INSTITUT FÜR STATISTIK
DER LUDWIG–MAXIMILIANS–UNIVERSITÄT MÜNCHEN

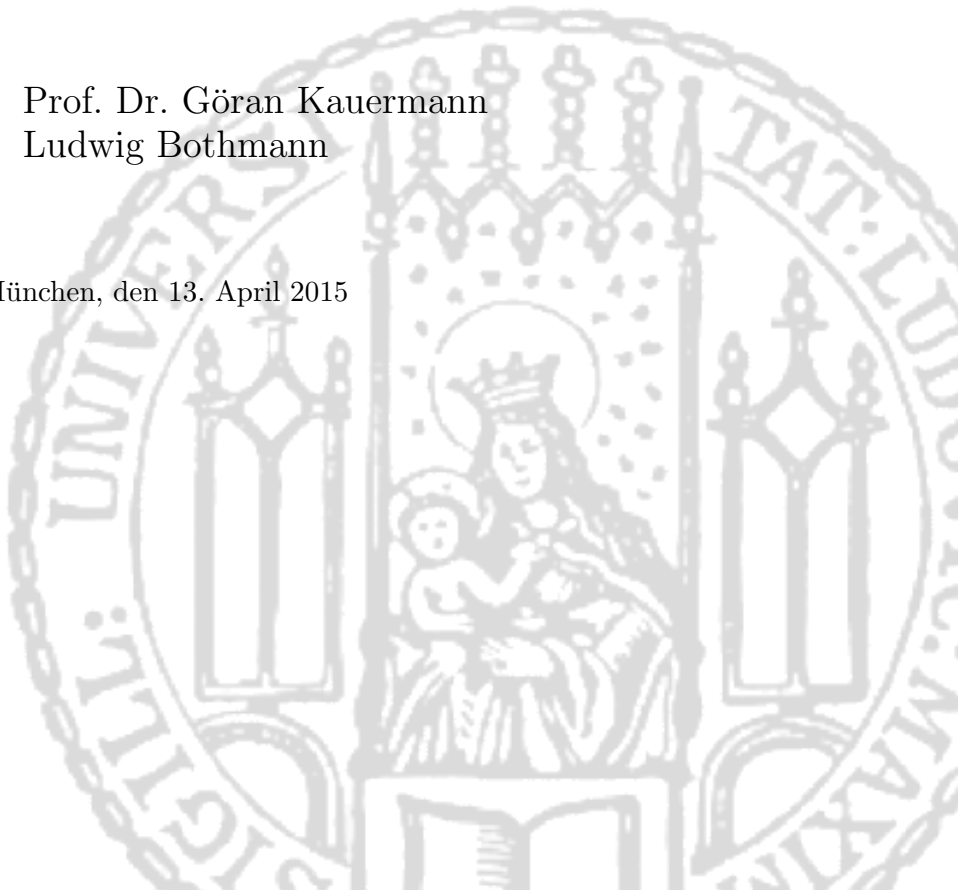
Masterarbeit

Räumliche Modelle
(Spatial Models)

Christine Jula

Betreuer: Prof. Dr. Göran Kauermann
Ludwig Bothmann

München, den 13. April 2015



Abstract

Grundlage vieler statistischer Modelle ist die Annahme unabhängiger Beobachtungen (gegeben den Kovariablen). Es wird davon ausgegangen, dass Beobachtungen unter identischen Bedingungen und unabhängig voneinander gewonnen werden und die Daten somit eine Zufallsstichprobe bilden. Eine häufige Charakteristik räumlicher Daten ist jedoch die Tatsache, dass sich räumlich nahe Beobachtungen ähnlicher sind als weit entfernte. Eine Folgerung daraus ist, dass diese Daten nicht dem Paradigma der Unabhängigkeit folgen und somit einer Modellierung von Abhängigkeiten bedürfen.

Die vorliegende Arbeit gibt einen Überblick über die Theorie zur Modellierung zweier räumlicher Datentypen: geostatistische Daten und Gitterdaten. Bei geostatistischen Daten liegt die räumliche Information stetig, in Form von Koordinaten vor, bei Gitterdaten ist sie auf eine abzählbare Menge an Regionen aggregiert. Es wird in beiden Fällen von einem zugrundeliegenden datengenerierenden stochastischen Prozess ausgegangen.

Die gebräuchlichste Methode der Geostatistik ist das *Kriging*. Mit Hilfe des sogenannten Variogramms werden hier räumliche Abhängigkeiten ausgedrückt und so die Schätzung von Werten an unbeobachteten Orten ermöglicht.

Gitterdaten hingegen werden mittels Markov-Zufallsfeldern modelliert. Die Ähnlichkeit bzw. Abhängigkeit zweier Regionen wird in diesem Fall über deren Nachbarschaftsverhältnis berücksichtigt.

Beide Ansätze lassen sich in die Theorie der Geoadditiven Modelle einbetten. Das Programmpaket *BayesX* (Umlauf et al. 2015) stellt die nötigen Funktionen zur Verfügung um eine Auswertung dieser Modelle in R (R Core Team 2014) durchzuführen.

Es wurden beispielhaft zwei Datensätze mit Hilfe der besprochenen Methoden analysiert. Bei der Auswertung von relativen Grünwerten aus Webcam-Bildern zur Bestimmung phänologischer Phasen konnte kein strukturierter räumlicher Effekt festgestellt werden. Es wurden deshalb unabhängige zufällige Effekte zur Modellierung verwendet. Möglicherweise könnte das Modell durch die Aufnahme weiterer Kovariablen (Regenfallmenge, Temperatur, Höhe des Standortes etc.) weiter verbessert werden.

Bei der Analyse der Herkunft von Studienanfängern an der LMU wurde ein starker regionaler Bezug bei der Studienortswahl festgestellt. Die Anzahl der Studienanfänger in den Kreisen nahm mit der Entfernung zur Hochschule stetig ab. Beim räumlichen Effekt zeigte sich ein erhöhter Erwartungswert in den südöstlichen Gebieten Bayerns. Außerdem scheint die Anziehungskraft der eigenen Hochschule in Landkreisen mit Universitätsstadt größer zu sein, als die der LMU.

Inhaltsverzeichnis

1. Einführung

> Everything is related to everything else, but near things are more related than distant things.
>
> *(W. Tobler (1970):* First law of geography*)*

Wir werden täglich mit räumlichen und räumlich-zeitlichen Daten konfrontiert. Dies geschieht im normalen Alltag im Fernsehen, in Zeitungen oder immer mehr auch auf mobilen Geräten in Form von Wetterkarten oder ähnlichem. Noch vor einigen Jahren war die handelsübliche Papierkarte das Mittel der Wahl um Standorte festzustellen. In der heutigen Zeit verfügt so gut wie jeder über einen GPS-Empfänger in Smartphone oder Tablet mit einer Ortsgenauigkeit von weniger als 10 Metern. Auch durch solche technischen Entwicklungen rücken räumliche Daten immer weiter in den Fokus des allgemeinen Interesses. Aber auch Statistiker müssen sich mit diesem Thema auseinandersetzen. Jede Beobachtung beschreibt eine Eigenschaft an einem bestimmten Ort im Raum zu einem bestimmten Moment der Zeit. Diese Tatsache wird in vielen Analysen außer Acht gelassen, da Ort und Zeitpunkt als nicht relevant betrachtet werden (Bivand et al. 2013). Diese Arbeit beschäftigt sich mit dem Fall, dass die räumliche Komponente von speziellem Interesse der Analyse ist.
Ziel der Analyse räumlicher Daten ist die Inferenz über Parameter, die den zugrundeliegenden datengenerierenden Prozess erklären bzw. die Vorhersage von Werten an unbeobacheten Orten im Raum (Interpolation).
Grundlage vieler statistischen Modelle ist die Annahme unabhängiger Beobachtungen (gegeben den Kovariablen). Es wird davon ausgegangen, dass Beobachtungen unter identischen Bedingungen und unabhängig voneinander gewonnen werden und die Daten somit eine Zufallsstichprobe bilden. Eine häufige Charakteristik räumlicher Daten ist jedoch die Tatsache, dass sich räumlich nahe Beobachtungen ähnlicher sind als weit entfernte. Eine Folgerung daraus ist, dass diese Daten nicht dem Paradigma der Unabhängigkeit folgen und somit einer Modellierung von Abhängigkeiten bedürfen. Diese Erkenntnis beschrieb Tobler (1970) als das "erste Gesetz der Geographie".
Die folgende Arbeit ist in drei thematische Abschnitte gegliedert. Zunächst wird in Kapitel 2 ein allgemeines Modell räumlicher stochastischer Prozesse aufgestellt. Dieses wird in den Kapiteln 4 und 5 auf die Spezialfälle der Geostatistik und der Gitterdaten bzw. dem Disease-Mapping (Kapitel 6) heruntergebrochen. Diese Kapitel stellen die jeweilige Theorie zur Modellierung der Datentypen vor. Eine Auswertung zu den einführenden Beispielen aus Kapitel 3 auf Basis der

vorgestellten Theorie ist in Kapitel 8 zu finden. In Kapitel 7 werden die zuvor beschriebenen Methoden in das Grundgerüst der *Geoadditiven Modelle* eingebaut. Außerdem wird in diesem Zusammenhang das Programmpaket BayesX (Umlauf et al. 2015) vorgestellt. Alle statistischen Analysen, die dieser Arbeit zugrunde liegen, wurden mit der Software R (R Core Team 2014) durchgeführt. Die Shapefiles zur Erstellung der Karten stammen aus der *GADM database of Global Administrative Areas* (2004) bzw. vom *Bundesamt für Kartographie und Geodäsie* (2011).

2. Räumliche stochastische Prozesse

Das folgende Kapitel motiviert ein allgemeines Modell für räumliche Daten. Hierbei wurde sich in Notation und Aufbau an Cressie (1993) gehalten. Um eine möglichst große Menge an Problemen mit diesem Modell angehen zu können, wird es innerhalb dieses Kapitels zunächst sehr allgemein und somit flexibel gehalten. Die zugrundeliegenden Daten können stetig oder diskret sein, punktuell oder räumlich aggregiert vorliegen und ihre Positionen können regulär, aber auch irregulär sein. Die notwendigen Einschränkungen für die einzelnen Datentypen werden dann in den nachfolgenden Kapiteln genauer besprochen.
Sei $s \in \mathbb{R}^d$ eine Lokation im d-dimensionalen euklidischen Raum und sei $Z(s)$ eine Zufallsvariable. Dann beschreibt

$$\{Z(s) : s \in D\} \tag{2.1}$$

ein multivariates Zufallsfeld (bzw. Zufallsprozess) mit der Indexmenge $D \subset \mathbb{R}^d$. Eine Realisation dieses Zufallsfeldes wird mit $\{z(s) : s \in D\}$ gekennzeichnet.
Cressie (1993) nimmt D dabei als zufällig an, um mit Hilfe dieses Modells auch Punktprozesse beschreiben zu können. Da diese nicht Teil dieser Arbeit sind, wird D hier als fest angenommen. Kapitel 4 und 5 beschäftigen sich mit folgenden Spezialfällen von (2.1) (vgl. Cressie (1993)):

- Kapitel 4: *Geostatistische Daten.* D ist eine feste Teilmenge von $\mathbb{R}^d$, welche ein d-dimensionales Rechteck positiven Volumens enthält; $Z(s)$ ist eine Zufallsvariable an der Stelle $s \in D$.
- Kapitel 5: *Gitterdaten.* D ist eine feste (reguläre oder irreguläre) Menge abzählbarer Punkte aus $\mathbb{R}^d$; $Z(s)$ ist eine Zufallsvariable an der Stelle $s \in D$.

Die in dieser Arbeit verwendeten Methoden sind auch auf den univariaten Fall der Zeitreihen anwendbar. Zeitreihen unterliegen generell der gleichen Theorie wie Räumliche Prozesse (in $\mathbb{R}^1$). Um diese jedoch abgrenzen zu können, wird in diesem Fall meist der Index t verwendet, sodass (2.1) umgeschrieben wird in

$$\{Z(t) : -\infty < t < \infty\}. \tag{2.2}$$

Dieser Fall ist in dieser Arbeit aber nicht von speziellem Interesse, sodass sich auf die Definition in (2.1) beschränkt werden kann.

3. Datentypen und einführende Beispiele

In der Literatur über räumliche Daten wird generell zwischen drei verschiedenen Datentypen unterschieden:

- Geostatistische Daten
- Gitter- bzw. Pixeldaten
- Punktprozesse

Diese werden nun einzeln genauer vorgestellt und anhand von Datenbeispielen erläutert. Punktprozesse sind nicht Teil dieser Arbeit und werden deshalb hier nur am Rande betrachtet.
In allen Fällen wird als Grundlage der Daten, wie in Kapitel 2 besprochen, ein Zufallsprozess

$$\{Z(s) : s \in D\}$$

angenommen.

3.1. Geostatistische Daten

Im Fall geostatistischer Daten variiert s stetig im d-dimensionalen Euklidischen Raum innerhalb der Indexmenge D (Region). In den meisten Anwendungen wird dies auf $\mathbb{R}^2$ und $\mathbb{R}^3$ eingeschränkt. Die Lokationen s bestehen dann aus stetigen x- und y-, bzw. x-, y- und z-Koordinaten, also:

$$\boldsymbol{s} = (s_x, s_y)^T \in \mathbb{R}^2 \quad \text{bzw.} \quad \boldsymbol{s} = (s_x, s_y, s_z)^T \in \mathbb{R}^3$$

mit den zugehörigen Daten $z(\boldsymbol{s}_1), \ldots, z(\boldsymbol{s}_n)$, an n vorgegebenen Lokationen $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$.

Beispiel: Phänologie

Ein wichtiges Thema unserer Zeit sind die Auswirkungen des Klimawandels. Ein wichtiger Indikator hierfür stellt die Phänologie, also die Studie im Jahresablauf periodisch wiederkehrender Naturereignisse dar. Es werden hier Eintrittszeiten biologischer Prozesse festgehalten, wie z.B. das erste Blühen von Pflanzen. Veränderungen im Zeitablauf können auf den Klimawandel zurückgeführt werden.

Eine beliebte Methode zur Bestimmung phänologischer Phasen stellt die Messung von Grünwerten auf Basis von Webcam-Bildern dar. Ansteigende Temperaturen und veränderte Lichtverhältnisse im Frühling geben das Startsignal zum Ergrünen der Vegetation. Dieser Anstieg spiegelt sich im Grünwert aufgenommener Webcam-Bilder wider.
Dhital (2011) sammelte Webcam-Bilder von 500 verschiedenen Stationen in Deutschland mit Vorliegen von Vegetation. Auf Grund von Qualitätsmängeln wurden nur 182 für die weitere Analyse ausgewählt. Es wurden vom 25. März bis zum 8. Juni 2011 täglich Bilder gespeichert und die Grünwerte extrahiert. Hierfür wurden für jede Station sogenannte ROIs (engl. "regions of interest") ausgewählt und eine Maske über das Bild gelegt (vgl. Abbildung 3.1).

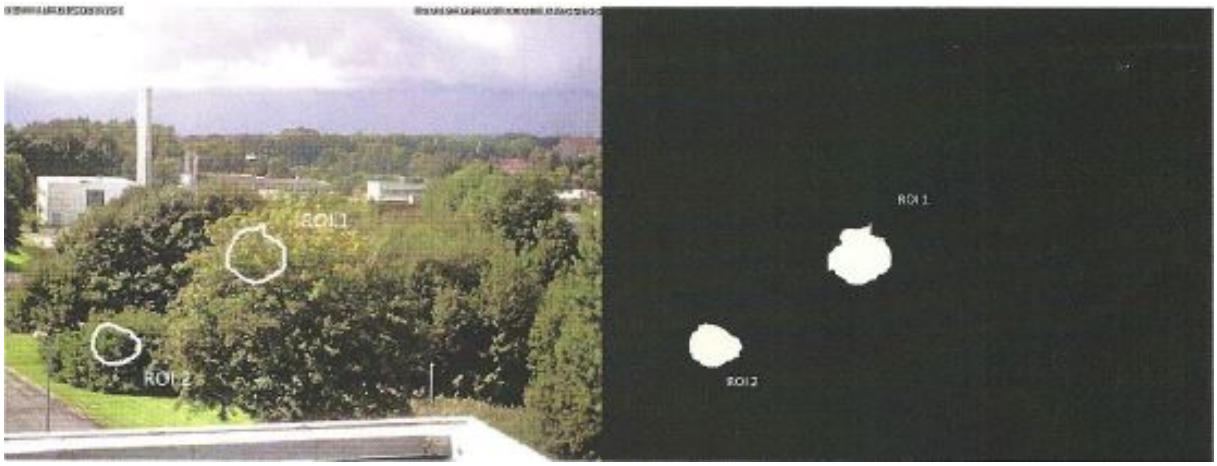


Abbildung 3.1.: Beispiel analysierter ROIs für die Webcam Clausthal-Zellerfeld am DOY 247; Quelle: Dhital (2011).

Das Zeitfenster entspricht den Tagen (DOY="day of year") 84 bis 159 des Jahres 2011. Das genaue Vorgehen kann in der Arbeit (Dhital 2011) nachgelesen werden.
Die entstandenen Daten wurden zur weiteren Analyse für diese Arbeit zur Verfügung gestellt. Für jede Station liegen die Koordinaten der Kamera und die gemessenen relativen Grünwerte vor. Der relative Grünwert berechnet sich aus dem Anteil des Grünwertes an der Summe der Rot-/Grün- und Blauwerte des jeweiligen Bildes, d.h.

$$\mathrm{relG} = \frac{\mathrm{G}}{\mathrm{G} + \mathrm{R} + \mathrm{B}}.$$

Abbildung 3.2 zeigt die gemessenen relativen Grünwerte an den Stationen beispielhaft für DOY 84. Abbildung 3.3 zeigt die aggregierten Daten über die Zeit. Es lässt sich ein Anstieg des relativen Grünwertes bis etwa zum DOY 120 erkennen. Die Tage 141 bis 143 fehlen aufgrund eines technischen Problems mit dem Server, auf dem die Bilder gespeichert wurden.

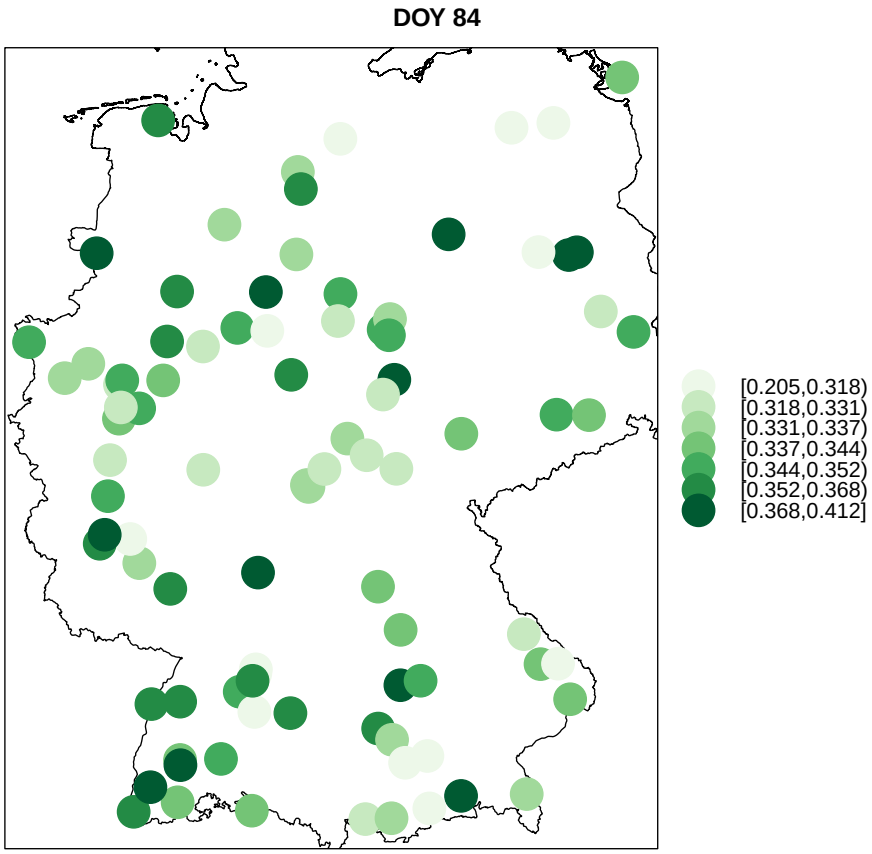


Abbildung 3.2.: rel. Grünwerte der einzelnen Stationen für DOY 84.

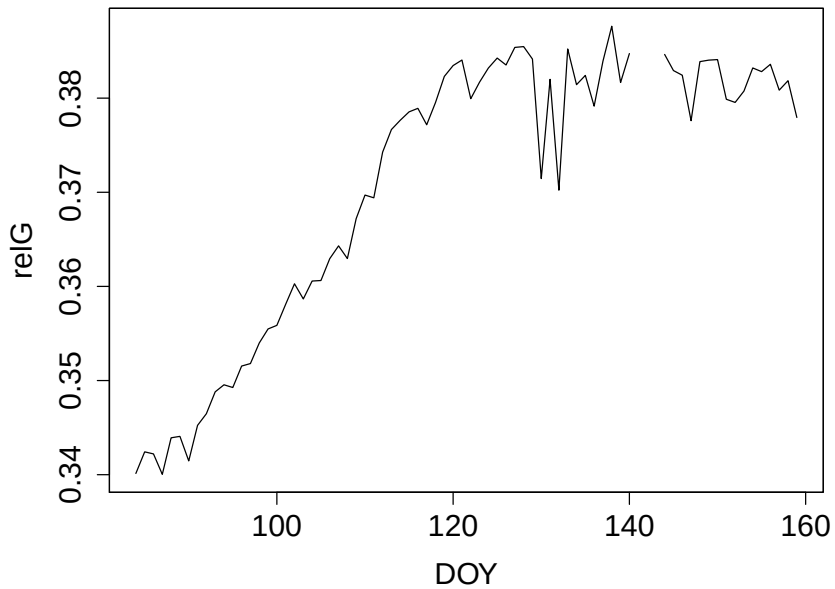


Abbildung 3.3.: Zeitreihe von DOY 84 bis 159.

Das Ziel der bisherigen Analyse war die Identifikation verschiedener phänologischer Zeitpunkte. Für den Frühling waren dies

- der Start der Wachstumszeit (SOS -"start of growing season") - Datum des Erscheinens erster Blätter,
- die Reife der Blätter (MAT - "maturity of the leaf") - Datum der vollen Reife des Laubs.

Zur Schätzung wurde die zweite Ableitung einer geglätteten Funktion über die relativen Grünwerte berechnet. SOS und MAT wurden dann auf den Tag des Maximums (SOS) bzw. Minimums (MAT) der zweiten Ableitung gesetzt.
Zweck der vorliegenden Arbeit ist der Einbezug räumlicher Strukturen in die Analyse.

3.2. Gitterdaten

Im Fall von Gitterdaten besteht die Indexmenge $D \subset \mathbb{R}^d$ aus einer abzählbaren Menge räumlicher Einheiten mit wohldefinierten Grenzen, in denen Daten beobachtet wurden. Die räumliche Information liegt diskret in Form eines räumlichen Indizes $s \in \{1, \ldots, S\}$ vor. Das Gitter kann dabei unregelmäßig (z.B. administrative Einheiten wie Landkreise) oder regelmäßig (z.B. Pixel in einem Bild) sein.

Beispiel: Studierendenzahlen der LMU

Im Rahmen eines Consulting Projektes des Instituts für Statistik in Kooperation mit der Stabstelle Strategie und Entwicklung der Ludwig-Maximilians-Universität München wurde die Entwicklung der Studierendenzahlen an der LMU im Vergleich zu anderen Hochschulen des Bundesgebiets betrachtet.
Grundlage der Analyse war die Statistik der Studenten [Erhebungsjahre: 2004-2011] der Statistischen Ämter des Bundes und der Länder.
Die Erhebung erfolgt über die Verwaltungsdaten der Hochschulen, welche für administrative Zwecke erhoben werden. Es handelt sich somit um eine Sekundärstatistik.
Enthalten sind u.a. soziodemografische Merkmale der Studierenden (Geschlecht, Geburtsdatum, Staatsangehörigkeit), Informationen zum Studium im Berichts- und im vorhergehenden Semester (Hochschule, Art der Einschreibung und des Studiums, angestrebte Abschlussprüfung, Studienfach), zu bereits vor dem Berichtssemester abgelegten Abschlussprüfungen und dem Erwerb der Hochschulzugangsberechtigung.
Der Datenzugang zur Studentenstatistik erfolgt über das Forschungsdatenzentrum München via On-Site-Nutzung am Gastwissenschaftlerarbeitsplatz. Dort "stehen PC-Arbeitsplätze bereit, an denen faktisch anonymisierte Einzeldaten in den geschützten Räumen der amtlichen Statistik von Gastwissenschaftlern analysiert werden können. Die faktische Anonymität wird hierbei nicht allein durch die Anonymisierung der Daten erreicht, sondern in Kombination mit einer Regulierung des Datenzugangs"(Statistische Ämter des Bundes und der Länder 2015).

Ein Teilbereich der Analyse stellte die Betrachtung des Einzugsraums der verschiedenen Universitäten dar. Als Indikator der Herkunft der Studierenden wurde der Ort, an dem die Hochschulzugangsberechtigung (HZB) erworben wurde, verwendet. Abbildung 3.4 zeigt die Herkunft der Studienanfänger an der LMU im Jahre 2004 mit Hilfe des *Standardized Incidence Ratio's* (SIR). Dieses setzt die beobachtete Anzahl an Studenten ins Verhältnis zu der erwarteten Anzahl (für eine genauere Erläuterung siehe Kapitel 6).

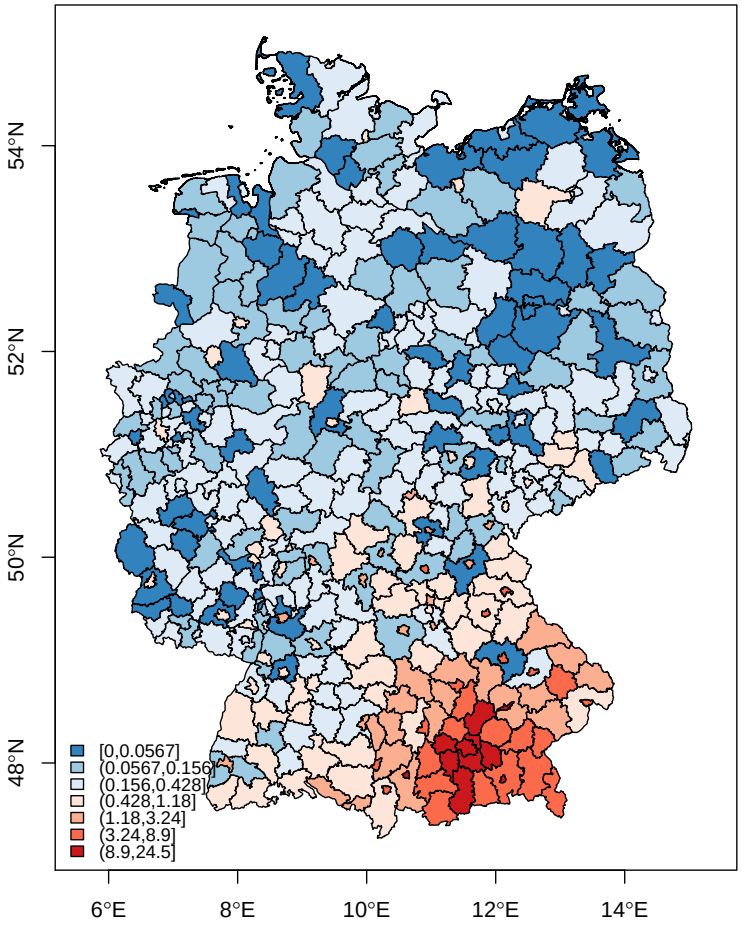


Abbildung 3.4.: Herkunft der Studierenden der LMU 2004 (Standardized Incidence Ratio - SIR).

Für diese Arbeit wurde dieses Thema noch einmal aufgegriffen und mit Hilfe von Modellen aus der räumlichen Statistik analysiert. Dabei wurde die Analyse am Beispiel der LMU für die Jahre 2004 und 2011 durchgeführt. Zusätzlich wurde ein Vergleich mit der Humboldt-Universität zu Berlin angestrengt. Eine gleichzeitige Modellierung aller vorhandenen Jahre war nicht möglich, da es im betrachteten Zeitraum immer wieder zu Gebietsreformen innerhalb des Bundesgebietes kam. Dadurch wurden die Grenzen zwischen den einzelnen Regionen geändert, sodass die vorliegenden Daten nicht in einem gemeinsamen Modell beschrieben werden können. Dieses Problem wird auch als "modifiable areal unit problem" bezeichnet und ist z.B. in Cressie (1996) näher beschrieben.

