

# **Methoden der räumlichen Statistik und Ökonometrie zur Konzentrationsanalyse.**

Eingereicht am 16. September 2020, als Abschlussarbeit  
zur Anerkennung des Grades **BSc Wirtschaftsmathematik**.

**Falko Krause Cauduro**  
**Matrikel: 2714676**

Fakultät 1: Mathematik, Informatik, Physik, Elektro- und Informationstechnik  
Brandenburgische Technische Universität Cottbus-Senftenberg

**Gutachter an der BTU: Prof. Dr. rer. nat. habil. Ralf  
Wunderlich**



*„[...] Since all models are wrong the scientist cannot obtain a “correct” one by excessive elaboration. On the contrary, following William of Occam he should seek an economical description of natural phenomena. Just as the ability to devise simple but evocative models is the signature of the great scientist so overelaboration and overparameterization is often the mark of mediocrity.“*

George E.P. Box, 1976

## Abstract

Es werden zunächst klassische Konzentrations- und Entropiemaße der deskriptiven räumlichen Statistik angewandt. Darauf folgt die Anwendung einer Clusteranalyse durch globale Indizes (Moran's I und Geary's c) und statistischer Tests auf Signifikanz, sowie spezialisierter lokaler Maße und Tests (Local Getis GI, Local Morans' I). Daran knüpfen räumliche Regressionsmodelle an, deren Parameter durch Inferenzverfahren ermittelt werden.

## Erklärung

Der Verfasser erklärt, dass er die vorliegende Arbeit selbstständig, ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt hat. Die aus fremden Quellen (einschließlich elektronischer Quellen) direkt oder indirekt übernommenen Gedanken sind ausnahmslos als solche kenntlich gemacht. Wörtlich und inhaltlich verwendete Quellen wurden entsprechend den anerkannten Regeln wissenschaftlichen Arbeitens zitiert. Die Arbeit ist nicht in gleicher oder vergleichbarer Form, auch nicht auszugsweise im Rahmen einer anderen Prüfung bei einer anderen Hochschule vorgelegt worden.

Sie wurde bisher auch nicht veröffentlicht.

Ich erkläre mich damit einverstanden, dass die Arbeit mit Hilfe eines Plagiatserkennungsdienstes auf enthaltene Plagiate überprüft wird.

Ort, Datum

Unterschrift des Verfassers



## **Danksagung**

Auf diesem Wege danke ich all jenen Personen, die mich während der Arbeiten an diesem umfangreichen Thema unterstützt haben.

Besonderer Dank gilt hierbei meinem Betreuer Herrn Prof. Dr. Ralf Wunderlich, der mir mit Rat und Fachwissen stets zur Seite steht sowie Herrn Prof. Dr. Carsten Hartmann für das bereitwillige Teilen seines Wissensschatzes und seiner Erfahrung. Ich schätze mich äußerst glücklich, von beiden Persönlichkeiten zu lernen.

Ganz besonders danke ich meiner Familie für ihre unbedingte Unterstützung und all die Energie und Inspiration.



# Inhaltsverzeichnis

<b>Abstract</b>	i
<b>Danksagung</b>	iii
<b>1 Einführung</b>	1
<b>2 Datenlage und GIS Konzepte</b>	4
2.1 GIS Software . . . . .	4
2.2 Projektionen und Koordinatensysteme . . . . .	5
2.3 Datengrundlage . . . . .	7
2.3.1 Merkmale und Attribute . . . . .	7
2.3.2 Raumeinheiten . . . . .	8
<b>3 Datentypen und Methodenklassen</b>	10
3.1 Besonderheiten räumlicher Daten . . . . .	10
3.2 Allgemeines Modell räumlicher Stochastik . . . . .	17
3.3 Räumliche Nachbarschaft . . . . .	22
<b>4 Räumliche Autokorrelation</b>	26
4.1 Theorie und Indizes räumlicher Autokorrelation . . . . .	26
4.2 Indizes und Tests der Autokorrelationsmessung . . . . .	28
4.2.1 Globale Autokorrelationsmaße . . . . .	30
4.2.2 Lokale Autokorrelationsmaße . . . . .	37
<b>5 Räumliche Regression</b>	41

5.1	Spatial Statistics – Allgemeine Räumliche Autoregression . . . . .	42
5.2	Simultaneous Autoregressive Models - SAR . . . . .	44
5.2.1	Räumlich fehlerbezogenes Modell - SEM . . . . .	46
5.2.2	Räumlich verzogenes Modell - SLM . . . . .	48
5.2.3	Kombiniertes Modell . . . . .	54
5.3	Conditional Autoregressive Models - CAR . . . . .	55
5.4	Parameterschätzung . . . . .	55
5.4.1	ML-Schätzung im klassischen Regressionsmodell . . . . .	56
5.4.2	ML-Schätzung im verallgemeinerten Regressionsmodell . . . . .	58
5.4.3	ML-Schätzung für SE-Modelle . . . . .	60
<b>6</b>	<b>Analyse und Auswertung</b>	<b>62</b>
6.1	Charakterisierung des Beobachtungsgebietes . . . . .	62
6.1.1	Quantilkarten ausgesuchter Attribute . . . . .	63
6.2	Globale Indizes . . . . .	65
6.2.1	Globaler Moran Index . . . . .	65
6.2.2	Monte-Carlo Simulation . . . . .	66
6.3	Lokale Indizes . . . . .	67
6.3.1	Lokaler Moran Index . . . . .	67
<b>7</b>	<b>Zusammenfassung und Wertung</b>	<b>69</b>
7.1	Anpassungen und Erweiterungen . . . . .	69
	<b>Bibliographie</b>	<b>70</b>

# Tabellenverzeichnis

3.1	Klassifikationsübersicht der räumlichen Stochastik mit Anwendungsbeispiele. Quelle: eigene Darstellung in Anlehnung an Cressie (1993) . . . . .	18
3.2	Buchkapitel gemäß Klassifikation durch Cressie (1993) . . . . .	22
4.1	Kritische Ablehnungsbereiche der Alternativhypotesen . . . . .	35



# Abbildungsverzeichnis

3.1	Randomisierte Bevölkerungsdichten . . . . .	11
3.2	Zoneneffekt . . . . .	14
3.3	Zoneneffekt . . . . .	15
3.4	Zoneneffekt . . . . .	16
3.5	Nachbarschaftsgraph . . . . .	24
3.6	Nachbarschaftssysteme . . . . .	24
3.7	Nachbarschaftssysteme . . . . .	25
6.1	Nachbarschaftsrelationen der Metropolregion . . . . .	63
6.2	Quantilkarten der Metropolregion. Li: Fläche Re: Arbeitnehmer je Bev. . .	64
6.3	Histogramme ausgewählter Attribute . . . . .	64
6.4	Lokale Autokorrelation . . . . .	68



# Kapitel 1

## Einführung

Ein Großteil aller uns umgebenden Informationen weist einen inhärent räumlichen Bezug auf, oft ohne dass wir diesen bewusst wahrnehmen. Die räumliche und zeitliche Orientierung sind fundamentale Achsen der Informationsverarbeitung, von makroskopischen Sternenformationen über mikroskopische Proteinfaltungen hin zu unserer persönlichen Mobilität im Alltag. Wir nutzen diese Achsen, um alle weiteren Informationen in Beziehungen zu setzen, zu transformieren und im neuen Kontext zu verwenden. Diese Arbeit gibt einen Einblick in die wissenschaftliche Analyse räumlicher Daten zum Verständnis komplexer naturwissenschaftlicher und sozialer Phänomene.

Es werden in dieser Arbeit Methoden der *raumdatenbezogenen Statistik* (engl. Spatial Statistics) untersucht und angewendet. Über die historischen Analysen mathematischer Zusammenhänge von Datenpunkten mit räumlichen Bezug zueinander wird in Kapitel 3 berichtet. Darüber hinaus hat die Anwendung auf reale Datensätze starken interdisziplinären Charakter. Der Forschungszweig der *Spatial Analysis* im geografischen Kontext mit *Geoinformationssystemen (GIS)* ist in dieser Hinsicht vergleichsweise jung. Die ersten bahnbrechenden Arbeiten stammen aus den späten 1950er Jahren und wurden grundlegend von der Entwicklung der Informationstechnologie (engl. „Geography Quantitative Revolution“) getrieben. Im Prinzip handelt es sich zu Beginn um die direkte Verknüpfung der Kartographie mit computergestützten Datenbanken. Als wichtige Wegbereiter sind Roger F. Tomlinson am *Canada Geographic Information System* sowie Howard T. Fisher als

Gründer des *Harvard Laboratory for Computer Graphics and Spatial Analysis* bekannt. Die Überschneidungen der Geographie und Geologie, Biologie, Hydrographie, Meteorologie und vielen weiteren Fachgebieten begünstigte die Verbreitung dieser neuen Methodologie und Technologie. Weitere Anwendungsbeispiele unter Implementierung ähnliche Algorithmen und statistischer Methoden finden sich in der Ökologie (e.g. makroskopische Verteilungen von Lebensräumen, mikroskopische Proteinfaltungen), Soziologie (Entwicklung von Migrationsströmen), Anthropologie (Ausbreitungsmuster historischer Siedlungen), Astronomie und Kosmologie (Systemformationen und Galaxiencluster) sowie Neurologie (Auswertung von Gehirnscans). In der Informatik entwickelt der Bereich *Spatial Data Mining* neue Algorithmen zur effizienten Verarbeitung großer Datenmengen mit Raumbezug und Nachbarschaftsbeziehungen.

Seit den frühen Tagen der Geoinformationstechnologie werden rasante Fortschritte erzielt. Durch technische Errungenschaften wie etwa LIDAR, Dronen oder Formationen von Minisatelliten werden immense Ströme an Geodaten generiert. Diese dienen etwa der landwirtschaftlichen Analyse von Feldern, der Sichtung illegaler Schifffahrts- oder Bebauungs- und Rodungsaktivitäten, der Fernüberwachung des Bewuchses von Bahntrassen und vielem mehr.

Klassische Industrien und Einsatzgebiete zur Auswertung von georeferenzierten Daten umfassen nunmehr die Agrar- und Forstwirtschaft, Bergbau- und Förderindustrien, Luft- und Raumfahrt, Navigation, Nautik und maritime Wirtschaft, Sicherheitstechnik- und Rüstungsindustrie, Transport- und Logistik, Anlagen- und Immobilienerfassung sowie Kataster und Stadtplanung, Infrastruktur und Instandhaltung von Schienen-, Telekommunikation-, Wasser-, Pipeline- und Stromnetzen, Katastrophenmanagement, Naturschutzgebiet- und Wildtierüberwachung und viele weitere.

Die Motivation dieser Bachelorarbeit liegt in einer möglichst breiten Aufgliederung vieler Aspekte, ohne umfangreiche Details und tiefere Analysen leisten zu können, wie sie etwa ein Handbuch bietet . Der Fokus liegt auf einer schnell und einfach lesbaren Aufbereitung

der Grundlagen räumlicher Statistik mit einer groben Einteilung wichtiger Teilgebiete, sodass dem Leser eine schnelle und effiziente Übersicht über einen großen Forschungszweig zu seiner eigenen Orientierung gegeben wird.

Zugleich wird durch die praktische Anwendung auf lokale Geodaten ein beispielhafter Einblick in den Prozess der Modellbildung und die Pipeline der Datenanalyse geliefert. Dieser Spagat zwischen theoretischer Abstrahierung und praktischer Anwendung soll einen Eindruck der Anwendbarkeit und Relevanz vermitteln und die Bandbreite an Anwendungsmöglichkeiten verdeutlichen. Dies möge den Leser zu eigenen Projekten motivieren und ihm zusätzliche Werkzeuge und Fähigkeiten zur Analyse der immensen Datensätze zu liefern, welche inzwischen auch frei verfügbar sind.

# Kapitel 2

## Datenlage und GIS Konzepte

Dieses Kapitel erläutert technische Voraussetzungen der Geoinformationssysteme (GIS) und ihrer räumlichen Datenanalyse, bevor im darauf folgenden Kapitel die Theorie eingeführt wird.

### 2.1 GIS Software

Der Markt für GIS-Anwendungen ist breit gefächert und stark segmentiert. Viele Anbieter sind auf Nischenlösungen spezialisiert und technische Neuerungen ändern die Marktverhältnisse fortwährend, was sowohl auf der Entwicklungs- als auch Nutzerseite solcher Software für ständigen Bedarf an Fachpersonal sorgt. Der seit den 80er Jahren etablierte ArcGIS-Anbieter ESRI sowie Open Source Alternativen wie QGIS, GRASS, und SAGA sind insbesondere in der akademischen Forschung und bei Behörden verbreitet. Dies betrifft zudem R und Python, welche durch Pakete bzw. Bibliotheken zu voll funktionsfähigen GIS-Systemen erweiterbar sind sich hervorragend mit anderen Softwarelösungen integrieren lassen. Gegenüber klassischen relationalen Datenbanksystemen werden spezialisierte GIS-Datenbanken benötigt, welche mit den speziellen Vektorgeometrien und Rasterdaten operieren können. PostGIS für PostgreSQL ist ein Beispiel für diese Entwicklung. Kommerzielle Lösungen wie Topobase und Map3D der Firma Autodesk, MicroStation von Bentley Systems sowie Geomedia von Intergraph finden vorwiegend als spezialisierte Industrielösungen Anwendung. Online-GIS mit Cloud-Anbindungen wie

Google Maps, HERE und OpenStreetMap erweitern das Spektrum an Angeboten insbesondere für den privaten Endnutzer.

Die Daten dieser Arbeit wurden zunächst übersichtsweise mit MS Excel untersucht sowie nach relevanten Kriterien zusammengefasst und schließlich mittels der Open-Source Scriptsprache R weiterverarbeitet, einheitlich archiviert und schließlich numerisch modelliert und ausgewertet. Zwar ist der Umgang mit vielen Daten innerhalb von MS Excel zunächst intuitiver und interaktiver möglich, jedoch stößt die Software bei größeren Datensätzen schnell an Ihre Grenzen. Insbesondere ist die Analyse außerhalb klassischer „Spreadsheets“ nur bedingt möglich und bietet nicht genug Funktionalität für komplexe Datenanalysen. So existiert etwa keine Anbindung zur Auswertung geografischer Daten. Hierfür müsste etwa auf PowerBI, einer weiteren Software im Microsoft-Katalog zurückgegriffen werden. Außerdem werden die Rohdaten nicht getrennt von der Modellierung und Analyse bearbeitet. Dies verletzt einen wichtigen Grundsatz für eine praktische und sichere Analyse in getrennten Dateien und verhält sich ähnlich zur Nutzung von LaTeX gegenüber Word.

R liefert eine etablierte Platform und mit dem zentralen *cran*-Register ein sehr stabiles System an Bibliotheken. Zu nennen sind insbesondere die Pakete *dplyr* und *ggplot2* aus dem Paketverbund *tidyverse* von Hadley Wickham unter aktiven Entwicklung. Derartige Pakete stellen einen großen Vorteil gegenüber vielen anderen Lösungen dar. Für die Raumdaten bieten sich *raster* als Paket für Rasterdaten sowie *sf* für Vektordaten an, welches das etablierte Paket *sp* ablöst. Einen guten Überblick bieten (Fischer and Getis, 2010, S. 53) sowie Bivand et al. (2013).

## 2.2 Projektionen und Koordinatensysteme

Wichtige Grundbegriffe und Konzepte der raumdatenbezogenen Analyse sind die räumlichen *Merkmale* (engl. features), *Träger* (engl. supports) und *Attribute* (engl. attributes). *Merkmale* bilden als Raumpunkte, Linien, Flächen und Volumen die Grundbausteine der In-

formationsverarbeitung. Räumliche *Träger* definieren die genaue Größe, Form und Orientierung der Merkmale als zusätzliche Referenzinformation. Punkte etwa weisen weder Größe, Fläche noch Orientierung auf und besitzen den kleinsten Träger aller Merkmale. Regionen hingegen unterscheiden sich in allen drei Dimensionen voneinander. *Attribute* wiederum stellen Messwerte und Zufallsvariablen dar, welche mit den räumlichen *Merkmalen* verknüpft werden. Für eine genaue Übersicht und weitere Details wird auf (Waller and Gotway, 2004, S.38) verwiesen.

Das mathematische Teilgebiet der *Geodäsie* behandelt die möglichst genaue Lokalisierung und Referenz räumlicher Merkmale auf der Erdoberfläche und bedient sich hierfür fundamentaler Konzepte aus Geometrie und Topologie. Eine essentielle Grundlage bei der Arbeit mit georeferenzierten Daten bilden geographische und projizierte Koordinatensysteme.

Ein *geographisches Koordinatensystem* (engl. geographic coordinate system) (**CGS**) definiert durch Längen- und Breitengrade ein perfektes sphärisches Koordinatensystem. Dieses Gradnetz (engl. graticule) lokalisiert Orte eindeutig auf einer Sphäre und mit Einschränkungen auf der imperfekten ellipsoiden Erdoberfläche.

Zur einfacheren Analyse und Darstellung werden diese dreidimensionalen, sphärischen Daten auf zweidimensionale Oberflächen wie Karten und Bildschirme projiziert. Diese Transformation geht zwangsläufig mit Verzerrungen von Form, Flächeninhalt, Distanzen und Richtungen einher. Aus diesem Grund bestehen zahlreiche verschiedene Projektionen (e.g. Mercator, Albers) mit einer eigenen Balance der Verzerrungen. Als Ausgangspunkt der Transformation bestehen verschiedene Standardellipsoide zur Approximation der Erdform. Die Orientierung und genaue Position des Ellipsoids relativ zur Erde wird als *geodätisches Datum* bezeichnet. Ein lokales Datum approximiert dabei einen kleinen Teilbereich der Erdoberfläche besonders gut. In dieser Analyse verwenden wir den Standard *WGS84* (World Geodetic System of 1984). Nach der Projektion werden Vektoren, Polygone und Zentroide erzeugt. (Waller and Gotway, 2004, S. 49) enthält hierzu detaillierte Berechnungsvorschriften.

## 2.3 Datengrundlage

Genutzt werden öffentlich verfügbare Daten zu Bevölkerungsgrößen, Arbeitsmarktsituation, Flächeneigenschaften und weiteren Wirtschaftsdetails. Die *shape*-Dateien zur Analyse von Vektordaten von Gemeinden und Bezirken sind durch das Geoportal Brandenburg und das OpenData Portal der Stadt Berlin frei verfügbar.

### 2.3.1 Merkmale und Attribute

Neben dieser Grundlage werden spezielle Daten des Unternehmensregisters über die Standorte von Unternehmen aus allen erfassten Wirtschaftsbranchen nach WZ-Systematik genutzt. Diese werden durch das *Amt für Statistik Berlin-Brandenburg* öffentlich bereitgestellt. Die Daten sind für Brandenburg auf Gemeindeebene und für Berlin auf Bezirkslevel aggregiert und liegen jahresweise zusammengefasst für einen Zeitraum von 2006 bis 2015 vor. Die Unternehmensstandorte des Registers liegen aus Datenschutzgründen nicht als punktgenaue GPS-Standorte oder Adressen vor, sondern geben nur die Gemeinde an, in welcher die Unternehmen angesiedelt sind. Auf diese räumliche Verdichtung der Daten und damit verbundene Probleme wird in Abschnitt 3.1 genauer eingegangen. Weiterhin sind die Unternehmensgrößen nach Anzahl der Arbeitnehmer in 5 allgemeinen Größenklassen angegeben. Zur besseren Berechnung wurde hieraus ein ganzzahliger Schätzwert der Unternehmensgröße abgeleitet, um eine Beschränkung auf Methoden kategorischer Variablen zu vermeiden. Weitere Merkmale über die Unternehmensgröße wie etwa Umsatz oder Flächennutzung fehlen hierbei. Am problematischsten erweist sich jedoch die völlige Anonymisierung der einzelnen Unternehmen im Zeitverlauf. Es ist unbekannt, welche Merkmalsausprägung ein einzelnes Unternehmen in anderen Jahren aufweist, oder ob es dort überhaupt existiert. Somit lassen sich einzelne Unternehmen weder örtlich noch zeitlich nachverfolgen und Aussagen über Standortwechsel können nur in allgemeiner Art und Weise getroffen werden.

Ein *Unternehmen* oder *Firma* ist gemäß des Thüringer Landesamts für Statistik eine rechtlich selbständige Einheit eines Gewerbebetriebs. Werden vom Unternehmen weitere

selbständige, meist in anderen Orten befindliche Einheiten betrieben, ohne dass sie ein Tochterunternehmen sind, werden diese *Zweigniederlassung* genannt.

In der Datenbank lassen sich sowohl Niederlassungen als auch Rechtliche Einheiten abrufen. Es geht in der folgenden Analyse um die Verteilung allgemeiner Standorte einer ganzen Branche im Raum und deren allgemeine zeitliche Entwicklung. Somit wird auf die Niederlassungen zugegriffen, welche zudem eine zahlenmäßig umfangreichere Datenbasis liefern.

Weitere Auswahlmöglichkeiten bestehen bei unterschiedlichen Tabellenformaten zur Speicherung der Informationen einzelner Gemeinden. Die Rohdaten wurden im für Datenbanken typischen LONG-Format heruntergeladen. Diese sind für Menschen unpraktisch und schlecht zu überschauen, für Programmfunctionen jedoch hervorragend geeignet und erlauben einen schnellen Datenabruft und effiziente Sortierung. Viele Merkmale werden jedoch redundant gespeichert und die Dateigröße wird somit aufgebläht. Datenbanken sind allerdings speziell für große Dateimengen ausgelegt und auf schnellen Abruf optimiert. Dieses Format wurde für die Analyse in R in ein WIDE-Format umgewandelt wie es für übersichtliche Spreadsheets und Tabellen üblich ist. Zudem passt es zu den Dataframes der Shapedateien und kann somit direkt verknüpft werden.

