole-Effekt und Unbeschränktheit, die in Daten vorliegen können, existiert weitaus mehr Hintergrund. Tiefer gehende Beschreibungen dieser Eigenschaften sowie Lösungen für diese werden in Webster und Oliver (2007) vorgestellt.

Als Basis für die Prädiktion wurden drei Kriging-Methoden vorgestellt, die bereits für viele Anwendungen gute Ergebnisse liefern. Neben diesen existieren nach Webster und Oliver (2007) weitere Methoden wie *Lognormal Kriging*, *Factorial Kriging*, *Indicator Kriging* oder *Disjunctive Kriging* für spezielle Datensituationen oder nach Wackernagel (1995) das *Block Kriging* wenn nicht eine konkrete Lokation, sondern ein Teilgebiet $v_0 \in D$ vorhergesagt werden soll.

Allgemein wird in der Arbeit ein Überblick über die Theorie geostatistischer Verfahren und ihrer Anwendung mit R gegeben. Für einen tieferen Einblick sei für die Theorie Webster und Oliver (2007) und für die Anwendung Bivand et al. (2013) empfohlen.

Literatur

Bivand, R. (2005). Interfacing GRASS 6 and R, *GRASS Newsletter* **3**: 11–16. **URL:** `http://grass.osgeo.org/newsletter/GRASSNews_vol3.pdf` [Letzter Zugriff: 13. Juni 2014].

Bivand, R. (2007). Using the R–GRASS Interface: Current Status, *OSGeo Journal* **1**: 36–38. **URL:** `http://www.osgeo.org/files/journal/final\_pdfs/OSGeo\_vol1\_GRASS-R.pdf` [Letzter Zugriff: 13. Juni 2014].

Bivand, R. (2013). *spgrass6: Interface between GRASS 6 and R.* Version: 0.8-3; **URL:** `http://CRAN.R-project.org/package=spgrass6` [Letzter Zugriff: 13. Juni 2014].

Bivand, R. S., Pebesma, E. und Gomez-Rubio, V. (2013). *Applied spatial data analysis with R, Second edition*, Springer, New York.

Diggle, P. J. und Ribeiro, P. J. (2007). *Model-based Geostatistics*, Springer, New York.

Fahrmeir, L., Kneib, T. und Lang, S. (2009). *Regression*, Statistik and ihre Anwendungen, Springer-Verlag Berlin Heidelberg.

Fahrmeir, L., Künstler, R., Pigeot, I. und Tutz, G. (2011). *Statistik: Der Weg zur Datenanalyse*, Springer-Verlag Berlin Heidelberg.

GRASS Development Team (1993). *Spearfish Sample Database.* **URL:** `http://grass.osgeo.org/uploads/grass/sampledata/spearDB.pdf` [Letzter Zugriff: 04. Juni 2014].

GRASS Development Team (2012). *Geographic Resources Analysis Support System (GRASS GIS) Software*, Open Source Geospatial Foundation, USA. Version 6.4.3; **URL:** `http://grass.osgeo.org` [Letzter Zugriff: 04. Juni 2014].

Hattermann, F. (2014). *Einführung in die Geostatistik (4), (5) und (6)*, Vorlesungsfolien, Potsdam Institute for Climate Impact Research. **URL:** `http://www.pik-potsdam.de/~fred/geostatistik/` [Letzter Zugriff: 07. Juni 2014].

Ligges, U. (2007). *Programmieren mit R*, Statistik und ihre Anwendungen, Springer-Verlag Berlin Heidelberg.

Navratil, G. (2006). *Ausgleichsrechnung II*, Vorlesungsskript, Institut für Geoinformation, Technische Universität Wien. **URL:** `ftp://ftp.geoinfo.tuwien.ac.at/navratil/Ausgleich2.pdf` [Letzter Zugriff: 04. Juni 2014].

Neteler, M. (2003). *GRASS-Handbuch: Der praktische Leitfaden zum Geographischen Informationssystem GRASS*, GDF Hannover, Gesellschaft für Datenanalyse und Fernerkundung. **URL:** `http://grass.osgeo.org/gdp/handbuch/neteler_grasshandbuch_v12.pdf` [Letzter Zugriff: 13. Juni 2014].

Neteler, M. und Mitasova, H. (2008). *Open Source GIS: A GRASS GIS Approach Third Edition*, Springer, New York.

Pebesma, E. J. (2004). Multivariable geostatistics in s: the gstat package, *Computers & Geosciences* **30**: 683–691. Version: 1.0-19.

Pebesma, E. J. (2014a). *gstat user's manual*. **URL:** `http://www.gstat.org/gstat.pdf` [Letzter Zugriff: 04. Juni 2014].

Pebesma, E. J. (2014b). *The meuse data set: a brief tutorial for the* `gstat` *R package*. **URL:** `http://cran.r-project.org/web/packages/gstat/vignettes/gstat.pdf` [Letzter Zugriff: 13. Juni 2014].

R Core Team (2014). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. Version: 3.1.0 beta; **URL:** `http://www.R-project.org` [Letzter Zugriff: 13. Juni 2014].

Schmid, V. und Feilke, M. (2012). *Räumliche Statistik*, Vorlesungsskript, LMU Institut für Statistik. **URL:** `http://bioimg.userweb.mwn.de/lehre/rs2012/script/rs.pdf` [Letzter Zugriff: 04. Juni 2014].

Wackernagel, H. (1995). *Multivariate Geostatistics*, Springer-Verlag Berlin Heidelberg.

Webster, R. und Oliver, M. A. (2007). *Geostatistics for Environmental Scientists*, Statistics in Practice, John Wiley & Sons, Ltd., England.

Eidesstaatliche Erklärung

Ich versichere hiermit, dass ich die vorliegende Bachelorarbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Die Arbeit war in dieser oder ähnlicher Form noch nicht Bestandteil einer Prüfungsleistung.

München, den 16. Juni 2014

_____

Hannah Otterbach