3.3. Punktprozesse

Bei der Analyse von Punktprozessen liegt das Interesse daran, wo ein Ereignis auftritt. Die Indexmenge D ist in diesem Fall zufällig und beschreibt die Menge $D = \{s_1, \ldots, s_n\}$, wobei $s_1, \ldots, s_n$ die Lokationen zufälliger Ereignisse darstellen. Die Daten $z(s)$ enthalten die Information, ob das Ereignis eingetreten ist, oder nicht.
Es könnten z.B. die Positionen von Pflanzen in einem bestimmten Ausschnitt betrachtet werden. Typische Fragestellungen bei der Analyse von Punktprozessen sind:

- Ist die Verteilung zufällig?
- Bilden sich Cluster? (Aggregation)
- Bilden sich reguläre Strukturen? (Abstoßung, Disaggregation)

Die genauere Betrachtung dieses Datentyps ist nicht Teil dieser Arbeit. Ein guter Überblick findet sich beispielsweise in Diggle (2003). Hinweise über die Auswertung mit Hilfe von R findet sich außerdem in Bivand et al. (2013).

4. Geostatistische Daten

Wie in Kapitel 2 beschrieben, lassen sich die Daten als Realisationen eines Zufallsprozesses $\{Z(s) : s \in D\}$ ausdrücken. Im Fall geostatistischer Daten variiert s stetig im d-dimensionalen Euklidischen Raum innerhalb der Indexmenge D (Region). In den meisten Anwendungen wird dies auf $\mathbb{R}^2$ und $\mathbb{R}^3$ eingeschränkt. Die Lokationen s bestehen dann aus stetigen x- und y-, bzw. x-, y- und z-Koordinaten, also:

$$\boldsymbol{s} = (s_x, s_y)^T \in \mathbb{R}^2 \quad \text{bzw.} \quad \boldsymbol{s} = (s_x, s_y, s_z)^T \in \mathbb{R}^3$$

mit den zugehörigen Daten $z(\boldsymbol{s}_1), \cdots, z(\boldsymbol{s}_n)$, an n vorgegebenen Lokationen $\boldsymbol{s}_1, \cdots, \boldsymbol{s}_n$.
Die Geostatistik beschäftigt sich also mit der Analyse von Zufallsfeldern $Z(\boldsymbol{s})$. Dabei sind typischerweise Messungen an einer limitierten Menge (von manchmal zufällig gewählten) Lokationen vorhanden und die Vorhersage von Z an nicht beobachteten Lokationen $\boldsymbol{s}_0$ wird benötigt.

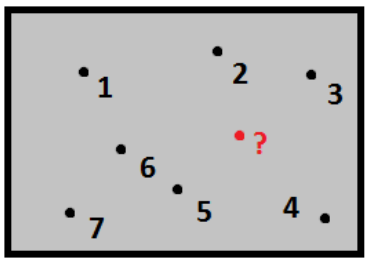


Abbildung 4.1.: Veranschaulichung des Ziels der geostatistischen Analyse.

Grundbaustein der Geostatistik ist das Verfahren des Kriging. Ziel des Verfahrens ist die Vorhersage von unbeobachteten Messwerten auf Basis der beobachteten Werte $Z(\boldsymbol{s}_1), \ldots, Z(\boldsymbol{s}_n)$. Dabei sollen Lokationen mit höherer räumlicher Korrelation zum Punkt $\boldsymbol{s}_0$ auch ein höheres Gewicht in der Berechnung bekommen. Der Berechnung der Gewichte wird somit ein geostatistisches Modell zugrunde gelegt. Gesucht ist also ein Schätzer

$$\hat{Z}(\boldsymbol{s}_0) = \sum_{i=1}^{n} w_i Z(\boldsymbol{s}_i).$$

Die folgenden Abschnitte motivieren die Grundlagen für das Verfahren des Kriging. Dabei wird zunächst der Gauß-Prozess definiert sowie die Annahmen der Stationarität und Isotropie eingeführt. Danach werden das (Semi-)Variogramm und die einzelnen Formen des Kriging vorgestellt.

4.1. Stationäre, räumliche Gaußprozesse

Für die Bestimmung der Kriging-Gewichte wird ein Modell der räumlichen Korrelation der Messstationen benötigt. Korrelationen werden im Normalfall, wenn mehrere Datenpaare $\{x, y\}$ vorhanden sind, aus dem Scatterplot geschätzt. Die räumliche Korrelation zweier Beobachtungen $z(\boldsymbol{s})$ an den Stellen $\boldsymbol{s}_1$ und $\boldsymbol{s}_2$ kann nicht geschätzt werden, da nur ein einzelnes Paar vorhanden ist (Bivand et al. 2013). An jedem Ort kann nur genau eine Erhebung durchgeführt werden. Erhobene geostatistische Daten stellen also eine unvollständige Stichprobe einer einzelnen Realisation des Zufallsprozesses Z dar (Cressie 1993). Dieses Problem wird in der Geostatistik damit gelöst, dass fehlende Messwiederholungen durch Werte an anderen Orten ersetzt werden. Dafür müssen die Werte jedoch der gleichen Grundgesamtheit entstammen. Es bedarf somit weiterer Annahmen über Z um eine Inferenz möglich zu machen (Cressie 1993). Diese werden im Folgenden erläutert. Dabei wird auf die Ausführungen in Schaeben et al. (2013, S. 28f) zurückgegriffen, welche eine gute Übersicht verschaffen.
Es wird im Weiteren vom einfachen Modell

$$Z(\boldsymbol{s}) = \mu(\boldsymbol{s}) + \epsilon(\boldsymbol{s})$$

ausgegangen.
Eine starke Annahme über die Wahrscheinlichkeitsstruktur wäre die der starken Stationarität. Diese liegt vor, wenn die endlich-dimensionalen Verteilungen verschiebungsinvariant sind. Dies bedeutet, dass jede der n Zufallsvariablen $Z(\boldsymbol{s}_1), \ldots, Z(\boldsymbol{s}_n)$ die gleiche Verteilung aufweist.
In der Geostatistik spielen jedoch meist nur die ersten zwei Momente der Verteilung eine Rolle, sodass man sich auf die Definition der schwachen Stationarität beschränken kann.
Das erste Moment entspricht dem Erwartungswert von $Z(\boldsymbol{s})$. Dieser ist abhängig von $\boldsymbol{s}$, d.h.:

$$E[Z(\boldsymbol{s})] = \mu(\boldsymbol{s})$$

Die verschiedenen zweiten Momente definieren Schaeben et al. (2013) wie folgt:

a) Varianzfunktion

$$Var[Z(\boldsymbol{s})] = E[(Z(\boldsymbol{s}) - \mu(\boldsymbol{s}))^2]$$

b) Kovariogramm

$$Cov[Z(\boldsymbol{s}), Z(\boldsymbol{s}')] = c(\boldsymbol{s}, \boldsymbol{s}') = E[(Z(\boldsymbol{s}) - \mu(\boldsymbol{s}))(Z(\boldsymbol{s}') - \mu(\boldsymbol{s}'))]$$

c) Variogramm (Varianz des Inkrements zweier Zufallsvariablen)

$$2\gamma(\boldsymbol{s}, \boldsymbol{s}') = Var[Z(\boldsymbol{s}) - Z(\boldsymbol{s}')]$$

Das Kovariogramm, sowie das Variogramm sind beide von den Punkten $\boldsymbol{s}$ und $\boldsymbol{s}'$ abhängig.

Nimmt man nun einen konstanten Erwartungswert

$$E[Z(\boldsymbol{s})] = \mu = const., \quad \forall \boldsymbol{s} \in D,$$

an und fordert weiterhin, dass das Kovariogramm nur von der Differenz (dem Abstandsvektor $\boldsymbol{h}$) zweier Punkte abhängt, nicht aber von deren genauen Lage im Raum, also dass gilt:

$$Cov[Z(\boldsymbol{s}), Z(\boldsymbol{s}')] = c(\boldsymbol{s} - \boldsymbol{s}') \quad \text{für alle } \boldsymbol{s}, \boldsymbol{s}' \in D$$

so liegt *schwache Stationarität* vor.

Definition: Schwache Stationarität

Der (räumliche) SP $Z = \{Z(\boldsymbol{s}), \boldsymbol{s} \in D\}$ heißt schwach stationär, wenn gilt

a) $E[Z(\boldsymbol{s})] = \mu = const.$,

b) $Cov[Z(\boldsymbol{s} + \boldsymbol{h}), Z(\boldsymbol{s})] = c(\boldsymbol{h}) \quad \forall \boldsymbol{s} \in D$

Unter der Annahme schwacher Stationarität hängt auch das Variogramm nur vom Abstandsvektor $\boldsymbol{h}$ ab, d.h.:

$$Var[Z(\boldsymbol{s} + \boldsymbol{h}) - Z(\boldsymbol{s})] = E[Z(\boldsymbol{s} + \boldsymbol{h}) - Z(\boldsymbol{s}))^2] = 2\gamma(\boldsymbol{h}).$$

Zusätzlich ergibt sich eine konstante Varianz

$$Var[Z(\boldsymbol{s})] = c(\boldsymbol{0}) = \sigma^2 = const \quad \forall \boldsymbol{s} \in D. \tag{4.1}$$

Es gilt außerdem die Beziehung

$$Var[Z(\boldsymbol{s} + \boldsymbol{h}) - Z(\boldsymbol{s})] = Var[Z(\boldsymbol{s} + \boldsymbol{h})] + Var[Z(\boldsymbol{s})] - 2Cov[Z(\boldsymbol{s} + \boldsymbol{h}), Z(\boldsymbol{s})] \tag{4.2}$$

Mit 4.1 und 4.2 lässt sich lässt sich das Variogramm $2\gamma(\boldsymbol{h})$ ausdrücken durch

$$2\gamma(\boldsymbol{h}) = 2(c(\boldsymbol{0}) - c(\boldsymbol{h})) = 2(\sigma^2(1 - \rho(\boldsymbol{h})),$$

wobei $\rho(\boldsymbol{h}) = \frac{c(\boldsymbol{h})}{c(\boldsymbol{0})}$ dem Korrelogramm entspricht. Kovariogramm und Variogramm stellen somit gleichwertige Beschreibungen der Autokorrelation dar.
Da nicht immer eine endliche Varianz existiert, wie sie die Annahme der schwachen Stationarität verlangt (vgl. (4.1)), wird in der Geostatistik meist die Form der *intrinsischen Stationarität* verwendet.

Die Hypothese lautet in diesem Fall:

Definition: Intrinsische Stationarität

Z intrinsisch stationär $\Leftrightarrow$

a) $E[Z(\boldsymbol{s})] = \mu = const. \quad \forall \boldsymbol{s} \in D,$

b) $\frac{1}{2} Var[Z(\boldsymbol{s}+\boldsymbol{h}) - Z(\boldsymbol{s})] = \gamma(\boldsymbol{h}) \quad \forall \boldsymbol{s} \in D$

Es gilt:

Satz:

Z stark stationär $\Rightarrow$ Z schwach stationär $\Rightarrow$ Z intrinsisch stationär

Unter den vorgestellten Hypothesen hängt das (Ko-)Variogramm vom Abstandsvektor $\boldsymbol{h}$ ab. Somit spielen die Länge und Richtung des Vektors zwischen zwei Punkten eine Rolle. Kann die Richtung vernachlässigt werden, spricht man von einem *isotropen* räumlichen Prozess. Bei einem *anisotropen* räumlichen Prozess ist die Korrelation hingegen richtungsabhängig.
Im Folgenden wird implizit (soweit nicht anders angegeben) ein isotroper Prozess angenommen, sodass der Abstandsvektor $\boldsymbol{h}$ durch den euklidischen Abstand $h = \|\boldsymbol{h}\|$ ersetzt wird.

Definition: Isotropie und Anisotropie

Sind $c(\boldsymbol{h}) = c(\|\boldsymbol{h}\|)$ bzw. $\gamma(\boldsymbol{h}) = \gamma(\|\boldsymbol{h}\|)$ nur Funktionen des euklidischen Abstands $\|\boldsymbol{h}\|$, so heißt Z bzw. c/γ isotrop; ansonsten anisotrop.

Unter der Annahme intrinsischer Stationarität können also mit Hilfe des Variogramms Aussagen über die räumliche Korrelation getroffen werden, da nicht mehr die genaue Lage der Messungen, sondern nur deren Abstand h eine Rolle spielt.

4.2. Variogramm

Im vorherigen Abschnitt wurde hergeleitet, dass das Variogramm das wichtigste Werkzeug der Geostatistik darstellt. Mit ihm lassen sich Aussagen über die räumliche Struktur des Zufallsprozesses treffen. Darum wird im Folgenden auf Eigenschaften und die Schätzung des Variogramms eingegangen.

4.2.1. Eigenschaften

Das Variogramm ist in der Regel eine monoton wachsende Funktion. Der Zusammenhang zweier Zufallsvariablen eines räumlichen Prozesses nimmt meist mit ihrem Abstand ab. Somit nimmt die Varianz der Differenz der beiden, also das Variogramm zu (vgl. Abbildung 4.2).

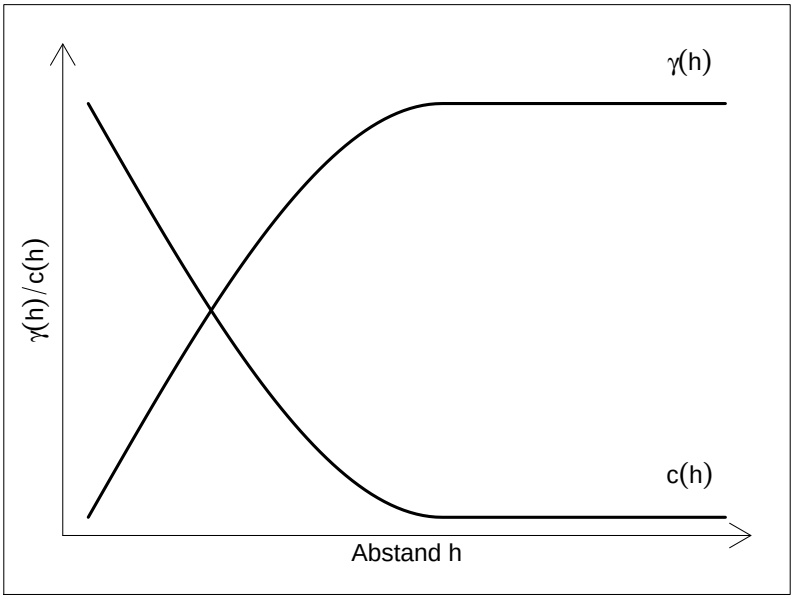


Abbildung 4.2.: Variogramm und Kovariogramm.

Das Verhalten des Variogramms im Ursprung informiert über die Stetigkeits- und Differenzierbarkeitseigenschaften des Prozesses $Z(\cdot)$. Die üblichen Fälle wurden von Matheron (1971, S.58) betrachtet und in Cressie (1993) noch einmal zusammengefasst:

1. $2\gamma(\cdot)$ ist stetig im Ursprung. Dann ist $Z(\cdot)$ L_2-stetig.
2. $2\gamma(h)$ ist d-mal differenzierbar in h=0. Dann ist $Z(\cdot)$ L_2-differenzierbar für alle $\boldsymbol{s} \in \mathbb{R}^d$.
3. $2\gamma(h)$ nähert sich nicht der 0, wenn sich h dem Ursprung nähert. Dann ist $Z(\cdot)$ nicht L_2-stetig und höchst irregulär. Diese Diskontinuität wird als Nugget-Effekt bezeichnet.
4. $2\gamma(\cdot)$ ist eine positive Konstante (außer am Ursprung, wo es 0 ist). Dann sind $Z(\boldsymbol{s})$ und $Z(\boldsymbol{s}')$ unkorreliert für alle $\boldsymbol{s} \neq \boldsymbol{s}'$, egal wie nah sie sich sind. $Z(\cdot)$ wird oft als *weißes Rauschen* (white noise) bezeichnet.

Das Variogramm lässt sich mit einigen Kennzahlen näher beschreiben. Diese werden im Folgenden vorgestellt. Zur Veranschaulichung dient Abbildung 4.3.

Nugget Effekt

Die Definition des Nugget-Effekts ist in der Literatur nicht einheitlich formuliert. Im Folgenden werden die Annahmen aus Cressie (1993) vorgestellt.
Definitionsgemäß gilt $\gamma(0) = 0$. Weiterhin wird c_0 mit

$$\gamma(h) \to c_0 > 0 \quad \text{für } h \to 0$$

als Nugget-Effekt bezeichnet. Dieser Ausdruck stammt von Matheron (1962) und ergibt sich aus der Hypothese einer "microscale variation" (kleine Klumpen/Nuggets), die eine Diskontinuität am Ursprung verursacht. Für einen L_2-stetigen Prozess ist dies mathematisch nicht möglich. Sobald also Stetigkeit des Phänomens auf der Mikroebene angenommen wird, kann $c_0 > 0$ nur einem Messfehler zugeschrieben werden. Im Folgenden bezeichnet c_{ME} die Messfehlervarianz (engl. measurement-error variance).
Matheron (1962) geht davon aus, dass die "microscale variation" nicht stetig ist und fügt zur Modellierung des Prozesses geringer Entfernungen einen White-Noise-Prozess zu einem Prozess stetiger Sample-Pfade hinzu. Da im Normalfall keine Daten für so nah beieinanderliegende Orte zur Verfügung stehen, kann diese Annahme nicht überprüft werden.
Die Varianz des White-Noise-Prozesses wird mit c_{MS} bezeichnet. Daraus ergibt sich dann

$$c_0 = c_{MS} + c_{ME}.$$

Sill

Der Schwellenwert (sill) entspricht dem höchsten Wert, welchem sich die Variogrammkurve asymptotisch annähert. Je größer der Abstand h zweier Punkte wird, desto niedriger wird die Korrelation, sodass

$$2\gamma(h) = 2\sigma^2(1 - \rho(h)) \to 2\sigma^2, \quad \text{für } h \to \infty$$

Range

Der Range entspricht dem Abstand h bei dem zwei Punkte im Raum mit einer größeren Entfernung als h als vernachlässigbar korreliert angesehen werden können. Dies kommt dem Abstand h gleich, bei dem die Kurve den Sill erreicht.

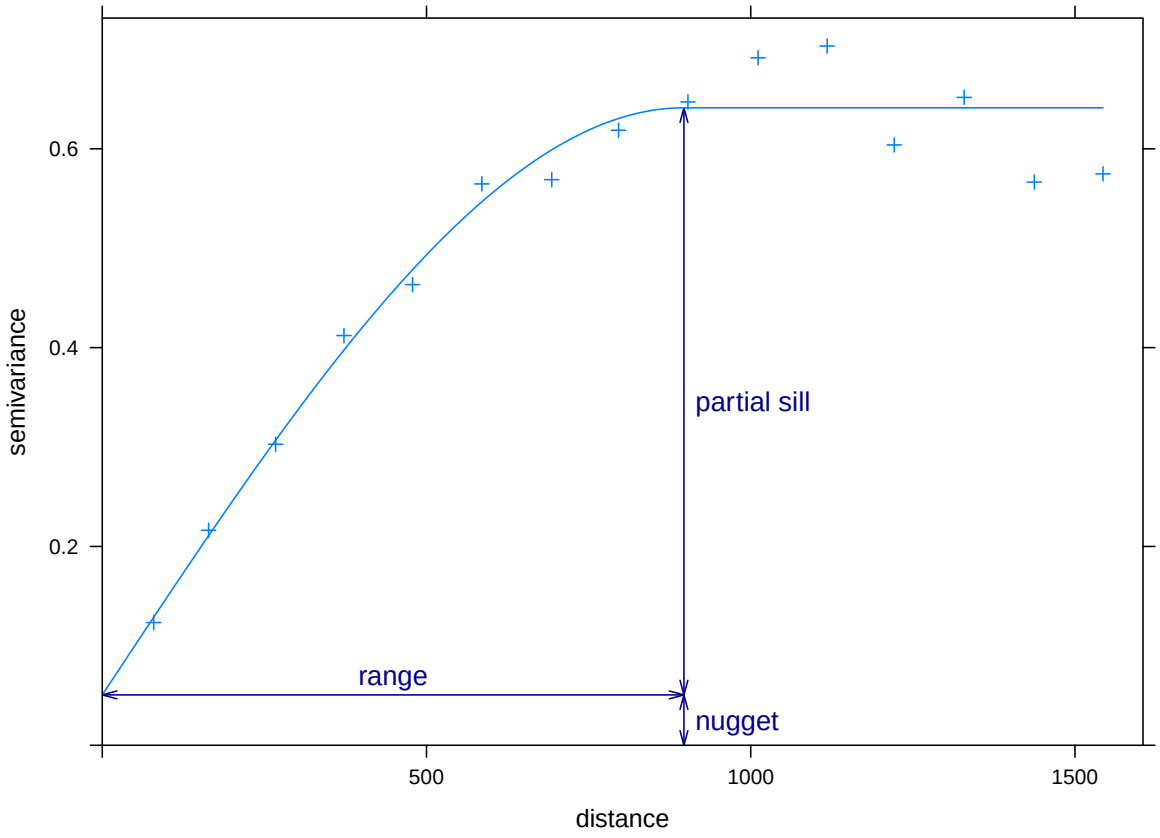


Abbildung 4.3.: Empirisches Variogramm und gefittetes Modell; Quelle: Bivand et al. (2013).

4.2.2. Typische Probleme

Anisotropie

Sobald die Abhängigkeit zwischen $Z(\boldsymbol{s})$ und $Z(\boldsymbol{s}')$ eine Funktion der Länge *und* der Richtung des Vektors $\boldsymbol{h}$ zwischen den beiden Punkten ist, liegt Anisotropie vor.
Bei der sogenannten geometrischen Anisotropie ist der Sill für alle Richtungen identisch, die Range unterscheidet sich jedoch. Dies kann durch eine lineare Transformation des Abstandsvektors $\boldsymbol{h}$ korrigiert werden (siehe Fahrmeir et al. (2009)). Dabei ersetzt man den euklidischen Abstand

$$\|\boldsymbol{s}_1 - \boldsymbol{s}_2\| = \sqrt{(\boldsymbol{s}_1 - \boldsymbol{s}_2)'(\boldsymbol{s}_1 - \boldsymbol{s}_2)}$$

durch

$$\sqrt{(\boldsymbol{s}_1 - \boldsymbol{s}_2)'\boldsymbol{R}(\psi)'\boldsymbol{D}(\delta)\boldsymbol{R}(\psi)(\boldsymbol{s}_1 - \boldsymbol{s}_2)},$$

wobei $\boldsymbol{R}(\psi)$ eine Rotationsmatrix mit Anisotropie-Winkel $\psi \in [0, 2\pi]$ bezeichnet, also

$$\boldsymbol{R}(\psi) = \begin{pmatrix} \cos(\psi) & \sin(\psi) \\ -\sin(\psi) & \cos(\psi) \end{pmatrix},$$

und $\boldsymbol{D}(\delta)$ eine Dehnungsmatrix mit Anisotropieverhältnis $\delta \leq 1$, d.h.

$$\boldsymbol{D}(\delta) = \begin{pmatrix} \delta^{-1} & 0 \\ 0 & 1 \end{pmatrix}.$$

Schwieriger wird es für die zonale Anisotropie. In diesem Fall ist der Sill für verschiedene Richtungen unterschiedlich.
Der gesamte Prozess zerfällt dabei in unabhängige Subprozesse (Cressie 1993)

$$Z(\boldsymbol{s}) = Z_1(\boldsymbol{s}) + Z_2(\boldsymbol{s}) + \cdots + Z_p(\boldsymbol{s}), \quad s \in D$$

mit einer Variogramm-Zerlegung

$$2\gamma(\boldsymbol{h}) = 2\gamma_1(\boldsymbol{h}) + 2\gamma_2(\boldsymbol{h}) + \cdots + 2\gamma_p(\boldsymbol{h})$$

Selbst wenn jeder dieser Subprozesse einer einfachen geometrischen Anisotropie folgt, kann es unmöglich sein, diese auf Basis des beobachteten Prozesses $Z(\cdot)$ zu identifizieren.
Anisotropie lässt sich durch die Berechnung direktionaler Variogramme aufzeigen. Dabei werden für verschiedene Winkelbereiche (Richtung+Toleranzbereich) Gruppen gebildet und separate Variogramme geschätzt.

Drift und Hole-Effekt

Aus dem Variogramm lassen sich noch weitere Abweichungen von den Annahmen erkennen.

In Abbildung 4.4 erreicht das Variogram nicht, oder nur scheinbar den Schwellenwert (Sill). Danach steigen die Werte weiter an. Der sogenannte *Drift-Effekt* deutet auf eine Verletzung der Stationaritätsannahme hin. Der Erwartungswert ist nicht konstant über das Untersuchungsgebiet.

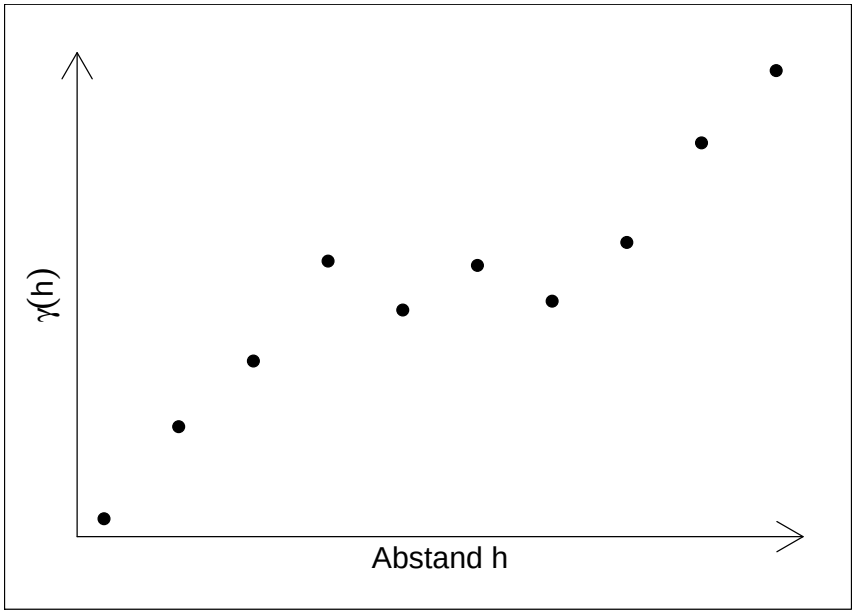


Abbildung 4.4.: Drift-Effekt.

In Abbildung 4.5 wird der Schwellenwert (Sill) scheinbar erreicht, danach fallen die Variogrammwerte wieder. Dies deutet auf regelmäßige Strukturen hin, bei denen sich die Werte in regelmäßigen Abständen wieder stärker ähneln. Bezeichnet wird dieser Effekt mit *Hole-Effekt*.

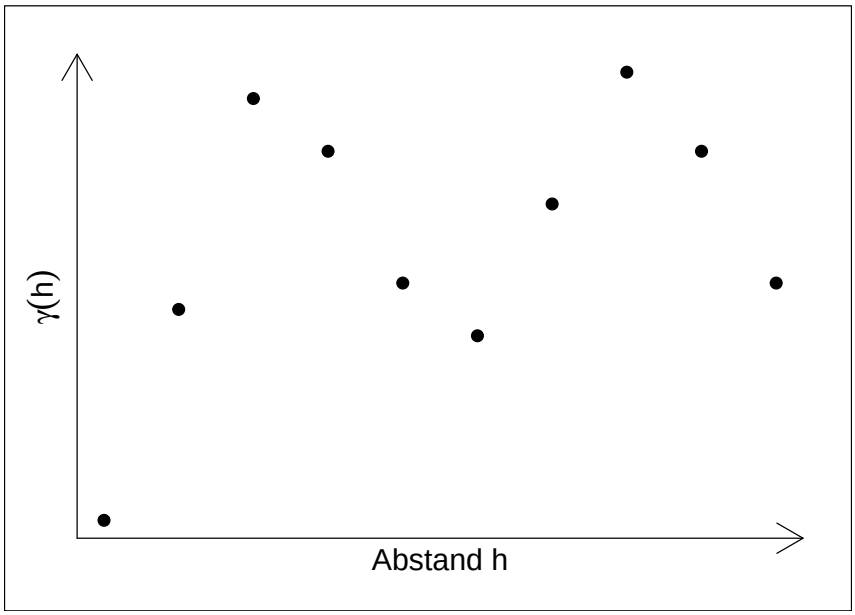


Abbildung 4.5.: Hole-Effekt.

4.2.3. Empirisches Variogramm

Ein empirischer Schätzer für das Variogramm ist, unter der Annahme eines konstanten Mittelwertes, gegeben durch (Matheron 1962)

$$2\hat{\gamma}(h) = \frac{1}{|N(h)|} \sum_{N(h)} (Z(\boldsymbol{s}_i) - Z(\boldsymbol{s}_j))^2 \tag{4.3}$$

mit $N(h) = \{(\boldsymbol{s}_i, \boldsymbol{s}_j) : ||\boldsymbol{s}_i - \boldsymbol{s}_j|| = h; i, j = 1, \ldots, n\}$ und $|N(h)|$ Anzahl verschiedener Paare in $N(h)$.
In den meisten Anwendungen sind die Daten irregulär, sodass $|N(h)|$ sehr klein wird und der Schätzer instabil ist. Es wird dann eine "Toleranzregion" um h spezifiziert und der Schätzer somit über Intervalle, anstatt über genaue Abstände berechnet.
Es ergibt sich so der Schätzer (vgl. Cressie (1993))

$$2\hat{\gamma}(h) = \text{ave}\left\{(Z(s_i) - Z(s_j))^2 : (i, j) \in N(h); h \in T(h)\right\}, \tag{4.4}$$

wobei $T(h)$ der Toleranzregion in über h entspricht und ave$\{\cdot\}$ einen möglicherweise gewichteten Durchschnitt bezeichnet. Es ergibt sich hieraus eine Treppenfunktion über eine definierte Anzahl an Abstandsintervallen.
Der entstehende Schätzer ist ähnlich wie bei der Histogrammschätzung abhängig von den gewählten Intervallen. Eine naheliegende Überlegung ist dann eine "moving average"-Schätzung analog zur Kerndichteschätzung.