### 2.3.2 Raumeinheiten

Die räumlich feinste Gliederung bietet die sogenannte „Raumabgrenzungen der Regionalstatistik auf Grenzen der administrativen Einheiten“. In Brandenburg bezeichnet dies die Gemeinden und in Berlin die Bezirke. In Berlin existieren seit 2016 zudem Daten auf Ebene der Kieze, sogenannte „Lebensweltlich orientierte Räume“ (LOR), als wichtigste kleinräumige Gliederung, deren räumliche Abgrenzung nach fachlichen Kriterien festgelegt wurde.

Derartige Grenzen werden aus administrativen und politischen Gründen im Zeitverlauf angepasst. Diese Änderungen müssen bei Betrachtungen über längere Zeiträume in die

Shape-Datei integriert werden, um einheitliche Referenzen und Formate bei der Datenanalyse zu gewährleisten. Es handelt sich jedoch bisher in Brandenburg und Berlin ausschließlich um Eingemeindungen. Damit treten nur Vergrößerungen der Partition aus Gemeinden im Zeitverlauf auf, bzw. Verfeinerungen der Partition bei einem Rückblick in vergangene Jahre, was den Aufwand erheblich vereinfacht. Beispiele für Eingemeindungen im Beobachtungszeitraum sind etwa die Integration von „Hornow-Wadelsdorf“ in die Stadt Spremberg ab 2016 sowie die Eingemeindung von „Madlitz-Wilmersdorf“ nach „Briesen (Mark)“ ab 2014.

# Kapitel 3

## Datentypen und Methodenklassen

Dieses Kapitel führt in die Theorie und Konzepte räumlicher Methoden ein und orientiert sich hierzu an den historischen Entwicklungen.

### 3.1 Besonderheiten räumlicher Daten

Die räumliche Statistik unterscheidet sich in einigen grundlegenden Punkten von der klassischen Statistik und erfordert zusätzliche Betrachtungen der Datengrundlage sowie die Anpassung bestimmter Grundannahmen und Testvoraussetzungen.

#### Räumliche Autokorrelation

Das Konzept der räumlichen Autokorrelation nimmt einen fundamentalen Platz in jeder Raumdatenanalyse ein und wird in Kapitel 4 tiefer behandelt. Abbildung 3.1 verdeutlicht das Konzept räumlicher Autokorrelation in optischer Weise. Die starke Präsenz von räumlicher Abhängigkeit verkompliziert die Analyse, da nicht von unabhängigen Daten ausgegangen werden kann. Diese Abbildung zeigt grob auf, welches Erscheinungsbild eine über Verwaltungskreise zufällig verteilte Bevölkerung im Vergleich zur realen Situation aufwiese, und macht das Konzentrationsmaß deutlich.

Diese Beobachtung stützt die grundlegende Annahme, dass räumlich benachbarte Beobachtungen einer Attributvariablen einander ähnlicher sind als weiter entfernte. Dieser

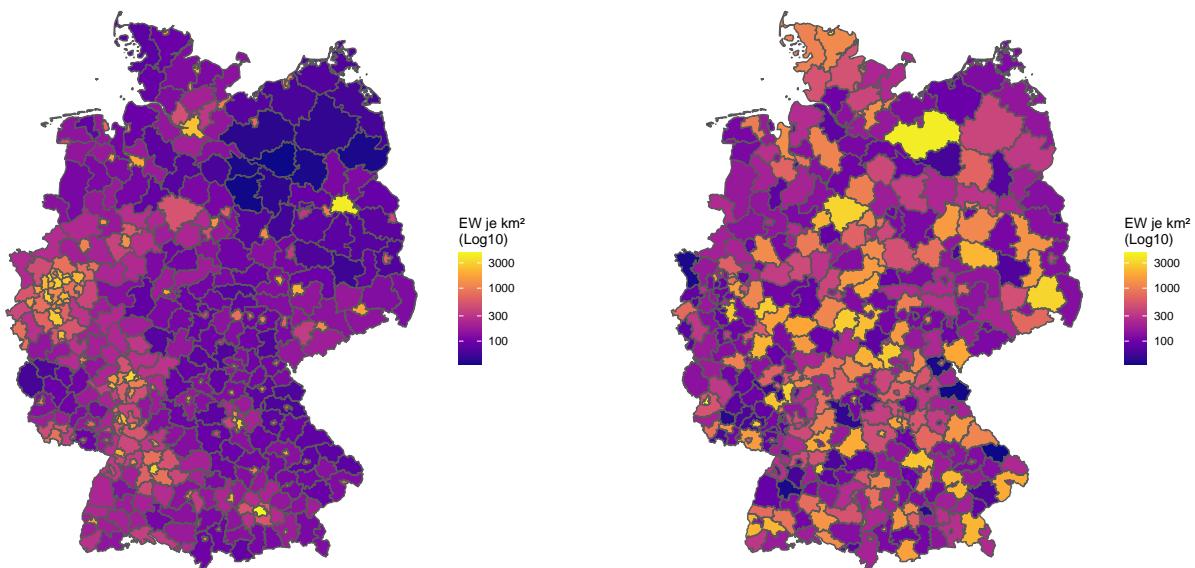


Abbildung 3.1: Bevölkerungsdichte je Kreis in 2018. Li: real Re: randomisiert

Umstand wird durch Toblers erstes Gesetz der Geographie ausgedrückt (Tobler, 1970, S.234):

*„[...] the first law of geography: everything is related to everything else, but near things are more related than distant things.“* Waldo R. Tobler

Trotz der simplen Präsentation und obgleich diese Formulierung nicht sonderlich erwähnenswert anmutet und zudem fraglich ist, ob tatsächlich „alles miteinander in Beziehung steht“ wie Tobler (2004) selbst anmerkt, so ist er der erste Wissenschaftler, der dieses grundlegende Konzept räumlicher Dynamik klar ausformulierte. Inwiefern zwischen einem universalen wissenschaftlichen Gesetz und einer beobachteten Regelmäßigkeit (engl. causality vs. regularity) unterschieden werden kann und ob in den Sozialwissenschaften von Gesetzen gesprochen werden kann, ist Gegenstand wissenschaftlicher Auseinandersetzung (e.g. Law of Utility Maximization) (Goodchild, 2004). Tobler sieht die Distanz selbst als Ursache und argumentiert somit gegen die Sicht, dass Distanz und räumliche Verteilung von Merkmalen als zwei Beobachtungen einer verborgenen Ursache nur „beobachtete Regelmäßigkeiten“ sind. Die Theorie fällt historisch in die Zeit der eingangs erwähnten Quantitativen Revolution der modernen Geographie, wird jedoch erst mit Aufkommen und Verbreitung geographischer Informationssysteme (GIS) seit den 1990er Jahren in größerem Umfang beachtet, gelehrt und referenziert (Sui, 2004).

Das Gesetz entspricht der menschlichen Intuition aus historischen Beobachtungen und ist fest verankert in wissenschaftlichen Paradigmen. In der Theorie basiert diese Gesetzmäßigkeit wiederum auf dem Konzept der *Distanzreibung* (engl. friction of distance) und daraus resultierenden Abklingfunktionen (engl. distance decay) für allgemeine räumliche Merkmale. Derartige Abklingfunktionen bilden die Grundlage für die Interpolationsmethoden der *Geostatistik*, wie in Kapitel 3.2 näher erläutert wird. Weitere Grundlage räumlicher Interpolation ist die Messung der Ähnlichkeit räumlich benachbarter Merkmale als Ziel der Theorie *räumlicher Autokorrelation* (engl. spatial autocorrelation). Die räumliche Position einer Punktmessung liefert selbst bereits implizit inhärente Zusatzinformationen über die Beobachtungswerte in der Umgebung des Beobachtungswertes. Diese Zusatzinformation stört somit die Grundannahme einer *unabhängig identisch verteilten* (i.i.d.) Grundgesamtheit, da Messungen an verschiedenen Positionen nicht unabhängig voneinander sind, insbesondere bei geringer Distanz zueinander. In Kapitel 4 werden globale und lokale Autokorrelationsindikatoren zur Detektion räumlich-zeitlicher hot- und coldspots und als Werkzeuge der Regression eingeführt.

Zunehmende Distanz erhöht zudem die Kosten der Interaktion zwischen Raumpunkten. Im Beispiel der sozio-ökonomischen Gravitationsmodelle beeinflussen der (teilweise in Transportkosten enthaltene) Energie-, Zeit und Materialaufwand für Mobilität die kulturelle, soziale und politisch-wirtschaftliche Distanz. Die Distanzreibung dient zudem der Erklärung historischer Wirtschaftsentwicklung, militärischer Dominanz eines Staates bis hin zu ökologischen Ausbreitungsmustern von Vegetationszonen und Tierarten. Das geografische Gesetz begründet auf Grundlage dieser Distanzeffekte die Beobachtung annähernd homogener Regionen und zusammenhängende Gebiete mit ähnlichen Merkmalen wie etwa die historische Bildung von Städten oder Vegetationszonen.

Geringere Bekanntheit erlangte Toblers zweites Gesetz: „*The phenomenon external to an area of interest affects what goes on inside*“. Hiermit weist er darauf hin, dass in praktischen Experimenten die Abmessungen des endlichen Beobachtungsraumes zu ei-

nem bestimmten Grad willkürlich gewählt sind und weiterhin durch Vorgänge über die Abgrenzung hinaus beeinflusst werden. Viele Phänomene weisen Autokorrelationen zwischen räumlichen Zonen auf, welche nicht im untersuchten Prozess begründet liegen. Wie erwähnt ist die fehlende Unabhängigkeit der Stichproben aufgrund der räumlichen Nähe problematisch. Eine wissenschaftliche Auseinandersetzung aus dem Jahr 1889 ist als "Galtons Problem" überliefert, in welcher Sir Francis Galton auf Fehlschlüsse hinweist, sobald Daten nicht mehr unabhängig voneinander sind. Er zeigte, dass positive räumliche Abhängigkeit den Informationsgehalt der Beobachtungen verringert, da sie durch andere Beobachtungen in der Nähe teilweise prognostiziert werden können. Dies reduziert den effektiven Stichprobenumfang. In solch einem Fall liegt den abhängigen Daten ein gemeinsamer Vorgang zugrunde, während eine räumliche Aufteilung in verschiedene Zonen voneinander unabhängige Daten suggeriert (Bivand et al., 2013, S. 264).

Im folgenden Abschnitt wird auf weitere Probleme durch Zerlegung und Skalierung eingegangen.

## Räumliche Aggregationseffekte

Räumliche Beobachtungsmerkmale können durch Skalierung ihre Gesetzmäßigkeiten ändern. So lässt sich die Länge einer Küstenlinie nicht eindeutig bestimmen, sondern wird aufgrund ihrer fraktalen Struktur mit zunehmendem Detailgrad in zunehmender Länge gemessen, ohne gegen einen endlichen Wert zu konvergieren. Daher wird nach Möglichkeit auf skaleninvariante Messgrößen zurückgegriffen.

Darüber hinaus treten Aggregationseffekte bei Änderungen der räumlichen Auflösung sowie Vergleichen zwischen verschiedenen Maßstäben auf. Dieses Problem ist als *Problem des Trägerwechsels* oder *Problem veränderlicher Bezugsregion* (engl. change-of-support problem) (**COSP**) bekannt und wurde durch Cressie (1996) formalisiert. Der räumliche Träger bezeichnet hierbei Form, Größe und Orientierung der räumlichen Merkmale, wie in Abschnitt 2.2 beschrieben. Eine Änderung der Merkmale des Trägers einer Variablen, etwa durch Aggregation, erzeugt eine neuartige Variable mit abweichenden statistischen und räumlichen Eigenschaften. Die Beziehung der räumlichen Variation beider Variablen

ist der Kern des COSP. (Waller and Gotway, 2004, S. 107) Das Problem manifestiert sich durch Instabilität statistischer Resultate bei Änderung der Aggregationsweise. Zum Beispiel ließe sich Bevölkerungseinkommen als räumliche Variable über dem Träger der punktgenauen Wohnadressen erfassen und ihre Variation betrachten. Eine Aggregation der Daten in Stadtviertel als weiteren möglichen Träger resultiert in einer anderen Variation der Einkommensvariablen.

Das Problem tritt für unterschiedliche Anwendungen in vielfältigen Varianten auf. Auf dem Gebiet der *Geostatistik* sind Bohrkernproben im Gestein nur punktweise praktisch möglich und einige  $\text{cm}^3$  groß, wodurch von einer Stichprobe an Bohrungen auf die Grundverteilung von Mineralien und Erzadern mit vielen  $\text{m}^3$  Volumen inter- und extrapoliert werden muss. In der Interpolation sind begrenzte Rohdaten in einer feinen Auflösung lückenhaft vorhanden und Inferenz wird in einer größeren Auflösung für die Umgebung der Bohrungen, über weiteren Lokationen oder Zonen benötigt.

Als Spezialfall des COSP für *Gitterdaten* addressiert das *Problem der veränderbaren Gebietseinheit* (engl. modifiable areal unit problem) (**MAUP**) die Aggregation von punktbasierten, kontinuierlichen Messgrößen über diskrete Raumzonen als zusätzliche Fehlerquelle. Abbildung 3.2 stellt Aggregationseffekte einer einfachen Zählvariablen schematisch dar.

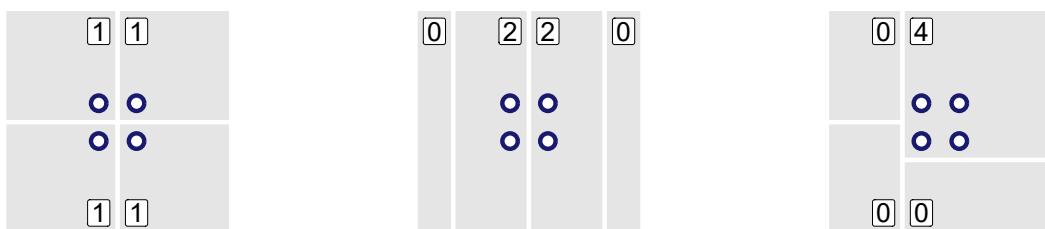


Abbildung 3.2: Einfacher Zoneneffekt für 3 Konfigurationen von 4 Zonen

Dies betrifft beispielsweise die Angabe von Bevölkerungsdaten, welche zwar theoretisch punktweise an adressgenauen Lokationen oder GPS-Raumpunkten erfasst werden können, aber aus Datenschutzerwägungen oder aufgrund der Verwaltungsmethodik ohne genaue Adresse erfasst werden und stattdessen auf Gemeindeebene zusammengefasst sind. Viele geografische Rohdaten sind nur als derartige Zählraten per Raumeinheit bzw. Zone

verfügbar. Diese Zonen sind typischerweise willkürlich, zeitlich veränderlich und haben oft keine intrinsische geografische Bedeutung (Fischer and Getis, 2010, S. 213). (Auch die Dichte der Population je Fläche ist theoretisch messbar in  $\epsilon$ -Umgebung von punktgenauen, stetigen Raumpunkten. Durch die Verdichtung auf Zonen, erlaubt sie aber zumindest den Vergleich verschieden großer Zonen.)

Dem Aggregationsproblem liegt im Detail das Zusammenspiel zweier zusammenhängender Effekte zugrunde. Zum einen kann eine feste Stichprobe über unterschiedlich feine Partitionen der Gebietsflächen verdichtet werden. Daraus resultiert ein Vergleich von Daten unter verschiedenen räumlichen Auflösungen (engl. spatial resolution) bzw. Maßstäben (engl. spatial scale) und wird als *Skalierungseffekt* (engl. scale or aggregation effect) bezeichnet. Dies beinhaltet auch das Problem unterschiedlicher Granularität innerhalb einer Auflösungsstufe. So stehen etwa einige großflächige (dünnbesiedelte) Gemeinden im direkten Vergleich mit sehr kleinen (dichtbesiedelten) Gemeinden. Werden hingegen bei fester Auflösung mit konstanter Anzahl der Raumeinheiten einzelne Grenzverläufe verschoben und Gebietseinheiten so zu einer neuen Konfiguration verformt, so werden resultierende Inkonsistenzen räumlicher Variation und Ergebnisschwankungen der Inferenz als *Zoneneffekt* (engl. grouping or zoning effect) bezeichnet. Theoretisch wäre z.B. eine Situation denkbar, in der ein administrativer Distrikt Teilgebiete an einen anderen übergibt oder für eine feste Anzahl Verwaltungszonen neue Grenzverläufe definiert werden. Abbildung 3.3 gibt schematisch den Zoneneffekt einer räumlichen Variable bei Mittelwertbildung wieder.

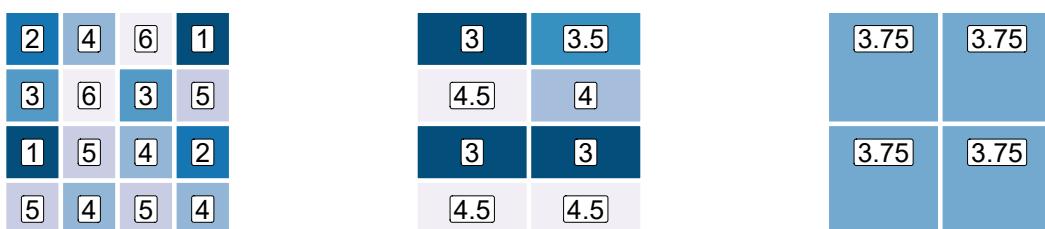


Abbildung 3.3: Zoneneffekt für 3 (bzw. 2) Konfigurationen von 4 (bzw. 8) Zonen

Ein weiteres Schema zeigt Abbildung 3.4.

Ein Datensatz kann somit durch unterschiedliche Aggregationskonfigurationen verschiedene Ergebnisse erzeugen. Zugleich muss innerhalb einer Analyse oft zwischen verschiedenen

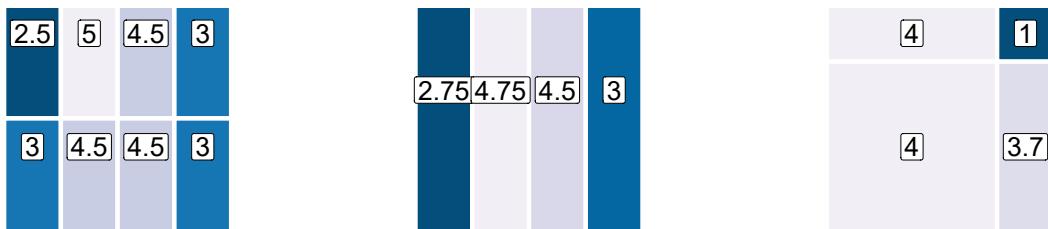


Abbildung 3.4: Zoneneffekt für 5 Konfigurationen von 4 Zonen

Aggregationsstufen skaliert werden. Dieser Effekt kann Fehler und Missinterpretationen verursachen, wird aber auch zur bewussten Manipulation eingesetzt. In der politischen Festlegung von Wahldistrikten sind insbesondere in den USA das *gerrymandering* und *political redistricting* als Wahlbeinflussung hochumstritten. Aber auch in unserem vorliegenden Forschungsprojekt sind die Grenzverläufe der Gemeinden kritisch zu betrachten. Die Analyse wird verzerrt, sobald der zugrundeliegende, datengenerierende Prozess nicht ausreichend exakt durch die Partition gedeckt ist und übereinstimmt. Eine Miskonfiguration des Modells erzeugt bereits eine räumliche Scheinkorrelation. Andererseits können auch räumliche Informationen in den Residuen verschwinden. Verwaltungsgrenzen sind oft willkürlich gesetzt und ändern sich im Zeitverlauf. Auch wenn diese Grenzen ex-ante vorgegeben sind, so weist dieser Bias auf die direkte Abhängigkeit der Ergebnisse vom Aggregationsvorgang hin. Die exakte Untersuchung von räumlichen Mustern und Kenngrößen wird durch derartige Aggregation behindert. Im Gegensatz zu systematischer Informationsverdichtung nach statistischer Methodik welche für eine sinnvolle Analyse notwendig ist, handelt es sich um inkonsistente Rohdatensätze bereits zu Analysebeginn.

Das Problem instabiler Resultate setzt sich in der Korrelationsmessung und Inferenz fort und betrifft darüber hinaus Regression, hierarchische Methoden und weitere Anwendungen (Waller and Gotway, 2004, S.106). Werden statistische Kennzahlen über eine bestimmte Konfiguration von Distrikten gebildet, so stimmen aus dieser Zusammenfassung gezogene Inferenzen nicht notwendig überein mit den Schlüssen, die eine andere Konfiguration der Distrikte liefert. Besonders bei der Ableitung von individuellen Kennzahlen aus aggregierten Gebietsdaten für einen einzelnen Datenträger muss mit Vorsicht agiert werden, selbst wenn die Kenngrößen gebietsübergreifend stabil sind und die Populationen

ähnlich groß eingeteilt sind (Waller and Gotway, 2004, S.104).

Der Begriff wurde durch Openshaw et al. (1979) geprägt, indem diese systematisch demonstrierten, wie sehr Korrelationskoeffizienten für einen festen Datensatz variieren können, indem nur die Raumpartition manipuliert wird. Über das zugrundeliegende Problem instabiler Inferenz wird jedoch bereits seit den 1930er Jahren publiziert (Gehlke and Biehl, 1934). Monmonier (2013) illustriert auf S. 218 derartige Effekte am berühmten Beispiel der Londoner Cholera-Karte des Mediziners John Snow aus dem Jahre 1856, deren historische Bedeutung in Kapitel 3.2 noch deutlich wird.

Das MAUP ist die geografische Manifestation des *Ökologischen Bias bzw. Kollektiven Fehlschlusses* (engl. ecological inference problem or fallacy), 1950 vom Statistiker William S. Robinson formuliert. Dieses Konzept bezeichnet den fehlerhaften Vergleich von Individuen verschiedener Gruppen auf Basis von Kenngrößen auf Gruppenlevel statt der gesamten Verteilungen der Messgrößen.

Eine generalisierte Lösungsmethode für dieses Problem konnte bisher nicht gefunden werden. Es existieren ausschließlich spezialisierte Ansätze. Für eine umfangreiche Übersicht siehe (Waller and Gotway, 2004, S.104) und (Haining, 2003, S. 150). Umfangreiche und tiefergehende Details geben Fischer and Nijkamp (2014) in Kapitel 59 (S.1157-1169) sowie Gelfand et al. (2010) im Kapitel 29 (S.517-536).