4.2.4. Theoretische Variogramme

Das bisher betrachtete empirische Variogramm dient als Näherung des theoretischen Variogramms. Die Anpassung eines parametrischen Modells an die Daten geschieht aus zwei Gründen:

- Die räumliche Interpolation (Kriging) benötigt Schätzer des Variogramms $\gamma(h)$ auch für Abstände h, die nicht in den Daten vorhanden sind.
- Die Vorhersage-Varianzen der geschätzten Werte müssen nicht-negativ sein (Bivand et al. 2013). Dies kann durch das empirische Variogramm nicht garantiert werden (siehe Cressie (1993) für genauere Betrachtung).

Ein gültiges Modell für die Semivarianz muss bedingt negativ-definit sein, d.h.

$$\sum_{i=1}^{m} \sum_{j=1}^{m} a_i a_j 2\gamma(\boldsymbol{s}_i - \boldsymbol{s}_j) \leq 0,$$

für jegliche endliche Anzahl an räumlichen Lokationen $\{\boldsymbol{s}_i : i = 1, \ldots, m\}$ und reelle Zahlen $\{a_i, i = 1, \ldots, m\}$ welche die Gleichung $\sum_{i=1}^{m} a_i = 0$ erfüllen.

Klassische parametrische Modelle werden nun im Folgenden vorgestellt:

Sphärisches Modell

$$\gamma(h) = \begin{cases} 0 & \text{für } h = 0, \\ c_0 + c_1 \left(\frac{3h}{2a} - \frac{1}{2} \left(\frac{h}{a} \right)^3 \right) & \text{für } 0 < h \leq a, \\ c_0 + c_1 & \text{für } h \geq a. \end{cases}$$

Exponentielles Modell

$$\gamma(h) = \begin{cases} 0 & \text{für } h = 0, \\ c_0 + c_1(1 - e^{-h/a}) & \text{für } h \neq 0. \end{cases}$$

Gauß'sches Modell

$$\gamma(h) = \begin{cases} 0 & \text{für } h = 0, \\ c_0 + c_1(1 - e^{-(h/a)^2}) & \text{für } h \neq 0. \end{cases}$$

Matern Modell

$$\gamma(h) = \begin{cases} 0 & \text{für } h = 0, \\ c_0 + c_1 \left[1 - \frac{1}{2^{\kappa-1}\Gamma(\kappa)} \left(\frac{h}{a} \right)^{\kappa} K_{\kappa} \left(\frac{h}{a} \right) \right] & \text{für } h \neq 0, \end{cases}$$

wobei K_κ der Bessel-Funktion entspricht und $\kappa > 0$. Die Darstellung dieser Familie ist nur mit Hilfe der modifizierten Bessel-Funktionen der Ordnung κ möglich. Diese sind nur numerisch auswertbar und lassen sich nicht explizit darstellen. Für $\kappa = 0.5, 1.5, 2.5, \ldots$ sind jedoch explizite Formen möglich (siehe z.B. Fahrmeir et al. (2009)). Für $\kappa = 0.5$ entspricht die Matern-Funktion dem Exponential-Modell.

Abbildung 4.6 zeigt die vorgestellten Modelle beispielhaft mit einem Nugget-Effekt von 0.1 und einem *partial sill* von 1. Für κ wurde 0.3 gewählt um den Unterschied vom Matern- zum Exponential-Modell erkennen zu können. Im Falle, dass das Variogramm nur asymptotisch den Sill erreicht (Exponential- und Gauß-Modell), wird der sogenannte effektive Range betrachtet. Dieser ist definiert als die Distanz, an der die Semivarianz 95% des Sills erreicht. Der effektive Einflussbereich entspricht, bei gefittetem Range a, $3a$ im Exponential- bzw. $\sqrt{3}a$ im Gauß-Modell.

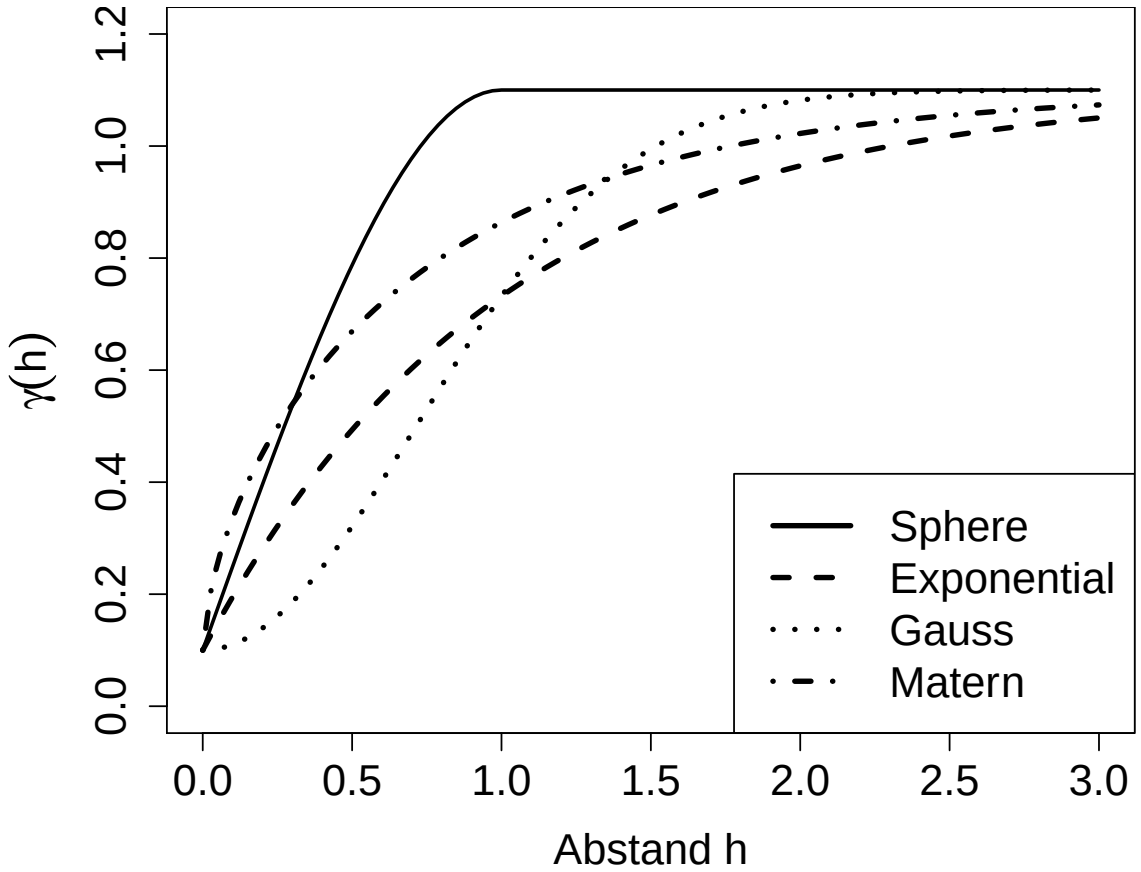


Abbildung 4.6.: Parametrische Variogramm-Modelle.

4.3. Kriging

Ziel der Geostatistik ist die Vorhersage bzw. Interpolation von Variablenwerten $Z(s_0)$ auf Basis von beobachteten Messwerten $Z(\boldsymbol{s}_1), \ldots, Z(\boldsymbol{s}_n)$ in der Nachbarschaft. Das klassische Instrument der Geostatistik hierfür ist das sogenannte *Kriging*. Der Name stammt von Matheron (1963), welcher das Verfahren nach D.G. Krige, einem südafrikanischen Bergbauingenieur benannte. Dieser entwickelte in den 1950er Jahren eine empirische Methode, um die Verteilung von Erzgehalten basierend auf Stichproben zu bestimmen (siehe Krige (1951)). Cressie (1990) bezeichnet die Methode auch als *spatial optimal linear prediction*, bei der der unbekannte Erwartungswert des Zufallsprozesses durch den besten linearen unverzerrten Schätzer (best linear unbiased estimator = BLUE) geschätzt wird. Optimal ist der Schätzer in Hinsicht auf die Minimierung des mittleren quadratischen Vorhersagefehlers (engl. *mean squared prediction error*)

$$\text{MSPE} = E\left[\left(Z(\boldsymbol{s}_0) - \hat{Z}(\boldsymbol{s}_0)\right)^2\right] = Var\left[Z(\boldsymbol{s}_0) - \hat{Z}(\boldsymbol{s}_0)\right].$$

Dieser wird auch mit Kriging-Varianz bezeichnet.
Es gibt unterschiedliche Formen des Kriging, die sich darin unterscheiden, welche Annahmen über den Erwartungswert getroffen werden. Diese werden nun im Folgenden vorgestellt.

4.3.1. Einfaches (simple) Kriging

Beim einfachen Kriging wird angenommen, dass μ bekannt ist. Diese Annahme ist in den meisten Anwendungen unrealistisch. Deshalb wird hier auf eine genauere Herleitung der Krige-Gleichungen verzichtet. Eine detaillierte Betrachtung hierzu findet sich in Cressie & Wikle (2011, Kap. 4.1.2).

4.3.2. Gewöhnliches (ordinary) Kriging

Beim gewöhnlichen Kriging wird der Erwartungswert als konstant, aber unbekannt angenommen. Es gilt das *constant-mean*-Modell

$$Z(\boldsymbol{s}) = \mu + \delta(\boldsymbol{s}), \quad \boldsymbol{s} \in D,$$

wobei $\mu \in \mathbb{R}$ unbekannt und $\delta(\cdot)$ ein zero-mean intrinsisch stationärer Prozess mit Variogramm $2\gamma(\cdot)$ ist. Es gilt also

$$E[Z(\boldsymbol{s})] = \mu, \quad \text{für alle } \boldsymbol{s} \in D.$$

Gesucht wird ein Schätzer

$$\hat{Z}(\boldsymbol{s}_0) = \sum_{i=1}^{n} \lambda_i Z(\boldsymbol{s}_i).$$

mit den Kriging-Gewichten $\lambda_1, \ldots, \lambda_n$, der den mittleren quadratischen Vorhersagefehler bzw. die sogenannte Kriging-Varianz minimiert.

Neben der Minimierung des MSPE soll Erwartungstreue des Schätzers gelten. Zu diesem Zweck wird die Nebenbedingung

$$\sum_{i=1}^{n} \lambda_i = 1 \tag{4.5}$$

eingeführt. Diese garantiert Erwartungstreue wegen

$$\begin{aligned} & E\left(\sum_{i=1}^{n} \lambda_i Z(\boldsymbol{s}_i)\right) - \mu \stackrel{!}{=} 0 \\ \Leftrightarrow \quad & \sum_{i=1}^{n} \lambda_i E(Z(\boldsymbol{s}_i)) - \mu = 0 \\ \Leftrightarrow \quad & \mu \cdot \sum_{i=1}^{n} \lambda_i - \mu = 0 \\ \Leftrightarrow \quad & \sum_{i=1}^{n} \lambda_i = 1. \end{aligned}$$

Unter der Annahme intrinsischer Stationarität lässt sich die Krige-Varianz über das Variogramm ausdrücken. Um zusätzlich die Nebenbedingung 4.5 zu berücksichtigen wird der Lagrange-Multiplikator m eingeführt. Somit ergibt sich der zu minimierende Term wie folgt (Herleitung siehe Cressie (1993)):

$$-\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \gamma(\boldsymbol{s}_i - \boldsymbol{s}_j) + 2 \cdot \sum_{i=1}^{n} \lambda_i \gamma(\boldsymbol{s}_0 - \boldsymbol{s}_i) - 2m\left(\sum_{i=1}^{n} \lambda_i - 1\right). \tag{4.6}$$

Ableiten von 4.6 nach $\lambda_1, \ldots, \lambda_n$ bzw. m und Nullsetzen ergibt die Krigegleichungen

$$-\sum_{j=1}^{n} \lambda_j \gamma(\boldsymbol{s}_i - \boldsymbol{s}_j) + \gamma(\boldsymbol{s}_0 - \boldsymbol{s}_i) - m = 0 \quad i = 1, \ldots, n \tag{I}$$

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

Aus diesem Gleichungssystem lassen sich nun die Gewichte $\lambda_1, \ldots, \lambda_n$ bestimmen.
Die Krigevarianz ist hier

$$\sigma^2_{ok}(\boldsymbol{s}_0) = \sum_{i=1}^{n} \lambda_i \gamma(\boldsymbol{s}_0 - \boldsymbol{s}_i) + m.$$

Sie hängt nur von den Gewichten, den Messlokationen und dem Variogramm, nicht aber von den eigentlichen Messwerten ab.

Beispiel

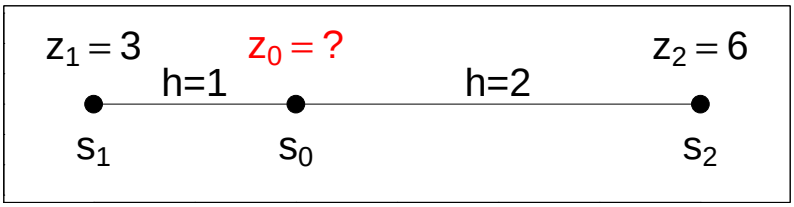


Abbildung 4.7.: Beispiel Krige-Gleichungen.

Annahme:	Es liegt ein isotroper, intrinsisch stationärer Prozess mit linearem Variogramm vor, d.h. $\gamma(h) = \|h\|$.
Geg:	$\boldsymbol{s}_1 = (1,0), \boldsymbol{s}_2 = (4,0), \boldsymbol{s}_0 = (2,0), z(\boldsymbol{s}_1) = 3, z(\boldsymbol{s}_2) = 6$
Ges:	$z(\boldsymbol{s}_0)$
Krigegleichungen:	

$$-3\lambda_2 + 1 - m = 0 \quad \text{(I)}$$
$$-3\lambda_1 + 2 - m = 0 \quad \text{(II)}$$
$$\lambda_1 + \lambda_2 = 1 \quad \text{(III)}$$

Daraus folgt: $\lambda_1 = \frac{2}{3}, \lambda_1 = \frac{1}{3}$ und somit $\hat{z}(\boldsymbol{s}_0) = \frac{2}{3} * 3 + \frac{1}{3} * 6 = 4$

4.3.3. Instationäre Methoden

Bisher wurde (intrinsische) Stationarität und somit ein konstanter Mittelwert angenommen. Dies ist in der Realität meist nicht gegeben, sodass $E[Z(s)]$ nicht länger als konstant, sondern als eine Linearkombination bekannter Funktionen $(f_0(\boldsymbol{s}), \ldots, f_p(\boldsymbol{s})), \boldsymbol{s} \in D$ angenommen wird.
Somit gilt die Annahme (vgl. Cressie (1993))

$$Z(\boldsymbol{s}) = \sum_{j=1}^{p+1} f_{j-1}(\boldsymbol{s})\beta_{j-1} + \delta(\boldsymbol{s}), \quad \boldsymbol{s} \in D,$$

wobei $\beta = (\beta_0, \ldots, \beta_p)' \in \mathbb{R}^{p+1}$ ein unbekannter Vektor von Parametern und $\delta(\cdot)$ ein zero-mean intrinsisch stationärer Zufallsprozess mit Variogramm $2\gamma(\cdot)$ ist.
Auch hier können analog zu vorher Kriging-Gleichungen aufgestellt werden. Auf diese soll hier nicht näher eingegangen werden. Eine ausführliche Herleitung findet sich in Cressie (1993).
Beim Universal-Kriging wird der Trend durch die Lagekoordinaten erklärt. Es handelt sich um ein instationäres Interpolationsverfahren. Ähnlich sind das *External-Drift*-Kriging und *Regression*-Kriging. Hier werden in beiden Fällen zusätzliche Hilfsvariablen verwendet um den Trend zu schätzen. Diese müssen sowohl für die Messpunkte, als auch an den Orten, für die interpoliert werden soll, bekannt sein. Dies stellt in der Praxis häufig ein Problem dar. Ein Vergleich der Methoden findet sich beispielsweise in Hengl et al. (2003).

4.4. Kriging als Basisfunktionenansatz

Häufig angewendet wird in der Statistik die nichtparametrische Glättung von Oberflächen. Im Folgenden soll diese Methode mit dem Kriging-Ansatz in Verbindung gebracht werden. Die Herleitung hierfür stammt aus Fahrmeir et al. (2009) und wurde an die bisherige Notation der Arbeit angepasst.
Fahrmeir et al. (2009) definieren das Modell

$$y(\boldsymbol{s}_i) = \boldsymbol{x}(\boldsymbol{s}_i)'\boldsymbol{\beta} + \delta(\boldsymbol{s}_i) + \epsilon(\boldsymbol{s}_i), \quad i = 1, \ldots, n$$

als klassisches geostatistisches Modell. Dabei sind

$\boldsymbol{x}(\boldsymbol{s}_i)'\boldsymbol{\beta}$	der durch Kovariablen $\boldsymbol{x}$ parametrisierte räumliche Trend,
$\delta(\boldsymbol{s}_i)$	ein stationärer Gauß-Prozess mit Erwartungswert 0, Varianz τ^2 und Korrelationsfunktion $\rho(h)$,
$\epsilon(\boldsymbol{s}_i)$	der übliche Fehlerterm, also $\epsilon(\boldsymbol{s}_i) \sim N(0, \sigma^2)$

In Matrixnotation lässt sich das Modell schreiben als

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\delta} + \boldsymbol{\epsilon},$$

wobei $\boldsymbol{\delta} = (\delta(\boldsymbol{s}_1), \ldots, \delta(\boldsymbol{s}_n))'$ die Werte des stationären Gaußprozess an den n verschiedenen beobachteten räumlichen Lokationen $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$ und $\boldsymbol{Z} = \boldsymbol{I}_n$ die n-dimensionale Einheitsmatrix bezeichnet. Die Kovarianzmatrix der Zielvariablen $\boldsymbol{y}$ setzt sich aus einem unkorrelierten Teil $\sigma^2\boldsymbol{I}_n$ und einem korrelierten Teil $\tau^2\boldsymbol{Z}\boldsymbol{R}\boldsymbol{Z}'$ zusammen, d.h.

$$Cov(\boldsymbol{y}) = \tau^2\boldsymbol{Z}\boldsymbol{R}\boldsymbol{Z}' + \sigma^2\boldsymbol{I}_n,$$

wobei die Kovarianzmatrix $\boldsymbol{R}$ der räumlichen Effekte gegeben ist durch

$$\boldsymbol{R} = (Corr(\delta(\boldsymbol{s}_i), \delta(\boldsymbol{s}_j)) = (\rho(\boldsymbol{s}_i - \boldsymbol{s}_j))$$

Durch eine Reparametrisierung des Modells zu

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{R} \cdot \boldsymbol{R}^{-1}\boldsymbol{\delta} + \boldsymbol{\epsilon} = \boldsymbol{X}\boldsymbol{\beta} + \tilde{\boldsymbol{Z}}\tilde{\boldsymbol{\delta}} + \boldsymbol{\epsilon},$$

mit $\tilde{\boldsymbol{Z}} = \boldsymbol{Z}\boldsymbol{R}$ und $\tilde{\boldsymbol{\delta}} = \boldsymbol{R}^{-1}\boldsymbol{\delta}$ erreichen Fahrmeir et al. (2009) eine äquivalente Modellformulierung mit veränderter Interpretation der Matrix $\tilde{\boldsymbol{Z}}$.
Deren Einträge lauten nun

$$\tilde{\boldsymbol{Z}}[i, j] = \rho(\boldsymbol{s}_i, \boldsymbol{s}_j).$$

Verwendet man nun die Korrelationsfunktion ρ wie eine Basisfunktion und die beobachteten Lokationen als Knoten, so zeigt sich eine äquivalente Formulierung zur Konstruktion der Desi-

gnmatrix bei Tensorprodukt-Splines. Somit lässt sich das geostatistische Modell für die einzelnen Beobachtungen schreiben als

$$y(\boldsymbol{s}_i) = \boldsymbol{x}(\boldsymbol{s}_i)'\boldsymbol{\beta} + f_{geo}(\boldsymbol{s}_i) + \epsilon(\boldsymbol{s}_i),$$

wobei

$$f_{geo}(\boldsymbol{s}_i) = \sum_{j=1}^{n} \tilde{\delta}_j B_j(\boldsymbol{s}_i),$$

dem räumlichen Effekt und

$$B_j(\boldsymbol{s}_i) = \rho(\boldsymbol{s}_i, \boldsymbol{s}_j)$$

den auf der Korrelationsfunktion basierenden Basisfunktionen entsprechen. Liegen isotrope Korrelationsfunktionen vor, erhält man radiale Basisfunktionen der Form $B_j(\boldsymbol{s}_i) = \rho(||\boldsymbol{s}_i, \boldsymbol{s}_j||)$.
Die Knoten entsprechen hier den beobachteten Lokationen und sind somit anders als im normalen Basisfunktionen-Ansatz im Vorhinein festgelegt.
Die gemeinsame Verteilung der räumlich korrelierten Effekte $\tilde{\delta}$ ist gegeben durch

$$\tilde{\delta} \sim \mathrm{N}(0, \tau^2 \boldsymbol{R}^{-1}),$$

sodass der Kriging-Ansatz einer Glattheits-Priori wie in den Penalisierungsansätzen der nichtparametrischen Regression entspricht.

5. Gitter- bzw. Pixeldaten

Dieses Kapitel beschäftigt sich mit der zweiten Art der drei in Kapitel 3 besprochenen Datentypen, den sogenannten *Gitterdaten* (engl. lattice).
Ausgangspunkt ist wieder ein stochastischer Prozess $\{Z(s); s \in D\}$. In diesem Fall ist die Indexmenge D eine abzählbare Sammlung räumlicher Orte an denen Daten beobachtet wurden. Die räumliche Information liegt somit diskret z. B. in Form einer Regionenvariable s vor. Das entstehende Gitter wird zusätzlich um eine Nachbarschaftsinformation (vgl. Abschnitt 5.1) ergänzt.
Es können drei Charakteristika von Gitterdaten unterschieden werden:

1. Handelt es sich um ein *reguläres* (z. B. Pixel) oder ein *irreguläres* (z. B. Bundesländer) Gitter?
2. Verweisen die Lokationen des Gitters auf Punkte oder Regionen?
3. Ist die Zielvariable diskret oder metrisch?

5.1. Nachbarschaften

Im Fall geostatistischer Daten lassen sich Beziehungen zwischen Lokationen über deren Abstand (z. B. mit Hilfe der euklidischen Distanz) zueinander definieren. Dies ist im Falle diskreter räumlicher Information nicht möglich. Im Folgenden wird darum das Konzept der Nachbarschaften eingeführt, um die räumliche Anordnung der Daten beschreiben zu können.
Nachbarschaften lassen sich auf verschiedene Weise konstruieren.
Sei beispielsweise

$$D = \{(x_i, y_i) : i = 1, \ldots, 100\}$$

ein Gitter über 100 Regionen, wobei x und y den Längen- und Breitengraden der jeweiligen Kreisstadt einer Region entsprechen.
Eine Möglichkeit ein Nachbarschaftssystem zu konstruieren, ist die Definition über die Entfernung der Zentroide bzw. hier der Kreisstädte. Beispielsweise können alle Regionen als Nachbarn einer Region i angesehen werden, deren Kreisstadt weniger als 50km von der Kreisstadt der Region i entfernt ist.
Eine Abwandlung hiervon ist die Konstruktion über die k-nächsten Nachbarn (engl. *k-nearest neighbour*). Dies führt in den meisten Fällen zu einem asymmetrischen Graphen, gewährleistet dafür aber, dass jedes Gebiet genau k Nachbarn besitzt.

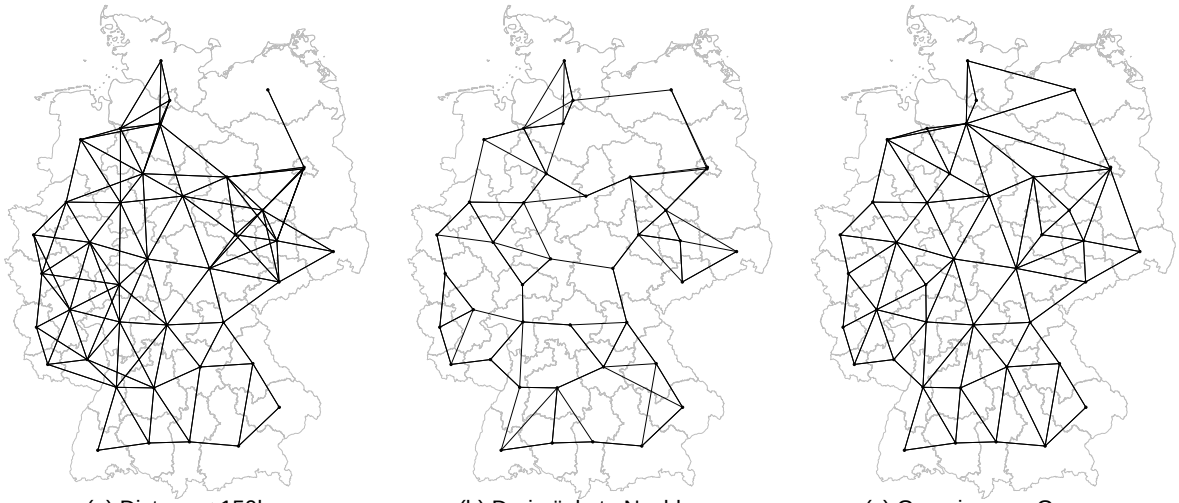
Eine weitere Möglichkeit besteht in der Betrachtung gemeinsamer Grenzen.
Es ergibt sich daraus das Nachbarschaftssystem:

$$\partial = \{N(s) : s \in D\},$$

wobei gilt:

$N(s)$ entspricht der Menge aller Nachbarn von s
$s \notin N(s)$
$v \in N(s) \Leftrightarrow s \in N(v)$
Alle $v \in N(s)$ heißen Nachbar von s. (Notation: $v \sim s$)

Abbildung 5.1 zeigt die Graphen der verschiedenen Nachbarschaftskriterien anhand der Regierungsbezirke in Deutschland.



(a) Distanz < 150km (b) Drei nächste Nachbarn (c) Gemeinsame Grenzen

Abbildung 5.1.: Nachbarschaftssysteme mit verschiedenen Kriterien.

Zusätzlich zur Wahl eines Nachbarschaftskriteriums kann es sinnvoll sein Nachbarn geeignet zu gewichten. Bivand et al. (2013, Kapitel 9.2.2) rät jedoch davon ab, weit von einer binären Repräsentation abzuweichen, wenn wenig über den räumlichen Prozess bekannt ist.
In Fahrmeir et al. (2009, Kapitel 7.2.4) werden folgende Strategien zur Definition der Gewichte vorgeschlagen:

- Gleiches Gewicht für alle Nachbarn
- Gewichte invers proportional zum Abstand der Zentroide
- Gewichte proportional zur Länge der gemeinsamen Grenze

Im Weiteren wird von einer Nachbarschaftsmatrix W ausgegangen, deren Einträge w_{ij} den Gewichten entsprechen. Dabei wird w_{ii} generell auf Null gesetzt. Oft erfolgt eine Standardisierung, indem die Einträge w_{ij} durch die Zeilensumme $\sum_j w_{ij} = w_{i+}$ geteilt werden.
Die Konstruktion von Nachbarschaftssystemen kann auch auf zweite oder höhere Nachbarn ausgeweitet werden. Hierfür können beispielsweise Distanzintervalle $(0, d_1], (d_1, d_2]$, usw. definiert werden. Alle ersten Nachbarn von i liegen dann innerhalb der Distanz d_1 von i. Alle zweiten Nachbarn liegen weiter entfernt von i als d_1, aber sind näher als d_2. In den Abbildungen 5.2 und 5.3 sind Nachbarschaftssysteme für reguläre und irreguläre Gitter auf Basis gemeinsamer Grenzen zu sehen. Dabei könnte das System für irreguläre Gitter ebenfalls ausgeweitet werden, indem zusätzlich diejenigen Regionen mit einbezogen werden, welche eine gemeinsame Grenze zu den ersten Nachbarn besitzen.

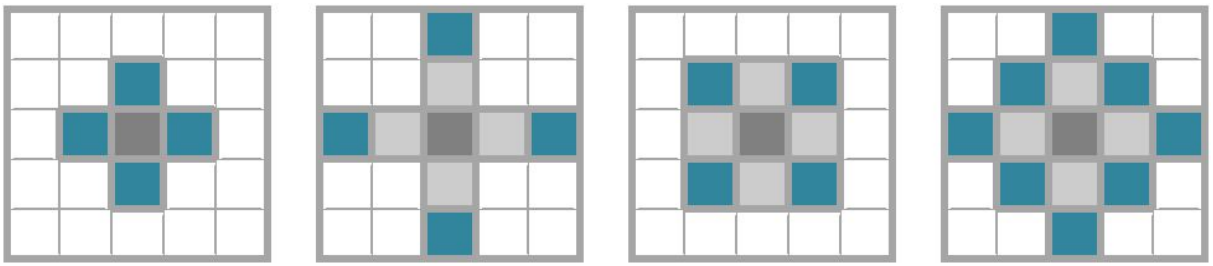


Abbildung 5.2.: Nachbarschaftssysteme auf regulären Gittern. V.l.n.r: Erste Nachbarn, zweite Nachbarn, zweite diagonale Nachbarn, zweite und diagonale Nachbarn.

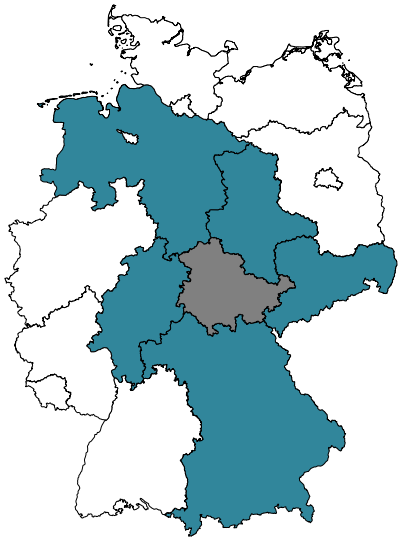


Abbildung 5.3.: Nachbarschaft erster Ordnung auf einem irregulären Gitter.

Analog zu W können dann Nachbarschaftsmatrizen $W^{(1)}, W^{(2)}$, usw. gebildet werden, welche die ersten bzw. zweiten Nachbarn enthalten.

5.2. Tests auf räumliche Autokorrelation

Bei der Analyse räumlicher Daten wird meist angenommen, dass sich Daten ähnlicher sind, je näher sich ihre räumlichen Lokationen sind. Dabei spricht man von positiver räumlicher Autokorrelation, also der Korrelation der Variable mit sich selbst. Räumliche Autokorrelation bezieht sich somit auf die Korrelation zwischen $Z(s_i)$ und $Z(s_j)$ zweier Punkte s_i und s_j. Zur Messung räumlicher Autokorrelation werden also zwei Informationen in Verbindung gebracht: Die Ähnlichkeit der Beobachtungen und die Ähnlichkeit der Lokationen.
In der Literatur wird zwischen globalen und lokalen Maßzahlen unterschieden. Globale Berechnungen fassen die räumlichen Abhängigkeiten über alle Daten zusammen, wohingegen lokale Statistiken (engl. *Local indicators of spatial association - LISA*) angeben in welchem Ausmaß die Anordnung der Werte um eine spezifische Lokation von räumlichem Zufall abweicht (Anselin et al. 2000).
Die bekanntesten Maßzahlen globaler, räumlicher Autokorrelation sind *Moran's I* und *Geary's c*. Beide können als Anpassung des Kreuzproduktes (vgl. Anselin (1995))

$$\sum_{i=1}^{n}\sum_{j=1}^{n} m_{ij}w_{ij}$$

ausgedrückt werden. Dabei enstpricht w_{ij} der Ähnlichkeit der Lokationen i und j (vgl. zuvor definierte Gewichte in der Nachbarschaftsmatrix) und m_{ij} der Ähnlichkeit der Beobachtung an den Stellen i und j.
Die beiden Maßzahlen unterscheiden sich in ihrer Definition von der Ähnlichkeit der Werte, also von m_{ij}. Moran's I basiert auf dem Produkt $(z_i-\bar{z})(z_j-\bar{z})$, wohingegen Geary's c die quadrierte Differenz $(z_i - z_j)^2$ verwendet.
Damit ergeben sich die Gleichungen (vgl. (Cliff & Ord 1981))

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j\neq i}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(z_i-\bar{z})(z_j-\bar{z})}{\sum_{i=1}^{n}(z_i-\bar{z})^2}$$

für Moran's I und

$$I = \frac{n-1}{2\sum_{i=1}^{n}\sum_{j\neq i}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(z_i-z_j)^2}{\sum_{i=1}^{n}(z_i-\bar{z})^2}$$

für Geary's c.
Diese Statistiken sind nicht direkt interpretierbar. Sie sind beide asymptotisch normalverteilt. Zum Testen eignen sich aber besser permutationsbasierte Tests, in denen die Beobachtungen zufällig den Lokationen zugewiesen werden.
Analog zum Semivariogramm in der Geostatistik lässt sich die räumliche Autokorrelation als Funktion der Distanz betrachten. Dazu wird die gewählte Statistik zur Messung räumlicher Autokorrelation, z.B. I, für jede Distanzklasse berechnet. Dies entspricht also der Berechnung von I auf Basis der Nachbarschaftsmatrizen $W^{(1)}, \ldots, W^{(q)}$ für die ersten bis q-ten Nachbarn.

Danach lassen sich die I_d gegen die Distanz d plotten.
Eine Annahme der Tests ist, dass kein systematischer Trend in den Daten vorliegt, bzw. dass dieser durch das Modell berücksichtigt wird. Eine Missspezifikation des Erwartungswertes kann unter Abwesenheit einer räumlichen Korrelation zu einer fälschlichen Signifikanz der Teststatistik führen. Es sollte somit zunächst von Priorität sein, dass alle relevanten Kovariablen in das Modell aufgenommen werden und auch deren funktionaler Einfluss richtig spezifiziert ist. Lokale Maßzahlen werden innerhalb dieser Arbeit nicht näher betrachtet, können aber beispielsweise in Waller & Gotway (2004) nachgelesen werden.

5.3. Penalisiertes KQ-Kriterium

Eine intuitive Möglichkeit diskrete, räumliche Information in ein Regressionsmodell mit aufzunehmen, stellen Fahrmeir et al. (2009, Kapitel 7) vor. Dabei wird jeder Region s ein eigener Koeffizient $f_{geo}(s) = \delta_s, s = 1, \dots, d$ zugewiesen. Um einen glatten Effekt zu erzielen und die sich dadurch ergebende, hohe Anzahl an Parametern zu verringern, wird analog zur Theorie der Splines, ein penalisiertes KQ-Kriterium eingeführt. Koeffizienten benachbarter Regionen sollten sich nicht allzu stark voneinander unterscheiden. Deshalb konstruieren Fahrmeir et al. (2009) einen Strafterm basierend auf den quadrierten Differenzen zwischen Parametern benachbarter Regionen, also

$$PKQ(\lambda) = \sum_{i=1}^{n}(y_i - f_{geo}(s_i))^2 + \lambda \sum_{s=2}^{d} \sum_{r \in N(s), r<s} (\delta_r - \delta_s)^2,$$

wobei $N(s)$, wie zuvor definiert, der Menge aller Nachbarn der Region s entspricht. Der Strafterm lässt sich auch umschreiben in $\lambda \boldsymbol{\delta}' \boldsymbol{K} \boldsymbol{\delta}$ mit

$$\boldsymbol{K}[s,r] = \begin{cases} -1, & s \neq r, s \sim r, \\ 0, & s \neq r, s \nsim r, \\ |N(s)|, & s = r. \end{cases} \tag{5.1}$$

Bei der Matrix $\boldsymbol{K}$ handelt es sich um eine Adjazenz- bzw. Nachbarschaftsmatrix, deren Einträge $\boldsymbol{K}[s,r]$ nur dann von Null verschieden sind, wenn s und r Nachbarn sind.
Werden Gewichte verwendet ändert sich $\boldsymbol{K}$ zu

$$\boldsymbol{K}[s,r] = \begin{cases} -w_{rs}, & s \neq r, s \sim r, \\ 0, & s \neq r, s \nsim r, \\ w_{s+}, & s = r, \end{cases} \tag{5.2}$$

mit den symmetrischen Gewichten $w_{sr} = w_{rs}$ und $w_{s+} = \sum_{r:r\sim s} w_{sr}$.

5.4. Markov-Zufallsfelder

Eine in der Literatur weiter verbreitete Herangehensweise ist die Bayesianische Modellformulierung. Diese führt zu den sogenannten *Markov-Zufallsfeldern* (MZF).
Markov Zufallsfelder definieren sich über die räumliche Markov-Eigenschaft, also die Gedächtnislosigkeit innerhalb eines ungerichteten Graphen. D.h. die bedingte Verteilung von δ_s gegeben alle übrigen Effekte $\delta_r, r \neq s$ hängt nur von den Nachbarn ab.

Definition: Markov-Zufallsfeld (MZF)

Sei $D = 1, \ldots, d$ die Menge aller Regionen, dann heißt $\boldsymbol{\delta} = \{\delta_s, s \in D\}$ *Markov-Zufallsfeld*, wenn die bedingte Verteilung von δ_s gegeben alle übrigen Effekte $\delta_r, r \neq s$ nur von den Nachbarn abhängt, d.h. wenn gilt:

$$p(\delta_s|\delta_r, r \neq s) = p(\delta_s|\delta_r, r \in N(s))$$

Eines der gebräuchlichsten Modelle in Bezug auf MZF zur Beschreibung räumlicher Effekte ist das sogenannte *Conditional Autoregressive Model (CAR)* oder auch *Gauß-Markov-Zufallsfeld (GMFZ)*.
Dieses wurde ursprünglich von Besag (1974) eingeführt und wird in Rue & Held (2005) ausführlich besprochen. Die bedingten Verteilungen entsprechen hier, wie der Name vermuten lässt, einer Normalverteilung.
Das übliche Vorgehen bei der Modellspezifikation ist die Vorgabe der bedingten Verteilungen $\delta_s|\delta_r, r \in N(s)$ und daraus die Herleitung der gemeinsamen Verteilung des Vektors $\boldsymbol{\delta}$. Da nicht jede Spezifikation zwingend zu einer gültigen gemeinsamen Verteilung führt, müssen zunächst einige theoretische Überlegungen angestellt werden. Nach Fahrmeir et al. (2009) ergibt sich das CAR-Modell jedoch direkt aus dem zuvor betrachteten penalisierten KQ-Kriterium. Das Modell lässt sich beschreiben durch

$$\delta_s|\delta_r, r \in N(s) \sim N\left(\sum_{r:r\sim s} \frac{1}{|N(s)|}\delta_r, \frac{\tau^2}{|N(s)|}\right).$$

Inhaltlich entspricht dieses Modell der gewünschten Eigenschaft, dass der Effekt einer Region s ähnlich zu denen ihren Nachbarn ist. Dabei steuert die Varianz τ^2, wie stark der einzelne Effekt vom Mittelwert der benachbarten Regionen abweichen darf.
Die gemeinsame Verteilung ist ebenfalls eine Normalverteilung mit

$$\boldsymbol{\delta} \sim N(0, \tau^2 \boldsymbol{K}^{-1}),$$

wobei die Präzisionsmatrix $\boldsymbol{K}$ der Strafmatrix (vgl. 5.1) aus dem PKQ-Kriterium entspricht, sodass beide Herangehensweisen zur äquivalenten Modellformulierung führen.

Wie zuvor können auch hier Gewichte für den Einfluss der Nachbarn eingeführt werden. Die bedingten Dichten werden dann erweitert zu

$$\delta_s|\delta_r, r \in N(s) \sim N\left(\frac{w_{sr}}{w_{s+}} \sum_{r:r\sim s} \delta_r, \frac{\tau^2}{w_{s+}}\right)$$

und die Matrix $\boldsymbol{K}$ gemäß 5.2.

6. Disease Mapping

Ein wichtiges und häufig angewendetes Gebiet der räumlichen Statistik ist das sogenannte *Disease Mapping*. Dabei werden Regionen mit erhöhtem Risiko einer Krankheit identifiziert. Die räumliche Verteilung der Krankheit kann zur Entdeckung bisher unbekannter Risikofaktoren führen.
Dieses Kapitel setzt sich mit der Analyse von Daten auseinander, welche in Bezug auf bestimmte Verwaltungsregionen aggregiert wurden. Im eigentlichen Kontext des Disease Mapping handelt es sich dabei um Krankheits- bzw. Sterbefälle innerhalb einer Region. Die vorgestellten Verfahren lassen sich jedoch auch auf andere Anwendungsgebiete übertragen (hier: Anzahl der Studenten an der LMU mit Herkunft aus einer bestimmten Region).
Im Gegensatz zu Kapitel 5 sind nun also Anzahlen von Interesse. Das Ziel der Analyse ist die Darstellung der räumlichen Verteilung des Risikos einer Krankheit über das in Regionen eingeteilte Untersuchungsgebiet. Zu diesem Zweck müssen die Daten die Personen unter Risiko und die aufgetretenen Fälle jeder Region enthalten. Im Folgenden bezeichnet P_i die Risikopopulation (engl. Population at risk) und O_i die Anzahl beobachteter Fälle (engl. observed cases) in der Region i. Die Notation orientiert sich damit an Bivand et al. (2013). Sämtliche Herleitungen finden sich außerdem in Banerjee et al. (2004).
Die Daten sind in vielen Fällen zusätzlich in verschiedene Strata unterteilt. Diese unterscheiden sich beispielsweise in Geschlecht oder Alter. Die Notation erweitert sich dann zu P_{ij} bzw. O_{ij} für die Population und die Anzahl der Fälle in Region i und Stratum j. Aufsummieren über die einzelnen Strata pro Region führt dann zu P_i und O_i. Die Gesamtpopulation und die gesamte Anzahl der Fälle werden mit P_+ bzw. O_+ bezeichnet.
Um eine Schätzung des Risikos zu erlangen, müssen die beobachteten Fälle mit einer *erwarteten* Anzahl an Fällen verglichen werden. Diese kann aus

$$E_i = P_i r_+,$$

mit $r_+ = \frac{O_+}{P_+}$, berechnet werden. r_+ entspricht also der Gesamt-Inzidenzrate.
Bei gruppierten Daten kann ähnlich vorgegangen werden. Für jedes Stratum j wird eine eigene Inzidenzrate $r_j = \frac{\sum_i O_{ij}}{\sum_i P_{ij}}$ berechnet. In diesem Fall ergibt sich die erwartete Anzahl an Fällen in Region i aus

$$E_i = \sum_j P_{ij} r_j.$$

Dieses Vorgehen wird als interne Standardisierung bezeichnet, da die beobachteten Daten ver-

wendet werden um die Referenzrate(n) zu berechnen, ohne dass dies berücksichtigt wird. Besser ist die externe Standardisierung, bei der bereits aus anderen Quellen bekannte Tabellen verwendet werden.

6.1. Traditionelle Modelle

Eine häufig verwendete Annahme der Statistik in Bezug auf Anzahlen ist die der Poissonverteilung. Im hier vorliegenden Kontext bedeutet dies, dass gilt

$$O_i|\theta_i \sim Po(E_i\theta_i).$$

Es wird also angenommen, dass die Anzahl der beobachteten Fälle in Region i poissonverteilt mit Erwartungswert $E_i\theta_i$ ist. Dabei entspricht θ_i dem wahren relativen Risiko in Region i.
Der Maximum-Likelihood-Schätzer von θ_i ergibt sich aus

$$\hat{\theta}_i \equiv SIR_i = \frac{O_i}{E_i}.$$

SIR_i wird als *Standardized Incidence Ratio* bezeichnet. Im Kontext des Disease Mapping wird dieses Verhältnis meist auch SMR (*Standardized Mortality Ratio*) genannt. Eine genauere Betrachtung dieser und anderer Normierungen findet sich in Waller & Gotway (2004).
Zur Visualisierung der räumlichen Variation des Risikos kann somit das SIR auf einer Karte abgetragen werden. Problematisch an dieser Vorgehensweise ist jedoch, dass die Standardabweichung der Schätzers $sd(SIR_i) = \sqrt{O_i}/E_i$ und somit proportional zu $1/E_i$ ist. Bei einer geringen Anzahl an erwarteten Fällen wird die Schätzung also sehr unsicher. Die Identifizierung einer Region als stark risikobehaftet könnte somit lediglich an ihrer geringen Einwohnerzahl liegen. Außerdem werden in dieser Analyse möglicherweise vorhandene räumliche Korrelationen nicht berücksichtigt.
Eine zweite Möglichkeit der Visualisierung ergibt sich aus der Betrachtung von p-Werten. Werte von SIR größer als 1 weisen darauf hin, dass mehr Fälle beobachtet, als in der Untersuchungspopulation erwartet wurden. Interessant für die Analyse ist somit die Hypothese

$$H_0 : \theta = 1 \quad \text{vs.} \quad H_1 : \theta > 1.$$

Unter der Nullhypothese gilt $O_i \sim Po(E_i)$, sodass sich der p-Wert für diesen Test aus (Banerjee et al. 2004)

$$P(X \leq O_i|E_i) = 1 - P(X < O_i/E_i) = 1 - \sum_{x=0}^{Y_i-1} \frac{E_i^x}{x!} e^{-E_i}$$

ergibt. Wird die Nullhypothese verworfen, so kann von einem signifikant erhöhtem Risiko in Region i ausgegangen werden. Alternativ können Konfidenzintervalle mit Hilfe der Poisson-Verteilung für SIR berechnet werden.

Wie zuvor ergeben sich auch hier die zwei genannten Nachteile: die p-Werte hängen von der erwarteten Anzahl an Fällen ab und mögliche räumliche Korrelationen werden nicht berücksichtigt. Um dem ersten Problem begegnen zu können, wurde von Clayton & Kaldor (1987) ein Verfahren vorgestellt, welches die SIR hin zu einem globalen Mittelwert schrumpft. Ausgangspunkt hierfür ist eine Annahme der Poissonverteilung, die in vielen Fällen angezweifelt werden muss: die Gleichheit von Erwartungswert und Varianz. Häufig liegt in realen Daten Überdispersion vor, d.h. die Varianz der Daten ist größer als ihr Erwartungswert. Ein Möglichkeit dies zu beachten ist die Verwendung der Negativen Binomialverteilung anstatt der Poissonverteilung (Bivand et al. 2013).
Diese lässt sich als gemischtes Modell formulieren. Dabei wird ein Zufallseffekt für jede Region angenommen, der einer Gamma-Verteilung mit Erwartungswert $\frac{\nu}{\alpha}$ und Varianz $\frac{\nu}{\alpha^2}$ folgt.
Das sogenannte *Poisson-Gamma*-Modell lässt sich also formulieren als

$$\begin{aligned} O_i|\theta_i, E_i &\sim Po(\theta_i E_i) \\ \theta_i &\sim Ga(\nu, \alpha) \end{aligned}$$

Die beobachteten Fälle O_i sind bedingt auf θ_i poissonverteilt mit Erwartungswert $\theta_i E_i$. Die O_i selbst sind somit negativ binomial-verteilt mit (Clayton & Kaldor 1987)

$$\begin{aligned} E(O_i) &= E_i \frac{\nu}{\alpha} \\ Var(O_i) &= E_i \frac{\nu}{\alpha} + E_i^2 \frac{\nu}{\alpha^2} \end{aligned}$$

Aufgrund der Konjugiertheit der Gamma-Priori zur Poisson-Likelihood ergibt sich für die Posteriori von θ_i wieder eine Gamma-Verteilung mit den Parametern $\nu + O_i$ und $\alpha + E_i$.
Der Posteriori-Erwartungswert von θ_i ist

$$\begin{aligned} E(\theta_i|O_i, E_i) &= \frac{\nu + O_i}{\alpha + E_i} \\ &= \frac{\alpha}{\alpha + E_i} \cdot \frac{\nu}{\alpha} + \frac{E_i}{\alpha + E_i} \cdot \frac{O_i}{E_i} \\ &= \left(1 - \frac{E_i}{\alpha + E_i}\right) \cdot \frac{\nu}{\alpha} + \frac{E_i}{\alpha + E_i} \cdot SMR_i. \end{aligned}$$

Somit ist der Punktschätzer ein gewichtetes Mittel aus dem datenbasierten SIR von Region i und dem Priori-Erwartungswert des relativen Risikos θ_i. Für Regionen mit kleinem E_i hat SMR_i also ein geringes Gewicht im Gegensatz zum Priori-Erwartungswert.
Da ν und α für alle Regionen gleich sind, wird Information von diesen geliehen um die Posteriori-Schätzer zu konstruieren. Dieses Konzept wird *borrowing strength* genannt.

6.2. Räumliche Modelle

Bisher wurden mögliche räumliche Effekte aus der Analyse außen vor gelassen. Dabei ist zu beachten, dass meist nicht die Zugehörigkeit zu einer Region selbst einen Effekt auf die abhängige Variable hat, sondern unbeobachtete, nicht durch die Daten erfasste Kovariablen mit räumlicher Struktur berücksichtigt werden sollen. Die räumliche Analyse kann somit auch Hinweise auf bisher unbekannte Risikofaktoren geben. Diese können eine räumliche Struktur aufweisen, oder nur lokal auftreten. Da in der Regel nicht bekannt ist, ob Einflussfaktoren eine räumliche Struktur mit sich bringen, schlugen Besag et al. (1991) ein Modell vor, welches sowohl strukturierte (räumlich korrelierte) als auch unstrukturierte (räumlich unkorrelierte) Effekte berücksichtigt, d.h. $f_{spat} = f_{str} + f_{unstr}$ (vgl. Fahrmeir et al. (2004)).

Besag, York, Mollie (BYM, 1991)

$$O_i|\theta_i \sim Po(\theta_i E_i)$$
$$\theta_i = \exp(\eta_i) = \exp(\beta_0 + f_{i,geo}(s_i) + b_i)$$

Dabei wird für den räumlich strukturierten Anteil f_{geo} ein GMZF (vgl. Kapitel 5.4) angenommen während b_i ein regionenspezifischer, zufälliger Effekt mit $b_i \overset{iid}{\sim} N(0, \nu^2)$ ist.
Damit ergibt sich für O_i ein log-lineares Poisson-Modell mit dem linearen Prädiktor

$$\eta_i = \beta_0 + f_{i,geo}(s_i) + b_i + \log(E_i),$$

wobei $\log(E_i)$ dem Offset entspricht. Dieser dient der Vergleichbarkeit der einzelnen Regionen.

7. Geoadditive Modelle - BayesX

In den meisten Anwendungen wird kein rein räumlicher Effekt geschätzt, sondern es liegen zusätzliche Kovariablen zur Erklärung der Zielvariablen vor. Die enstehende Modellklasse wird unter dem Begriff *Geoadditive Regression* geführt. Der Prädiktor des (generalisierten) additiven Modells η_i^{add} bestehend aus nichtparametrischen, glatten Funktionen und linearen Effekten wird in diesem Fall um einen räumlichen Effekt f_{geo} erweitert, d.h.

$$\eta_i = \eta_i^{add} + f_{geo}(s_i) = f_1(z_{i1}) + \ldots + f_q(z_{iq}) + f_{geo}(s_i) + \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}.$$

Liegt also, zusätzlich zu den Werten der Zielvariablen und metrischen oder kategorialen Kovariablen, zu jeder Beobachtung i eine räumliche Information, in Form einer Lokationsvariablen s_i vor, handelt es sich um eine Problemstellung der *geoadditiven Regression*.
Noch allgemeiner formulieren Fahrmeir et al. (2009) die Klasse der *(generalisierten) strukturiert-additiven Regressions (STAR) Modelle*. Für den strukturiert-additiven Prädiktor

$$\eta_i^{strukt} = f_1(v_{i1}) + \ldots + f_q(v_{iq}) + \boldsymbol{x}_i'\boldsymbol{\beta}$$

sind verschiedene Kombinationsmöglichkeiten von Funktionen mit unterschiedlicher Struktur möglich. Als Beispiele sind in Fahrmeir et al. (2009) folgende Typen genannt:

$f_1(v_1) = f_1(z_1)$,	$v_1 = z_1$,	nichtlinearer Effekt von z_1
$f_2(v_2) = f_{geo}(s)$,	$v_2 = s$,	räumlicher Effekt der Lokationsvariablen s.
$f_3(v_3) = \gamma_i u$,	$v_3 = (u, i)$,	individuenspezifischer zufälliger Effekt von u.
$f_4(v_4) = f(z)x$,	$v_4 = (z, x)$,	mit z variierender Effekt von x.
$f_5(v_5) = f_{1\|2}(z_1, z_2)$,	$v_5 = (z_1, z_2)$,	nichtlineare Interaktion zwischen z_1 und z_2.

Ein mächtiges Programmpaket für die Schätzung solcher Modelle stellt **BayesX** (Belitz et al. 2015) dar. Die Funktionalitäten dieses Programms stehen dem User über ein eigenständiges Programm oder aber auch über die Schnittstelle des Pakets **R2BayesX** (Umlauf et al. 2015) in R zur Verfügung.
Die Schätzung der Parameter ist in **BayesX** über drei unterschiedliche Inferenzkonzepte möglich:

- Volle Bayes-Inferenz basierend auf MCMC-Simulationstechniken
- Inferenz basierend auf der Repräsentation als gemischtes Modell
- Penalisierte Likelihood-Schätzung inklusive Variablenselektion

Für diese Arbeit wurde auf die Schätzung über die Repräsentation als gemischtes Modell zurückgegriffen. Die Grundidee dieses Ansatzes soll im Folgenden basierend auf Fahrmeir et al. (2004) und Fahrmeir et al. (2009) vorgestellt werden. Dort finden sich auch Informationen über die volle Bayes-Inferenz. Andere Quellen für den Einstieg in die hier nicht angesprochenen Ansätze bieten beispielsweise Brezger & Lang (2006) (Volle Bayes-Inferenz) und Belitz & Lang (2008) (Variablenselektion).
Falls die Funktionen $f_1, \ldots, f_q$ durch Basisfunktionenansätze modelliert werden, kann der strukturierte Prädiktor in Matrixform folgendermaßen dargestellt werden

$$\boldsymbol{\eta}^{strukt} = \boldsymbol{V}_1\boldsymbol{\gamma}_q + \ldots + \boldsymbol{V}_q\boldsymbol{\gamma}_q + \boldsymbol{X}\boldsymbol{\beta}. \tag{7.1}$$

Dabei entsprechen die $\boldsymbol{V}_j$ geeignet definierten Designmatrizen und $\boldsymbol{\gamma}_j$ den Koeffizientenvektoren. Ein Überblick zu den in der Arbeit verwendeten Termen findet sich in Tabelle 7.1.
Die generelle Form der Priori für $\boldsymbol{\gamma}_j$ ist

$$p(\boldsymbol{\gamma}_j|\tau_j^2) \propto exp\left(-\frac{1}{2\tau_j^2}\boldsymbol{\gamma}_j'\boldsymbol{K}_j\boldsymbol{\gamma}_j\right),$$

wobei $\boldsymbol{K}_j$ der jeweiligen Strafmatrix entspricht. In den meisten Fällen hat diese keinen vollen Rang, sodass die Priori teilweise uneigentlich ist, d.h. es gilt $\text{rg}(\boldsymbol{K}_j) > 0$, aber nicht $\text{rg}(\boldsymbol{K}_j) = \dim(\boldsymbol{\gamma}_j)$).
Um den strukturiert-additiven Prädiktor (7.1) als GLMM darzustellen werden die Regressionskoeffizienten $\boldsymbol{\gamma_j}, j = 1, \ldots, p$ in einen penalisierten und einen nicht penalisierten Teil zerlegt.
Bezeichne im Folgenden $d_j = \dim(\boldsymbol{\gamma}_j)$ die Dimension des j-ten Koeffizientenvektors und $r_j = \text{rg}(\boldsymbol{K}_j)$ den Rang der korrespondierenden Strafmatrix. Dann definieren Fahrmeir et al. (2004) die Zerlegung

$$\boldsymbol{\gamma}_j = \boldsymbol{V}_j^{unp}\boldsymbol{\gamma}_j^{unp} + \boldsymbol{V}_j^{pen}\boldsymbol{\gamma}_j^{pen},$$

mit den $d_j \times (d_j - r_j)$ bzw. $d_j \times r_j$ dimensionalen Designmatrizen $\boldsymbol{V}_j^{unp}$ und $\boldsymbol{V}_j^{pen}$.
Durch eine geeignete Wahl der Designmatrizen (genauere Betrachtung siehe Fahrmeir et al. (2004)) kann errreicht werden, dass der Parametervektor $\boldsymbol{\gamma}_j^{unp}$ als Vektor fester Effekte und $\boldsymbol{\gamma}_j^{pen} \sim \text{N}(\boldsymbol{0}, \tau_j^2\boldsymbol{I})$ als Vektor zufälliger Effekte aufgefasst werden kann.
Der Prädiktor (7.1) lässt sich damit umschreiben in

$$\boldsymbol{\eta}^{strukt} = \sum_{j=1}^{q}\boldsymbol{V}_j\boldsymbol{\gamma}_j + \boldsymbol{X}\boldsymbol{\beta} = \sum_{j=1}^{q}(\boldsymbol{V}_j\boldsymbol{V}_j^{unp}\boldsymbol{\gamma}_j^{unp} + \boldsymbol{V}_j\boldsymbol{V}_j^{pen}\boldsymbol{\gamma}_j^{pen}) + \boldsymbol{X}\boldsymbol{\beta} = \tilde{\boldsymbol{X}}\boldsymbol{\gamma}^{unp} + \tilde{\boldsymbol{V}}\boldsymbol{\gamma}^{pen},$$

wobei $\tilde{\boldsymbol{X}}_j = \boldsymbol{V}_j\boldsymbol{V}_j^{unp}$ und $\tilde{\boldsymbol{V}}_j = \boldsymbol{V}_j\boldsymbol{V}_j^{unp}$.