## 3.2 Allgemeines Modell räumlicher Stochastik

Die Grundkomponenten der mehrdimensionalen Lageanalyse sind räumlich verteilte Positionen  $\{s_1, s_2, \dots, s_n\}$  mit an diesen Positionen beobachteten Werten  $\{Y(s_1), Y(s_2), \dots, Y(s_n)\}$  als Realisierungen des Attributes  $Y$ . Stetige oder diskrete Attributwerte sind an regulär oder irregulär angeordneten Positionen in einem stetigen oder diskreten Raum möglich. Das Standardwerk von Cressie (1993) liefert ab S.8 eine flexible theoretische Struktur.

Sei  $s \in \mathbb{R}^d$  eine generische, d.h. anwendungsneutrale Position im  $d$ -dimensionalen euklidischen Raum und  $Y(s)$  eine Zufallsgröße an Position  $s$ . Unter Variation von  $s$  über

Indexmenge bzw. Definitionsbereich (engl. domain)  $D \subseteq \mathbb{R}^d$  wird der stochastische Prozess

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

als multivariates Zufallsfeld bzw. Zufallsprozess beschrieben.

Liegt der Fokus auf einer festen, endlichen Menge räumlicher Positionen  $\{s_1, s_2, \dots, s_n\} \subseteq \mathbb{R}^d$ , dann ist  $(Y(s_1), Y(s_2), \dots, Y(s_n))^\top$  ein Zufallsvektor, dessen multivariate Verteilung die räumlichen Abhängigkeiten der Beobachtungsvariable  $Y$  widerspiegelt. Eine beobachtete Realisierung dieses Zufallsfeldes wird mit  $\{y(s) : s \in D\}$  ausgedrückt.

Eine nützliche Unterteilung von praktisch konstruierbaren Prozessen in 3 Hauptklassen wird durch Cressie (1993) gegeben und in Tabelle 3.1 zusammengefasst. Diese Einteilung orientiert sich an der historischen Entwicklung in Abhängigkeit verschiedener Untersuchungsobjekte und unterscheidet bezüglich der Modellierung und ihrer Dimensionen.

	POSITION - UNABHÄNGIG, EXOGEN	
	DISKRETE VARIATION	STETIGE VARIATION
KEIN ATTRIBUT	-	<b>Punktprozesse</b> Krankheitsausbreitungen (ab ca. 1960)
ATTRIBUT - ABHÄNGIG, ENDOGEN	<b>Gitterdaten</b> Ertrag der Landwirtschaft (ab ca. 1910)	<b>Geostatistik</b> Verteilung Bodenschätzungen (ab ca. 1970)

Tabelle 3.1: Klassifikationsübersicht der räumlichen Stochastik mit Anwendungsbeispielen. Quelle: eigene Darstellung in Anlehnung an Cressie (1993)

In dieser Arbeit werden wir ausschließlich Gitterdaten modellieren. Zur Ergänzung werden im Folgenden wichtige Grundlagen der Punktprozesse und Geostatistik angerissen, soweit es das Verständnis der Gitterdaten unterstützt. Für weitere Details wird die passende Fachliteratur aufgegliedert.

## Punktprozesse

Für räumliche *Punktprozesse* (engl. Spatial Point Patterns) ist die Indexmenge der Prozesspositionen  $D = \{s_1, s_2, \dots, s_n\}$  stochastisch. Die relevante Variable sind hier die Lo-

kationen der Ereignisse selbst. Die Positionen bestimmter Ereignisse werden erfasst und typischerweise treten Fragen nach der Regularität der realisierten Punkteverteilung im Raum auf. Diese können etwa lokal geclustert oder ohne erkennbare Muster zufällig verteilt sein. Zusätzlich können die Positionen mit Markierungsvariablen  $(y(s_1), y(s_2), \dots, y(s_n))$  als Kovariablen (engl. Covariates) verknüpft und als markierter räumlicher Punktprozess (engl. marked spatial point process) untersucht werden. Typischerweise werden Punktprozesse auf Zufälligkeit ihrer Verteilung, auf Cluster oder reguläre Strukturen wie Abstoßungen hin untersucht. In der Ökologie kann beispielsweise die Verteilung von zwei Arten auf Konkurrenzverhalten untereinander oder etwa externe Faktoren für Verbreitungsmuster hin untersucht werden. In der Epidemiologie wird oft untersucht, ob eine Krankheit an bestimmten Orten geclustert ist und sich räumlich und zeitlich ausbreitet. Als frühestes Beispiel wird die berühmte Cholera-Karte des Mediziners Dr. John Snow angeführt, welcher 1854 in London die Epidemie auf eine vergiftete Trinkwasserpumpe zurückführen konnte. Er erkannte im Wasser lebende Mikroorganismen als Cholera-Erreger und widerlegte so die verbreitete Annahme von „üblichen Dünsten“ als Übertragungsweg. Dies führte zu langfristigen Verbesserungen des Abwassersystems und begründete zugleich die Gebiete der Epidemiologie und Räumlichen Analyse. Inzwischen wurden beide Gebiete zum modernen *Disease Mapping* weiterentwickelt. Wichtige Verfahren sind die *Nächste-Nachbarn-Klassifikation* (engl. Nearest-Neighbor methods) sowie Maße räumlicher Homogenität wie *Ripley's K- und L-Funktion*.

## Geostatistik

Stetige Raumdaten sind charakteristisch für das Gebiet der *Geostatistik* (engl. geostatistics) und ihrer Interpolationsmethoden. Hierbei wird  $D \subseteq \mathbb{R}^d$  angenommen. Teilmenge  $D$  enthält ein  $d$ -dimensionales Rechteck positiven Volumens. In den meisten Anwendungsfällen genügt ein zwei- oder dreidimensionales Beobachtungsvolumen. Hierbei wird eine Position  $s$  etwa durch

$$s = (s_x, s_y)^\top \in \mathbb{R}^2 \quad \text{oder} \quad s = (s_r, s_\varphi, s_\omega)^\top \in \mathbb{R}^3$$

über stetige Polar- oder kartesische Koordinaten beschrieben. Zufallsvariable  $Y(s)$  an der Stelle  $s \in D$  nimmt im Gegensatz zu Punktprozessen an jeder möglichen Position  $s$  im Raum relevante Werte an. Daher variiert auch der räumliche Index  $s$  stetig (engl. continuous spatial variation) über Indexmenge  $D$  (Gelfand et al., 2010, S.17). Es werden Punktstichproben aus einer stetigen räumlichen Verteilung gezogen, z.B. Temperaturmessungen an verschiedenen Stellen eines Wasserbehälters. Daraus werden per Interpolation und Inferenz Rückschlüsse über unbeobachtete Stellen abgeleitet. Der spezialisierte Prozess wird somit als stetig parametrisierter Prozess (engl. Continuous Parameter Stochastic Process) bezeichnet. Ursprünglich zur Bodenerkundung und Schätzung von Erzreserven entwickelt, wird räumliche Variabilität sowohl durch Trends im globalen Maßstab über das gesamte Untersuchungsgebiet als auch durch Korrelation im lokalen Maßstab modelliert.

Toblers geografisches Autokorrelationsgesetz aus Kapitel 3.1 begründet eine der *räumlichen Interpolation* (engl. spatial interpolation) zugrundeliegende Abklingfunktionen. Neben einfachen deterministischen Interpolationsmethoden der *Inversen Distanzgewichtung* (engl. inverse distance weighting - IDW) (Shepard 1968) sind Verfahren statistischer Natur wie etwa das *Kriging* zentraler Bestandteil der Geostatistik. Dieses wurde durch Danie G. Krige ab 1951 in Johannesburg als Interpolationsmethode im Bergbau empirisch entwickelt. Die zugrundeliegende *Theorie der regionalisierten Variable* (engl. regionalized variable theory - RVT) wurde 1963 durch Georges Matheron formalisiert. Ein großer Vorteil gegenüber deterministischen Methoden ist die Erfassung räumlicher Varianz unter Einsatz von *Semivariogrammen*. Zu beachten sind hierbei *Nugget-Effekt*, *Sill* und *Range*, *Drift* und *Hole-Effekt*. Weitere Grundlagen sind räumliche Gauß-Prozesse und ihre Momente, Annahmen der *Stationarität* sowie *Isotropie* bzw. *Anisotropie*.

## Gitterdaten

Die Klasse der *Gitterdaten* (engl. lattice data) definiert die Indexmenge  $D$  als feste Menge abzählbar vieler Raumeinheiten aus dem  $\mathbb{R}^d$  mit wohldefinierten Abgrenzungen vonein-

ander. Diese Einheiten sind durch zusätzlich konstruierte Nachbarschaftsinformationen im regulären oder irregulären Gitter angeordnet. Die unabhängige Beschreibung einzelner Punkte in beliebig feiner Raumauflösung wird hier nicht angestrebt. Stattdessen erfolgt eine Aggregation innerhalb fester Grenzen. Jeder Raumpunkt wird eindeutig auf ein Aggregationsobjekt abgebildet. Der Raum ist durch die Aggregationseinheiten, etwa in Form einer Partition, hinreichend beschrieben und räumliche Informationen durch diskrete Indizes  $s \in \{1, \dots, S\}$  ausreichend codiert. Zumeist wird der gesamte Beobachtungsraum vollständig zerlegt, sodass kein Bereich offen bleibt. Die Aggregationsobjekte wiederum bestehen aus einzelnen geometrischen Objekten oder umfassen diese vollständig, wodurch etwa Löcher entstehen können. So kann bei der Aggregation über Landmassen eine Landesgrenze mehrere Inseln umfassen oder eine Gemeinde einen See umschließen. Diese komplexen Grenzbeziehungen werden im GIS durch Polygonzüge effektiv und effizient verwaltet, sodass der Nutzer sich auf die Attributsvariable konzentriert und alle räumlichen Informationen in der Nachbarschaftsmatrix abgelegt sind.

Administrative Einheiten wie durch Landesgrenzen abgegrenzte Nationalstaaten beschreiben unregelmäßige Gitter. Eine pixelierte Bilddatei hingegen wird durch das regelmäßiges Gitter eines Pixelrasters beschrieben. Die Repräsentation der Gitterstellen durch Regionen wird in der Fachliteratur als *Gebietsdaten* (engl. areal data) bezeichnet. Eine Repräsentation durch Punkte ist ebenfalls denkbar, etwa durch Zentroide der Regionen. Diese Transformation führt wieder teilweise in die Theorie der Punktmuster.

$Y(s)$  ist hier ebenfalls Zufallsvariable an der Stelle  $s \in D$  mit Realisierungen  $y(s_1), y(s_2), \dots, y(s_n)$ . Eine übliche Anwendung ist die Zusammenfassung von Wahlergebnissen in lokalen, administrativen Einheiten. Besonders durch Fernerkundung (engl. remote sensing) und Satellitenbildanalyse, in denen die Erdoberfläche in rechteckige Bildelemente von beispielsweise 56m x 56m zerlegt wird, lassen sich landwirtschaftliche Vegetationsdaten, Wetterdaten usw. erheben. Unter Projektionsannahmen der Erdkrümmung werden diese Daten als reguläres Gitter im  $\mathbb{R}^2$  mit Rechteckzentroiden als Gitterpunkten modelliert. Das Gittermodell kann auf die Konstruktion eines ungerichteten Graphen generalisiert werden. Die

Positionen werden durch Knoten repräsentiert, von denen einige über Kanten verknüpft sind, deren Verbindungskanten die Nachbarschaftsinformationen repräsentieren.

Die obige Einteilung in drei Fachgebiete wurde bisher in zahlreichen Werken der Fachliteratur übernommen. Tabelle 3.2 liefert eine Übersicht. Eine Ausnahme bilden Fischer and Getis (2010), welche sich auf übergreifende Konzepte wie Autokorrelation, Clustering und Filtering konzentrieren und auf Modelldetails verschiedener Anwendungsbereiche erst innerhalb dieser Konzepte eingehen. Dieses Werk bietet sich somit besonders zur Erfassung interdisziplinärer Zusammenhänge und als Nachschlagewerk an.

FACHLITERATUR	PUNKTPROZESSE	GEOSTATISTIK	GITTERDATEN
<b>Cressie (1993)</b>	8	2 - 3	6 - 7
<b>Waller and Gotway (2004)</b>	5 - 6	8	7, 9
<b>Schabenberger 2005</b>			
<b>Bivand et al. (2013)</b>	7	8	9 - 10
<b>Anselin and Rey (2010)</b>	6 - 9		5, 10 - 13
<b>Gelfand et al. (2010)</b>	16 - 22	2 - 11	12 - 15
<b>Fischer and Getis (2010)</b>			
<i>Exploratory Analysis</i>	B.2.4	B.2.4	B.2.5
<i>Autocorrelation</i>	B.3 (S.266-267)	B.3 (S.265-266)	B.3 (S.255-265)
<i>Global Clustering</i>	B.4.2 (286-289)		B.4.2 (280-285)
<i>Local Clustering</i>	B.4.3 (293-298)	B.4.3 (293-298)	B.4.3 (289-298)
<i>Filtering</i>		B.5	B.5
<i>Kriging</i>		B.6	

Tabelle 3.2: Buchkapitel gemäß Klassifikation durch Cressie (1993)

In dieser Arbeit liegt der Fokus auf Gitterdaten. Jedoch überlappen sich einige Konzepte. Die drei Methodenklassen sind untereinander zum Teil verknüpft über die Konstruktion der Nachbarschaft bestimmter Positionen untereinander. Außerdem ist Autokorrelation ist für alle drei Gebiete essentiell und wird im darauf folgenden Kapitel eingeführt.

### 3.3 Räumliche Nachbarschaft

Für die räumliche Analyse insbesondere von Zonen ist ein zusätzlicher Problempunkt, diese Zonen in adäquate Beziehungen zueinander zu setzen, da keine direkte Entfernungsmetrik wie bei den Punktdaten zu Verfügung steht. Ein Nachbarschaftssystem muss manuell

erstellt und mit Vorsicht justiert werden, da weitere Analysen sensibel auf die Konfigurationen reagieren. Ein Ansatz ist die Kontiguität bzw. Angrenzung (engl. contiguity), um aneinander angrenzende Gebietspolygone mit gemeinsamer Grenze als Nachbarn zu deklarieren. Die Elemente der Menge aller Nachbarn von  $s_i$

$$N_i = N(s_i) = \{v \in D : v \text{ teilt Grenze mit } s_i\}, \quad i = 1, \dots, n$$

stehen in Nachbarschaftsrelation zueinander (Notation:  $v \sim s_i$ ) und erfüllen  $s_i \notin N(s_i)$  sowie  $s_i \in N(v) \Leftrightarrow v \in N(s_i)$ . Die Indexmenge  $D_N = \{(i; N_i) : i = 1, \dots, n\}$  bildet ein räumliches Gitter, repräsentierbar durch einen Graphen. Für administrative Verwaltungszonen wie Gemeinden sind keine exakten Positionsdaten definiert. Die räumlichen Abhängigkeiten zwischen Gemeinden werden zwischen Zonen modelliert (Cressie, 1993, S. 385).

Sei die Indexmenge eines Gitters über  $n$  Regionen durch  $\mathcal{D} = \{(i; x_i, y_i) : i = 1, \dots, n\}$  gegeben. Hierbei bilden Koordinaten  $x_i$  die Längengrade und  $y_i$  die Breitengrade des Zentroides der Region  $i$ . Nun werden über die Entferungen (bzw. Entfernungsbänder) der Zentroide räumliche Beziehungen definiert. So können etwa alle Regionen als Nachbarn der Region  $i$  definiert werden, deren Zentroide weniger als 100 km vom Zentroid der Region  $i$  entfernt liegen. Eine weitere Abwandlung von diesem Konzept ist die Konstruktion der  $k$  nächsten Nachbarn (engl. k-nearest neighbour). Zwar resultiert aus diesem Kriterium meist einen asymmetrischer Nachbarschaftsgraph, es wird aber gesichert, dass jedem Gebiet exakt  $k$  Nachbarn zugewiesen sind. Abbildung 3.5 illustriert über die Konstruktion als Graph die unterschiedlichen Nachbarschaftssysteme mittels der Zentroide der Bundesländer.

Nach der Konstruktion der Nachbarschaftsrelation können einzelne Nachbarn geeignet gewichtet werden. Bivand et al. (2013), Kapitel 9.2.2 raten jedoch bei geringer Kenntnis über den zugrundeliegenden räumlichen Prozess davon ab, von einer binären Repräsentation abzuweichen.

Ein solches System der Nachbarschaftsrelation ist erweiterbar auf zweite und weitere Stu-

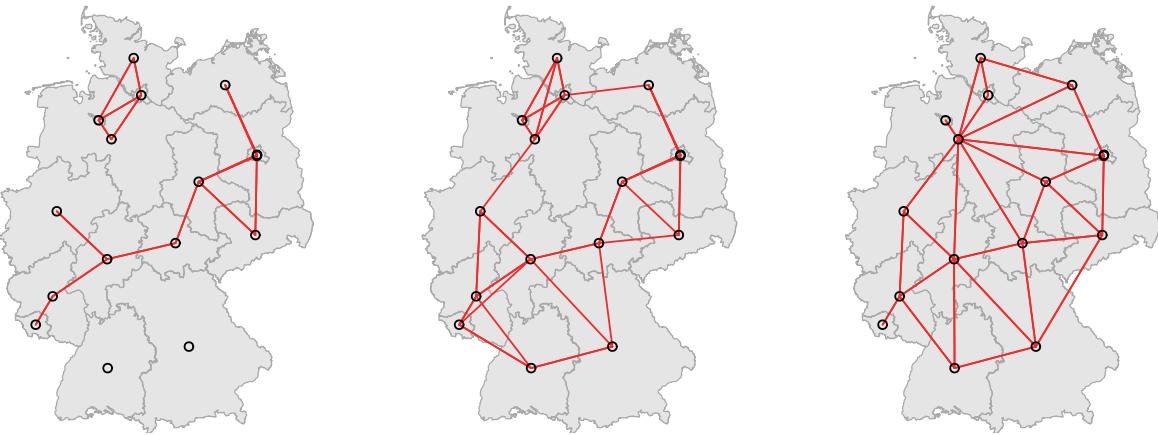


Abbildung 3.5: Nachbarschaftsgraph auf irregulärem Gitter am Bsp der Bundesländer. V.l.n.r.: Distanz  $< 260\text{km}$ , Drei nächste Nachbarn, Gemeinsame Grenze

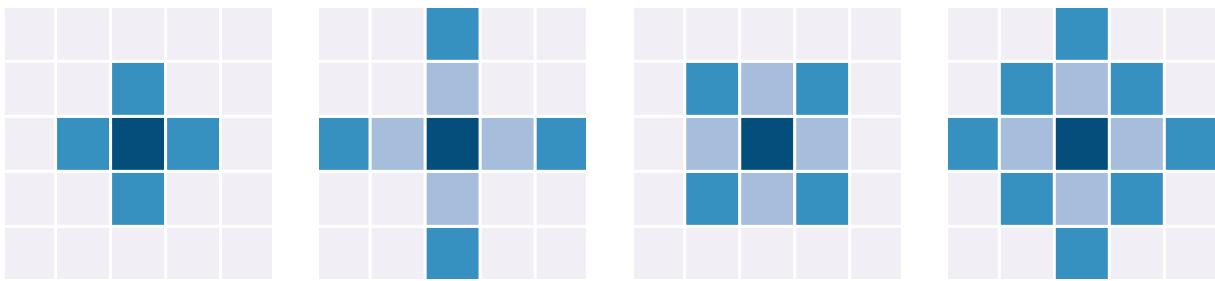


Abbildung 3.6: Nachbarschaftssystem auf regulärem Gitter. V.l.n.r.: Erste, zweite, diagonale, zweite und diagonale Nachbarn

fen der Nachbarschaft. Über Distanzintervalle  $(0, d_1]$ ,  $[d_1, d_2]$ , ... usw. definieren sich erste Nachbarn innerhalb der Distanz  $d_1$  von  $i$ . Zweite Nachbarn sind innerhalb des Entfernungsbandes von  $d_1$  bis  $d_2$  lokalisiert.

In Abbildung 3.6 sind verschiedene Nachbarschaftssysteme für ein reguläres Raster-Gitter schematisch bezüglich gemeinsamer Grenzen dargestellt. In Abbildung 3.7 steht die Anwendung auf ein irreguläres Gitter der Gebiete deutscher Bundesländer mit  $s_i = \text{Thüringen}$ .

Nachbarschaftsmatrix  $W$  speichert im folgenden alle Nachbarschaftsinformationen. Die Einträge  $w_{ij}$  entsprechen den Gewichten. Dabei werden Diagonaleinträge  $w_{ii}$  auf Null gesetzt. In einer Standardisierung werden die Einträge der Zeile  $i$  durch Zeilensumme  $\sum_{j=1}^n w_{ij}$  geteilt. Analog zur Erweiterung des Nachbarschaftssystems kann die Nachbarschaftsmatrix  $W_1$ ,  $W_2$  usw. mit Nachbarschaftsrelationen ersten bzw. zweiten Grades usw. gebildet werden.