Die Designmatrizen und Vektoren sind dabei wie folgt zusammengesetzt

$$\begin{aligned}\tilde{\boldsymbol{V}} &= (\tilde{\boldsymbol{V}}_1\tilde{\boldsymbol{V}}_2\cdots\tilde{\boldsymbol{V}}_q)\\ \boldsymbol{\gamma}^{pen} &= ((\boldsymbol{\gamma}_1^{pen})',\ldots,(\boldsymbol{\gamma}_q^{pen})')'\\ \tilde{\boldsymbol{U}} &= (\tilde{\boldsymbol{U}}_1\tilde{\boldsymbol{U}}_2\cdots\tilde{\boldsymbol{U}}_p\boldsymbol{U})\\ \boldsymbol{\gamma}^{unp} &= ((\boldsymbol{\gamma}_1^{unp})',\ldots,(\boldsymbol{\gamma}_q^{unp})'\boldsymbol{\beta}')'\end{aligned}$$

Dadurch ergibt sich ein GLMM mit festen Effekten $\boldsymbol{\gamma}^{unp}$ und zufälligen Effekten $\boldsymbol{\gamma}^{pen} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Lambda})$, wobei $\boldsymbol{\Lambda} = \mathrm{diag}(\tau_1^2,\ldots,\tau_1^2,\ldots,\tau_q^2,\ldots,\tau_q^2)$, sodass sich die üblichen Schätzverfahren dieser Methode verwenden lassen.

Termtyp	Designmatrix $\boldsymbol{V}$	Strafmatrix $\boldsymbol{K}$
P-Spline	Basisfunktionen ausgewertet an den Beobachtungen.	$\boldsymbol{K} = \boldsymbol{D}_k'\boldsymbol{D}_k$, mit $\boldsymbol{D}_k$ Differenzenmatrix k-ter Ordnung
2D-P-Spline	2D-Basisfunktionen ausgewertet an den Beobachtungen.	$\boldsymbol{K} = \boldsymbol{I}\otimes\boldsymbol{K}_1 + \boldsymbol{K}_2\otimes\boldsymbol{I}$ mit Einheitsmatrix $\boldsymbol{I}$ und Strafmatrizen $\boldsymbol{K}_1$ und $\boldsymbol{K}_2$ wie für univariate P-Splines.
Kriging	Auf der Korrelationsfunktion basierende Basisfunktionen.	$\boldsymbol{K} = \boldsymbol{R}$, mit Korrelationsmatrix $\boldsymbol{R}$.
Markov-Zufallsfeld	0/1 Inzidenzmatrix, die Beobachtungen und Regionen verknüpft.	$\boldsymbol{K}$ = Nachbarschaftsmatrix.
Zufällige Konstante	0/1 Inzidenzmatrix, die Beobachtungen und Cluster verknüpft	$\boldsymbol{K} = \boldsymbol{I}$, mit Einheitsmatrix $\boldsymbol{K} = \boldsymbol{I}$.

Tabelle 7.1.: Übersicht über verwendete Modellterme mit zugehöriger Design- und Strafmatrix; Quelle: in Anlehnung an Fahrmeir et al. (2009, Tab. 8.2)

8. Auswertung

8.1. Beispiel: Phänologie

8.1.1. Deskriptive Analyse

Zur Analyse stehen, außer dem gemessenen relativen Grünwert, lediglich die Koordinaten der jeweiligen Webcam, sowie der Tag der Messung zur Verfügung. Es liegen 12864 Beobachtungen verteilt auf 182 Stationen und 73 Tage des Jahres 2011 vor.

Rel. Grünwerte

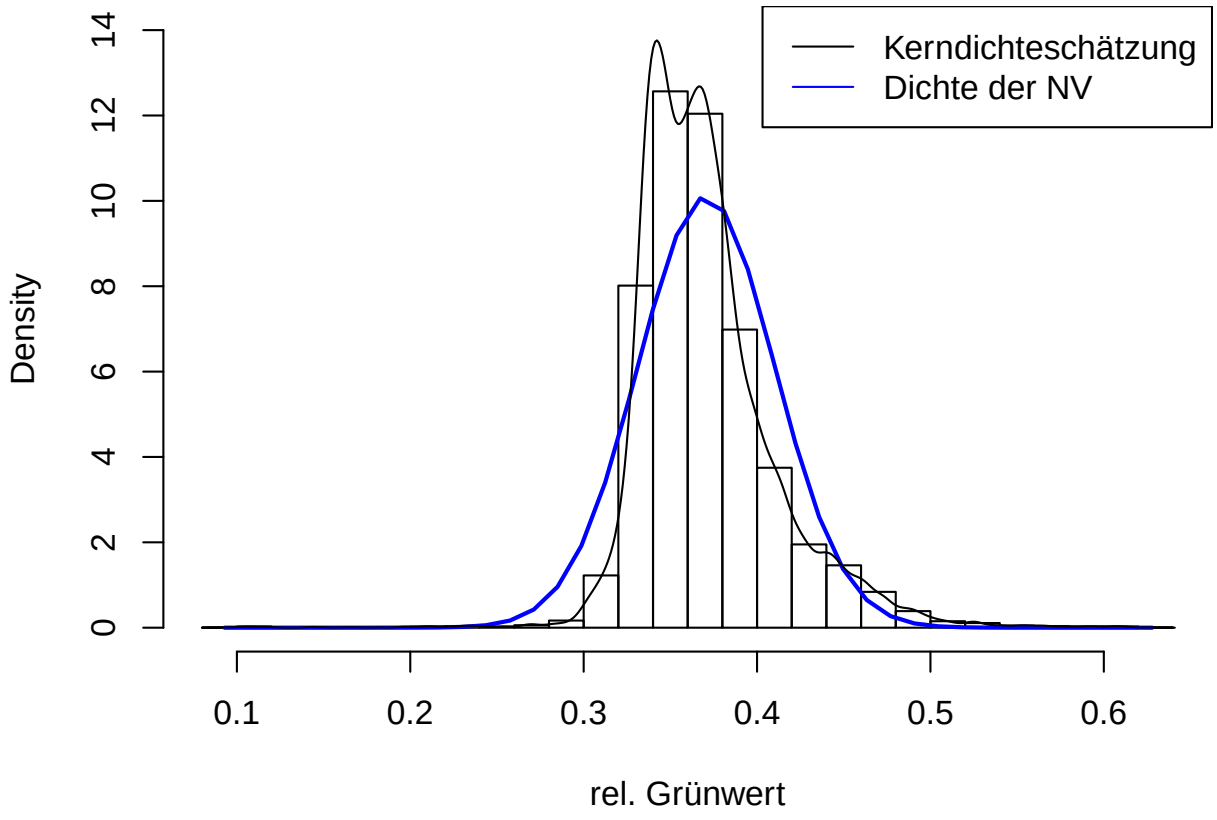


Abbildung 8.1.: Histogramm und Kerndichteschätzung.

Die relativen Grünwerte schwanken zwischen 0.09 und 0.63, wobei 75% der Daten zwischen 0.33 und 0.41 liegen. Es liegt eine schwach linkssteile Verteilung der Werte vor, die jedoch nicht stark von der Normalverteilung abweicht (vgl. Abbildung 8.1).

Web-Cams

Die ausgewählten Web-Cams liegen alle innerhalb Deutschlands zwischen dem 6. und 15. Längen- und dem 47. und 54. Breitengrad. Wie man auf der Karte in Abbildung 8.2 erkennen kann, stehen für den Norden Deutschlands nur sehr wenige Web-Cams zur Verfügung. In den Alpen, im Bayrischen Wald, am Bodensee, sowie in der Schwäbischen Alp und dem Schwarzwald liegen hingegen vergleichsweise viele Messstellen vor. Insgesamt scheinen die Webcams vor allem in größeren Höhenlagen oder um Seen und große Städte angesiedelt zu sein. Da das Klima und somit die Phänologie vermutlich mit der Höhe der Messstation in Zusammenhang steht, wäre diese Variable für die Analyse von hohem Interesse. Diese steht jedoch nicht zur Verfügung.

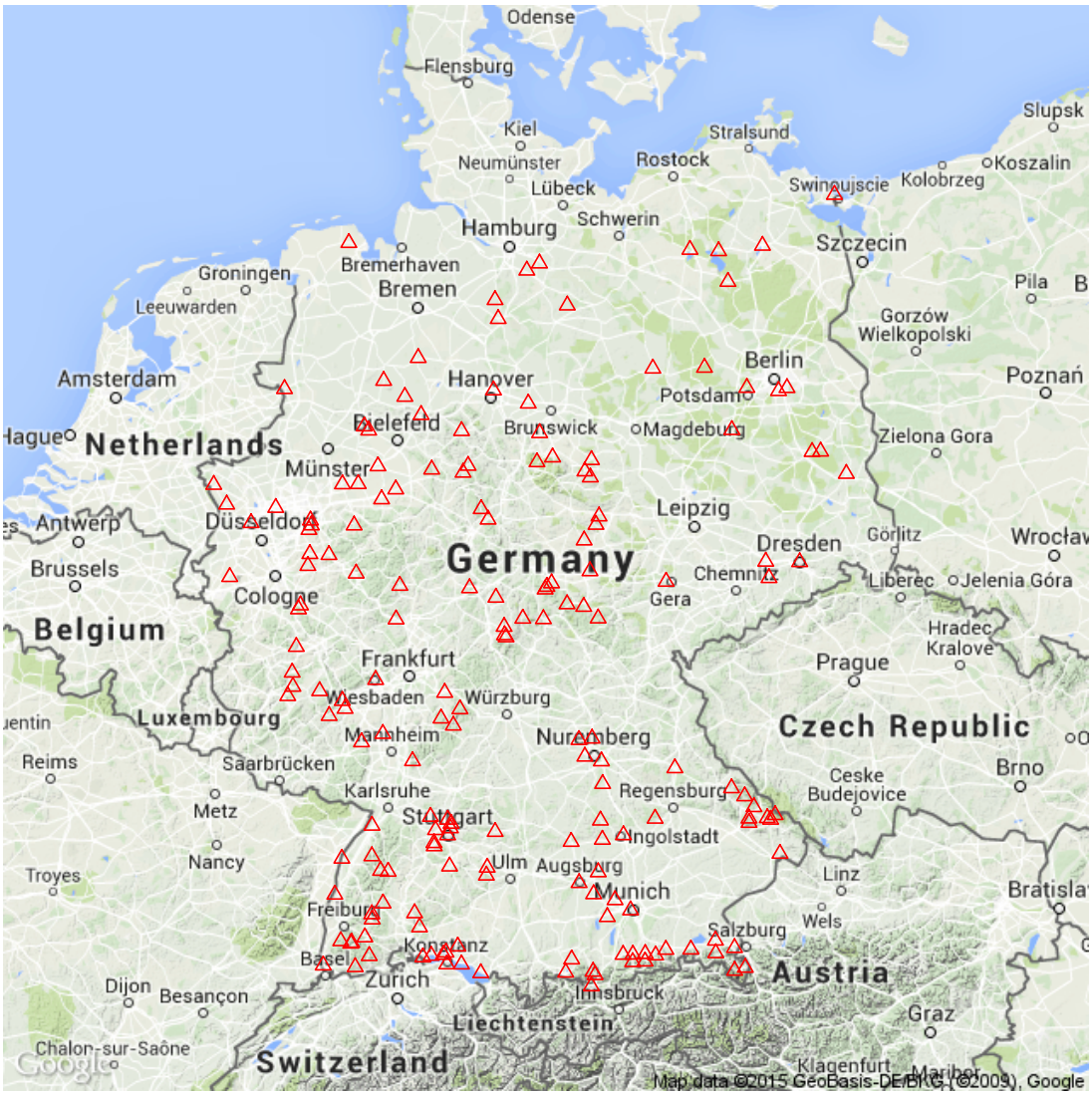


Abbildung 8.2.: ausgewählte Web-Cams in Deutschland.

Zeitpunkte

Für die Analyse stehen die Tage 84 bis 159 des Jahres 2011 zur Verfügung. Zwischen Tag 141 und 143 kam es zu einem Ausfall des Servers, sodass diese im Datensatz fehlen. Über die Zeit hinweg, lässt sich ein Anstieg in den aggregierten relativen Grünwerten erkennen (vgl. Abbildung 8.3).

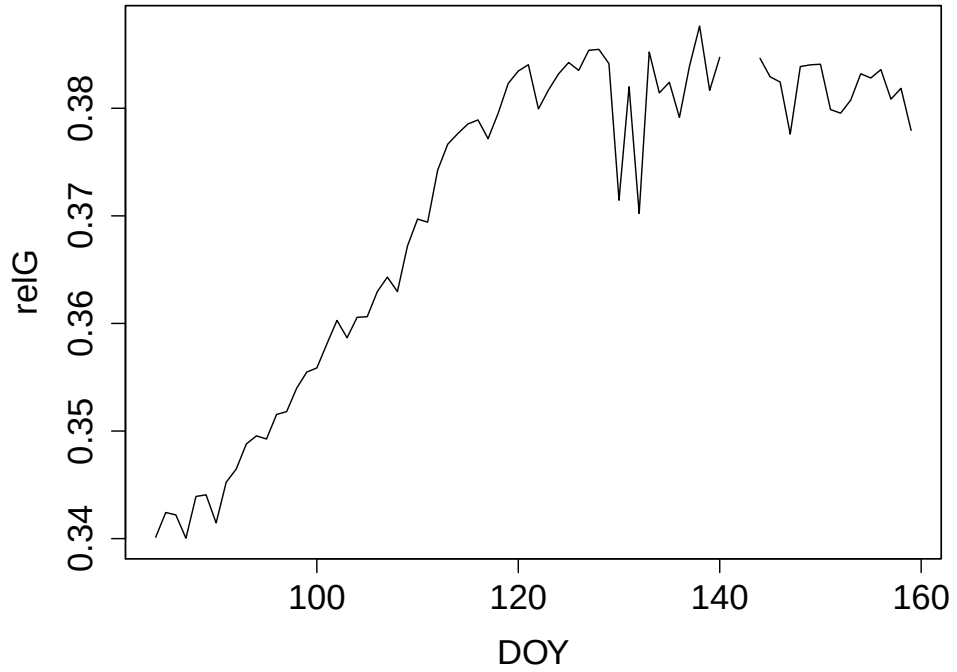


Abbildung 8.3.: Zeitreihe von DOY 84 bis 159.

Verlauf für einzelne Stationen

Abbildung 8.4 zeigt die Verläufe der einzelnen Stationen. Um einen besseren Vergleich zu ermöglichen, wurde der Bereich der rel. Grünwerte auf 0.3 bis 0.5 beschränkt. Es sind dadurch nur weniger als 2% der Daten nicht sichtbar.
Es zeigt sich auch einzeln betrachtet für die meisten Messstationen ein Anstieg in den relativen Grünwerten. Um den Vergleich der Verläufe in Bezug auf SOS und MAT zu erleichtern, wurden die Daten skaliert und in Abbildung 8.5 abgetragen. Es lässt sich für die meisten Stationen die gleiche funktionale Form im Verlauf der Grünwerte erkennen. Der Grünwert liegt auf einem Grundniveau, bis er ab dem Erscheinen der ersten Blätter (SOS) bis zur vollständigen Reife des Laubs (MAT) auf einen Maximalwert ansteigt. Es zeigen sich jedoch Phasenverschiebungen, Veränderungen in der Differenz zwischen SOS und MAT sowie in der Amplitude (vgl. Abbildung 8.6).

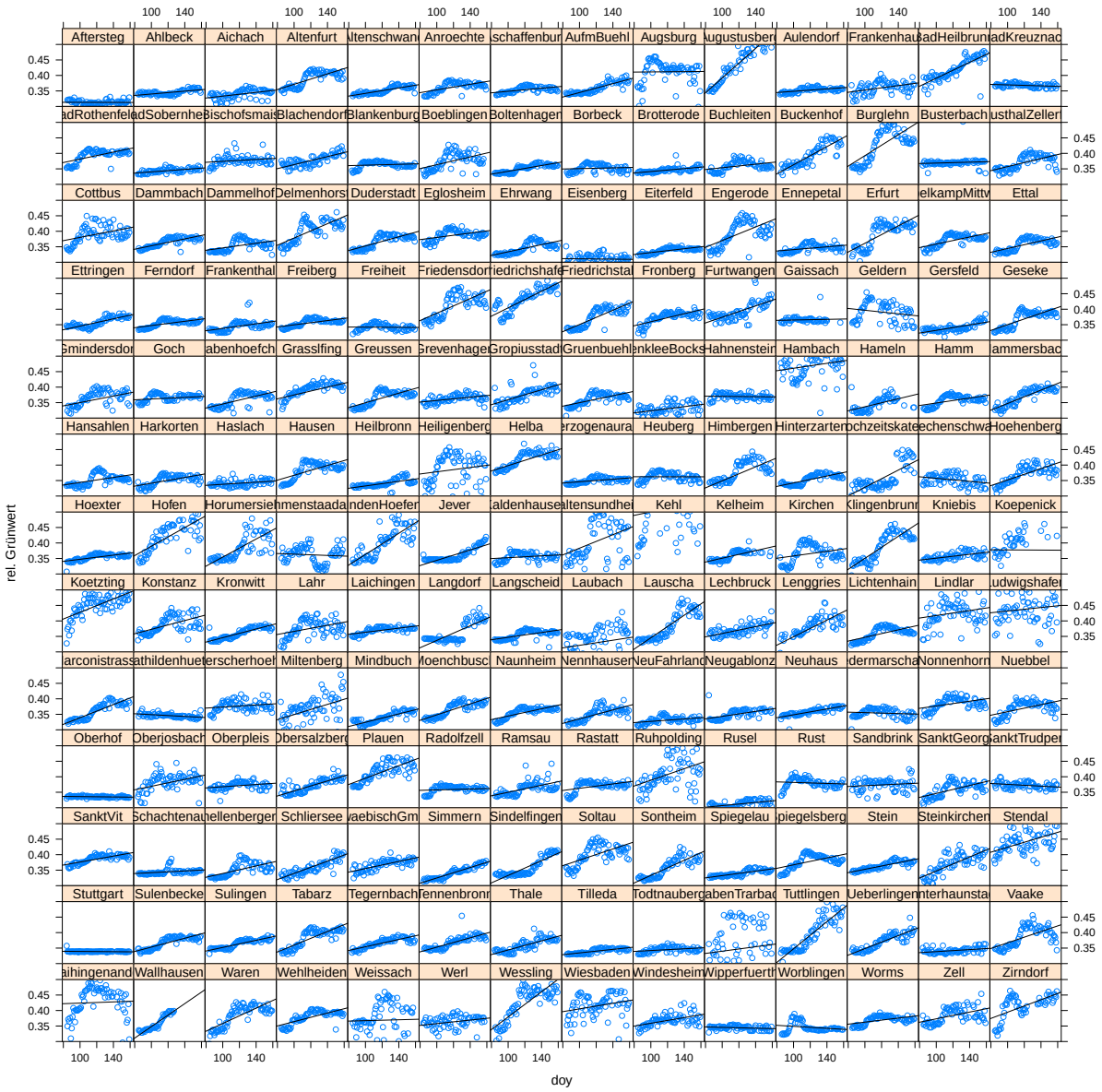


Abbildung 8.4.: Verläufe der einzelnen Stationen (nach Alphabet) über die Zeit.

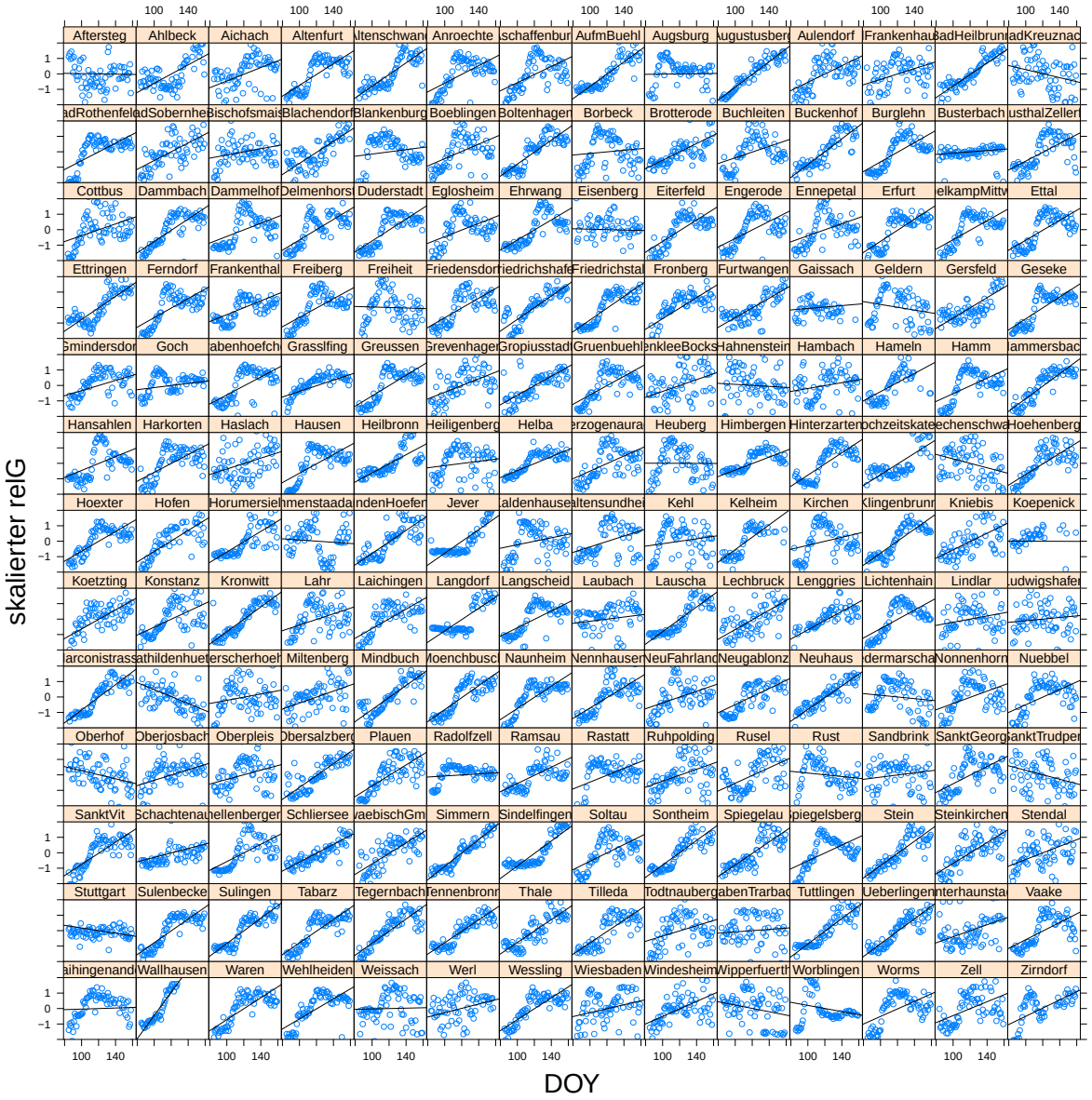


Abbildung 8.5.: Verläufe der skalierten einzelnen Stationen (nach Alphabet) über die Zeit.

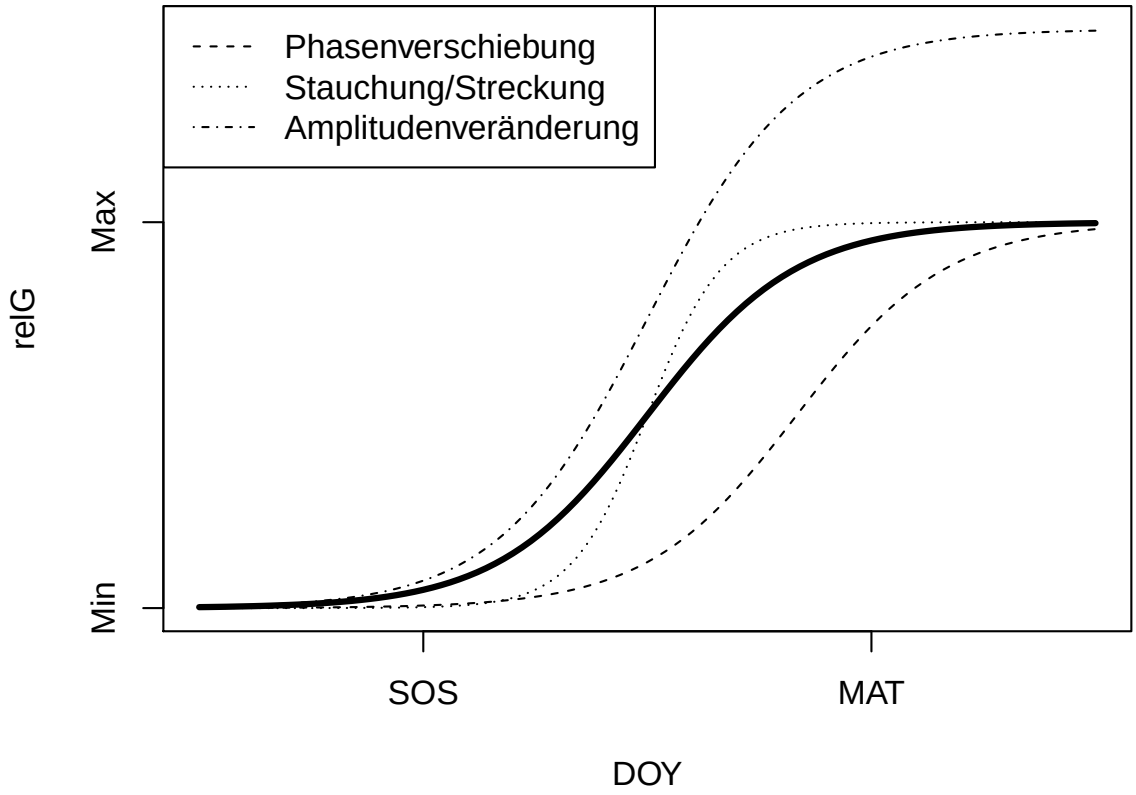


Abbildung 8.6.: Theoretischer Anstieg des relativen Grünwertes mit möglichen Veränderungen

8.1.2. Modellvarianten

Um eine mögliche räumliche Heterogenität optimal zu modellieren, wurden unterschiedliche Modelle betrachtet und mittels AIC verglichen. Die verschiedenen Varianten sind in Tabelle 8.1 zusammengefasst. Der zeitliche Effekt wurde in allen Modellen durch einen P-Spline mit 20 inneren Knoten und einem Strafterm basierend auf Differenzen zweiter Ordnung modelliert. Der strukturierte räumliche Effekt wurde mit zweidimensionalen P-Splines (Model 2), einer zweidimensionalen Funktion auf Basis des Krigingansatzes (Modell 3) bzw. einem Markov-Zufallsfeld (Modell 4) geschätzt. Bei der Modellierung durch das MZF wurden diejenigen Beobachtungen als benachbart betrachtet, welche weniger als 120km voneinander entfernt lagen. Zusätzlich zum strukturierten räumlichen Effekt wurden teilweise unabhängige zufällige Effekte mit aufgenommen. In der Schätzung der Modelle mit zufälligen Effekten (Modelle 2, 3 und 4) kam es zu Konvergenzproblemen innerhalb des Algorithmus. Dies liegt an dem sehr kleinen Varianzparameter der räumlich strukturierten Oberfläche. Da dadurch das Maximum der marginalen Likelihood am Rand des Parameterraums liegt, versagt der Fisher-Scoring Algorithmus. Dies führt zu einem Abbruch der Schätzung kleiner Varianzen in BayesX (Belitz et al. 2015, Kap. 6.2).
Die geschätzten strukturierten räumlichen Effekte der Modelle 2(.1) und 3(.1) finden sich in den Abbildungen 8.7-8.10. Darin zeigt sich die geringe Varianz der Oberflächen in den Modellen mit zufälligen Effekten. Es dominiert in diesen Fällen der unstrukturierte Effekt. Ohne die zufälligen Effekte ergibt sich für die Modellierung durch den P-Spline eine sehr unruhige Modellierung. Hier scheinen zufällige Effekte sinnvoller zu sein. Auch die Betrachtung der AIC's (vgl. Abbildung 8.11) lässt darauf schließen, dass eine Modellierung ohne strukturierten räumlichen Effekt gerechtfertigt ist. Es kann somit zur traditionellen Analyse logitudinaler Daten mit Random-Intercept übergegangen werden. Der geschätzte zeitliche Effekt dieses Modells (Modell 1) ist in Abbildung 8.12 zu sehen. Es ist ein Anstieg des rel. Grünwertes zu erkennen. Dieser stagniert etwa am DOY 120 (MAT). Der genaue Anfang der Wachstumszeit (SOS) ist hingegen nicht direkt erkennbar.

Modell	*formula*						
0	relG	$\sim$	s(doy)				
1	relG	$\sim$	s(doy)			+	b_i
2	relG	$\sim$	s(doy)	+	P-Spline	+	b_i
2.1	relG	$\sim$	s(doy)	+	P-Spline		
3	relG	$\sim$	s(doy)	+	Kriging	+	b_i
3.1	relG	$\sim$	s(doy)	+	Kriging		
4	relG	$\sim$	s(doy)	+	MZF	+	b_i

Tabelle 8.1.: Modellvarianten - s(doy) entspricht P-Spline-Modellierung des zeitlichen Effekts.

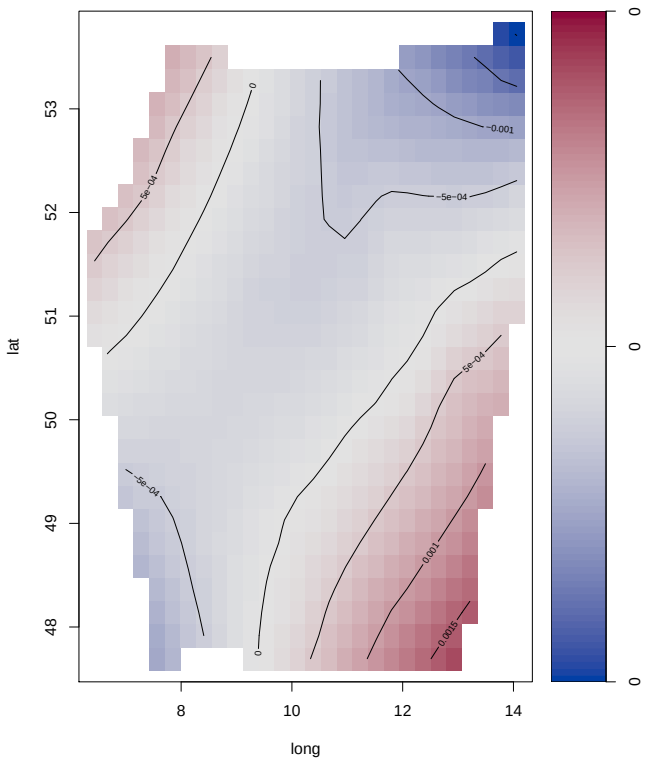


Abbildung 8.7.: Geschätzter strukturierter räumlicher Effekt (P-Spline, Modell 2).

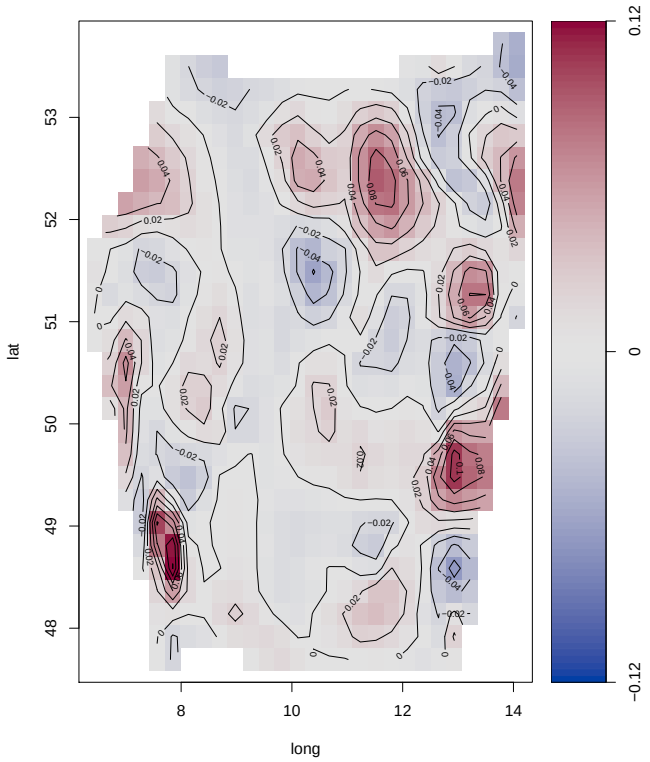


Abbildung 8.8.: Geschätzter strukturierter räumlicher Effekt (P-Spline, Modell 2.1).

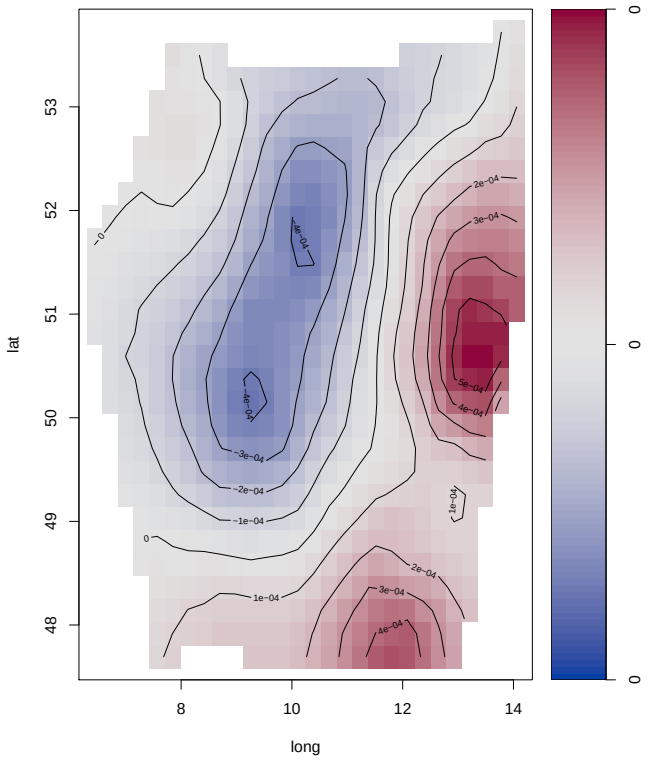


Abbildung 8.9.: Geschätzter strukturierter räumlicher Effekt (Kriging, Modell 3).

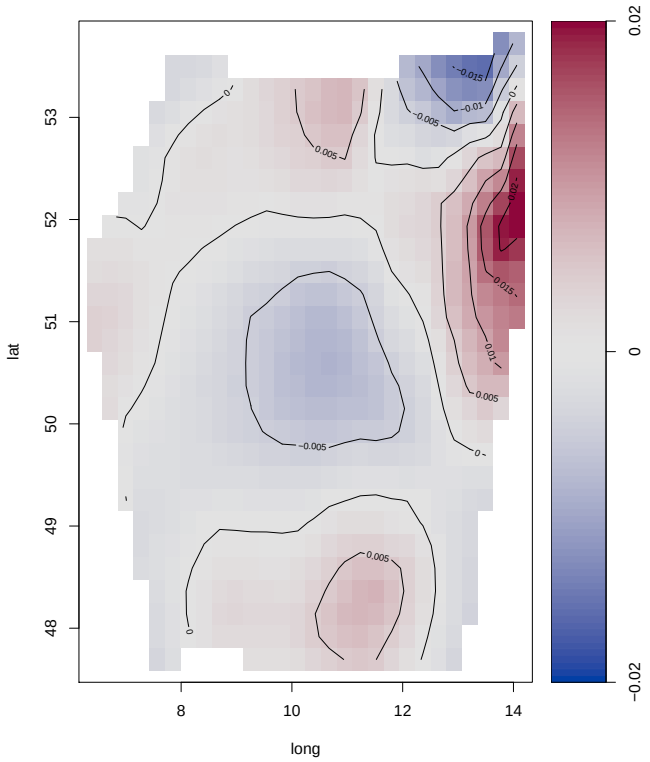


Abbildung 8.10.: Geschätzter strukturierter räumlicher Effekt (Kriging, Modell 3.1).

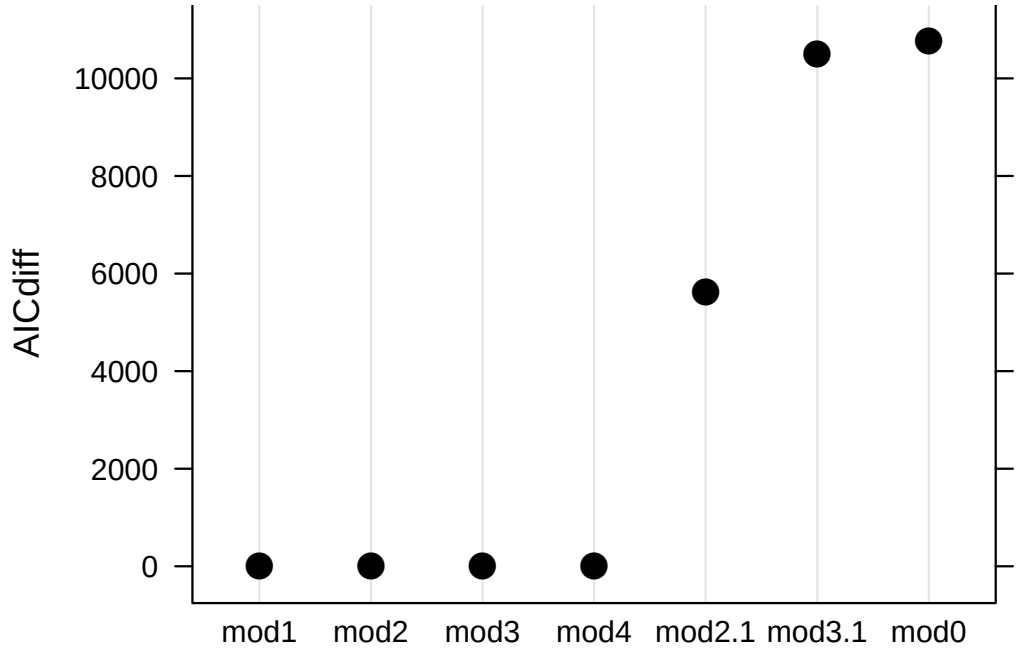


Abbildung 8.11.: AIC-Differenzen zum besten Modell für die verschiedenen Modellvarianten.

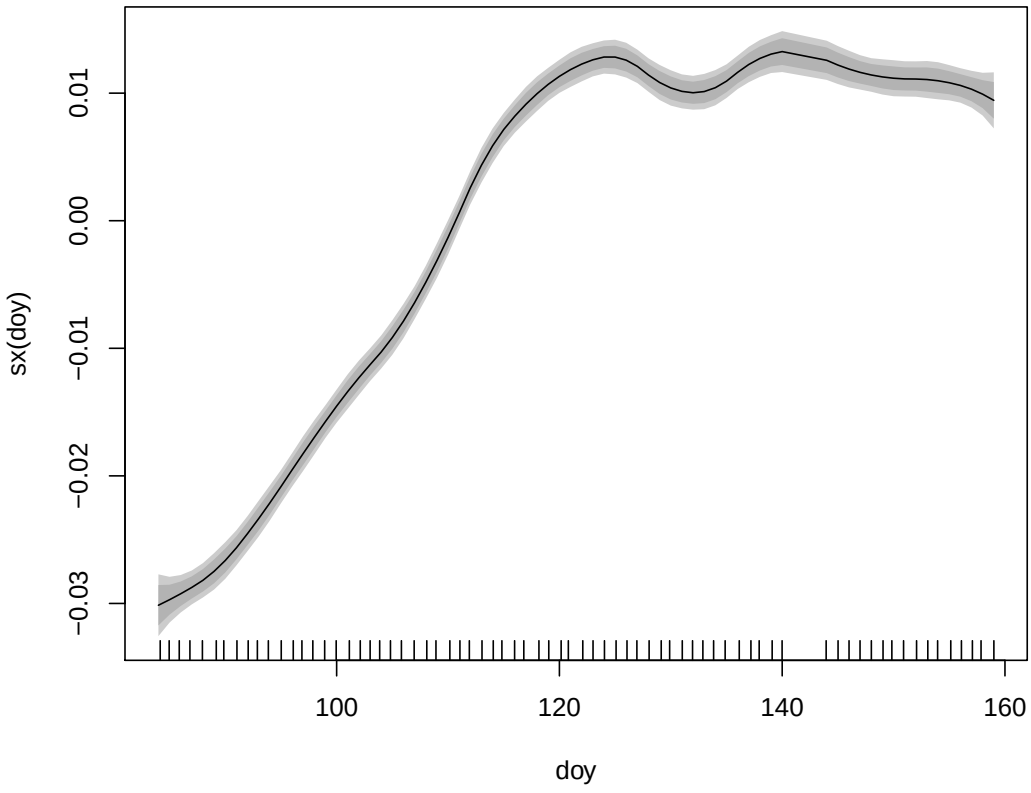


Abbildung 8.12.: Geschätzter zeitlicher Effekt (Modell 1).

8.2. Beispiel: Hochschulen

Nun sollen die in Kapitel 3.2 vorgestellten Daten analysiert werden.
Für das Jahr 2004 ist die Herkunft von 7483 Studienanfängern der Ludwig-Maximilians-Universität (LMU) dokumentiert. Im Jahr 2011 sind es bereits 11001 Studienanfänger. Die meisten Studierenden stammen dabei in beiden Jahren aus dem Stadtgebiet München. Es gibt 58 (2004) bzw. 18 (2011) Landkreise in denen kein Studienanfänger seine Hochschulzugangsberechtigung (HZB) erworben hat. Für Gesamtdeutschland ergibt sich eine Inzidenzrate von 9.1 (2004) bzw. 13.7 Studenten pro 100000 Einwohner (2011).
Die "Population unter Risiko"wurde durch die Einwohnerzahl der einzelnen Kreise definiert. Eine bessere Herangehensweise wäre die Betrachtung der Personen mit HZB. Diese Zahlen liegen jedoch nicht vor, sodass von einer homogenen Verteilung über ganz Deutschland ausgegangen werden muss.

8.2.1. Standard-Inzidenzraten (SIR) und p-Werte

2004

Die berechneten Standard-Inzidenzraten sind in Abbildung 8.13 visualisiert. Dabei fallen vor allem die erhöhten Raten um München herum sowie in einigen Städten Bayerns auf. Die minimale (größer Null) bzw. maximale Standard-Inzidenzrate ist 0.02 bzw. 24.48. Die maximale Rate tritt in Rosenheim auf. Hier wird jedoch nur eine Anzahl von 5.47 Studenten erwartet. Die Standardabweichung $sd_i = \sqrt{O_i}/E_i$ in diesem Kreis liegt bei 2.11 und nimmt damit den höchsten beobachteten Wert ein. An diesem Beispiel zeigt sich die Unsicherheit der Schätzung der SIR bei einer geringen Anzahl erwarteter Fälle.
Abbildung 8.14 zeigt die berechneten p-Werte für den Test auf $\theta_1 = 1$ auf Basis der Poisson- bzw. der Negativ-Binomialverteilung. Auch diese sprechen für erhöhte Raten im näheren Umfeld der LMU. Dabei sind die p-Werte der Negativ-Binomialverteilung wie erwartet höher, da hier eine größere Varianz zugelassen wird.

2011

Für 2011 ergeben sich sehr ähnliche Werte, sodass hier auf eine genauere Betrachtung verzichtet wird. Die SIR für dieses Jahr sind in Abbildung 8.15 abgetragen. Hier zeigt sich noch etwas klarer der Zusammenhang mit der Distanz von der LMU als in 2004.

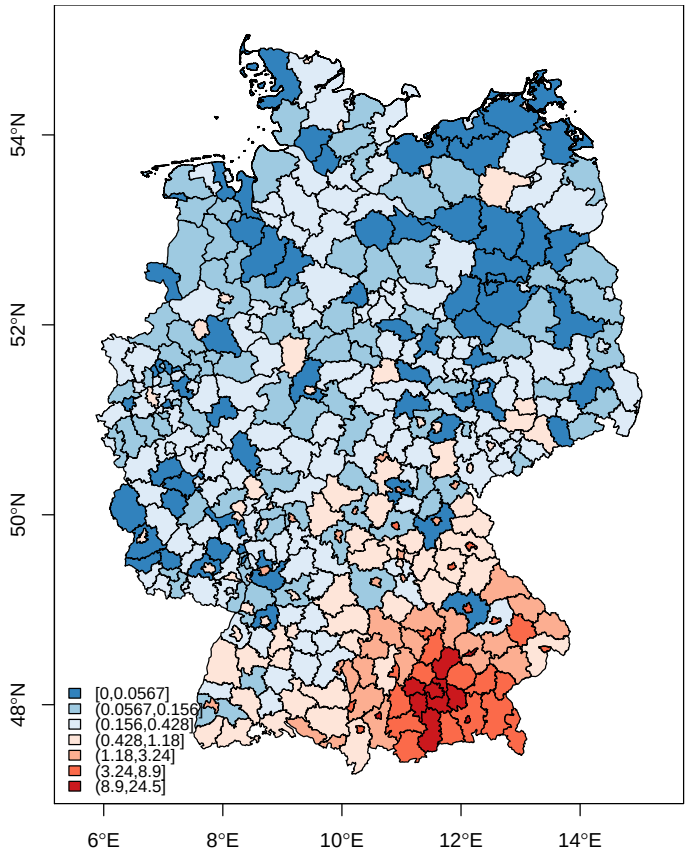


Abbildung 8.13.: Herkunft der Studienanfänger der LMU 2004 (Standardized Incidence Ratio - SIR).

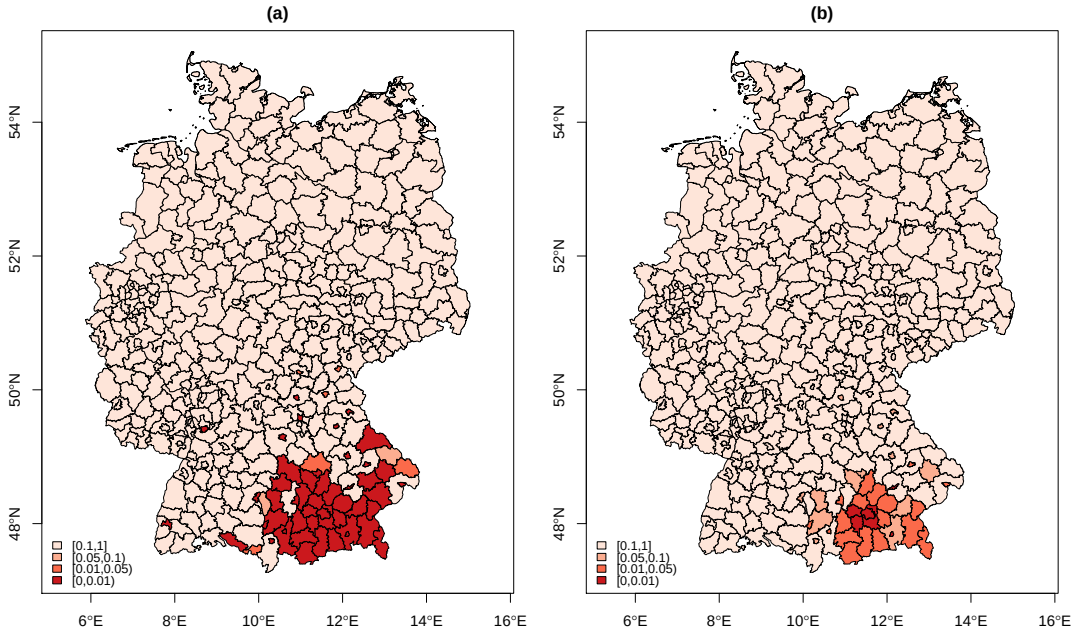


Abbildung 8.14.: p-Werte aus Basis der (a) Poisson- bzw. (b) Negativ-Binomialverteilung für das SIR der Studienanfänger an der LMU 2004.

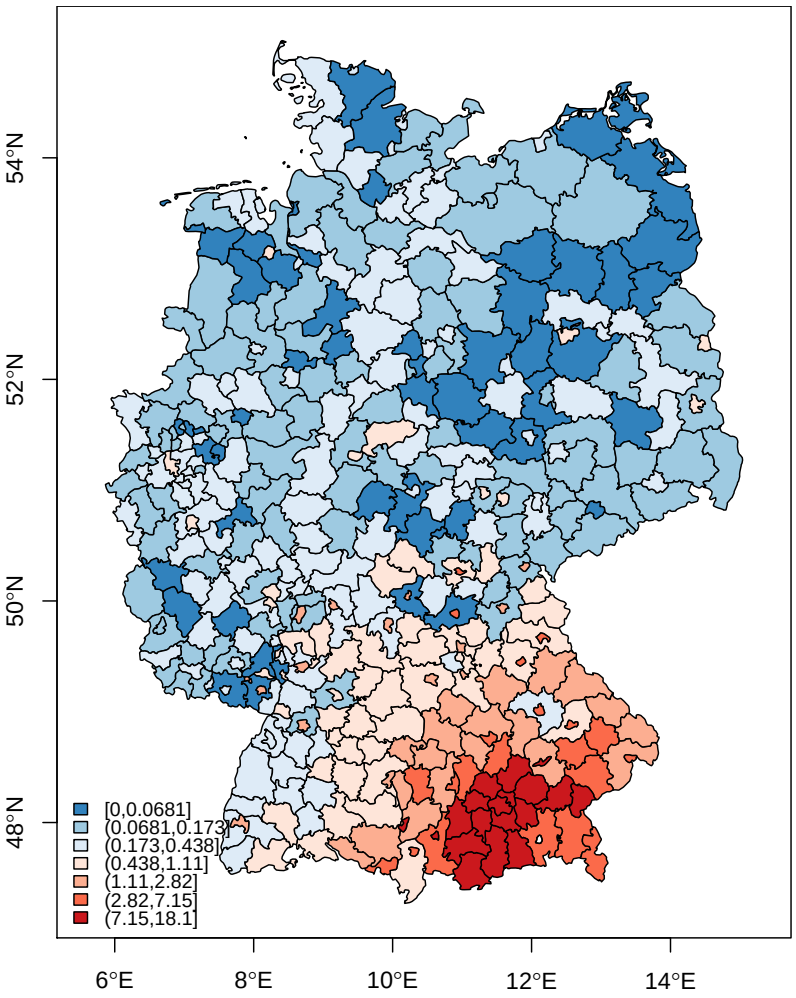


Abbildung 8.15.: Herkunft der Studienanfänger der LMU 2011 (Standardized Incidence Ratio - SIR).

8.2.2. Räumliche Modelle

Für die folgenden Analysen wurde eine Nachbarschaftsmatrix auf Basis gemeinsamer Grenzen erstellt. Zu diesem Zweck wurde die Insel Rügen aus dem Datensatz für 2004 entfernt, da diese keine Verbindung zu anderen Kreisen besitzt. In 2011 war dies nicht nötig, da es in diesem Jahr aufgrund von Gebietsreformen zu einer Eingliederung des Kreises in den Kreis Nordvorpommern kam. In Abbildung 8.16 ist der Graph der Nachbarschaften abgebildet.

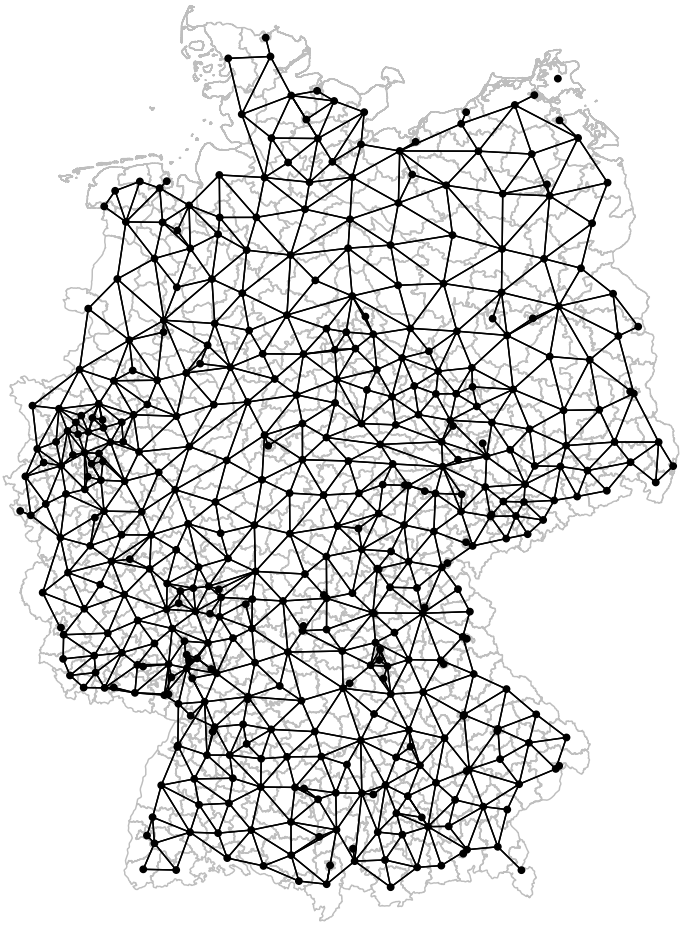


Abbildung 8.16.: Nachbarschaften auf Basis gemeinsamer Grenzen.

Um festzustellen, ob räumliche Autokorrelation vorliegt wurde Moran's I auf Basis der SIR berechnet. Hierbei konnte auch nach Berücksichtigung vorliegender Überdispersion in beiden Jahren eine räumliche Autokorrelation festgestellt werden.

Markov-Zufallsfelder

Aus diesem Grund wurde im nächsten Schritt ein Modell mit strukturierten und unstrukturierten Effekten gemäß Kapitel 6.2 geschätzt.

2004

Die geschätzten Effekte sind in Abbildung 8.17 visualisiert. Der räumlich strukturierte Effekt überwiegt dabei deutlich den unstrukturierten Effekt. Es zeigt sich, wie bereits erwartet, ein positiver Effekt in der Nähe der LMU. Außerdem ergeben sich in beiden Termen höhere Werte für die Städte im Gegensatz zu den Landkreisen. Dies könnte an einer erhöhten Anzahl an Personen mit HZB in den Städten liegen. Für einen Vergleich mit den Standard-Inzidenzraten

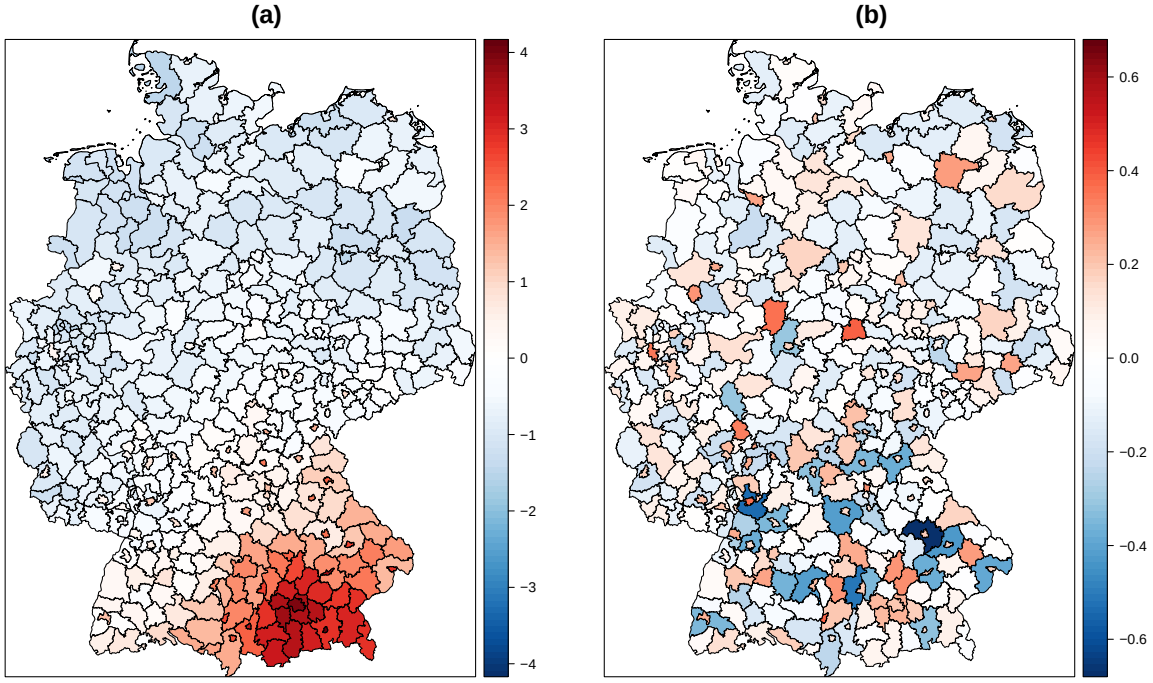


Abbildung 8.17.: Schätzungen des strukturierten (a) und des unstrukturierten (b) räumlichen Effekts; Modell ohne weitere Kovariablen 2004.

wurden die geschätzten Werte für $\hat{\theta}_i = \exp(\hat{\beta}_0 + \hat{f}_{i,geo} + \hat{b}_i)$ in Abbildung 8.18 visualisiert. Die Werte scheinen mit wachsender Entfernung von der LMU zu sinken. Aus diesem Grund wurde im nächsten Schritt die Distanz des jeweiligen Zentroiden eines Kreises zur LMU als nichtlinearer Effekt mit in das Modell aufgenommen. Außerdem wurde eine Dummy-Variable zur Unterscheidung zwischen Landkreisen und Kreisfreien Städten eingeführt. Daraus ergibt sich der lineare Prädiktor

$$\eta_i = \beta_0 + \beta_1 Landkreis_i + f(distance_i) + f_{i,geo}(s_i) + b_i + \log(E_i), \tag{8.1}$$

Die geschätzte Funktion für die Distanz ist in Abbildung 8.19 zu sehen. Für die Schätzung wurde ein P-Spline mit 20 inneren Knoten und einem Strafterm basierend auf Differenzen zweiter Ordnung verwendet. Die Anzahl der Studierenden nimmt, wie bereits in den vorherigen Grafiken ersichtlich, mit wachsender Distanz zur LMU immer weiter ab. Erst ab etwa 500km stagniert die Kurve, sodass hier kein Unterschied mehr durch weitere Distanzen entsteht.