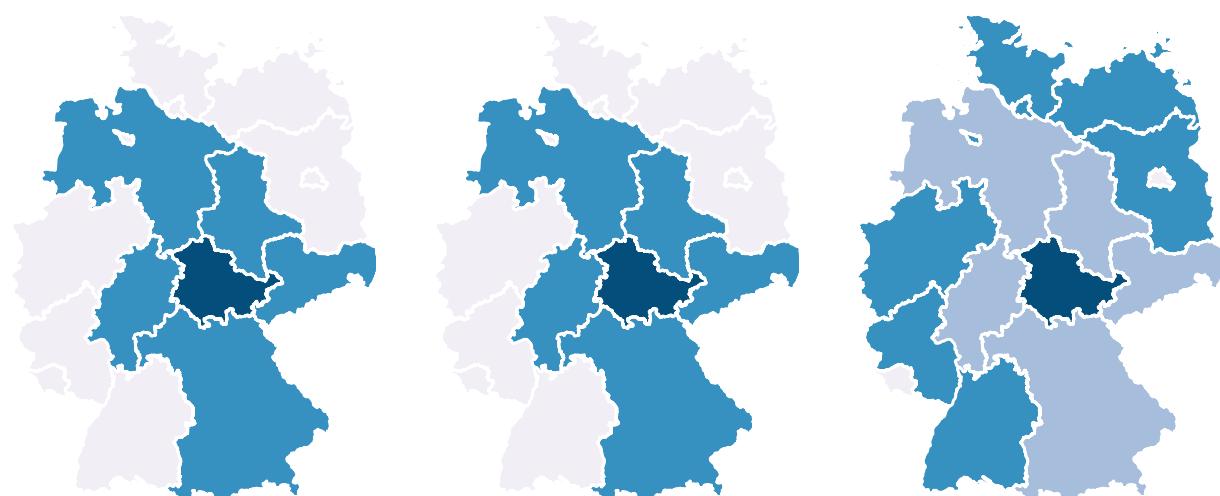


Abbildung 3.7: Nachbarschaftssystem auf irregulärem Gitter am Bsp der Bundesländer.  
V.l.n.r.: Erste, zweite, diagonale

# Kapitel 4

## Räumliche Autokorrelation

Ein wichtiges Konzept der räumlichen Statistik ist die *räumliche Abhängigkeit* und daraus resultierende *räumliche Autokorrelation* (engl. spatial dependence and autocorrelation). Diese kann zum einen als Störfaktor gesehen werden, da statistische Tests verkompliziert werden. Bei räumlicher Autokorrelation liegt eine Abhängigkeit der Störgrößen des ökonometrischen Modells von regionalen Untersuchungseinheiten vor. Dies führt bei Kleinst-Quadrat-Schätzungen zu Verzerrungen der Regressionskoeffizienten oder Ungültigkeit der Signifikanztests. Zum anderen liefert Autokorellation zusätzliche Information und Möglichkeiten zur räumlichen Interpolation. Während Trendprognosen eine zeitliche Autokorrelation benötigen, ermöglicht räumliche Autokorrelation, bzw. räumliche Persistenz eine distanzabhängige Interpolation. In diesem Kapitel werden Eigenschaften und Berechnung derartiger Statistiken untersucht.

### 4.1 Theorie und Indizes räumlicher Autokorrelation

Die Korrelation (engl. correlation) beschreibt funktionale Beziehungen zwischen zwei Variablen und liefert ein Maß für den Stärkegrad der Abhängigkeit zwischen den Variablenpaaren. Der *Bravais-Pearson'sche Korrelationskoeffizient*, auch Produkt-Moment-Korrelationskoeffizient oder Pearson's r genannt, erfasst den linearen Zusammenhang zweier annähernd normalverteilter Variablen und ist zur Messung der Abhängigkeit metrischer und dichotomer Daten geeignet. Im Gegensatz dazu sind *Spearman's rho* und *Kendall's*

*tau* als nichtparametrische Rangkorrelationskoeffizienten auf ordinalskalierten Daten ohne Verteilungsannahmen anwendbar.

Autokorrelation misst die Korrelation eines Merkmals mit sich selbst und wird zwischen zeitlich oder räumlich benachbarten Beobachtungen erfasst. Bei Messungen eines Beobachtungsobjektes im Zeitverlauf ist es allgemein betrachtet wahrscheinlich, dass zeitlich nahe beieinander liegende Beobachtungen ähnlichere Messwerte liefern als solche mit größerem zeitlichen Abstand. Messwerte ändern sich stetig im Zeitverlauf, wenn auch nicht streng monoton.

Zeitliche Autokorrelation misst in der Zeitreihenanalyse den Grad der Assoziation zwischen Signalfolgen oder Folgen von Zufallsvariablen in Einheiten bestimmter Zeitabstände, welche meist äquidistant auftreten und durch Lags gesteuert werden. Eine Folge wird zeitlich verschoben und zu sich selbst in Beziehung gesetzt. Auf diese Art wird eine signifikante Beziehung zwischen zeitlich versetzten Folengliedern gesucht. Ergebnisse werden in der *Autokorrelationsfunktion (ACF)* zusammengefasst. Die Anwendung auf zwei verschiedene Folgen wird hingegen als Kreuzkorrelation bezeichnet.

Während temporale Autokorrelation sich auf eindimensionale Nachbarschaft mit zwei möglichen Verlaufsrichtungen bezieht, wird räumliche Autokorrelation zwischen räumlichen Objekten mit mindestens 2 Dimensionen gemessen, deren räumliche Distanz nicht immer eindeutig bestimmt ist. Diese Objekte sind etwa Punkte, Rasterzellen oder Gebiete in komplexen Polygonzügen wie etwa Gemeindegrenzen (Fischer and Getis, 2010, S.260). Analog zu den Einführungsbeispielen existieren Koeffizienten und Indizes zur Erfassung einer Autokorrelation von räumlich verteilten Messungen. In der räumlichen Analyse liegt ebenfalls die Annahme zugrunde, dass räumlich benachbarte Beobachtungen einer Variablen ähnlicher sind als weiter entfernte. Dieses Konzept wird durch *Toblers erstes Gesetz der Geographie* aus Abschnitt 3.1 begründet. Bekannt ist der Zusammenhang als positive räumliche Autokorrelation, also als eine Korrelation der Variablen mit sich selbst. Es werden somit zwei Informationen verknüpft: die Ähnlichkeit der beobachteten Messwerte mit der Ähnlichkeit ihrer Positionen.

Aus positiver Autokorrelation folgt im Umkehrschluss nicht notwendig eine negative Autokorrelation weiter entfernter Lokationen, welche trotzdem oft in realen Daten beobachtet werden kann. Es gibt darin eine begrenzte Entfernungsspanne um Beobachtungspunkte, innerhalb welcher positive Autokorrelation mit Nachbarn besteht. Für größere Radien verschwindet diese oder schlägt in negative Autokorrelation um.

Daten aus Punktprozessen und Geostatistik lassen sich direkt über die euklidische Norm räumlich in Beziehung setzen. Für Gitterdaten (engl. lattice data) liegen jedoch diskrete Raumbeziehungen vor, und die räumliche Anordnung der Daten wird anders kodiert. Die folgenden Konzepte sind insbesondere für derartige Gitterdaten relevant, wie im Abschnitt 4.3 definiert. Im Gitter können sowohl Zonen als auch Punkte angeordnet werden. Die räumliche Information liegt hier diskret z. B. in Form einer Regionenvariable  $s$  vor. Dieses Raumgitter muss um Nachbarschaftsinformationen in Form der Nachbarschaftsmatrix aus Abschnitt 3.3 ergänzt werden, um für das weitere Vorgehen die Nachbarschaftsrelationen paarweise eindeutig zu definieren.

## 4.2 Indizes und Tests der Autokorrelationsmessung

Räumliche Autokorrelation vergleicht parweise die Ähnlichkeit von Merkmalswerten  $y_i = Y(s_i)$  mit  $y_j = Y(s_j)$  in Bezug zur Nähe ihrer Lokationen  $s_i$  und  $s_j$  und leitet die tendenzielle, funktionale Beziehung ab. Aufgrund vielfältiger Anwendungsfälle wurde eine breite Palette an Indizes entwickelt, deren bekannteste Vertreter *Moran's I* und *Geary's C* im folgenden untersucht werden. Alle Autokorrelationsmaße teilen sich als Spezialisierungen die folgende Grundform eines Kreuzproduktes. Die Repräsentation einer räumlichen Statistik als Kreuzprodukt

$$\Gamma_{ij} = \sum_{i=1}^n \sum_{j=1}^n W_{ij} Y_{ij} \quad (4.1)$$

von  $n$  georeferenzierten Beobachtungen erlaubt die flexible Entwicklung zahlreicher Sonderformen. Hierbei repräsentieren die Gewichte  $w_{ij}$  der Nachbarschaftsmatrix  $W$  die räumliche Beziehung einer jeden Lokation  $i$  zu allen anderen Standorten  $j$ . (Fischer and Getis, 2010, S. 259)

In der Attributsmatrix  $Y$  werden die Realisierungen  $y_j = Y(s_j)$  an verschiedenen Positionen  $s_j$  um  $s_i$  herum in Beziehung gesetzt. Die Werte interagieren wahlweise paarweise additiv ( $y_i + y_j$ ), multiplikativ ( $y_i \cdot y_j$ ), differenziert ( $y_i - y_j$ ) oder dividiert ( $y_i/y_j$ ). (Fischer and Getis, 2010, S. 262) Ähnlichkeit der Attributswerte  $y_i$  und  $y_j$  kann durch folgende Maße ausgedrückt werden:

$$U_{ij} = (y_i - \mu)(y_j - \mu) \quad (4.2)$$

$$U_{ij} = (y_i - \bar{y})(y_j - \bar{y}) \quad (4.3)$$

$$U_{ij} = |y_i - y_j| \quad (4.4)$$

$$U_{ij} = (y_i - y_j)^2 \quad (4.5)$$

Praktisch relevante Beispiele sind aufgrund guter Berechenbarkeit und Interpretierbarkeit das Kreuzprodukt der Kovarianzform  $(y_i - \bar{y})(y_j - \bar{y})$  für Morans Index sowie quadrierte Differenzen  $(y_i - y_j)^2$  für Gearys c. Voraussetzung ist ein räumlich konstanter Erwartungswert  $\mathbb{E}[Y(s)] = \mu$ . Ist  $\bar{y}$  konsistent für  $\mu$ , dann ist auch  $\mathbb{E}[(y_i - \bar{y})(y_j - \bar{y})]$  konsistenter Schätzer von  $\text{Cov}(y_i, y_j)$  (Schabenberger and Gotway, 2005, S. 21).

Somit gibt  $Y$  eine spezifische Beziehung zwischen Beobachtungswerten an Punkt  $i$  zu solchen an allen anderen Orten  $j$  wieder. Weisen  $W$  und  $Y$  ähnliche Strukturen auf (d.h. sowohl große als auch kleine Werte liegen in den gleichen  $(i, j)$ -Zellen), so liegt ein hoher Grad positiver räumlicher Autokorrelation vor. Bei gegenläufig übereinstimmenden Werten in den gleichen  $(i, j)$ -Zellen liefert  $\Gamma$  negative Autokorrelation. Für eine sinnvolle Bewertung muss  $Y$  den räumlichen Bezug wiedergeben und  $W$  die räumliche Struktur sinnvoll repräsentieren. Weist auch nur eine der beiden Matrizen eine zufällige Verteilung auf, so liefert  $\Gamma$  ein Maß von Null. Nachbarschaftsmatrix  $W$  muss daher vorsichtig spezifiziert werden und die real vorliegenden, räumlichen Strukturen widerspiegeln. Abschnitt 3.3 lieferte mögliche Formen und Konstruktionsdetails für  $W$ . (Fischer and Getis, 2010, S.259)

Zahlreiche Maße und Tests sind auf dieser Basis etabliert und unterscheiden sich voneinander in der Struktur der Matrizen  $W$  und  $Y$  sowie ihren Skalierungsfaktoren. Es können globale sowie lokale Maße und Tests abgeleitet werden. Globale Statistiken aggregieren die Autokorrelation über die Gesamtheit der Daten. Alle Elemente der  $W$  und  $Y$  Matrix mit ihren räumlichen Assoziationen werden gemeinsam ausgewertet und fließen in einen einzelnen Ausgabewert zur Bewertung der Korrelationssituation im gesamten Beobachtungsgebiet ein. Lokale Maße (engl. Local indicators of spatial association) (**LISA**) bewerten Autokorrelation hingegen mit Fokus auf einzelne räumliche Teilgebiete. Somit beeinflusst nur eine Reihe der  $W$  und  $Y$  Matrix die Berechnung direkt. (Fischer and Getis, 2010, S. 262) Sie geben für die Nachbarschaftswerte um eine feste Position den Grad der Abweichung von einer zufälligen Verteilung an. Sie erlauben die Zerlegung globaler Indikatoren (wie Moran's I) in die Anteile, welche die Messwerte einzelner Lokationen beitragen. (Anselin, 1995, S.94)

In der Geostatistik drückt das Semi-Variogram den Grad der räumlichen Autokorrelation in einem Datensatz aus. Analog hierzu wird eine Statistik als distanzabhängige Funktion betrachtet, indem ihre I-Werte für verschiedene Distanzbänder auf Basis der Nachbarschaftsmatrizen  $W(1), W(2), \dots, W(q)$  für Nachbar ersten bis q-ten Grades abgetragen werden.

#### 4.2.1 Globale Autokorrelationsmaße

Ziel globaler Autokorrelationsindizes ist die Zusammenfassung des Ausmaßes, in welchem ähnliche Beobachtungen in räumlicher Nähe zueinander vorkommen. Als Summation über das gesamte Beobachtungsgebiet werden keine individuellen Cluster ermittelt, sondern auf eine allgemeine, gebietsübergreifende Clusterung getestet.

##### **Gamma**

Als Basisfunktion aller weiteren Maße und Tests wurde diese Statistik durch Ausdruck 4.1 eingeführt. Durch die Randomisierung der Werte der Y Matrix in einer Reihe von Simu-

lationsläufen wird die Hülle einer Referenzverteilung erzeugt. Liegt der tatsächliche Wert im kritischen Bereich, so liegt signifikante Autokorrelation vor. Details liefern (Fischer and Getis, 2010, S. 262) und (Schabenberger and Gotway, 2005, S. 15).

### Join-count

Darüber hinaus existiert der *Join-Count* Index für nominal klassifizierte Daten ... Ein Anwendungsbeispiel sind Klassifizierungen von Raumeinheiten nach Landnutzungstypen. Es werden hier paarweise die Klassen der räumlich benachbarten Einheiten verglichen und auf statistisch signifikante Häufungen geprüft. Genaue Ausführungen finden sich in (Schabenberger and Gotway, 2005, S. 19) sowie (Fischer and Getis, 2010, S.263).

### Globaler Moran Index

Der globale Moran Index (engl. global Moran's I) nach Patrick A. P. Moran (1950) misst als Statistik den Grad sowohl positiver als auch negativer Autokorrelation. Er ist als Produktmoment wie der Pearson Korrelationskoeffizient strukturiert. Jedoch wird, unter zusätzlichen Integration der räumlichen Information aus Matrix  $W$ , die Autokorrelation einer Variablen mit sich selbst gemessen statt einer klassischen Korrelation zweier Variablen miteinander.

**Def. 4.2.1.1 (Moranscher Index):** Analog zur Pearson Statistik erfolgt eine Skalierung. Der Vorfaktor  $n / \left( \sum_{i=1}^n \sum_{j=1}^n w_{ij} \sum_{i=1}^n (y_i - \bar{y})^2 \right)$  ergibt die folgende Statistik zur Beziehung eines Datenpunktes (e.g. Einkommen) mit seinen Umgebungswerten: (Moran, 1950, S.22)

$$I = \frac{n}{\sum_{i=1}^n \sum_{j=1}^n w_{ij}} \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^n (y_i - \bar{y})^2}, \quad i \neq j \quad (4.6)$$

Hierbei sind:

- $y_i$  = i-ter Beobachtungswert der räumlichen Variable  $Y$
- $w_{ij}$  = Distanzgewichte der räumlichen Beobachtungen
- $n$  = Anzahl Beobachtungen der Stichprobe

Matrix  $W$  kann in beliebiger Form aufgestellt werden und beschränkt die räumliche Interpretation nicht. Somit ist der Test sehr flexibel, aber auch auch anfällig gegen Ausreißer und Fehlspezifikationen der Nachbarschaftsbeziehungen. Matrix  $Y$  in Form einer Kovarianzmatrix misst die Abweichungen der Beobachtungen vom Mittel aller Beobachtungen. (Fischer and Getis, 2010, S. 263)

Es ist Konvention, keine Selbstassoziation zu erfassen ( $i$  ist ungleich  $j$ ). Durch Standardisierung der Gewichte können Ausgabewerte des Index auf die Spanne  $[-1,1]$  eingeschränkt werden. Die lokale Variante als Weiterentwicklung der Moran I Statistik wird in Abschnitt 4.2.2 behandelt. Zuvor wird im folgenden Teilabschnitt die Nutzung der I-Statistik zur statistischen Inferenz erläutert.

### Globaler Gearyscher Index

Bekannt als Gearysches Nachbarschaftsverhältnis oder Gearyscher Index (engl. Geary's  $c$ , Geary's Contiguity Ratio) (Geary, 1954). Geary (1954) Der Unterschied zur Moran Statistik liegt in der Berechnung der Attribute durch das Kreuzprodukt  $Y_{ij} = (y_i - y_j)^2$  auf Basis der Differenzen benachbarter Beobachtungen. Die Gewichtsmatrix  $W$  kann wie im Moran Index verwendet werden und erlaubt die gleiche Flexibilität. Geary's  $c$  ist jedoch sensitiver bezüglich lokaler Autokorrelation.

$$c = \frac{n-1}{2 \sum_{i=1}^n \sum_{j=1}^n w_{ij}} \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij} (y_i - y_j)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \quad i \neq j \quad (4.7)$$

Weitere Unterschiede zum Moran Index sind eine unverzerrte Varianzschätzung per Faktor  $n-1$  statt  $n$  und die Quadrierung negativer Differenzen im Term  $(y_i - y_j)^2$ . Die Ausgabewerte liegen im Bereich  $[0,2]$ . Ein Wert von 1 impliziert, dass keine räumliche Autokorrelation vorliegt, während der Wert 0 eine perfekt positive und Wert 2 perfekt negative Autokorellation signalisieren. (Fischer and Getis, 2010, S. 265)

### Statistische Inferenz und Tests

Analog zur allgemeinen Mantel Statistik ist Inferenz für die I und c Statistik auf Basis von Permutationstests, Monte-Carlo Tests oder auf asymptotischen Verteilungen von I

und c basierten Approximationstests möglich.

Von der Voraussetzung einer Unabhängigkeit der Stichprobenwerte muss gegenüber der klassischen Statistik abgewichen werden. Geprüft wird an dessen Stelle die Hypothese einer zufälligen bzw. unkorrelierten Verteilung der Stichprobengrößen. Wird diese Nullhypothese (selbst für große  $\alpha$ ) nicht abgelehnt, was tendenziell auf eine zufällige Verteilung hindeutete, darf jedoch nie eine Unabhängigkeit der Stichprobenwerte gefolgert werden. Die Voraussetzung identisch verteilter Stichprobengrößen mit konstanten Erwartungswerten und Varianzen über dem gesamten Stichprobenumfang bleibt bestehen, und wird durch Begriffe wie *Stationarität* und *Populationshomogenität* angedeutet. Der Moran Index reagiert sensibel auf Ausreißer in der Stichprobe. Für den Geary Index gilt dies umso mehr, da eine Quadrierung der Differenzen erfolgt.

Es sind zur Nullhypothese zwei grundlegende Verteilungsannahmen über die zufällige Verteilung der Y-Werte möglich. (Schabenberger and Gotway, 2005, S.22) sowie (Fischer and Getis, 2010, S. 264) weisen auf die Konsequenzen in der Ableitung von Erwartungswert und Varianz der Indizes hin. Die erste Annahme einer Normalverteilung (engl. gaussianity assumption) sieht die beobachteten Y-Werte  $y_i$  als Realisierung einer für jede Raumeinheit eigenen Gaußschen Zufallsvariablen mit identischen Mittelwerten  $\mu$  und Varianzen  $\sigma^2$ . In der zweiten Annahme einer Randomisierung (engl. randomness assumption) besteht die Annahme einer zufälligen Positionierung durch randomisierte Zuordnung der Y-Realisierungen  $y_i$  zu den einzelnen Gitterlokalisationen. Diese Annahme interpretiert Y-Werte als Realisierungen einer gleichverteilten Zufallsvariablen und sieht alle Realisierungen als gleich wahrscheinlich an. (Bivand et al., 2013, S. 280) weist auf die Berechnungsunterschiede hin. Die Randomisierungsannahme führt gegenüber der einfacheren Normalverteilungsannahme einen zusätzlichen Korrekturterm der Kurtosis ein. Entspricht die Kurtosis der Attributwertverteilung der einer Normalverteilung, dann stimmen die Testergebnisse unter beiden Annahmen überein. Um so mehr die Verteilung von einer Normalverteilung abweicht, desto mehr kompensiert die Randomisierungsannahme durch Anheben der Varianz.