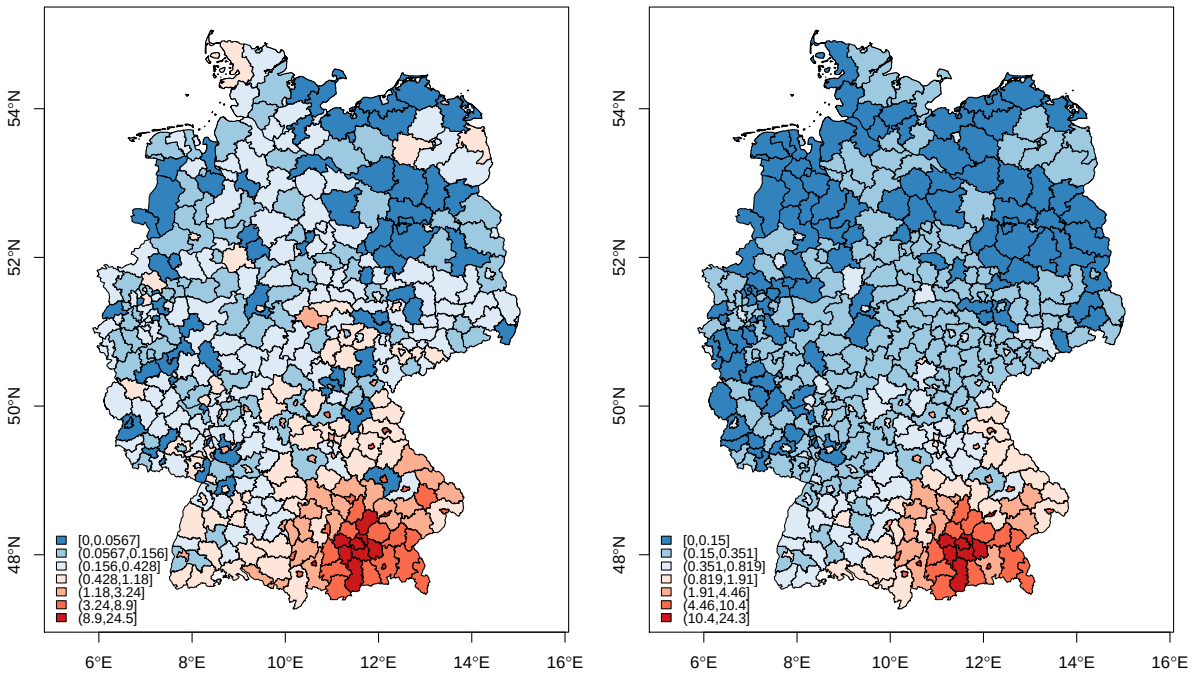


Abbildung 8.18.: Geschätzte Standard-Inzidenzraten: (a) $\hat{\theta}_i = O_i/E_i$ (b) $\hat{\theta}_i = \exp(\hat{\beta}_0 + \hat{f}_{i,geo} + \hat{b}_i)$.

Die Variable *Landkreis* hat einen signifikanten, negativen Einfluss. Aus den Landkreisen stammen im Erwartungswert um den Faktor $\exp(\beta_1) = \exp(-0.9260) = 0.396$ weniger Studienanfänger als aus Kreisfreien Städten. In Abbildung 8.20 sind erneut die Schätzungen der strukturierten und unstrukturierten räumlichen Effekte abgetragen. Auffällig sind hier hohe Werte in den süd-östlichen Gebieten Ober- und in den östlichen Teilen Niederbayerns. Bei den unstrukturierten Effekten fallen vor allem die Kreise Regensburg, Augsburg, Heidelberg und Nürnberg ins Auge. Hier scheint die Anziehungskraft der eigenen Universitäten höher zu sein, als die der LMU.

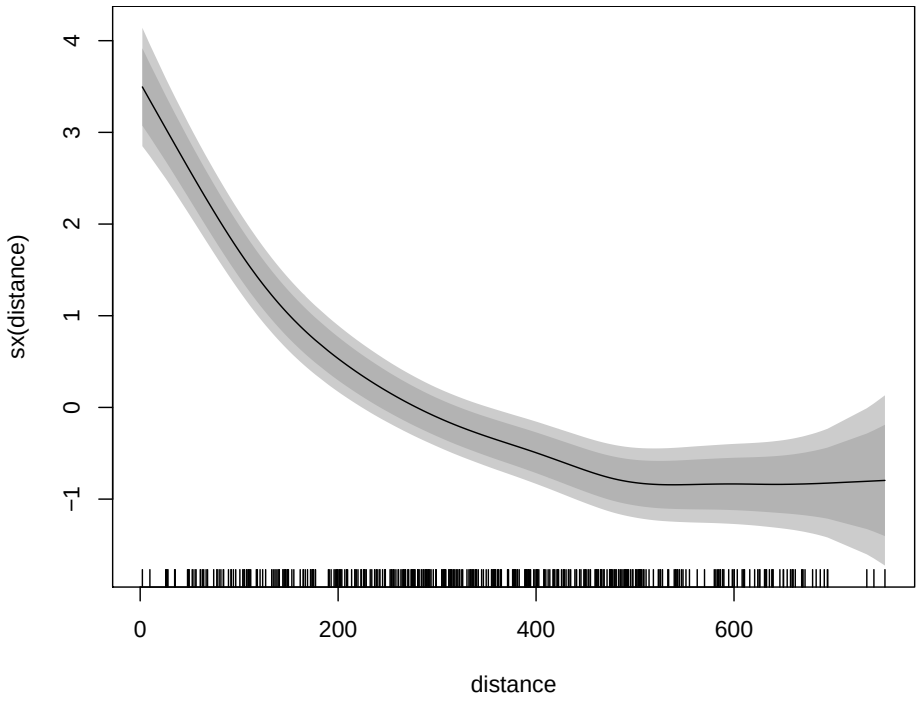


Abbildung 8.19.: Geschätzter Effekt für die Distanz der Zentroiden zur LMU 2004.

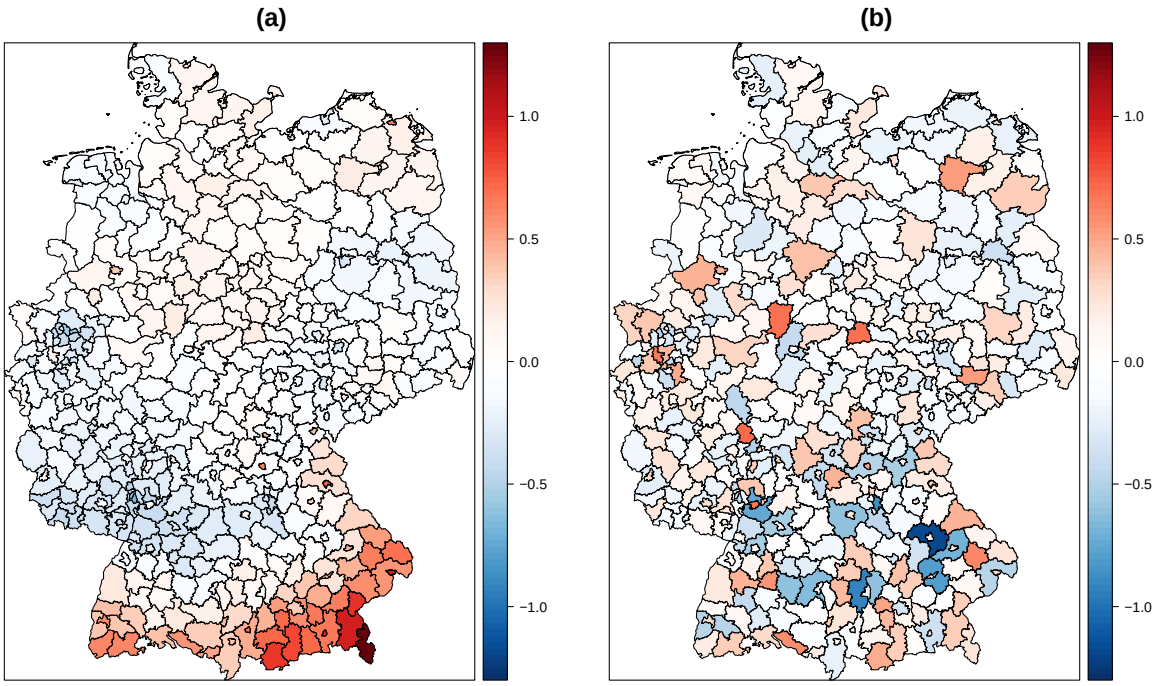


Abbildung 8.20.: Schätzungen des strukturierten (a) und des unstrukturierten (b) räumlichen Effekts; Modell mit Kovariablen 2004.

2011

Die gleichen Modelle wurden für 2011 geschätzt. Im Modell ohne Kovariablen (vgl. Abbildung 8.21) werden stark negative Effekte für einige Gebiete Brandenburgs und Sachsen-Anhalts geschätzt. Diese sind sowohl im unstrukturierten wie auch im strukturierten Effekt dominant.

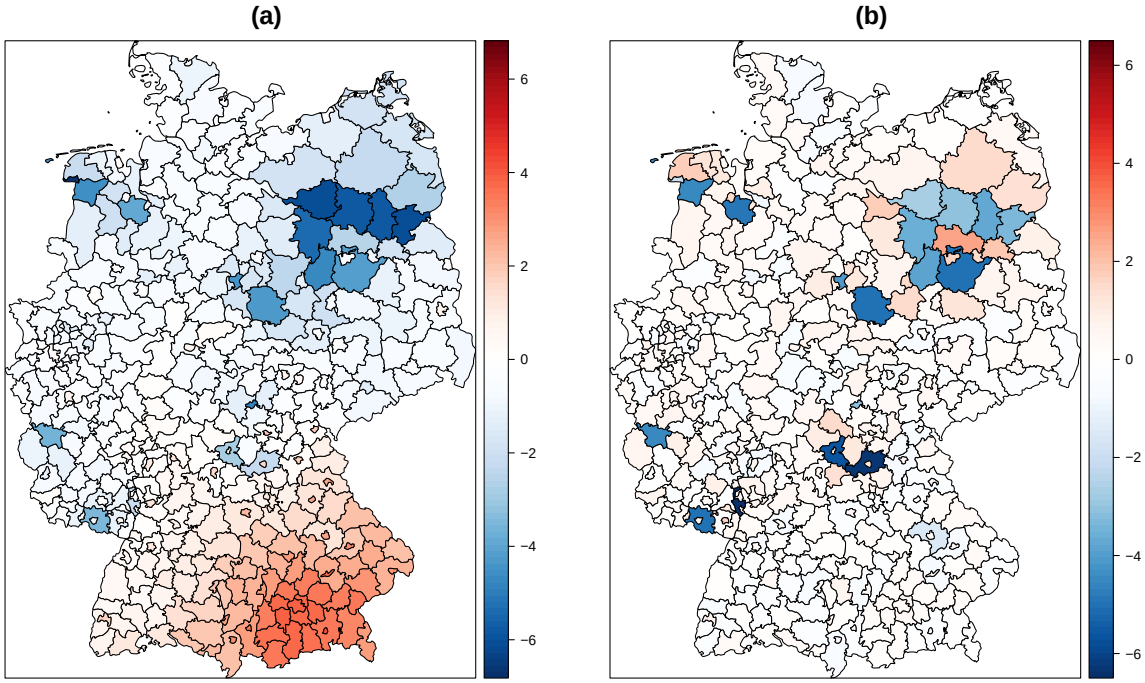


Abbildung 8.21.: Schätzungen des strukturierten (a) und des unstrukturierten (b) räumlichen Effekts; Modell ohne weitere Kovariablen 2011.

Im Modell mit Kovariablen wird eine ähnliche Funktion für den Einfluss der Distanz wie in 2004 geschätzt (vgl. Abbildung 8.22). Auch die Variable *Landkreis* hat erneut einen signifikanten, negativen Einfluss. Aus den Landkreisen stammen in diesem Fall um den Faktor $\exp(\beta_1) = \exp(-0.7221) = 0.486$ weniger Studienanfänger als aus Kreisfreien Städten. Der Anteil von Studenten aus den Landkreisen ist also im Vergleich zu 2004 angestiegen. Die räumlichen Effekte im Modell mit Kovariablen sind sehr ähnlich zu denen in 2004. Es kommen mehr Studienanfänger aus Gebieten süd-östlich von München als durch die pure Betrachtung der Distanz zu erwarten wäre. Interessant ist jedoch, dass hingegen weniger Studierende für die neuen Bundesländer erwartet werden. Bei den unstrukturierten Effekten hebt sich erneut der Kreis Regensburg deutlich ab. Auch hier zeigen Kreise mit eigener Universität eher einen negativen Effekt. Ausnahmen sind dabei Berlin, Hamburg und Göttingen mit einem vergleichsweise starkem positiven Effekt.

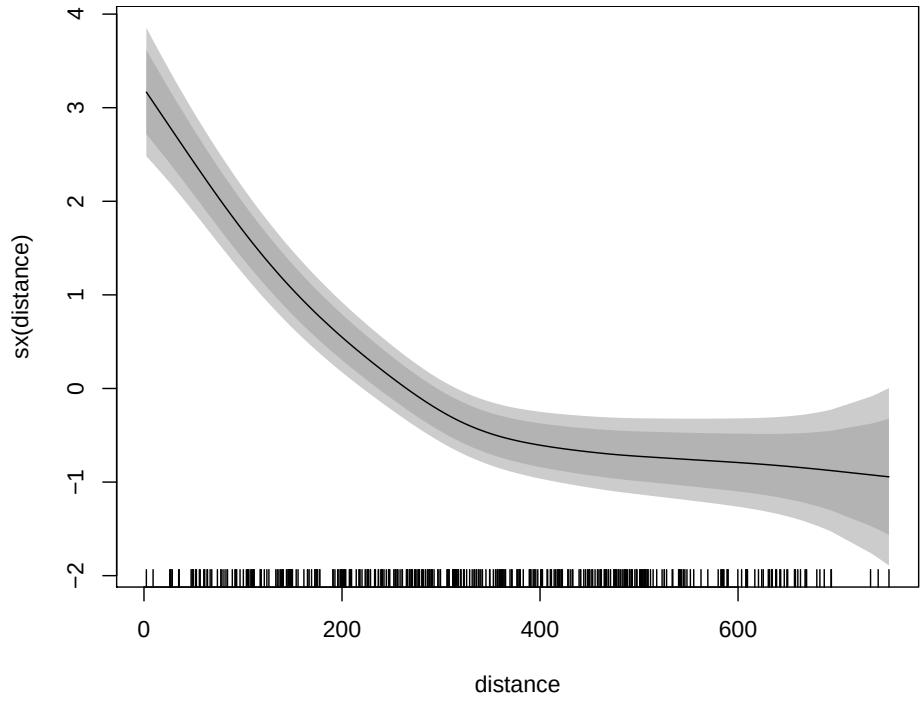


Abbildung 8.22.: Geschätzter Effekt für die Distanz der Zentroiden zur LMU 2011.

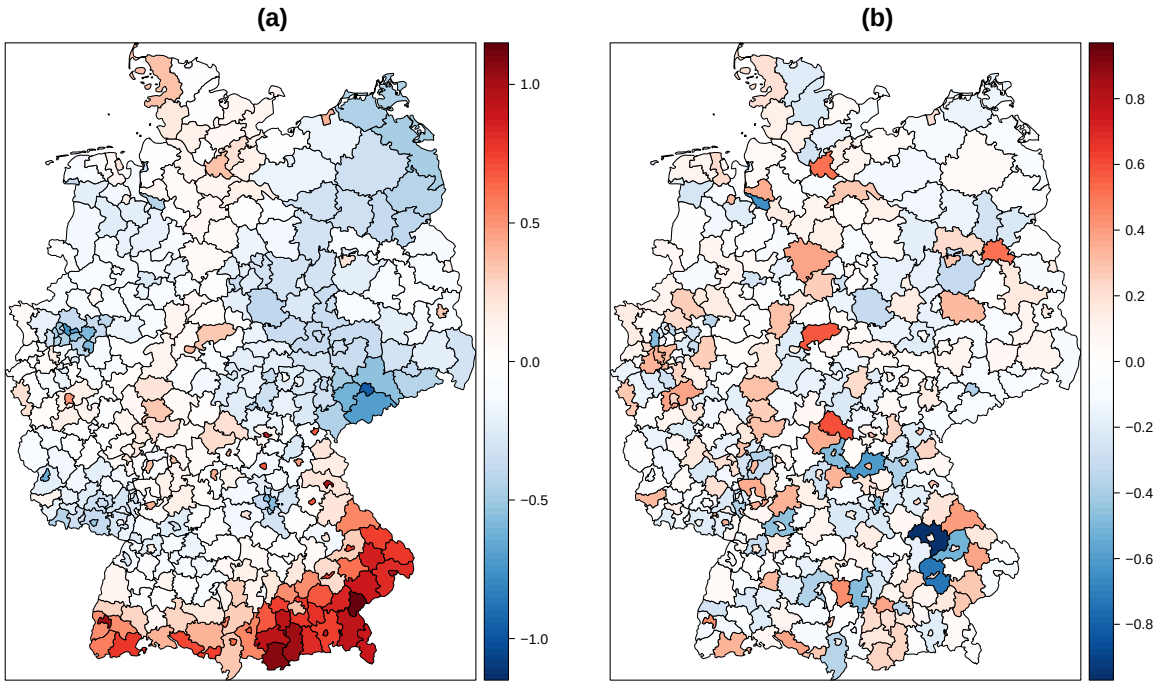


Abbildung 8.23.: Schätzungen des strukturierten (a) und des unstrukturierten (b) räumlichen Effekts; Modell mit Kovariablen 2011.

Kriging

Eine weitere Möglichkeit den räumlichen Effekt zu schätzen, ist das Regression-Kriging über die Zentroide der Kreise. Das Vorgehen wird im folgenden Abschnitt am Beispiel der Daten für die LMU in 2004 erläutert.
Die besondere Art der Aufteilung in Kreise und Kreisfreie Städte führt dabei in Deutschland zu einem Problem, das eine sinnvolle Schätzung des Variogramms bei kleinen Distanzen unmöglich macht. In vielen Fällen liegt eine kreisfreie Stadt innerhalb eines Kreises (vgl. Abbildung 8.24), sodass der Zentroid der Stadt und der des Kreises sehr nahe beieinander liegen. Dies führt zu einer Überschätzung der Semivarianz bei kleinen Distanzen. Aus diesem Grund wurden die Kreisfreien Städte dem jeweils nächsten Landkreis hinzugefügt.

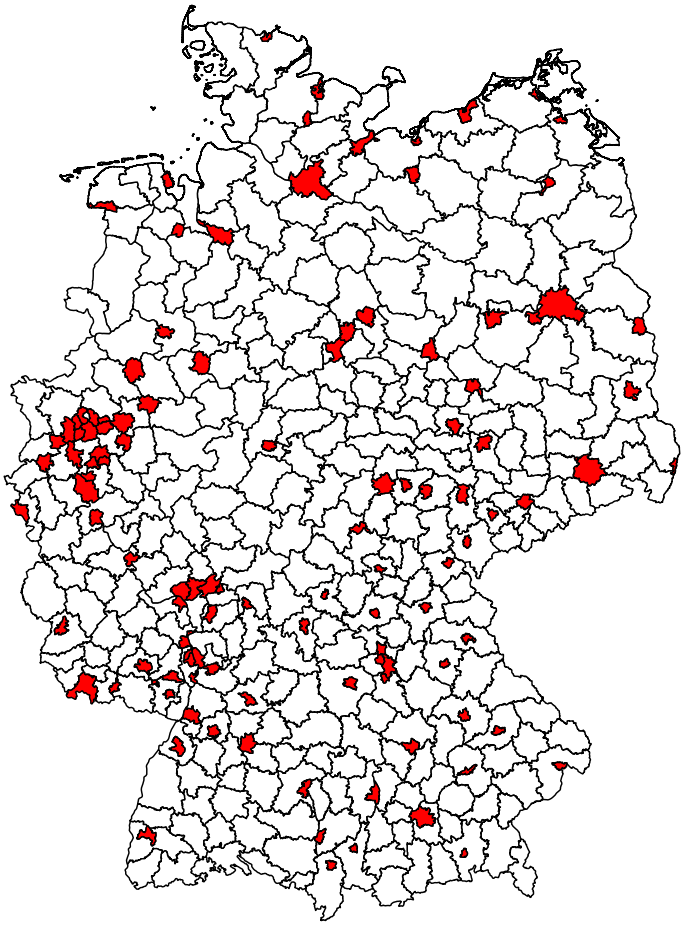


Abbildung 8.24.: Kreisfreie Städte in Deutschland.

Im nächsten Schritt wurde der Trend, der durch die Distanz zur LMU entsteht, über ein Quasi-Poisson-Modell herausgerechnet. Zur Schätzung eines räumlichen Effekts wurden dann die stan-

dardisierten Pearson-Residuen

$$r_i^{P*} := \frac{1}{\sqrt{E_i}} \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}} = \frac{1}{\sqrt{E_i}} r_i^P$$

berechnet (vgl. Abbildung 8.25). Die Standardisierung durch die erwartete Anzahl an Studierenden E_i aus Region i folgt dabei der gleichen Logik, wie die Verwendung eines Offsets bei der Poisson-Regression. Um einen sinnvollen Schätzer für den Nugget-Effekt zu erhalten,

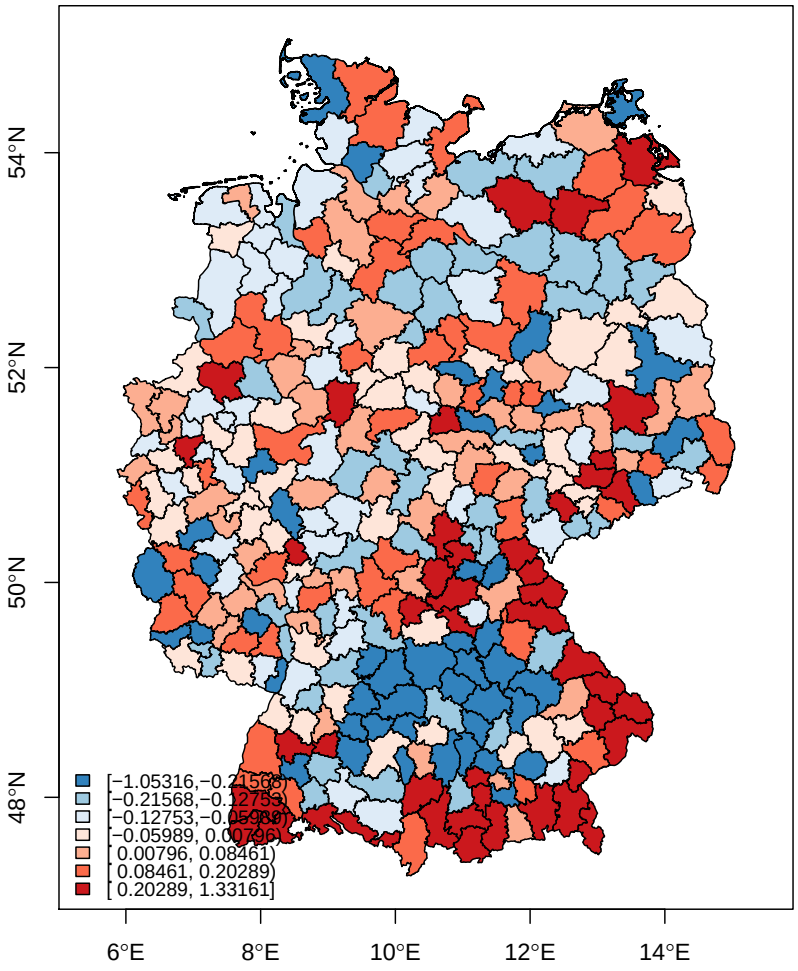


Abbildung 8.25.: Standardisierte Pearson Residuen.

wurden außerdem einige Hotspots (Beobachtungen mit sehr verschiedenen Nachbarn) aus der Variogramm-Analyse ausgeschlossen. Das geschätzte Variogramm ist in Abbildung 8.26 abgebildet. Zur Schätzung des räumlichen Effekts wurde dann ein Grid über Deutschland gelegt und gewöhnliches Kriging auf die Residuen durchgeführt. Das Ergebnis ist in Abbildung 8.27 zu sehen. Es deckt sich mit den geschätzten strukturierten Effekten mittels eines Markov-Zufallsfeldes im vorherigen Abschnitt.

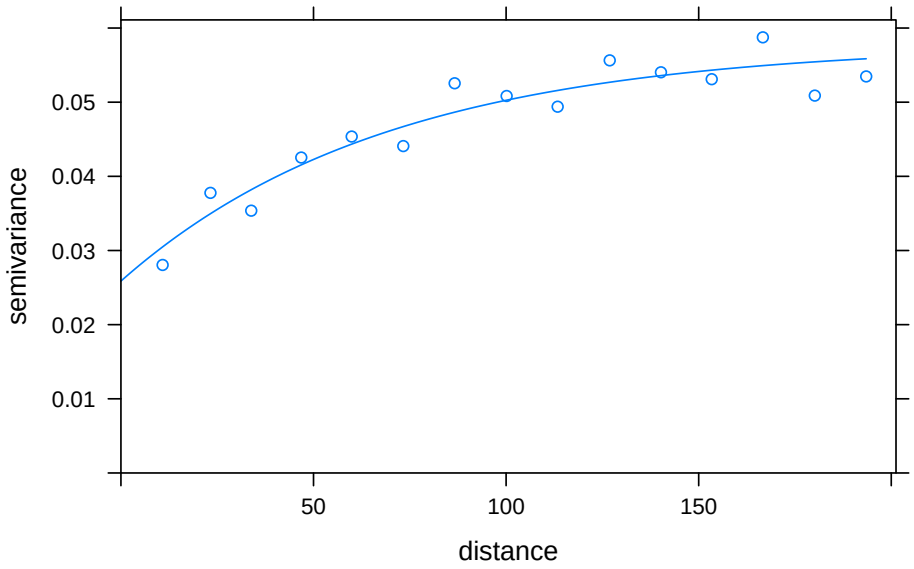


Abbildung 8.26.: Geschätztes Variogramm; Matern, psill=0.03, range=69.45, kappa= 0.5.

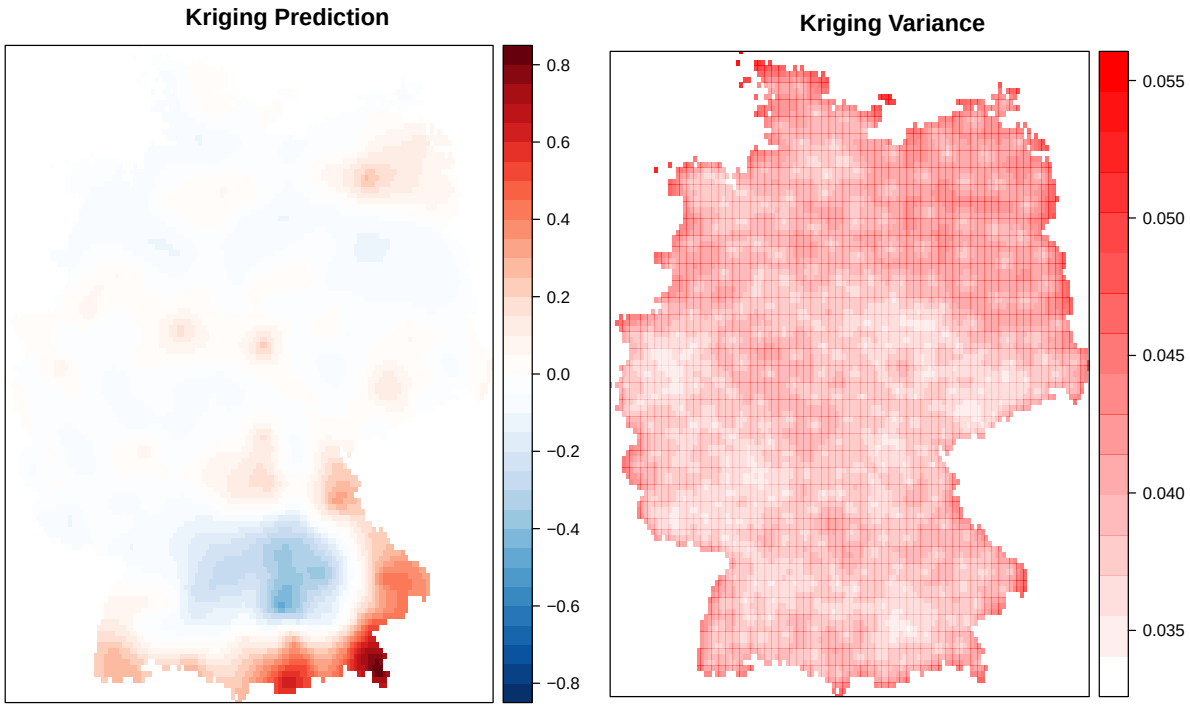


Abbildung 8.27.: Kriging Vorhersage und Varianz auf die standardisierten Pearson-Residuen.