Beide Annahmen der Normalverteilung (Subscript  $\mathcal{N}$ ) und Randomisierung (Subscript  $\mathcal{R}$ ) resultieren im gleichen Erwartungswert der Indizes,

$$\mathbb{E}_{\mathcal{N}}[I] = \mathbb{E}_{\mathcal{R}}[I] = -\frac{1}{(n-1)} \quad (4.8)$$

$$\mathbb{E}_{\mathcal{N}}[c] = \mathbb{E}_{\mathcal{R}}[c] = 1 \quad (4.9)$$

während die Varianzen je nach Annahme voneinander abweichen und ihre Berechnung komplex ist:

Als Korrelationsindex ist der Moran I-Wert eine dimensionslose Größe. Der realisierte Ausgabewert  $\hat{I}$  einer Stichprobe hat allein keine direkte Aussagekraft, sondern ist erst im Bezug zu Verteilungsmomenten der mathematischen Statistik  $I$  interpretierbar. Gilt  $\hat{I} > \mathbb{E}[I]$ , dann weisen räumlich verbundene Lokationen tendenziell ähnliche Attributswerte auf. Der Stärkegrad dieser positiven Autokorellation steigt mit dem Abstand  $|\hat{I} - \mathbb{E}[I]|$  an. Für  $\hat{I} < \mathbb{E}[I]$  hingegen weisen räumlich verbundene Lokationen tendenziell unterschiedliche Attributswerte auf. Die Interpretation der c Statistik von Geary verhält sich genau umgekehrt. Für Werte  $\hat{c} > \mathbb{E}[c]$  liefern räumlich benachbarte Positionen tendenziell unterschiedliche Werte und vice versa für  $\hat{c} < \mathbb{E}[c]$ . [Schabenberger Gotway 2005, p.22]

Ein Hauptgrund für die Popularität der Statistiken ist die asymptotische Normalverteilung der Ausgabewerte mit zunehmendem  $n$  (Cliff and Ord 1973). Um die Signifikanz der Abweichung vom Erwartungswert zu bewerten, wird der realisierte Index-Ausgabewert standardisiert und durch z-Werte der Standardnormalverteilung bewertet.

$$T(Y(s_1), \dots, Y(s_n)) = \frac{\hat{I} - \mathbb{E}[I]}{\sqrt{Var(I)}} \sim \mathcal{N}(0, 1)$$

Diese Pivottestgröße liefert im Bezug zur Standardnormalverteilung eine Wahrscheinlichkeit für das Auftreten der beobachteten Realisierung der Statistik I unter der Nullhypothese. Die Nullhypothese geht von zufällig verteilten Y-Werten  $y_i$  aus, sodass keine Autokorellation bezüglich der gewählten räumlichen Gewichte vorliegt.

$$\text{Einseitiger Test: } (H_0 : \hat{I} \leq \mathbb{E}[I]) \quad (H_1 : \hat{I} > \mathbb{E}[I]) \quad (4.10)$$

$$\text{Einseitiger Test: } (H_0 : \hat{I} \geq \mathbb{E}[I]) \quad (H_1 : \hat{I} < \mathbb{E}[I]) \quad (4.11)$$

$$\text{Zweiseitiger Test: } (H_0 : \hat{I} = \mathbb{E}[I]) \quad (H_1 : \hat{I} \neq \mathbb{E}[I]) \quad (4.12)$$

Zumeist wird dieser Test in einer einseitigen Form angewendet, deren Alternativhypothese signifikant größere Realisierungen von  $I$  erfasst. Untersucht wird demnach auf eine mögliche Konzentration von Merkmalen mittels positiver Korrelation. Eine Messung von Dispersion mittels negativer Korrelation ist weniger relevant und Realisierungen in den negativen Verteilungsändern werden ignoriert. Der zweiseitige Test ist daher ebenfalls praktisch irrelevant, da beide Verteilungsänder fundamental verschiedene Situationen beschreiben. (Goodchild, 1986, S.26)

Dank der asymptotischen Normalverteilung der  $I$ -Statistik lässt sich ein kritischer Bereich über den z-score der Standardnormalverteilung bilden.

KRITISCHER BEREICH $K^*$	ALTERNATIVHYPOTHESE
$(-\infty, -z_{1-\alpha/2}) \cup (z_{1-\alpha/2}, \infty)$	zweiseitig: $H_1 : \hat{I} \neq \mathbb{E}[I]$
$(-\infty, -z_{1-\alpha})$	einseitig: $H_1 : \hat{I} < \mathbb{E}[I]$
$(z_{1-\alpha}, \infty)$	einseitig: $H_1 : \hat{I} > \mathbb{E}[I]$

Tabelle 4.1: Kritische Ablehnungsbereiche der Alternativhypotesen

Eine Ablehnung der Nullhypothese im Fall  $t \in K^*$  impliziert zu einem zuvor fixierten Konfidenzniveau (e.g. 95%) das Vorliegen signifikanter räumlicher Autokorrelation.

Voraussetzung für die korrekte Signifikanz im Einsatz als Test ist die Abwesenheit systematischer Trends. Getestet werden ausschließlich angepasste Trend-Modelle (engl. centering mean model) mit konstanten Mittelwerten über das gesamte Untersuchungsgebiet. Verbleibende Autokorrelationsmuster nach der Mittelung sind auf räumliche Beziehung innerhalb der Gewichtsmatrix zurückzuführen. In numerischen Testläufen lassen sich etwa räumlich unkorrelierte Zufallswerte für jede Lokation konstruieren. Durch Plotten derart zufälliger „Rauschkarten“ wie in Abbildung 3.1 ließe sich kein Muster erkennen.

Wird in diese Daten ein einfacher linearer Trend eingeführt, wie etwa eine Abnahme der Bevölkerungsdichte in West-Ost Richtung, so darf aus den Testwerten dieses Modells keine Autokorrelation abgeleitet werden. Im Westen lokalisierte Werte liegen trendbedingt im Durchschnitt über dem globalen Mittelwert und im Osten gelegene Werte darunter. Es liegt jedoch hierdurch nur eine globale, lineare Verschiebung des ursprünglichen „Rauschbildes“ vor, welches für lokale Nachbarschaftsbeziehungen weiterhin keinerlei Korrelationsmuster erkennen ließe. Trotzdem würden die globalen  $c$  und  $I$  Statistiken auf derartige globale Trends ansprechen und signifikante Autokorrelation suggerieren, da sie ursprünglich für homogene Mittelwerte und Varianzen ausgelegt sind. (Schabenberger and Gotway, 2005, S.22) liefern eine aufschlussreiche Beispielrechnung.

Da eine Annahme der Stationarität des Erwartungswert insbesondere über große Beobachtungsgebiete oft unrealistisch ist, werden zwei Auswege verfolgt.

1. Ein systematischer Trend wird durch Kovariablen eines linearen Trendmodells (engl. mean model) erfasst und korrekt spezifiziert, bevor ein Test der Residuen mit Erwartungswert Null gültige Aussagen liefert. (Bivand et al., 2013, S.277) Für das Modell  $Y(s) = X(s)\beta + e$  werden Fehlerquadrate minimiert (engl. ordinary least squares - OLS) um  $\beta$  zu schätzen. Die Kontrolle erfolgt mittels Moran oder Geary Statistik über die Residuen  $\hat{e}_i = y_i - x^t(s_i)\hat{\beta}$ . Weitere Details folgen im nächsten Kapitel 5 zur Regression (Schabenberger and Gotway, 2005, S.23).
2. Selbst unter global heterogenem Erwartungswert kann es sinnvoll sein, von konstanten Erwartungswerten auszugehen. Unter dieser Annahme wird die Berechnung der Autokorrelation lokalisiert. Dieser LISA-Ansatz lokalisierter Indikatoren wird im folgenden Abschnitt 4.2.2 eingeführt.

Aufgrund der Sensitivität der Tests bezüglich der Verteilungsannahmen und der genauen Konstruktion der Nachbarschaftsbeziehungen kann auf Simulationen als Ergänzung zurückgegriffen werden.

## Monte-Carlo Simulation

In der Monte-Carlo Simulation oder bootstrap-permutationsbasierten Tests werden realisierte Attributwerte in vielen Zufallsläufen neu auf die Lokationen randomisiert verteilt. Im Gegensatz zur Simulation lokaler Statistiken liegen im globalen Fall genug Beobachtungen für eine sehr große Anzahl Permutationen zur ausreichenden Wiederholung der Simulation ohne Gefahr von Wiederholungen vor. Somit werden Inferenzfehler minimiert. Jedoch beseitigen sie nicht die Abhängigkeit der globalen Statistik von Autokorrelationen aller möglichen Quellen und liefern keine Anhaltspunkte zum Entstehungsprozess der Daten. Die Simulationsanwendung darf nicht von Situationen ablenken, in denen bessere Modellanpassungen oder genauere Spezifikationen der untersuchten Variablen nötig sind. (Bivand et al., 2013, p. 278) Die praktische Anwendung auf eine reale Datengrundlage in Kapitel 6.2 liefert weitere Details zur Berechnung und Interpretation.

### 4.2.2 Lokale Autokorrelationsmaße

Globale Autokorrelationsmaße aggregieren über die einzelnen lokalen Nachbarschaftsbeziehungen der räumlichen Einheiten. Die Statistik für Matrizen  $M_2$  von Mantel kann als Summe der einzelnen Beiträge  $m_i$  aller Datenpunkte  $s_i$  geschrieben werden.

$$M_2 = \sum_{i=1}^n \sum_{j=1}^n w_{ij} u_{ij} = \sum_{i=1}^n m_i , \quad m_i = \sum_{j=1}^n w_{ij} u_{ij} \quad (4.13)$$

Die Beiträge  $m_i$  werden hierbei als lokales Autokorrelationsmaß der Lokation  $s_i$  interpretiert. In gleicher Weise lassen sich Spezialisierungen von  $M_2$  wie die globalen Autokorrelationsindizes in ihre lokalen Einzelbestandteile zerlegen und lokale Tests konstruieren (Schabenberger and Gotway, 2005, S.23).

Mit ihrer Hilfe sollen zum einen räumliche Anhäufungen und Konzentrationen (engl. Cluster), also Beobachtungen mit sehr ähnlichen Nachbarn bezüglich des Attributs, ermittelt werden. Zum anderen werden lokale Hochpunkte bzw. stationäre Schwerpunkte mit außergewöhnlichen Spitzenwerten (engl. Hotspots) erfasst, deren Beobachtungen sehr von

denen ihrer Nachbarn abweichen.

Durch Anselin (1995) wurde eine spezielle Klasse lokaler Indikatoren zur Messung räumlicher Beziehung (engl. Local Indicators of spatial association) (**LISA**) eingeführt. Durch Zerlegung globaler Statistiken wie Moran's I und Geary's c sollen besonders signifikante Beobachtungen sowie Ausreißer gefunden werden. Besonderheiten für ein LISA-Maß sind:

1. Messung des Grades räumlicher Autokorellation im Umkreis einer Lokation.
2. Berechnung für jede einzelne Lokation ist getrennt möglich.
3. Summe über alle Lokationen ist proportional zu einem globalen Maß.

### **Lokaler Gearyscher Index**

Die  $n$  lokalen Einzelindizes  $c_i = c(s_i)$  werden für jede Lokation  $s_i$  ermittelt.

$$c_i = \frac{\sum_{j=1}^n w_{ij} (y_i - y_j)^2}{1/n \sum_{i=1}^n (y_i - \bar{y})^2}, i \neq j, \text{ für } j \text{ innerhalb d von i} \quad (4.14)$$

In Abhängigkeit des Proportionalitätsfaktors  $\gamma$  werden die Einzelindizes zum zum globalen Geary Index  $c$  aufsummiert.

$$\gamma = \frac{2n}{(n-1)} \sum_{i=1}^n \sum_{j=1}^n w_{ij}$$

### **Lokaler Moran Index**

Die  $n$  lokalen Einzelindizes  $I_i = I(s_i)$  werden für jede Lokation  $s_i$  ermittelt und proportional über  $\gamma$  zum globalen Moran Index  $I$  aufsummiert. Die lokalen Moran-Statistiken  $I_i = I(s_i)$  sind als Kreuzprodukt skalierter Attributswerte  $y_i = Y(s_i)$  zwischen einer Lokation  $s_i$  und ihren Nachbarn  $s_j$  konstruiert und definiert als

$$I_i = \frac{(y_i - \bar{y}) \sum_{j=1}^n w_{ij} (y_j - \bar{y})}{1/n \sum_{j=1}^n (y_j - \bar{y})^2}, i \neq j, \text{ für } j \text{ innerhalb d von i} \quad (4.15)$$

Der Proportionalitätsfaktor beträgt

$$\gamma = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n w_{ij} \sum_{i=1}^n (y_i - \bar{y})^2$$

Der Erwartungswert beträgt unter der Randomisierungsannahme nicht vorhandener Autokorrelation.

$$\mathbb{E}_r[I_i] = \frac{-1}{n-1} \sum_{j=1}^n w_{ij}$$

Die Statistik misst somit die Stärke des Beitrags zur globalen Autokorrelation einer Beobachtung  $y_i$  in Bezug zu ihren Nachbarn  $y_j$ . Die Interpretation erfolgt analog zum globalen Fall. Für  $I_i > \mathbb{E}[I_i]$  weisen mit die Raumeinheit  $s_i$  verbundenen Positionen ähnliche Werte zu  $y_i$  auf. Diese lokal positive Autokorrelation weist auf Cluster hin und bedeutet, dass derartige hohe Werte an derartigen Positionen von weiteren hohen Werten umgeben sind, und geringe Werte überwiegend von weiteren geringen Werten. Für  $I_i < \mathbb{E}[I_i]$  haben mit Position  $s_i$  verbundene Lokationen überwiegend abweichende Werte. Hohe Werte sind von geringen Werten umgeben und umgekehrt, was auf Hot-Spots und Cold-Spots hinweist. Es liegt dann lokal negative Autokorrelation vor. Ein stark negativer Wert deutet zudem auf Ausreisser hin (Schabenberger and Gotway, 2005, S. 24).

Zur Visualisierung signifikanter räumlicher Cluster oder Hot-Spots werden nicht die lokalen  $I_i$ -Werte direkt visualisiert, sondern ihre Werte aus dem Moran Scatter Plot. In diesem Plot wird jeweils ein standardisierte Attribut  $y_i$  gegen den standardisierten Durchschnittswert der räumlichen Nachbarschaft dieser Beobachtung (engl. spatial lag) gemäß Gewichtsmatrix  $W$  geplottet. In dieser Abhängigkeit ihrer Nachbarschaftsbeziehungen (engl. spatial neighbourhood association) werden die Werte in vier Klassen gesammelt. Möglich sind hierbei hotspots, coldspots sowie 2 Arten von clustern. Die statistisch signifikanten Einheiten werden in der Karte hervorgehoben und nach ihrer Konzentrationsklasse farblich repräsentiert. Diese Karten werden als LISA-maps bezeichnet. Genaue Details und Beispielrechnungen sind in Abschnitt 6.3 ausgeführt.

Wie auch bei globalen Statistiken kann für lokale Statistiken die Abweichung vom Erwartungswert unter verschiedenen Annahmen getestet werden. Möglich sind die Annahme der Normalverteilung, analytischer Randomisierung, numerischer Sattelpunkt Approximation und exakte Methoden.

Für Autokorrelationstests werden zunächst die Momente der  $I_i$ -Verteilung genutzt. Die

genauen Verteilungseigenschaften der Autokorrelationsstatistiken sind nicht eindeutig. Gauß-Approximationen funktionsieren für lokale Statistiken weniger gut als für die globalen Maße. Die Anzahl der Nachbarn jeder Beobachtung ist für gewöhnlich klein, wodurch die Annahme der Normalverteilung problematisch ist. Als zweite Alternative wird die bedingte Randomisierung präferiert, da eine sonst vorhandene globale Autokorrelation die Interpretation der  $I_i$  beeinflusst (Fischer and Getis, 2010, S. 271). Für diese Randomisierung schlägt Anselin (1995) daher zusätzliche Bedingung vor, sodass Attributwert der Stelle  $i$  nicht permutiert wird. Die Implementierung der Permutation kann im lokalen Fall beschleunigt werden, da nur Nachbarwerte permutiert werden. Jedoch weisen Besag und Newell (1991) sowie Waller und Gotway (2004) darauf hin, dass bei heterogenen Erwartungswerten und Varianzen, die Randomisierungsannahme unzulässig ist. Einigen Fällen macht die Idee der Permutation wenig Sinn. Daher wird der Einsatz von Monte-Carlo Tests empfohlen. Unter Einsatz numerischer Methoden werden Sattelpunkt-Approximation oder exakte Wahrscheinlichkeitswerte ermittelt.

Lokalisierte Autokorrelationsmaße werden vor allem als exploratives Werkzeug empfohlen, sind jedoch mitunter schwierig zu interpretieren. Es wird explizit abgeraten, sie zur statistischen Bestätigung zu nutzen. Auf einem Gitter mit  $n$  Lokationen werden  $n$  lokale Maße ermittelt, welche jedoch nicht einzeln auf Autokorrelation getestet werden können (engl. multiplicity problem). Selbst unter Abwesenheit von Autokorellation sind die  $I_i$  korreliert, sobald dieselben Lokationen mehrfach involviert sind. (Schabenberger and Gotway, 2005, S. 25).

Das Aufspüren lokaler Autokorrelation bei gleichzeitig vorliegender globaler Autokorrelation ist anspruchsvoll. Die Ergebnisse lokaler Abhängigkeitstests sind daher als bedingt durch globale Autokorrelation zu betrachten und werden allgemein durch zugrundeliegende Prozesse, welche die Daten generieren, beeinflusst. Dieser indirekte Einfluss besteht womöglich entlang einer Bandbreite an globaler bis lokaler Größenordnung außerhalb der konkreten Beobachtungsreichweite. Das Aufspüren lokaler und globaler räumlicher Autokorrelation ist kein Selbstzweck für sich, sondern ein Zwischenschritt im komplexen Konstruktionsprozess eines adequaten Modells.

# Kapitel 5

## Räumliche Regression

Regressionsmodelle stellen die primäre Klasse zur räumlichen Modellanalyse dar. Zunächst bietet sich die Klasse der *multiplen linearen Regression* (MLR) an. Sie dient mit  $k$  unabhängigen Regressoren der Erklärung einer abhängigen Zielgröße (Regressand). Dies stellt eine Spezialisierung des *allgemeinen linearen Modells*, auch *multivariates lineares Regressionsmodell* dar, welches mehrere korrelierte (abhängige) Zielgrößen modelliert. Die multiple lineare Regression setzt aber in der Grundvariante als *klassisches multiples lineares Modell (KL-Modell)* identisch und unabhängig verteilte (i.i.d.) Störterme voraus. Das *klassisch lineare Modell der Normalregression (KLN-Modell)* geht zusätzlich noch von einer Normalverteilung der Störgrößen aus. Da wir jedoch Heteroskedastizität und Autokorrelation berücksichtigen, nutzen wir in dieser Arbeit das *verallgemeinerte multiple lineare Modell (VL-Modell)* mit einer erweiterten Varianz-Kovarianzmatrix der Fehlerterme. (Heteroskedastizität liegt z.B. vor, wenn die Fehlerterme verschiedener Zielgrößen unterschiedliche Varianzen aufweisen.)

$Y$  stellt einen Vektor aus Beobachtungswerten dar und wird im Regressionsmodell als zu erklärende, abhängige Variable modelliert und als Regressand, endogene Variable oder Zielgröße (engl. regressand, endogenous variable, response variable) bezeichnet.

$X$  besteht hingegen als Matrix aus insgesamt  $k$  n-dimensionalen Spaltenvektoren als den erklärenden, unabhängigen Variablen, welche auch als Regressoren, exogene Variablen oder Kovariate (engl. regressors, exogenous variables, explanatory variables, covariates) bezeichnet werden.

Die Fehlerterme bzw. Störgrößen (engl. error term, disturbance term, noise) werden mit  $\varepsilon$  bezeichnet.

Die *gewöhnliche Methode der kleinsten Quadrate (GKQ-Methode)* (engl. ordinary least squares - OLS)) dient als Standardverfahren der Ausgleichsrechnung zur Schätzung der unbekannten Modellparameter  $\beta$ ,  $\sigma^2$  und  $\rho$  im KL-Modell. Dagegen dient die Verallgemeinerte Kleinste-Quadrate-Schätzung (**VKQ**-Methode) (engl. generalized least squares - GLS) der Parametrisierung des VL-Modells.

Der Unterschied in der räumlichen Struktur zwischen stetigen Punktlokationen der Geostatistik und diskreten Raumeinheiten als Partition des Gesamtraumes in der Gitterdatenanalyse spiegelt sich auch in den Regressionsmodellen wieder. Der Georegression liegt die Stationarität der Kovarianz im Raum als fundamentale Annahme zugrunde. Die Verteilungsmodellierung der Residuen erfolgt mittels Variogram und zugehörigen Kovariogrammen. Im Gegensatz dazu wird im folgenden die Annahme der Stationarität ersetzt durch räumlich autoregressive Annahmen auf Grundlage der Gewichtsmatrix.

Peter Whittle forschte ab 1949 an der Zeitreihenanalyse und übernahm eine analoge Theorie für stationäre Gaußprozesse auf räumliche Prozesse Whittle (1954). (Gani, 1986, S. 187) enthält eine persönliche Zusammenfassung aus Whittles Sicht und seiner Zusammenarbeit mit Matern, Bartlett und weiteren bekannten Namen der Stochastik.

## 5.1 Spatial Statistics – Allgemeine Räumliche Autoregression

Im folgenden werden räumliche Autoregressionsmodelle (engl. spatial autoregressive models) eingeführt. Wie in Kapitel 4 erläutert, ist die Annahme unabhängig identisch verteilter Beobachtung aufgrund inhärenter räumlicher Abhängigkeiten nicht gerechtfertigt. Zudem muss der Einfluss räumlicher Autokorrelation getrennt werden von den impliziten Unterschieden der multivariaten Verteilung.

Im folgenden wird die relevante Menge der Raumeinheiten  $\{R_1, \dots, R_n\}$  durch ihre Indizes  $i = 1, \dots, n$  vereinfacht repräsentiert. Insbesondere wird es sich immer um die Stichpro-

beneinheiten handeln. Als lineares Grundmodell für Abhängigkeiten zwischen einzelnen Einheiten dient

$$Y_i = \beta_0 + \sum_{j=1}^k \beta_j x_{ij} + \nu_i, \quad i = 1, \dots, n \quad (5.1)$$

mit  $Y_i$  als relevantes Attribut für jede Raumeinheit  $i$  und  $(x_{ij} : j = 1, \dots, k)$  als Menge der Erklärungsattribute von  $i$  welche den Wert  $Y_i$  beeinflussen. Der Hauptunterschied zur Georegression ist die Modellierung von  $\nu_i$  über ein eigenes explizites lineares Model statt als zufälligen Fehlerterm  $\varepsilon_i$ .

Korrelierte Beobachtungen werden in Matrixschreibweise über Zusammenhang (5.2) modelliert.

$$Y = \mathbf{X}^\top \beta + \nu \quad (5.2)$$

Die Störgröße  $\nu$  wird als multinomial verteilter Zufallsvektor aus (nicht notwendig normalverteilten) Fehlertermen mit  $\mathbb{E}[\nu] = 0$  angenommen, sodass

$$\mathbb{E}[Y] = \mathbf{X}\beta$$

gilt. Während die Störterme  $\nu$  im klassischen Regressionsmodell als i.i.d angenommen werden, benötigen wir nunmehr eine Möglichkeit, räumlich auftretende Korrelationsstrukturen in dieses Modell zu integrieren. Zum einen lässt sich der Fehlerprozess als *raumstrukturell autoregressiver Prozess* (engl. spatially autoregressive process - SAR)

$$\nu = \rho \mathbf{W}\nu + \varepsilon$$

erfassen. Zum anderen ist eine Integration als *raumstruktureller Gleitmittelprozess* (engl. spatially moving average process - SMA)

$$\nu = \rho \mathbf{W}\varepsilon + \varepsilon$$

denkbar mit Gewichtsmatrix  $\mathbf{W}$ , Autoregressionsparameter  $\rho$  und  $\varepsilon$  als Vektor von i.i.d.-verteilten Störgrößen.

Die Struktur der Korrelation zwischen Gebieten wird durch eine Varianzmatrix  $V$  erfasst. Der Trend  $Y = \mathbf{X}^T \beta$  wird in linearer Abhängigkeit von unabhängigen Erklärungsvariablen  $\mathbf{X}$  bzw. mehreren Kovariablen (engl. covariate, control variable)  $X_1, \dots, X_k$  modelliert. Liegt als Output keine hinreichend normale, multivariate Verteilung vor, kann eine Transformation der abhängigen Reaktionsvariable (auch Zielgröße oder Regressand) helfen. Es folgen verschiedene Ansätze von Korrelationsstrukturen, verwirklicht in den Modellklassen der Simultaneous Autoregressive Models (SAR) und Conditionally Autoregressive Models (CAR).

## 5.2 Simultaneous Autoregressive Models - SAR

Das Hauptziel ist die Modellierung der Kovarianzstruktur von  $\nu$  unter Berücksichtigung der räumlichen Abhängigkeiten zwischen den Instanzen. Die räumliche Gewichtsmatrix  $W = [w_{ij} : i, j = 1, \dots, n]$  erfasst die diskrete Raumstruktur und räpresentiert oft ein Maß räumlicher Nähe. Größere Werte  $w_{ij}$  bedeuten hierbei größere Nähe und Einfluss von  $i$  und  $j$  aufeinander. Jedes Residuum  $\nu_i$  wird durch die Residuen der Nachbargebiete  $j$  mit positiven paarweisen Raumgewichten  $w_{ij}$  beeinflusst. Diesen Einfluss räpresentiert das lineare Teilmodell

$$\nu_i = \sum_{j=1; j \neq i}^n \alpha(w_{ij}) \nu_j + \varepsilon_i \quad (5.3)$$

mit  $\alpha(w_{ij})$  als Einflussfunktion und  $\varepsilon_i$  als Anteil des Residuums, welcher nicht durch andere Einheiten beeinflusst wird. Da die Gewichtsmatrix bereits große Flexibilität durch die Spezifikation der Raumstruktur bietet, wird die Funktion  $\alpha$  auf die einfachste denkbare Form eines räumlich unabhängigen Skalierungsfaktors  $\rho$  reduziert. Folglich resultiert aus (5.3) der Zusammenhang

$$\nu_i = \rho \sum_{j=1}^n w_{ij} \nu_j + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad i = 1, \dots, n \quad (5.4)$$

als lineares Regressionsmodell (ohne Interzeptor) für jedes Residuum  $\nu_i$  im Regress auf seine Nachbarn  $\nu_j$  mit Koeffizienten  $\rho w_{ij}$ . Aufgrund dieser Regression der Residuen auf „sich selbst“ wird dieses Teilmodell als *Räumlich Autoregressives Modell der Residuen* (engl. spatial autoregressive model of residual dependencies) bezeichnet Whittle (1954), inspiriert von den Begrifflichkeiten der Zeitreihentheorie. Selbstregression durch individuelle Residuen wird durch Voraussetzung  $w_{ii} = 0$  verhindert. Alternativ kann für die Summation auch  $i \neq j$  vorausgesetzt werden.

Nehmen wir für die intrinsischen Restkomponenten  $\varepsilon_i$  eine *i.i.d.* Verteilung gemäß

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2), i = 1, \dots, n \quad (5.5)$$

an, so gilt  $\mathbb{E}[\varepsilon_r] = 0$  sowie  $\mathbb{E}[\varepsilon_r \varepsilon_s] = 0$  und  $Var(\varepsilon_r) = \sigma_r^2$ . Parameter  $\rho$  steuert den Grad der Regression. Für  $\rho = 0$  wird jedes Residuum  $e_i$  auf seine eigene intrinsische Komponente  $\varepsilon_i$  reduziert und räumliche Abhängigkeiten verschwinden ganz. In diesem Fall wird Modell (5.2) zu einem einfachen linearen Regressionsmodell reduziert. Für große  $\rho$  werden die (positiven und negativen) räumlichen Abhängigkeiten dominanter. Daher wird  $\rho$  auch als *Parameter räumlicher Abhängigkeit* bezeichnet. Außerdem muss für beliebige Paare  $ij$  und  $kh$  mit positiven räumlichen Gewichten  $w_{ij}, w_{kh} > 0$  und einem nicht verschwindenden Parameter  $\rho \neq 0$ ,

$$\frac{\rho w_{ij}}{\rho w_{kh}} = \frac{w_{ij}}{w_{kh}}$$

gelten. Somit ist die relative Stärke ihrer räumlichen Dependenz ausschließlich durch ihre Gewichte bestimmt. Das Modell bietet somit eine natürliche Zuordnung der Parameter, da  $\rho$  das allgemeine Niveau räumlicher Abhängigkeit bestimmt und die räumliche Struktur der Gewichtsmatrix  $W$  ihre relative Stärke zwischen individuellen Paaren räumlicher Einheiten.

Zusammen formen (5.4) und (5.5) das *Autoregressive Modell der Residuen* nach Whittle

(1954) in Matrixschreibweise,

$$\nu = \rho \mathbf{W} \nu + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_n), \quad \text{diag}(\mathbf{W}) = 0 \quad (5.6)$$

Entwickelt wurde diese Form durch Ord (1975), welcher auch vom räumlich autoregressiven Prozess erster Ordnung sprach. Die Modellgleichung  $\nu = \rho \mathbf{W} \nu + \varepsilon$  lässt sich direkt nach  $\varepsilon$  umstellen und ergibt somit  $(\mathbb{I}_n - \rho \mathbf{W})\nu = \varepsilon$ . Sofern Inverse  $(\mathbb{I}_n - \rho \mathbf{W})^{-1}$  existiert, lässt sich die folgende Lösung für  $\nu$  in reduzierter Form erhalten

$$\nu = (\mathbb{I}_n - \rho \mathbf{W})^{-1} \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_n), \quad \text{diag}(\mathbf{W}) = 0 \quad (5.7)$$

### 5.2.1 Räumlich fehlerbezogenes Modell - SEM

Die Integration der autoregressiven Residuen aus (5.6) in das ursprüngliche Regressionsmodell (5.2) der Form  $Y = \mathbf{W}\beta + \nu$  liefert nun das Gesamtmodell

$$Y = \mathbf{X}\beta + \nu, \quad \nu = \rho \mathbf{W}\nu + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_n), \quad \text{diag}(\mathbf{W}) = 0 \quad (5.8)$$

und wird als *räumlich fehlerbezogenes Modell* (engl. spatial error model - SEM) mit räumlich autokorrelierten Fehlertermen bzw. raumstruktureller Autokorrelation im Fehlerterm bezeichnet. Diese Form stellt die direkte Umsetzung des Räumlich Autoregressiven Modells auf die Fehlerterme dar. Abweichend vom klassischen Regressionsmodell werden die Störgrößen nicht als i.i.d angenommen, sondern folgen einem räumlich autokorrelierten Prozess. Zugrunde liegt die Annahme, dass alle räumlichen Abhängigkeiten in den unbeobachteten Fehlerterminen  $\nu$  liegen, woher auch der Name röhrt. Wir gehen davon aus, mit den Regressoren in  $\mathbf{X}$  nicht die gesamte räumliche Variabilität erklären zu können bzw. die fehlenden erklärenden Faktoren nicht zu kennen und daher durch  $\nu$  abzudecken.

#### Simultanitätsstruktur

Alternativ erhalten wir mit der Umstellung  $\nu = Y - \mathbf{X}\beta$  bzw.  $\nu_i = Y_i - \sum_{j=1}^k \beta_j x_{ij}$  mit der Form  $Y = \mathbf{X}\beta + \rho \mathbf{W}(Y - \mathbf{X}\beta) + \epsilon$  das *Simultan Spezifizierte Autoregressive Modell* (engl.

simultaneous autoregressive model -SAR). Die Bezeichnung der „simultanen Spezifikation“ bezieht sich dabei auf die Anwendung der Modellgleichung auf jede Lokation  $i$  gemäß

$$Y_i = \beta_0 + \sum_{j=1}^k \beta_j x_{ij} + \rho \sum_{h=1}^n w_{ih} \left[ Y_h - \sum_{j=1}^k \beta_j x_{hj} \right] + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad i = 1, \dots, n \quad (5.9)$$

unter simultaner Variabilität von  $Y_i$  und  $Y_h$ . Derartige Interdependenzen sorgen für zusätzliche Komplexität und grenzen diese Modellklasse von den *Bedingt Autoregressiven Modellen* (engl. conditional autoregressive models - CAR) im nächsten Abschnitt ab.

In Matrixschreibweise lässt sich das Modell in die Kurznotation

$$Y = \mathbf{X}\beta + \rho\mathbf{W}(Y - \mathbf{X}\beta) + \epsilon \quad \Rightarrow \quad (\mathbb{I}_n - \rho\mathbf{W})(Y - \mathbf{X}\beta) = \varepsilon$$

umwandeln, wobei  $(\mathbb{I}_n - \rho\mathbf{W})$  für ein wohldefiniertes Modell regulär sein muss. (Einige Quellen wie (Cressie, 1993, S.440), (Waller and Gotway, 2004, S.363) und (Bivand et al., 2013, S. 293) fassen  $\rho\mathbf{W}$  unter einer Matrix  $\mathbf{B}$  zusammen.)

Daraus wird die Varianz-Kovarianzmatrix von  $Y$  durch

$$\Sigma_Y = \text{Var}(Y) = (\mathbb{I}_n - \rho\mathbf{W})^{-1} \Sigma_\varepsilon (\mathbb{I}_n - \rho\mathbf{W}^\top)^{-1}$$

abgeleitet und mit  $\mathbb{E}[Y] = \mathbf{X}\beta$  unterliegt  $Y$  einer multivariaten Normalverteilung. Oftmals wird  $\Sigma_\varepsilon = \sigma^2 \mathbb{I}_n$  gesetzt und somit die Varianz-Kovarianzmatrix zu

$$\Sigma_Y = \text{Var}(Y) = \sigma^2 (\mathbb{I}_n - \rho W\mathbf{W}^{-1} (\mathbb{I}_n - \rho\mathbf{W}^\top)^{-1})$$

vereinfacht. Hierbei sind asymmetrische Gewichtsmatrizen  $W$  zulässig, diese werden aber in der Praxis oft als symmetrisch angenommen (Bivand et al., 2013, S. 294).

Um das Modell als Instanz des allgemeinen linearen Regressionsmodells zu formulieren, wird zunächst der Sammelterm

$$B_\rho := \mathbb{I}_n - \rho\mathbf{W} \quad (5.10)$$

eingeführt und System (5.7) durch

$$\nu = (\mathbb{I}_n - \rho \mathbf{W})^{-1} \varepsilon = B_\rho^{-1} \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_n), \quad \text{diag}(\mathbf{W}) = 0 \quad (5.11)$$

zusammengefasst. Durch das Invarianztheorem multi-normaler Verteilungen folgt aus der multi-normalität von  $\varepsilon$  dieselbe Verteilung für  $\nu$  mit der Kovarianz

$$\begin{aligned} \text{cov}(\nu) &= \text{cov}(B_\rho^{-1} \varepsilon) = B_\rho^{-1} \text{cov}(\varepsilon) (B_\rho^{-1})^\top \\ &= B_\rho^{-1} (\sigma^2 \mathbb{I}_n) (B_\rho^{-1})^\top = \sigma^2 B_\rho^{-1} (B_\rho^\top)^{-1} = \sigma^2 (B_\rho^\top B_\rho)^{-1} = \sigma^2 V_\rho \end{aligned}$$

und der räumlichen Kovarianzstruktur  $V_\rho$ , welche durch

$$V_\rho := (B_\rho^\top B_\rho)^{-1} \quad (5.12)$$

alle räumlichen Aspekte der Kovarianz definiert. Hiermit wird SEM-Modellgleichung (5.8) durch

$$Y = \mathbf{X}\beta + \nu, \quad \nu \sim \mathcal{N}(0, \sigma^2 V_\rho) \quad (5.13)$$

umgeschrieben als allgemeines lineares Regressionsmodell.

Eine dritte Darstellungsform entsteht durch Eliminierung aller simultaner Beziehungen  $\nu = \rho \mathbf{W} \nu + \varepsilon$  nach Einsetzen von  $B_\rho$  und ergibt

$$Y = \mathbf{X}\beta + B_\rho^{-1} \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_n) \quad (5.14)$$

die reduzierte Form von Modell (5.8).

### 5.2.2 Räumlich verzogenes Modell - SLM

Ein alternatives Modell entsteht durch die Annahme, dass die autoregressiven Beziehungen zwischen den abhängigen Variablen selbst auftreten. Es wird als *räumlich verzogenes Modell* (engl. spatial lag model) (SLM) bezeichnet. Die zugrundeliegenden räumlichen Beziehungen werden auch hier durch die Gewichtsmatrix  $\mathbf{W}$  repräsentiert. Die Grund-

gleichung (5.1) wird durch Einsetzen von  $\nu_i = \rho \sum_h w_{ih} Y_h + \varepsilon_i$  angepasst zu

$$Y_i = \beta_0 + \sum_{j=1}^k \beta_j x_{ij} + \rho \sum_{h=1}^n w_{ih} Y_h + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad i = 1, \dots, n \quad (5.15)$$

Der autoregressive Term  $\rho \sum_{h=1}^n w_{ih} Y_h$  spiegelt mögliche Abhängigkeiten des Wertes  $Y_i$  von den Werten  $Y_h$  anderer Einheiten wieder. Neben den Erklärungsvariablen aus  $\mathbf{X}$  wird der Modelloutput auch indirekt durch die Werte in der räumlichen Nähe beeinflusst.

### Simultanitätsstruktur

Da die Residuen als unabhängig angenommen werden, scheint das obige Modell der gewöhnlichen Kleinste-Quadrate-Methode (KQM, engl. OLM) zu entsprechen, mit dem zusätzlichen Term  $\rho(\sum_{h=1}^n w_{ih} Y_h)$ , wobei der unbekannte räumliche Abhängigkeitsparameter  $\rho$  als ein einfacher beta-Parameter erscheint. Dies ist jedoch keineswegs der Fall, denn die  $Y_h$  sind Zufallsvariablen und erscheinen zudem auf beiden Seiten des Systems in der typischen „Simultanitätsstruktur“. Das heißt,  $Y_i$  wird auch in den Gleichungen für  $Y_h$  auftreten wann immer  $w_{hi} > 0$  ist. Es handelt sich somit keineswegs um einen einfachen Term eines KQM-Models. Dies zeigt sich insbesondere in der Matrixnotation

$$Y = \mathbf{X}\beta + \rho \mathbf{W}Y + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_n)$$

welche wir weiter umformen. Durch Gruppierung der Y-Terme erhalten wir

$$\begin{aligned} Y - \rho \mathbf{W}Y &= \mathbf{X}\beta + \varepsilon \Rightarrow (\mathbb{I}_n - \rho \mathbf{W})Y = \mathbf{X}\beta + \varepsilon \\ \Rightarrow B_\rho Y &= \mathbf{X}\beta + \varepsilon \Rightarrow Y = B_\rho^{-1} \mathbf{X}\beta + B_\rho^{-1} \varepsilon \end{aligned}$$

und leiten die reduzierte Form

$$Y = B_\rho^{-1} \mathbf{X}\beta + B_\rho^{-1} \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_n) \quad (5.16)$$

ab. In dieser Form ist ersichtlich, dass der räumliche Verzugsterm (engl spatial lag term)  $\rho \mathbf{W}Y$  nicht nur ein simpler Regressionsteil ist, sondern das Modell grundlegend vom KQM abweicht.

Zudem kann wenn auch das Modell als Sonderform der Linearen Regression formuliert werden, wenngleich auch nicht so direkt wie im Fall der SEM. Der räumliche Parameter  $\rho$  wird hierzu als bekannte Größe behandelt und das SLM auf gegebenes  $\rho$  bedingt. Hierfür betrachten wir

$$X_\rho = B_\rho^{-1} \mathbf{X}$$

als transformierten Datensatz. Zudem nutzen wir analog zum SEM den Sammelterm  $B_\rho$  und die Kovarianzstruktur  $V_\rho$  gemäß

$$B_\rho = \mathbb{I}_n - \rho \mathbf{W} \quad \text{sowie} \quad V_\rho := (B_\rho^\top B_\rho)^{-1}$$

um die Form

$$Y = X_\rho \beta + u, \quad u \sim \mathcal{N}(0, \sigma^2 V_\rho) \quad (5.17)$$

zu erreichen. Hier ist  $\rho$  nicht mehr nur unbekannter Parameter in der Kovarianzmatrix  $V_\rho$ , sondern tritt ebenso in  $X_\rho$  auf. Während obiger Ausdruck also die Anwendung der gewöhnlichen KQ-Methode auch auf räumlich verzogene Modelle (SLM) erlaubt, so ist die Anwendung eingeschränkter als für räumliche Fehlermodelle (SEM).

### Interpretation der Beta-Koeffizienten

Ein weiterer wichtiger Unterschied zwischen SLM und SEM ist die Interpretation der Beta-Koeffizienten. Eine angenehme Eigenschaft der gewöhnlichen KQM ist die einfache Interpretation ihrer Beta-Koeffizienten. Aus der gewöhnlichen KQM für die Anzahl Arbeitnehmer in der Logistik  $Y_i$  für Gemeinde  $i$  mit Arbeitslosigkeit als j-tem Regressor  $x_{ij}$  im Modell

$$Y_i = \beta_0 + \sum_{j=1}^k \beta_j x_{ij} \quad \text{mit } \varepsilon_i \sim \mathcal{N}(0, \sigma^2), i = 1, \dots, n$$

wird in einem negativen Koeffizienten  $\beta_j$  resultieren. Die auf realisierte Daten aller Attribute einer Gemeinde bedingte Erwartung

$$\mathbb{E}[Y_i|x_{i1}, \dots, x_{ik}] = \beta_0 + \sum_{j=1}^k \beta_j x_{ij}, \quad i = 1, \dots, n \quad (5.18)$$

gibt für  $\beta_j$  im Mittel den (erwarteten) Abfall an Logistikarbeiternehmern der Gemeinde  $i$  je zusätzlichem arbeitslosen Bewohner der Gemeinde an. Der bedingte Erwartungswert kann direkt für jede Gemeinde einzeln gebildet werden, da Attribute anderer Gemeinden hier keinen Einfuss haben. Diese Grenzänderungen können auch als partielle Ableitungen in der Form

$$\frac{\partial}{\partial x_{ji}} \mathbb{E}(Y_i|x_{i1}, \dots, x_{ij}, \dots, x_{ik}) = \beta_j, \quad i = 1, \dots, n, j = 1, \dots, k \quad (5.19)$$

ausgedrückt werden. Sie korrespondieren exakt zum  $\beta_j$  Koeffizienten für Variable  $x_j$  und erlauben eine direkte Interpretation.

Der Nachteil dieser KQM ist jedoch die gänzliche Vernachlässigung räumlicher Abhängigkeiten zwischen Gemeinden, daher wurde es zu Beginn des Kapitels weiterentwickelt zum SEM-Modell (5.13) der Form

$$Y_i = \beta_0 + \sum_{j=1}^k \beta_j x_{ij} + e_i \quad \text{mit } (e_i, \dots, e_n) \sim \mathcal{N}(0, V_\rho)$$

mit  $\mathbb{E}[e_i] = 0$  für alle  $i$ , wodurch weiterhin Gleichungen (5.18) und (5.19) gelten. Somit ist die Interpretation der  $\beta$  Koeffizienten auf das SEM übertragbar, während räumliche Abhängigkeiten einer bestimmten Art (zwischen Residuen) integriert sind. Werden jedoch räumliche Abhängigkeiten zwischen den Beschäftigungszahlen  $Y_i$  der Gemeinden vermutet und über ein SLM modelliert, so ist die Interpretation aufgrund der Simultanitätsstruktur weitaus komplexer. Dies wird ersichtlich aus der reduzierten Form (5.16) in Verbindung mit der Wellen- bzw. Riffelzerlegung aus Anhang A<sup>1</sup>.