8.2.3. Vergleich mit der Humboldt-Universität zu Berlin

Interessant ist nun, ob die Erkenntnisse aus den Modellen für die LMU auch auf andere Universitäten übertragbar sind. Zu diesem Zweck wurden die Daten der Humboldt-Universität zu Berlin (HUB) im Jahr 2011 für die weitere Analyse herangezogen. In Abbildung 8.28 sind die SIR visualisiert. Es zeigt sich auch hier ein starker regionaler Bezug bei der Studienplatzwahl. Aus diesem Grund wurde direkt das Modell mit Kovariablen (Modell 8.1) auf die Daten an-

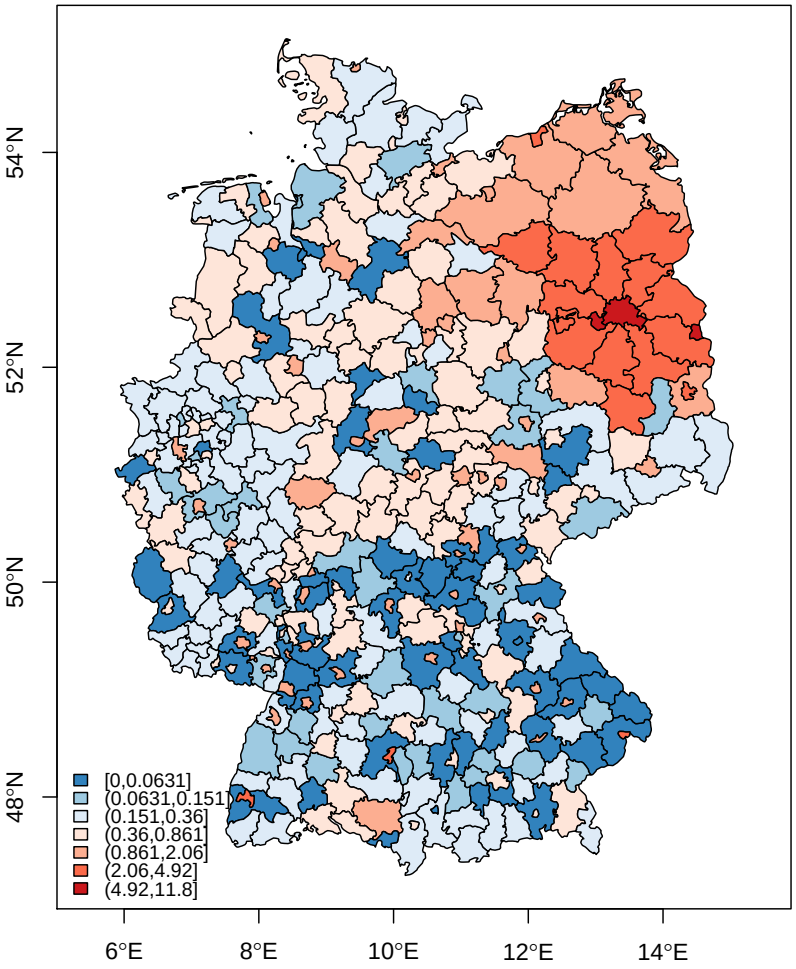


Abbildung 8.28.: Herkunft der Studienanfänger der HUB 2011 (Standardized Incidence Ratio - SIR).

gewendet. Der strukturierte Effekt wurde erneut durch ein Markov-Zufallsfeld geschätzt. Die Ergebnisse sind in den Abbildungen 8.29 und 8.30 abgetragen. Für die Variable *Landkreis* ergab sich ein signifikanter, negativer Einfluss. Aus den Landkreisen stammen in diesem Fall um den Faktor $\exp(\beta_1) = \exp(-0.9698) = 0.379$ weniger Studienanfänger als aus Kreisfreien Städten. Die Funktion für die Distanz zur HUB ist ähnlich im Verlauf wie die bei der LMU. Im räumlich strukturierten Effekt erkennt man positive Effekte für die Gebiete nördlich von Berlin. Dies ist vermutlich darauf zurückzuführen, dass hier im näheren Umfeld nur Universitäten in

Rostock und Greifswald zur Verfügung stehen. Negative Effekte werden hingegen in der Nähe der Universitätsstädte Leipzig, Chemnitz und Dresden geschätzt.
Insgesamt weisen die Daten auf einen starken regionalen Bezug bei der Wahl des Studienortes hin. Interessant wäre ein Vergleich mit weiteren Universitäten des Bundesgebiets. Dabei wäre es außerdem sinnvoll von der einfachen Distanz zu einer Definition über die Erreichbarkeit (via Auto, öffentliche Verkehrsmittel etc.) überzugehen.

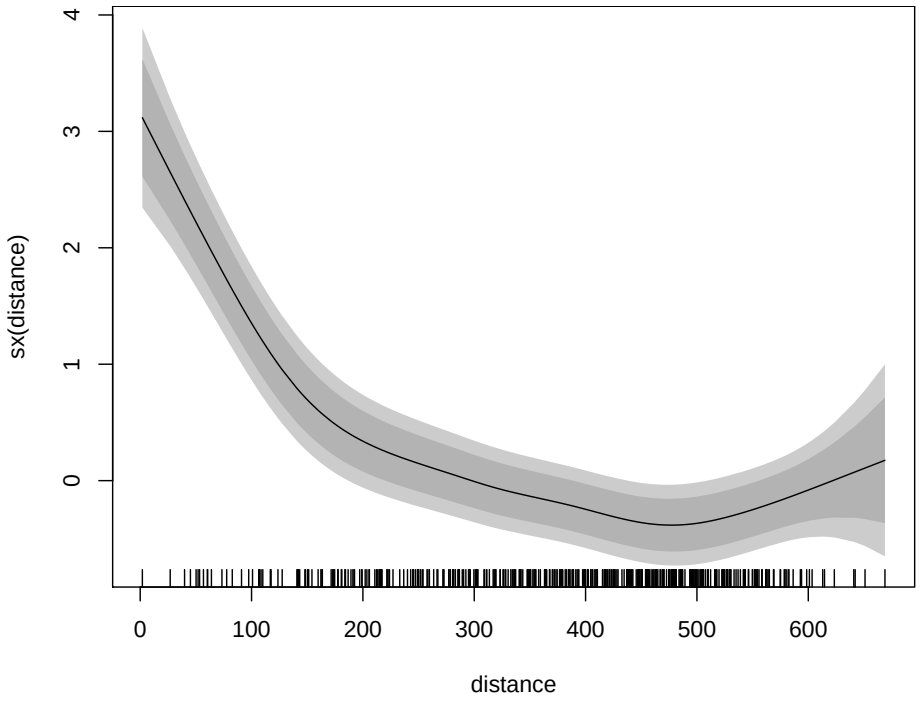


Abbildung 8.29.: Geschätzter Effekt für die Distanz der Zentroiden zur HUB 2011.

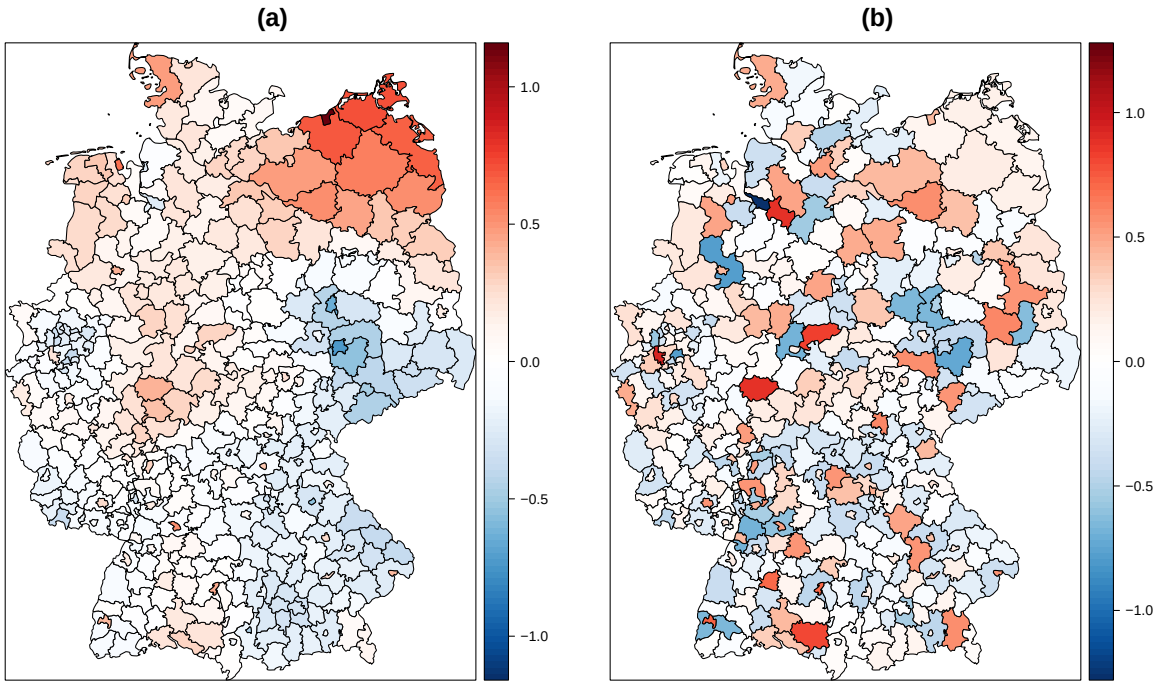


Abbildung 8.30.: Schätzungen des strukturierten (a) und des unstrukturierten (b) räumlichen Effekts; Modell mit Kovariablen HUB 2011.

9. Zusammenfassung und Ausblick

Diese Masterarbeit beschäftigte sich mit der Modellierung räumlicher Abhängigkeiten. Grundlage vieler statistischer Modelle ist die Annahme unabhängiger Beobachtungen (gegeben den Kovariablen). Eine häufige Charakteristik räumlicher Daten ist jedoch die Tatsache, dass sich nahe Beobachtungen ähnlicher sind als weit entfernte. Eine Folgerung daraus ist, dass diese Daten nicht dem Paradigma der Unabhängigkeit folgen und somit einer Modellierung von Abhängigkeiten bedürfen.
Die vorliegende Arbeit gibt einen Überblick über die Theorie zur Modellierung zweier räumlicher Datentypen: geostatistische Daten und Gitterdaten. Bei geostatistische Daten liegt die räumliche Information stetig, in Form von Koordinaten vor, bei Gitterdaten ist sie auf eine abzählbare Menge an Regionen aggregiert. Es wird in beiden Fällen von einem zugrundeliegenden, datengenerierenden stochastischen Prozess

$$\{Z(s) : s \in D\}$$

ausgegangen. Dabei entspricht $s \in \mathbb{R}^d$ einer Lokation im d-dimensionalen euklidischen Raum und D einer Indexmenge mit $D \subset \mathbb{R}^d$.
Die gebräuchlichste Methode der Geostatistik ist das *Kriging*. Mit Hilfe des sogenannten Variogramms werden hier räumliche Abhängigkeiten ausgedrückt und so die Schätzung von Werten an unbeobachteten Orten ermöglicht.
Gitterdaten hingegen werden mittels Markov-Zufallsfeldern modelliert. Die Ähnlichkeit bzw. Abhängigkeit zweier Regionen wird in diesem Fall über deren Nachbarschaftsverhältnis berücksichtigt.
Beide Ansätze lassen sich in die Theorie der Geoadditiven Modelle einbetten. Das Programmpaket *BayesX* (Umlauf et al. 2015) stellt die nötigen Funktionen zur Verfügung um eine Auswertung dieser Modelle in R (R Core Team 2014) durchzuführen.
Es wurden beispielhaft zwei Datensätze mit Hilfe der besprochenen Methoden analysiert. Bei der Auswertung von relativen Grünwerten aus Webcam-Bildern zur Bestimmung phänologischer Phasen konnte kein strukturierter räumlicher Effekt festgestellt werden. Es wurden deshalb unabhängige zufällige Effekte zur Modellierung verwendet. Möglicherweise könnte das Modell durch die Aufnahme weiterer Kovariablen (Regenfallmenge, Temperatur, Höhe des Standortes etc.) weiter verbessert werden. Dem Verlauf des relativen Grünwertes liegt zudem an jeder Station die gleiche funktionale Form zu Grunde. Der Grünwert liegt auf einem Grundniveau bis er ab dem Erscheinen der ersten Blätter (SOS) bis zur vollständigen Reife des Laubs (MAT) auf

einen Maximalwert ansteigt. Es zeigen sich jedoch Phasenverschiebungen, Veränderungen in der Differenz zwischen SOS und MAT sowie in der Amplitude (vgl. Abbildung 9.1).

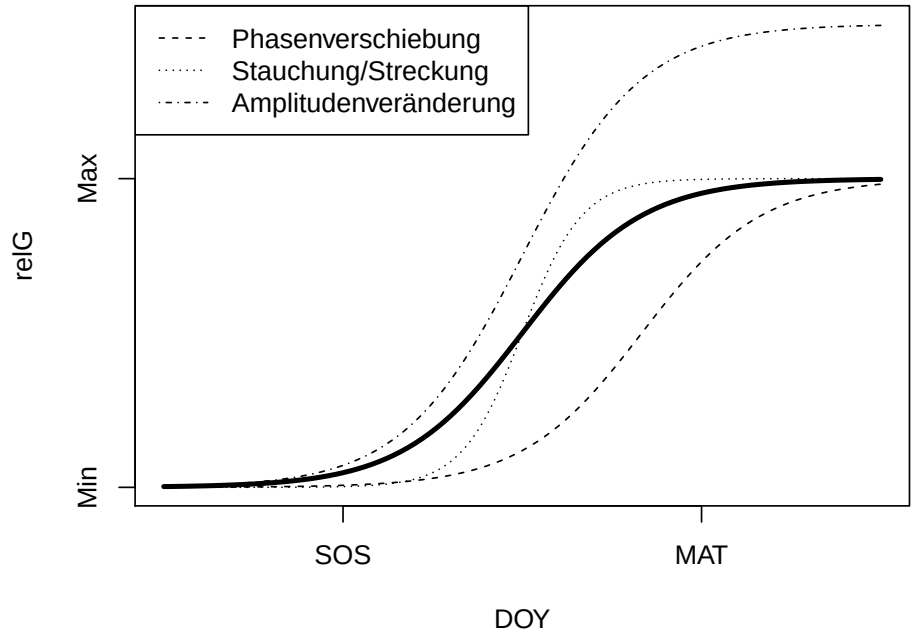


Abbildung 9.1.: Theoretischer Anstieg des relativen Grünwertes mit möglichen Veränderungen

Sinnvoll wäre hier über diese Arbeit hinaus eine Analyse auf Basis funktionaler Daten und einer Time-Warping-Funktion (siehe z.B. Silverman & Ramsay (2005)). Möglicherweise lässt sich dadurch auch doch noch ein räumlicher Effekte identifizieren.
Bei der Analyse der Herkunft von Studienanfängern an der LMU wurde ein starker regionaler Bezug bei der Studienortswahl festgestellt. Die Anzahl der Studienanfänger in den Kreisen nahm mit der Entfernung zur Hochschule stetig ab. Beim räumlichen Effekt zeigte sich ein erhöhter Erwartungswert in den südöstlichen Gebieten Bayerns. Außerdem scheint die Anziehungskraft der eigenen Hochschule in Landkreisen mit Universitätsstadt größer zu sein, als die der LMU. Interessant wäre ein weiterer Vergleich mit anderen Universitäten des Bundesgebiets. Lohnend wäre womöglich eine Gegenüberstellung mit zentraler gelegenen Universitäten. Dabei wäre es außerdem sinnvoll von der einfachen Distanz zur einer Definition über die Erreichbarkeit (via Auto, öffentliche Verkehrsmittel etc.) überzugehen.

A. Datenaufbereitung

Vor der eigentlichen Auswertung mussten zunächst die vorliegenden Daten geeignet aufbereitet werden. Auf der beigelegten CD befinden sich die fertigen Datensätze. Der folgende Abschnitt dient der Illustration der nötigen Vorbereitungsschritte.

A.1. Phänologie

Es wurden drei Datensätze zur Verfügung gestellt.

- Metadaten zu den Stationen: Name, Website, Geokoordinaten (Grad, Minuten, Sekunden)
- Rohdaten: Station, filename, meanR, meanG, meanB, relR, relG, relB
- Datenqualität: Beeinträchtigungen, Distanz des Baums zur Kamera, Typ des Baums, ROI, ausgewählte Tageszeit

Zunächst wurden die Geokoordinaten (Längen- und Breitengrad) in Dezimalgrad umgerechnet, um eine Verwendung in R zu ermöglichen. Dies geschieht nach der Formel

$$Koord_{dez} = Grad + Minute/60 + Sekunde/3600.$$

Außerdem konnte aus der File-Bezeichnung der Rohdaten das Datum der jeweiligen Messung extrahiert werden.
Mit Hilfe des Qualitätsdatensatzes wurden die Stationen im nächsten Schritt auf 182 verbleibende eingeschränkt. Es wurde jeweils nur eine ROI (Region of Interest) pro Standort ausgewählt. Außerdem wurde die Art des betrachteten Baums innerhalb der ROI auf den Typ 1 festgelegt. Leider fehlt hier eine ausreichende Dokumentation der Kodierung. Es wird aber davon ausgegangen, dass es sich dabei um Laubbäume handelt. In den Fällen in denen mehrere Kameras pro Station verzeichnet waren, wurde diejenige ausgewählt, die eine geringere Distanz zum Aufnahmeobjekt aufweist.
Auch die Kodierung der Störungen ist nicht dokumentiert. Es wurden deshalb alle Stationen gelöscht, die einen Eintrag enthielten. Um welche Störung es sich dabei handelte, konnte nicht festgestellt werden. Außerdem kam es in einigen Fällen zu technischen Problemen, die dazu führten, dass keine neuen Bilder gespeichert, sondern das letzte Bild mehrfach in den Datensatz mit aufgenommen wurde. Hier wurden die duplizierten Beobachtungen gelöscht.

A.2. Hochschulen

Aus der Arbeit am Forschungsdatenzentrum konnten Datensätze über die Anzahl der Studienanfänger mit HZB aus den einzelnen Kreisen Deutschlands erstellt werden. Dabei kam es zu Sperrungen bei geringen Fallzahlen (< 3) innerhalb eines Kreises. Diese fehlenden Werte wurden im Nachhinein zufällig mit 1 oder 2 ersetzt. Zur Analyse standen danach der Name des Kreises, die Kreiskennziffer, sowie die Anzahl Studienanfänger zur Verfügung.
Für die räumliche Analyse mussten diese Daten einem Shapefile als Metadaten hinzugefügt werden. Dabei mussten für die Jahre 2004 und 2011 unterschiedliche Datensätze verwendet werden, da es in der Zwischenzeit zu Gebietsveränderungen im Bundesgebiet gekommen ist. Das Shapefile für 2004 stammt aus der *GADM database of Global Administrative Areas* (2004). Die Karte für 2011 konnte vom *Bundesamt für Kartographie und Geodäsie* (2011) bezogen werden.
Für 2004 waren im Shapefile keine Kennziffern vorhanden, sodass die Daten anhand der Kreisbezeichnung zugeordnet werden mussten. Da diese nicht einheitlich sind, mussten sie zunächst angepasst werden um ein Matching zu ermöglichen. Für 2011 konnte dies stark vereinfacht mit Hilfe der Kennziffern durchgeführt werden.
Um die erwartete Anzahl an Studienanfängern pro Kreis (E_i) berechnen zu können, musste zunächst die "Population unter Risiko" (P_i) bestimmt werden. Da keine Daten zur Verteilung der Personen mit HZB im Bundesgebiet vorliegt, wurde diese als homogen angenommen und die Population unter Risiko durch die Einwohnerzahl der Kreise geschätzt. Diese stehen über die GENESIS-Online Datenbank des Statistischen Bundesamtes zur Verfügung.
Die erwartete Anzahl an Studienanfängern in Kreis i wurde dann gemäß

$$E_i = P_i r_+,$$

mit $r_+ = \frac{O_+}{P_+}$, berechnet (vgl. Kapitel 6).
Außerdem wurde die (Great-Circle-)Distanz der Zentroiden der einzelnen Kreise zur LMU (Koordinaten: N48° 9' 2.484" E11° 34' 49.296") berechnet.

B. Inhalt der CD-Rom

Auf beigelegter CD-Rom finden sich folgende Ordner und Dateien:

- Im Ordner `Daten/` befinden sich folgende Unterordner:
  - `Hochschulen/`:
    - `Germany_grid.Rdata`: Grid über Deutschland.
    - `[lmu, hub][04, 11]_[geo, agg].Rdata`: SpatialPolygonsDataFrame mit der Anzahl Studienanfänger für die LMU bzw. HUB für die Jahre 2004 und 2011.
  - `Phenology/`: SpatialPointsDataFrame mit den rel. Grünwerten an den Koordinaten der einzelnen Stationen.
  - `Deutschlandkarten/`:
    - SpatialPolygonsDataFrame Deutschlands, der Bundesländer, Regierungsbezirke und Kreise Deutschlands (2004) im .Rdata-Format.
    - Kreise (2011) als Shapefile.
- Im Ordner `Abbildungen/` befinden sich die in der Arbeit abgebildeten Grafiken.
- Im Ordner `Output/` befinden sich die Modelloutputs im .Rdata-Format:
  - `relG_mod[0, 1, 2, 2.1, 3, 3.1, 4].Rdata`: Modelle gemäß Tabelle 8.1.
  - `model[04, 11]_[, 1]_[, hub].Rdata`: Modelle der Hochschulen.
- Im Ordner `RCode/` befinden sich folgende Dateien:
  - `prepare_[hochschulen, phenology].R`: Code zur Erstellung der fertigen Datensätze (nicht lauffähig).
  - `helpfunctions.R`: Hilfs-Funktionen für andere Auswertungen.
  - `descriptive_phenology.R`: Deskriptive Auswertung der Phänologie-Daten.
  - `analysis_[hochschulen, phenology].R`: Code zur Analyse in Kapitel 8.
  - `graphics_[chap4, chap5].R`: Code für die Grafiken in Kapitel 4 und 5.
  - `sub_kriging_hochschulen.R`: Zusätzlicher Code zur Analyse der Hochschuldaten mittels Kriging.

- Im Ordner `Masterarbeit/` befinden sich folgende Dateien:
  - `Masterarbeit.pdf`: Die Masterarbeit zum Druck formatiert.
  - `Masterarbeit_elektronisch.pdf`: Die Masterarbeit als elektronische Version.

C. Eidesstattliche Erklärung

Hiermit versichere ich, Christine Jula, die vorliegende Masterarbeit selbstständig und lediglich unter Benutzung der angegebenen Quellen und Hilfsmittel verfasst zu haben.

München, den 13. April 2015

Christine Jula

Literaturverzeichnis

Anselin, L. (1995), 'Local indicators of spatial association—lisa', *Geographical Analysis* **27**(2), 93–115.
URL: *http://dx.doi.org/10.1111/j.1538-4632.1995.tb00338.x*

Anselin, L., Cohen, J., Cook, D., Gorr, W. & Tita, G. (2000), 'Spatial analyses of crime', *Criminal justice* **4**(2), 213–262.

Banerjee, S., Carlin, B. & Gelfand, A. (2004), *Hierarchical Modeling and Analysis for Spatial Data*, Chapman & Hall/CRC Monographs on Statistics & Applied Probability, Taylor & Francis.
URL: *http://books.google.de/books?id=A_R4AgAAQBAJ*

Belitz, C., Brezger, A., Kneib, T., Lang, S. & Umlauf, N. (2015), *BayesX: Software for Bayesian Inference in Structured Additive Regression Models*. Version 1.0.
URL: *http://www.BayesX.org/*

Belitz, C. & Lang, S. (2008), 'Simultaneous selection of variables and smoothing parameters in structured additive regression models', *Computational Statistics & Data Analysis* **53**(1), 61–81.

Besag, J. (1974), 'Spatial interaction and the statistical analysis of lattice systems', *Journal of the Royal Statistical Society. Series B (Methodological)* **36**(2), pp. 192–236.
URL: *http://www.jstor.org/stable/2984812*

Besag, J., York, J. & Mollié, A. (1991), 'Bayesian image restoration, with two applications in spatial statistics', *Annals of the Institute of Statistical Mathematics* **43**(1), 1–20.
URL: *http://dx.doi.org/10.1007/BF00116466*

Bivand, R. S., Pebesma, E. J. & Gomez-Rubio, V. (2013), *Applied Spatial Analysis with R*, UseR! Series, Springer.

Brezger, A. & Lang, S. (2006), 'Generalized structured additive regression based on bayesian p-splines', *Computational Statistics and Data Analysis* **50**(4), 967 – 991.
URL: *http://www.sciencedirect.com/science/article/pii/S0167947304003214*

Bundesamt für Kartographie und Geodäsie (2011). Stand: 13. April 2015.
URL: *http://www.geodatenzentrum.de*

Clayton, D. & Kaldor, J. (1987), 'Empirical bayes estimates of age-standardized relative risks for use in disease mapping', *Biometrics* **43**(3), pp. 671–681.
URL: *http://www.jstor.org/stable/2532003*

Cliff, A. & Ord, J. (1981), *Spatial Processes: Models & Applications*, Pion.
URL: *http://books.google.de/books?id=Mi0OAAAAQAAJ*

Cressie, N. (1990), 'The origins of kriging', *Mathematical Geology* **22**(3), 239–252.
URL: *http://dx.doi.org/10.1007/BF00889887*

Cressie, N. (1993), *Statistics for spatial data*, Wiley series in probability and mathematical statistics: Applied probability and statistics, J. Wiley.
URL: *http://books.google.de/books?id=4SdRAAAAMAAJ*

Cressie, N. A. (1996), 'Change of support and the modifiable areal unit problem'.

Cressie, N. & Wikle, C. (2011), *Statistics for Spatio-Temporal Data*, CourseSmart Series, Wiley.
URL: *http://books.google.de/books?id=-kOC6D0DiNYC*

Dhital, S. (2011), Use of public internet-connected webcam images for monitoring plant phenology in germany, Master's thesis, Technische Universität München.

Diggle, P. (2003), *Statistical Analysis of Spatial Point Patterns*, Mathematics in biology, Arnold.
URL: *http://books.google.de/books?id=fnFhQgAACAAJ*

Fahrmeir, L., Kneib, T. & Lang, S. (2004), 'Penalized structured additive regression for space-time data: a bayesian perspective', *Statistica Sinica* **14**(3), 731–762.

Fahrmeir, L., Kneib, T. & Lang, S. (2009), *Regression*, Statistik und ihre Anwendungen, Springer.
URL: *http://books.google.de/books?id=rZi6eqYXkdYC*

Forschungsdatenzentrum München (n.d.). Stand: 13. April 2015.
URL: *http://www.forschungsdatenzentrum.de/standorte/l_muenchen.asp*

GADM database of Global Administrative Areas (2004). Stand: 13. April 2015.
URL: *http://gadm.org/home*

Hengl, T., Heuvelink, G. B. & Stein, A. (2003), 'Comparison of kriging with external drift and regression-kriging', *Technical note, ITC* **51**.

Krige, D. (1951), 'A statistical approach to some basic mine valuation problems on the witwatersrand', *Journal of the Chemical, Metallurgical and Mining Society of South Africa* **52**(6), 119–139.

Matheron, G. (1962), *Traité de géostatistique appliquée. 1 (1962)*, Vol. 1, Editions Technip.

Matheron, G. (1963), 'Principles of geostatistics', *Economic Geology* **58**(8), 1246–1266.

Matheron, G. (1971), *The Theory of Regionalized Variables and Its Applications*, Centre de Morphologie Mathématique Fontainebleau: Les cahiers du Centre de Morphologie Mathématique de Fontainebleau, École national supérieure des mines.
URL: *http://books.google.de/books?id=TGhGAAAAYAAJ*

R Core Team (2014), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *http://www.R-project.org/*

Rue, H. & Held, L. (2005), *Gaussian Markov Random Fields: Theory and Applications*, Vol. 104 of *Monographs on Statistics and Applied Probability*, Chapman & Hall, London.

Schaeben, H., Akin, H. & Siemes, H. (2013), *Praktische Geostatistik: Eine Einführung für den Bergbau und die Geowissenschaften*, Hochschultext, Springer Berlin Heidelberg.
URL: *https://books.google.de/books?id=shmnBgAAQBAJ*

Silverman, B. & Ramsay, J. (2005), *Functional Data Analysis*, Springer.

Statistische Ämter des Bundes und der Länder (2015), 'Datenzugang'. Stand: 13. April 2015.
URL: *http://www.forschungsdatenzentrum.de/datenzugang.asp*

Statistisches Bundesamt, Wiesbaden (2015), 'GENESIS-Online Datenbank'.
URL: *https://www-genesis.destatis.de/genesis/online*

Studentenstatistik (n.d.). Stand: 13. April 2015.
URL: *http://www.forschungsdatenzentrum.de/bestand/studenten/index.asp*

Tobler, W. R. (1970), 'A computer movie simulating urban growth in the detroit region', *Economic Geography* **46**, pp. 234–240.
URL: *http://www.jstor.org/stable/143141*

Umlauf, N., Adler, D., Kneib, T., Lang, S. & Zeileis, A. (2015), 'Structured additive regression models: An R interface to BayesX', *Journal of Statistical Software* **63**(21), 1–46.
URL: *http://www.jstatsoft.org/v63/i21/*

Waller, L. & Gotway, C. (2004), *Applied Spatial Statistics for Public Health Data*, Wiley Series in Probability and Statistics, Wiley.
URL: *http://books.google.de/books?id=OuQwgShUdGAC*