---

<sup>1</sup>Soweit  $\rho$  und  $W$  Konvergenzbedingung  $|\rho| < 1/\lambda_W$  erfüllen

$$\begin{aligned}\mathbb{E}[Y|X] &= B_\rho^{-1} X \beta = (\mathbb{I}_n - \rho W)^{-1} X \beta = (\mathbb{I}_n + \rho W + \rho^2 W^2 + \dots) X \beta \\ &= X \beta + \rho W X \beta + \rho^2 W^2 X \beta + \dots\end{aligned}$$

Um hierauf partielle Ableitungen anzuwenden, müssen zuerst *alle* Attribute für *alle* Gemeinden spezifiziert werden. Eine getrennte Darstellung für  $Y_i$  muss im Detail entwickelt werden. Da nun Interaktionen zwischen einzelnen Gemeinden modelliert werden, sind für jede Gemeinde  $i$  die partiellen Ableitungen  $\partial/\partial x_{lj}$  bezüglich weiterer Gemeinden ( $l \neq i$ ) und Attribute  $j$  von  $\mathbb{E}[Y_i|x_{1i}, \dots, x_{ji}, \dots, x_{ki}]$  relevant und müssen im Regressionskoeffizienten berücksichtigt werden. Zuvor traten im SEM keine solchen Effekte zwischen Gemeinden auf und diese partiellen Ableitungen verschwanden für  $l \neq i$  einfach. Im SLM jedoch resultiert eine Erhöhung von  $x_{ij}$ , z.B. der Arbeitslosigkeit in Gemeinde  $i$ , nicht nur direkt in einer Verringerung der Logistikbeschäftigten innerhalb der gleichen Gemeinde  $i$ , sondern sorgt auch indirekt für Beschäftigungsveränderungen (positiv wie negativ) in allen anderen Gemeinden  $l \neq i$ . Im Umkehrschluss sorgen diese Änderungen in  $l$  wiederum zu indirekten Rückkopplungseffekten in  $i$ . Diese räumlichen Riffel- bzw. Welleneffekte führen zu komplexen Zwischenabhängigkeiten, welche die einzelnen Beta-Koeffizienten beeinflussen. Zur Trennung dieser Effekte zerlegen wir die Matrixprodukte der reduzierten Form in Summationen der einzelnen Vektoroperationen. Zur besseren Übersicht nutzen wir kurzzeitig für beliebige  $n \times m$  Matrizen  $A = [a_{ij} : i = 1, \dots, n, j = 1, \dots, m]$  die Notation  $A(i, j) = a_{ij}$  für einzelne Matrixelemente und  $A(\cdot, j)$  für die  $j$ -te Spalte in  $A$ . Zudem kürzen wir  $C := B_\rho^{-1}$  ab.

$$\begin{aligned}\mathbb{E}[Y|X] &= CX\beta = C \sum_{j=1}^k \beta_j X(\cdot, j) \\ &= \sum_{j=1}^k \beta_j \left[ \sum_{h=1}^n X(h, j) C(\cdot, h) \right] = \sum_{h=1}^n \sum_{j=1}^k X(h, j) \beta_j C(\cdot, h)\end{aligned}$$

In dieser Doppelsumme stellt die innere Summe für ein festes  $h$  eine Linearkombination aus

Spaltenvektoren dar. Die äußere Summe wendet darüber eine zweite Linearkombination für alle  $h$  an und ergibt einen finalen Spaltenvektor an Y-Werten. Somit lässt sich eine einzelne Zeile  $i$  aus  $\mathbb{E}[Y|X]$  schreiben als

$$\mathbb{E}[Y_i|X] = \sum_{h=1}^n \sum_{j=1}^k X(h, j) \beta_j C(i, h)$$

Hier von kann nun direkt die partielle Ableitung

$$\frac{\partial}{\partial x_{ij}} \mathbb{E}[Y_i|X] = \beta_j C(i, i)$$

gebildet werden. Im Gegensatz zur partiellen SEM-Ableitung (5.19) hängt dieser Grenzeffekt nicht nur von  $\beta_j$  ab, sondern auch vom  $i$ -ten Diagonaleintrag der Matrix  $C$  bzw.  $B_\rho^{-1}$ , welcher explizit die Form

$$B_\rho^{-1}(i, i) = 1 + \rho W(i, i) + \rho^2 W^2(i, i) + \dots = 1 + \rho^2 W^2(i, i)$$

nimmt, da die Diagonaleinträge  $W(i, i)$  auf Null gesetzt werden. Andererseits sind  $\rho^2 W^2(i, i)$  und höhere Ordnungen positiv, und der Effekt jedes  $\beta_j$  wird durch diese räumlichen Effekte aufgebläht, wie oben beschrieben. Da nun  $Y_i$  auch durch Änderungen in  $Y_h$  beeinflusst wird, folgt für Attribut  $j$  in Gemeinde  $h$  die Grenzänderung

$$\frac{\partial}{\partial x_{hj}} \mathbb{E}[Y_i|X] = \beta_j C(i, h)$$

auf Erwartungswert  $\mathbb{E}[Y_i|X]$ . Der Gesamteffekt aller Gemeinden auf  $\mathbb{E}[Y_i|X]$  durch Attribute aus anderen Gemeinden wird als *indirekte Effekte* bezeichnet. Analog wird der Gesamteffekt aller Attribute in der gleichen Raumeinheit bzw. Gemeinde  $i$  (mittels totalem Differential?) mit *direkten Effekten* benannt. (Waller and Gotway, 2004, S.335)

### 5.2.3 Kombiniertes Modell

Während der Entwicklung des SL-Modells stellt sich die Frage, warum alle unbeobachteten Faktoren als räumlich unabhängig behandelt werden. Es ist praktisch sehr wohl möglich sowohl zwischen den Y-Variablen als auch den Residuen  $\varepsilon$  räumlich autoregressive Abhängigkeiten zu messen. Wird nun zur Unterscheidung mit  $M$  eine eigene Gewichtsmatrix und  $\lambda$  ein zusätzlicher Abhängigkeitsparameter für die Fehlerkomponente eingeführt, so lassen sich beide Modelle gemäß

$$Y = \rho WY + X\beta + \nu, \quad \nu = \lambda M\nu + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_n)$$

kombinieren und in die zugehörige reduzierte Form

$$\begin{aligned} (\mathbb{I}_n - \rho W)Y &= X\beta(\mathbb{I}_n - \lambda M)^{-1}\varepsilon \\ \Rightarrow Y &= (\mathbb{I}_n - \rho W)^{-1}X\beta + (\mathbb{I}_n - \rho W)^{-1}(\mathbb{I}_n - \lambda M)^{-1}\varepsilon \end{aligned}$$

bringen. Dieses Modell wurde durch Kelejian and Prucha (2010) als Räumlich Autoregressives Modell mit autoregressiven Störungen (engl. Spatial Autoregressive Model with Autoregressive disturbances) bezeichnet und mit SARAR(1,1) abgekürzt. In unserer Arbeit wird dieses Modell zur Konstruktion von Vergleichstests zwischen SLM und SEM als Instanzen derselben Modellklasse dienen. Daher beschränken wir uns auf den Spezialfall  $M = W$  und erhalten

$$Y = \rho WY + X\beta + \nu, \quad \nu = \lambda W\nu + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_n) \quad (5.20)$$

als Kombiniertes Modell in der reduzierten Form:

$$Y = (\mathbb{I}_n - \rho W)^{-1}X\beta + (\mathbb{I}_n - \rho W)^{-1}(\mathbb{I}_n - \lambda W)^{-1}\varepsilon$$

Somit resultiert für gegebene Gewichtsmatrix  $W$  aus Kombinationsmodell (5.20) mit  $\rho = 0$  das SE-Modell und mit  $\lambda = 0$  das SL-Modell. Zwar ist dieses Modell wohldefiniert und

kann prinzipiell zur simultanen Schätzung sowohl von  $\rho$  als auch  $\lambda$  herangezogen werden. Jedoch können die Ergebnisse instabil sein, da beide Parameter auf die gleiche Matrix  $W$  bezogen ähnliche Rollen spielen.

## 5.3 Conditional Autoregressive Models - CAR

Während dieses Modell ein zum SEM ähnliches Konzept aufweist, liegt aus statistischer Sicht ein fundamentaler Unterschied vor. Statt in unserem Beispiel die gemeinsame Verteilung  $(Y_1, \dots, Y_n)$  der Arbeitnehmer aller Gemeinden zu modellieren, ist dieser Ansatz ausgerichtet auf die einzelnen bedingten Verteilungen einer jeden Beschäftigungszahl  $Y_i$ , gegeben alle anderen. Der Vorteil ist die Vermeidung einer komplexen Simultanitätsstruktur. Alle univariaten bedingten Verteilungen, welche von multinomial verteilten Zufallsgrößen abgeleitet werden, wiederum selbst normal.

Die reduzierte Form der SE-Modellgleichung wird angepasst

$$\begin{aligned} Y = X\beta + B_\rho^{-1}\varepsilon &\Rightarrow Y - X\beta = B_\rho^{-1}\varepsilon \Rightarrow B_\rho(Y - X\beta) = \varepsilon \\ &\Rightarrow (\mathbb{I}_n - \rho W)(Y - X\beta) = \varepsilon \\ &\Rightarrow Y - X\beta - \rho W(Y - X\beta) = \varepsilon \\ &\Rightarrow Y = X\beta - \rho W(Y - X\beta) + \varepsilon \end{aligned}$$

## 5.4 Parameterschätzung

Wir nutzen die bekannt *Methode größter Plausibilität* (engl. Maximum-Likelihood-Methode - ML). Für gegebene Stichprobenbeobachtung  $y$  und Parametervektor  $\theta$  aus  $\Theta \subseteq \mathbb{R}$  ist der ML-Schätzer durch

$$f(y|\hat{\theta}) = \max_{\theta} f(y|\theta)$$

definiert. Die korrespondierende Likelihoodfunktion  $l(\cdot|y)$  wird

$$l(\theta|y) := f(y|\theta)$$

verdeutlicht die Abhängigkeit vom unbekannten Parameter  $\theta$  für bekannte  $y$ . Da Wahrscheinlichkeitsdichten nichtnegative Werte annehmen und die Log-Likelihoodfunktion  $L(\theta|y) := \log [l(\theta|y)]$  stets monoton wachsend bezüglich  $l(\theta|y)$  ist, kann der ML-Schätzer  $\hat{\theta}$  auch durch die log-Likelihoodbedingung

$$L(\hat{\theta}|y) = \max_{\theta} L(\theta|y) = \max_{\theta} \log(f(y|\theta))$$

beschrieben werden.

#### 5.4.1 ML-Schätzung im klassischen Regressionsmodell

Wir gehen vom KLN-Modell mit normalverteilten Störgrößen aus und sichern somit eine multivariate Normalverteilung der Zielgröße  $Y \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2 \mathbb{I}_n)$ . Wir nutzen

$$\bullet \frac{1}{\sqrt{\det(\sigma^2 \mathbb{I}_n)}} = \frac{1}{\sqrt{(\sigma^2)^n \det(\mathbb{I}_n)}} = \frac{1}{\sqrt{(\sigma^2)^n}} \quad \text{sowie} \quad \bullet (\sigma^2 \mathbb{I}_n)^{-1} = \frac{1}{\sigma^2} \mathbb{I}_n^{-1} = \frac{\mathbb{I}_n}{\sigma^2}$$

und integrieren diese Vereinfachungen für gegebene Stichprobe  $\mathbf{y}$  und gesuchten Parametervektor  $\theta = (\beta_0, \beta_1, \dots, \beta_k, \sigma^2)$  mit  $\Sigma = \sigma^2 \mathbb{I}_n$  in die Dichtefunktion

$$\begin{aligned} f(\mathbf{y}|\beta, \sigma^2) &= \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} \exp \left( -\frac{1}{2} (\mathbf{y} - \mathbf{X}\beta)^\top \Sigma^{-1} (\mathbf{y} - \mathbf{X}\beta) \right) \\ &= \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \exp \left( -\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta) \right) \end{aligned}$$

der Zielverteilung. Daraus folgt die log-Likelihoodfunktion der GKQ-Methode durch

$$L(\theta|\mathbf{y}) = \log(f(\mathbf{y}|\theta)) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta)$$

Zur Ermittlung der ML-Schätzer  $\hat{\theta} = (\hat{\beta}, \hat{\sigma}^2)$  erfolgt die Maximierung von XX. Für beliebige  $\sigma^2$  wird L bezüglich  $\beta$  maximiert durch Minimieren des letzten Terms, auch als *mittlerer quadratischer Fehler (MQF)*

$$\text{MQF}(\beta) := (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta) = \mathbf{y}^\top \mathbf{y} - 2\mathbf{y}^\top \mathbf{X}\beta + \beta^\top \mathbf{X}^\top \mathbf{X}\beta \quad (5.21)$$

bezeichnet. Diese quadratische Form in der Unbekannten  $\beta$  liefert für das Minimierungsproblem das Optimalitätskriterium erster Ordnung in der Form

$$0 \stackrel{!}{=} \nabla \text{MQF}(\beta) = -2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\beta \Rightarrow \mathbf{X}^\top \mathbf{X}\beta = \mathbf{X}^\top \mathbf{y}$$

und nutzt die Symmetrie der quadratischen Matrix  $\mathbf{X}^\top \mathbf{X}$ . Hat  $X$  den vollen Rang  $k+1$ , d.h. es treten keine Kolinearitäten zwischen Spalten auf, so ist  $\mathbf{X}^\top \mathbf{X}$  regulär. Somit ergibt

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \quad (5.22)$$

die eindeutige Lösung für den GKQ-Schätzer von  $\beta$  zu Stichprobenrealisierungen  $y$  von  $Y$ . Die Hesse-Matrix

$$H_{MQF}(\beta) = \nabla(-2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\beta) = 2\mathbf{X}^\top \mathbf{X}$$

liefert durch positive Definitheit der Matrix  $\mathbf{X}^\top \mathbf{X}$  für reguläre  $X$  das Optimalitätskriterium zweiter Ordnung des Minimierers  $\hat{\beta}$ . (BEWEIS?-Für Matrix mit vollem Rang folgt positive Definitheit der quadratischen Form) Der Stichprobenumfang  $n$  muss hierbei deutlich über der Anzahl Parameter  $k+1$  liegen. Eine Verzerrung (engl. bias) des Schätzers  $\hat{\beta}$

$$\mathbb{E}[\hat{\beta}] = \mathbb{E}[(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}] = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbb{E}[Y] = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X}\beta = \beta$$

liegt nicht vor. Diese Eigenschaft der Unverzerrtheit ist insbesondere unabhängig von  $\text{cov}(\varepsilon)$ . Es ist nur die korrekte Spezifikation des linearen Trendes  $X\beta$  über  $\mathbb{E}[\varepsilon] = 0$  notwendig.

Die GKQ-Methode ist nicht direkt erweiterbar auf  $\sigma^2$ . Statt dessen wird der Minimierer  $\hat{\beta}$  der MQF aus Gleichung (5.21) substituiert um die reduzierte Funktion

$$L_c(\sigma^2|y) \equiv L(\hat{\beta}, \sigma^2|y) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} (y - X\hat{\beta})^\top (y - X\hat{\beta})$$

abzuleiten. Diese wird als konzentrierte Likelihood-Funktion von  $\sigma^2$  bezeichnet. Maximieren dieser Funktion bezüglich Parameter  $\sigma^2$  liefert

$$\begin{aligned} 0 &\stackrel{!}{=} \frac{d}{d\sigma^2} L_c(\sigma^2|y) = -\frac{n}{2} \left( \frac{1}{\sigma^2} \right) + \frac{1}{2} \left( \frac{1}{(\sigma^2)^2} \right) (y - X\hat{\beta})^\top (y - X\hat{\beta}) \\ &= -n + \frac{1}{\sigma^2} (y - X\hat{\beta})^\top (y - X\hat{\beta}) \end{aligned}$$

und schließlich den exakten ML-Schätzer

$$\hat{\sigma}^2 = \frac{1}{n} (y - X\hat{\beta})^\top (y - X\hat{\beta})$$

Der Maximierer  $\hat{\sigma}^2$  muss zudem beim Einsetzen in die zweite Ableitung von  $L_c$  negative Werte liefern. Der ML-Schätzer von  $\sigma^2 = \text{var}(\varepsilon) = \mathbb{E}[\varepsilon^2]$  lässt sich zudem über die Residuen ausdrücken

$$\hat{\sigma}^2 = \frac{1}{n} \hat{\varepsilon}^\top \hat{\varepsilon} = \frac{1}{n} \sum_{i=1}^n \hat{\varepsilon}_i^2$$

### 5.4.2 ML-Schätzung im verallgemeinerten Regressionsmodell

Im Anschluss erfolgt die Verallgemeinerung auf das VLN-Modell mit normalverteilten Störgrößen und resultiert somit in der verallgemeinerten multivariaten Normalverteilung der Zielgröße  $Y \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2 \mathbf{V})$ . Wir nutzen

$$\bullet \frac{1}{\sqrt{\det(\sigma^2 \mathbf{V})}} = \frac{1}{\sqrt{(\sigma^2)^n \det \mathbf{V}}} \quad \text{sowie} \quad \bullet (\sigma^2 \mathbf{V})^{-1} = \frac{\mathbf{V}^{-1}}{\sigma^2}$$

und erstellen erneut für gegebene Stichprobe  $\mathbf{y}$  und Parametervektor  $\theta = (\beta_0, \beta_1, \dots, \beta_k, \sigma^2)$  mit  $\Sigma = \sigma^2 \mathbf{V}$  in die Dichtefunktion

$$\begin{aligned} f(\mathbf{y}|\beta, \sigma^2) &= \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} \exp\left(-\frac{1}{2} (\mathbf{y} - \mathbf{X}\beta)^\top \Sigma^{-1} (\mathbf{y} - \mathbf{X}\beta)\right) \\ &= \frac{1}{\sqrt{(2\pi\sigma^2)^n \det \mathbf{V}}} \exp\left(-\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\beta)^\top \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\beta)\right) \end{aligned}$$

der Zielverteilung. Daraus folgt die log-Likelihoodfunktion der VKQ-Methode durch

$$L(\theta|\mathbf{y}) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2} \log(\det \mathbf{V}) - \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\beta)^\top \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\beta)$$

In Erweiterung zur klassischen Regression wird  $\hat{\beta}$  für beliebige  $\sigma^2$  die Funktion  $L(\beta, \sigma^2|\mathbf{y})$  maximieren, indem deren quadratischer Anteil  $(\mathbf{y} - \mathbf{X}\beta)^\top \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\beta)$  minimiert wird. Diese quadratische Form lässt sich als gewichtetes KQ-Problem interpretieren. Die Cholesky-Zerlegung für  $\mathbf{V}$  ist durch das Cholesky-Theorem gesichert und

$$\mathbf{V} = \mathbf{T}\mathbf{T}^\top \Rightarrow \mathbf{V}^{-1} = (\mathbf{T}^\top)^{-1} \mathbf{T}^{-1} = (\mathbf{T}^{-1})^\top \mathbf{T}^{-1}$$

erlaubt es, den quadratischen Term

$$\begin{aligned} (\mathbf{y} - \mathbf{X}\beta)^\top \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\beta) &= (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{T}^{-1})^\top \mathbf{T}^{-1} (\mathbf{y} - \mathbf{X}\beta) \\ &= (\mathbf{T}^{-1}\mathbf{y} - \mathbf{T}^{-1}\mathbf{X}\beta)^\top (\mathbf{T}^{-1}\mathbf{y} - \mathbf{T}^{-1}\mathbf{X}\beta) \\ &= (\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\beta)^\top (\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\beta) \end{aligned}$$

in eine reduzierte Form zu bringen. Diese entspricht der klassischen Version unter Transformationen  $\mathbf{T}^{-1}\mathbf{y}$  und  $\mathbf{T}^{-1}\mathbf{X}$ . Daraus ergibt sich der ML-Schätzer des KLN Modells

$$\begin{aligned} \hat{\beta} &= (\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^\top \tilde{\mathbf{y}} = [(\mathbf{T}^{-1}\mathbf{X})^\top (\mathbf{T}^{-1}\mathbf{X})]^{-1} (\mathbf{T}^{-1}\mathbf{X})^\top (\mathbf{T}^{-1}\mathbf{y}) \\ &= [\mathbf{X}^\top (\mathbf{T}^\top)^{-1} \mathbf{T}^{-1}\mathbf{X}]^{-1} \mathbf{X}^\top (\mathbf{T}^\top)^{-1} \mathbf{T}^{-1}\mathbf{y} = [\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X}]^{-1} \mathbf{X}^\top \mathbf{V}^{-1} \mathbf{y} \end{aligned}$$

unter Rücktransformation der Cholesky-Zerlegung. Für den ML-Schätzer  $\hat{\sigma}^2$  finden wir analog zur KLN

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n} (\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\hat{\beta})^\top (\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\hat{\beta}) = \frac{1}{n} (\mathbf{T}^{-1}\mathbf{y} - \mathbf{T}^{-1}\mathbf{X}\hat{\beta})^\top (\mathbf{T}^{-1}\mathbf{y} - \mathbf{T}^{-1}\mathbf{X}\hat{\beta}) \\ &= \frac{1}{n} (\mathbf{y} - \mathbf{X}\hat{\beta})^\top (\mathbf{T}^{-1})^\top (\mathbf{T}^{-1}) (\mathbf{y} - \mathbf{X}\hat{\beta}) \\ &= \frac{1}{n} (\mathbf{y} - \mathbf{X}\hat{\beta})^\top \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\hat{\beta})\end{aligned}$$

und interpretieren die ML-Schätzung aus VLN-Modellen als direkte Erweiterung aus der KLN-Modellierung.

### 5.4.3 ML-Schätzung für SE-Modelle

Bezüglich einer Gewichtsmatrix  $\mathbf{W}$  ist die Matrix  $\mathbf{B}_\rho = (\mathbb{I}_n - \rho\mathbf{W})^{-1}$  gegeben und definiert die räumliche Kovarianzstruktur  $\mathbf{V}_\rho = (\mathbf{B}_\rho^\top \mathbf{B}_\rho)^{-1} = \mathbf{B}_\rho^{-1} (\mathbf{B}_\rho^{-1})^\top$  aus Gleichung XX. Somit ist das SE-Modell ein Spezialfall der VLN-Modellform mit  $\mathbf{V}_\rho$  als Sonderform von  $\mathbf{V}$ . In  $\mathbf{V}_\rho$  ist im Gegensatz zu  $\mathbf{V}$  jetzt  $\rho$  als zusätzlicher Parameter enthalten. Der Parametervektor wird erweitert zu  $\theta = (\beta_0, \beta_1, \dots, \beta_k, \sigma^2, \rho)$  und mit

$$L(\theta|\mathbf{y}) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2} \log(\det \mathbf{V}_\rho) - \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\beta)^\top \mathbf{V}_\rho^{-1} (\mathbf{y} - \mathbf{X}\beta)$$

resultiert daraus die log-Likelihoodfunktion. Bedingt auf ein gegebenes  $\rho$  folgen erneut wie im VLN-Modell die ML-Schätzer

$$\hat{\beta}_\rho = (\mathbf{X}^\top \mathbf{V}_\rho^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{V}_\rho^{-1} \mathbf{y} \quad \text{sowie} \quad \hat{\sigma}_\rho^2 = \frac{1}{n} (\mathbf{y} - \mathbf{X}\hat{\beta}_\rho)^\top \mathbf{V}_\rho^{-1} (\mathbf{y} - \mathbf{X}\hat{\beta}_\rho)$$

mit Index  $\rho$  als Ausdruck der parametrischen Abhängigkeit. Die explizite Form der Schätzer als geschlossene darstellbare Funktion in Abhängigkeit von  $\rho$  wird in die Likelihoodfunktion eingesetzt um mit  $L_c(\rho|\mathbf{y}) \equiv L(\hat{\beta}_\rho, \hat{\sigma}_\rho^2, \rho|\mathbf{y})$  die *konzentrierte Log-Likelihoodfunktion*

$$L_c(\rho|\mathbf{y}) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\hat{\sigma}_\rho^2) - \frac{1}{2} \log(\det \mathbf{V}_\rho) - \frac{1}{2\hat{\sigma}_\rho^2} (\mathbf{y} - \mathbf{X}\hat{\beta}_\rho)^\top \mathbf{V}_\rho^{-1} (\mathbf{y} - \mathbf{X}\hat{\beta}_\rho)$$

für festes  $\rho$  zu erhalten, wie es zuvor in der KLN für  $\sigma^2$  angewendet wurde. Hiervon lässt sich der Schlussterm durch

$$-\frac{1}{2\hat{\sigma}_\rho^2} \left( \mathbf{y} - \mathbf{X}\hat{\beta}_\rho \right)^\top \mathbf{V}_\rho^{-1} \left( \mathbf{y} - \mathbf{X}\hat{\beta}_\rho \right) = -\frac{1}{2\hat{\sigma}_\rho^2} [n \hat{\sigma}_\rho^2] = -\frac{n}{2}$$

auf eine Konstante vereinfachen. Zur weiteren Reduktion wird unter Anwendung der Eigenschaften von Determinaten und Inversen die Form

$$\begin{aligned} \det \mathbf{V}_\rho &= \det \left( (\mathbf{B}_\rho^\top \mathbf{B}_\rho)^{-1} \right) = \det (\mathbf{B}_\rho^{-1}) \det \left( (\mathbf{B}_\rho^\top)^{-1} \right) \\ &= (\det \mathbf{B}_\rho)^{-1} (\det \mathbf{B}_\rho^\top)^{-1} = (\det \mathbf{B}_\rho)^{-2} \end{aligned}$$

ausgenutzt. Beide Vereinfachungen bringen uns schließlich mit

$$L_c(\rho|\mathbf{y}) = -\frac{n}{2} [1 + \log(2\pi) + \log(\hat{\sigma}_\rho^2)] + \log(\det \mathbf{B}_\rho)$$

die vereinfachte konzentrierte Log-Likelihoodfunktion für  $\rho$ . Da  $L_c(\rho|\mathbf{y})$  eine hinreichend glatte Funktion in einer Variablen darstellt, lässt sich diese Funktion mittels Liniensuchverfahren maximieren, um Schätzer  $\hat{\rho}$  zu ermitteln. Dieser wird dann in die Gleichung für  $\hat{\beta}_\rho$  und für  $\hat{\sigma}_\rho^2$  eingesetzt. Für große Stichprobengrößen  $n$  wird eine Spektralzerlegung zur Berechnung von  $\det \mathbf{B}_\rho$  genutzt. (Bivand et al., 2013, S. 300) Für zusätzliche Details wird auf (Cressie, 1993, S.440) verwiesen.

# Kapitel 6

## Analyse und Auswertung

Im folgenden werden *Gitterdaten* (engl. lattice data) analysiert, wie im Abschnitt 3.2 eingeführt. Die räumliche Information liegt diskret in Form einer Regionenvariable  $s$  vor.

### 6.1 Charakterisierung des Beobachtungsgebietes

Das Beobachtungsgebiet zeichnet sich durch große Heterogenität der Raumeinheiten bezüglich vieler Merkmale aus. Viele Gemeinden sind großräumig abgegrenzt und weisen einen Bruchteil der Einwohner aus den kleinflächigen, dicht besiedelten Berliner Bezirken auf. Die Raumeinteilung ist somit wenig einheitlich in der Fläche, welche selbst autokorreliert ist. Die dünn besiedelten ländlichen Gemeinden des Brandenburger Umlandes werden zunächst ignoriert.

Die Anzahl der Lagereimitarbeiter wird geschätzt und liegt aggregiert auf Gemeindeebene vor. Exakte Unternehmensstandorte oder Adressdaten sind nicht vorhanden.

Die Bevölkerung auf Ebene der Gemeinden wird durch statistische Schätzungen der Ämter jährlich aktualisiert fortgeschrieben und durch einen Zensus alle 10 Jahre überprüft und für die Zukunft angepasst. Der nächste Zensus ist für 2021 angesetzt. Distanzen zwischen Gemeinden werden über Zentroide gemessen.

### 6.1.1 Quantilkarten ausgesuchter Attribute

Wir definieren zunächst Nachbarschaften der Gemeinde-Polygone und nutzen die Queen-Definition: Angrenzende Polygone welche mind. 1 Knoten teilen (queen=TRUE) stehen in räuml. Beziehung

Als Teilgebiet wird die Metropolregion Berlin-Brandenburg betrachtet. Alle Zentroide im Umkreis von 13 km des eigenen Zentrums einer Region werden als Nachbarn deklariert. Jedem Nachbarschaftspolygon wird ein Gewicht zugewiesen. style="W" weist jedem Nachbarn identische Gewichte zu ( $1/\text{Anzahl Nachbarn}$ ). Reihenstandardisierung bewertet Beobachtungen mit wenigen Nachbarn stärker. Problematischerweise haben Randgebiete tendenziell weniger Nachbarn und ihre lag-Werte basieren auf weniger Polygonen. Sie sind somit anfälliger für Über- oder Unterschätzung der tatsächlichen Autokorrelation. Hierfür gibt es robustere Alternativen.

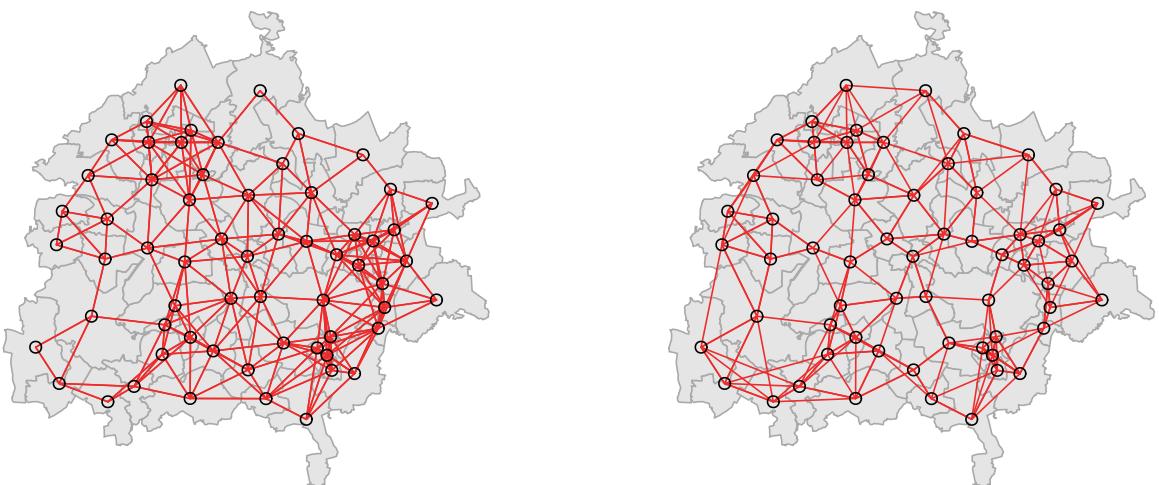


Abbildung 6.1: Nachbarschaftsrelationen: Li: 13km Radius. Re: 5 nächste Zentroide.

Diese Nachbarschaftsinformationen fließen direkt in die Nachbarschaftsmatrix  $W$  ein. Einige Pakete realisieren dies auch als Nachbarschaftsliste.

Als mögliche Attribute zur Realisierung der  $Y$ -Matrix stehen verschiedene Merkmale zur Verfügung, welche später auch in die Regression einfließen. Zur explorativen Analyse nutzen wir in Abbildung 6.2 das *tmap*-Paket, welches die automatische Einteilung per Quantil-Klassifikationsschema anbietet.

Zugleich sollen die Verteilungseigenschaften der *Arbeitnehmer per Einwohner* geprüft wer-

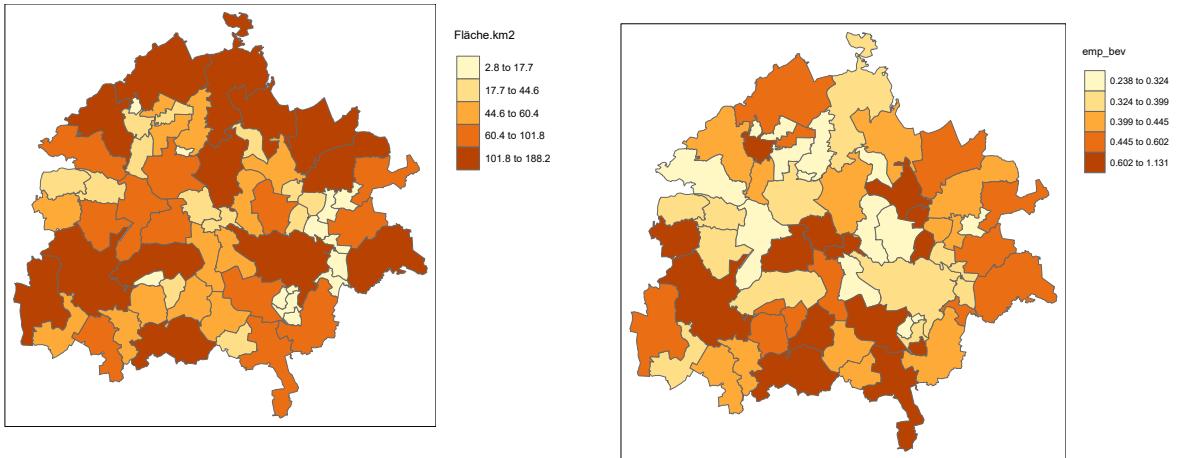


Abbildung 6.2: Quantilkarten der Metropolregion. Li: Fläche Re: Arbeitnehmer je Bev.

den.

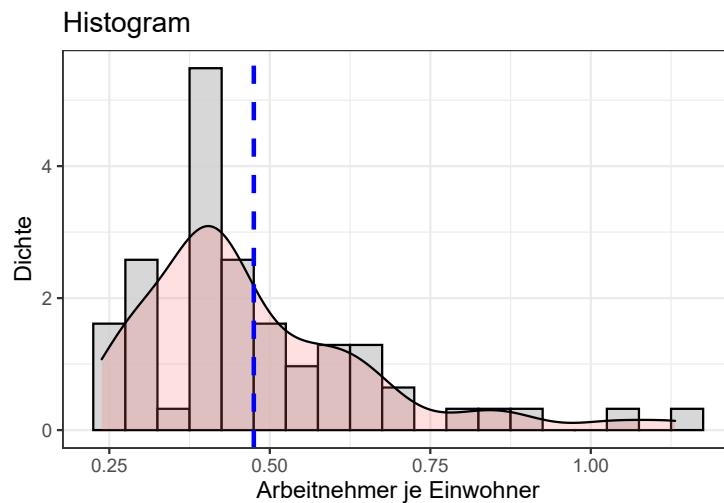


Abbildung 6.3: Histogramme, emp. Dichte und Mittelwert

Eine Normalverteilung scheint für dieses Attribut nicht gegeben. Insbesondere Ausreißer großer Werte verzerren die Verteilung. Dies betrifft etwa die beiden Gemeinden Schönefeld und Großbeeren, in denen nach Datenlage mehr Menschen arbeiten als leben. Hierfür müssen Transformation oder Ausreißerbereinigung durchgeführt werden

## 6.2 Globale Indizes

### 6.2.1 Globaler Moran Index

Der Standard Morantest unter der Normalverteilungsannahme entspricht dem Morantest der Regressionsresiduen eines Modells, welches lediglich mit einem Intercept ausgestattet wird. Dies zeigt, dass zusätzliche Kovariablen eingeführt werden können, um Missspezifikationen auszubügeln. Mit der gleichen Konstruktion kann zudem eine Sattelpunkt-Apprximation statt der analytischen Normalverteilungsannahme (Tiefeldorf 2002) sowie exakter Tests verwendet werden. Diese Methoden sind jedoch numerisch aufwendig, für moderate Datensätze wie im vorliegenden Fall jedoch noch praktikabel. Wie die Anwendung in Kapitel 4 zeigt, sind die Unterschiede für globale Tests gering, insbesondere falls Anzahl an Lokationen nicht klein ist. Für lokale Maße sind die Unterschiede zwischen Normalverteilungsannahme und Sattelpunktapproximation und exaktem test hingegen ausgeprägter (Bivand et al., 2013, S. 280).

```

1 > moran.test(DF$emp_beve, listw = lw, randomisation=TRUE)
2 # Output:
3 Moran I test under randomisation
4
5 data: DF$emp_beve      weights: lw
6
7 Moran I statistic standard deviate = 0.16349, p-value = 0.4351
8 alternative hypothesis: greater
9 sample estimates:
10 Moran I statistic      Expectation      Variance
11      -0.007611493     -0.016393443     0.002885485

```

Es gilt  $p\text{-value} > \alpha = 0.05$ . Die Nullhypothese kann nicht abgelehnt werden und es wird keine Autokorrelation erkannt. Die Sattelpunktapproximation liefert ähnliche Werte.

```

1 > lm.morantest.sad(lm(emp_beve ~ 1, SPDF), listw = lw)
2 # Output
3 Saddlepoint approximation for global Moran's I (Barndorff-Nielsen
  formula)

```

```

4
5 Saddlepoint approximation = 0.24664, p-value = 0.4026
6 alternative hypothesis: greater
7 sample estimates:
8 Observed Moran I
9 -0.007611493

```

Der exakte Test liefert:

```

1 > lm.morantest.exact(lm(emp_beve ~ 1, SPDF), listw = lw)
2
3 Exact standard deviate = 0.24433, p-value = 0.4035
4 alternative hypothesis: greater
5 sample estimates:
6 [1] -0.007611493

```

Der Geary-Test liefert:

```

1 > geary.test(SPDF$emp_beve, listw = lw, randomisation=TRUE)
2
3 Geary C test under randomisation
4
5 data: SPDF$emp_beve
6 weights: lw
7
8 Geary C statistic standard deviate = -0.1539, p-value = 0.5612
9 alternative hypothesis: Expectation greater than statistic
10 sample estimates:
11 Geary C statistic      Expectation      Variance
12          1.009477840        1.000000000       0.003792393

```

## 6.2.2 Monte-Carlo Simulation

```

1 > set.seed(1234)
2 > MC<- moran.mc(SPDF$emp_beve, listw = lw, nsim=999)
3 # Output
4 Monte-Carlo simulation of Moran I
5

```

```

6 data: SPDF$emp_bev
7 weights: lw
8 number of simulations + 1: 1000
9
10 statistic = -0.0076115, observed rank = 585, p-value = 0.415
11 alternative hypothesis: greater

```

Zur Abschätzung der Signifikanz randomisieren wir die Attributs-Werte über alle Gemeinden und implementieren keinerlei räumliche Autokorrelationsstruktur. Zu jeder Permutation wird ein Regressionsmodell angepasst und der Regressionsanstieg als Wert einer Moran I Statistik erfasst. Die Verteilung aller Werte gibt uns die erwartete Verteilung unter Annahme der Nullhypothese aus, dass Attributs-Werte zufällig über die Gemeinden verteilt sind. Letztlich wird der tatsächliche Moran I-Wert mit der simulierten Verteilung verglichen.

```

1 n <- 1599L    # Define the number of simulations
2 I_r <- vector(length=n)  # Create an empty vector
3
4 for (i in 1:n){
5   # Randomly shuffle income values
6   x <- sample(SPDF$emp_bev, replace=FALSE)
7   # Compute new set of lagged values
8   x.lag <- lag.listw(rwm, x)
9   # Compute the regression slope and store its value
10  M_r      <- lm(x.lag ~ x)
11  I_r[i] <- coef(M_r)[2]
12 }

```

## 6.3 Lokale Indizes

### 6.3.1 Lokaler Moran Index

Je stärker der lokale Moran Index vom Erwartungswert abweicht und je größer der Z-Score bzw. je kleiner der p-Value sind, desto unwahrscheinlicher ist die Annahme von

nicht autokorrelierten, zufallsverteilten Daten.

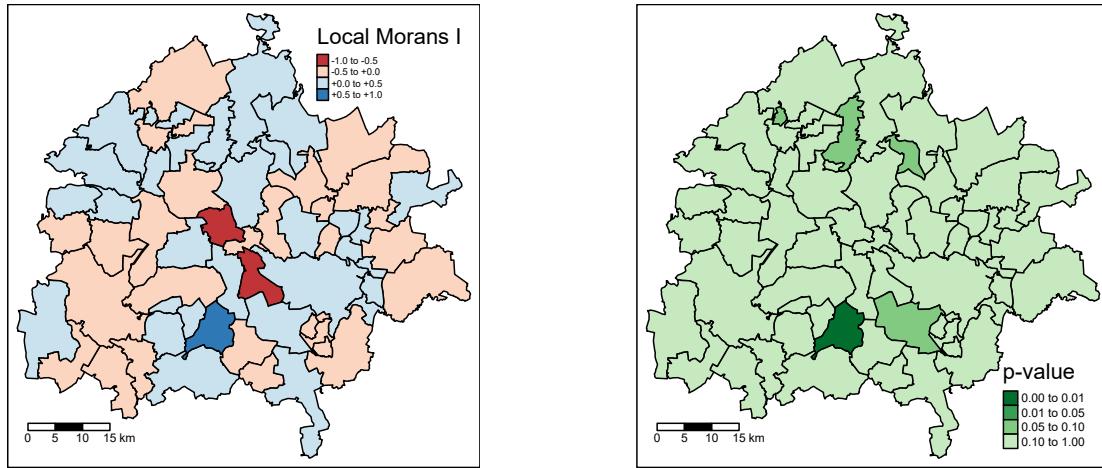


Abbildung 6.4: Lokaler Moran Index: Li: Art der Korrelation. Re: Signifikanz.

Zur detaillierten Untersuchung lokaler Autokorrelation eignet sich das Moran Scatterplot. Die Untersuchungsvariable wird auf der x-Achse (Abszisse) dargestellt und die räumlich gewichteten, mittleren Nachbarwerte (engl. spatially lagged values) mittels y- Achse (Ordinate). Der globale Moran Index wird als linearer Zusammenhang zwischen beiden Größen durch den Anstieg einer hervorgehobenen Geraden dargestellt. Der Plot wird zudem an Position der beiden Erwartungswerte in vier Quadranten zerlegt. Diese klassifizieren alle lokalen Beobachtungswerte entweder nach Clustern niedriger Werte (engl. Low-Low), nach Hot-Spots niedriger Werte(engl. Low-High), Hot-Spots hoher Werte (engl. High-Low) oder Clustern hoher Werte (engl. High-High).

# Kapitel 7

## Zusammenfassung und Wertung

Es wurden die Clustereffekte zunächst nur für Daten aus 2014 untersucht. Im nächsten Schritt kann eine räumliche Zeitreihenanalyse folgen, um Konzentrationsvorgänge bestimmter Wirtschaftszweige wie der Logistik dynamisch zu beurteilen.

Für detaillierte Analysen zum aktuellen Stand kann in Berlin auf die neue Einteilung nach LOR verfeinert werden, da Wirtschaft und Bevölkerung in den Berliner Bezirken vielfach stärker konzentriert ist als in Gemeinden. Zudem muss ein adäquates Regressionsmodell formuliert werden, nachdem aussagekräftige Kovariablen identifiziert wurden.

### 7.1 Anpassungen und Erweiterungen

Um die Aussagekraft der Ergebnisse zu verbessern und auf zusätzliche Bereiche zu erweitern, werden zusätzliche Maßnahmen zur weiteren Verarbeitung empfohlen. Zur Untersuchung der Autokorrelation wurden mit Moran's I und Geary's c zwei klassische Maße verwendet. Neben ihrer weiten Verbreitung und dem hohen Bekanntheitsgrad profitiert der Anwender von den zahlreichen Untersuchungen der genauen Eigenschaften dieser Maße und ihrer statistischen Tests. Weitere Hilfsmittel wie der Autokorrelationstest von Kelejian-Robinson stehen zur Verfügung. Zur weiteren Untersuchung wird jedoch auch der Einsatz moderner Indizes empfohlen, um weitere Forschungsentwicklungen in die eigene Analyse zu integrieren. Wichtige Vertreter sind die G-Statistik von Getis und Ord (Anselin and Rey, 2010, Kapitel 10). Auch die H-Statistik bietet weitere Vorteile.

# Literaturverzeichnis

Anselin, L.

1995. Local Indicators of Spatial Association—LISA. *Geographical Analysis*, 27(2):93–115. Publisher: John Wiley & Sons, Ltd.

Anselin, L. and S. J. Rey, eds.

2010. *Perspectives on Spatial Data Analysis*, number 1430-9602 in Advances in Spatial Science, 1 edition. Springer-Verlag Berlin Heidelberg.

Bivand, R. S., E. Pebesma, and V. Gómez-Rubio

2013. *Applied Spatial Data Analysis with R*, volume 10 of *Use R!*, 2 edition. Springer-Verlag New York.

Cressie, N. A. C.

1993. *Statistics for Spatial Data*, 1 edition. New York: John Willy and Sons. Inc., New York.

Fischer, M. M. and A. Getis, eds.

2010. *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications*, 1 edition. Springer-Verlag Berlin Heidelberg.

Fischer, M. M. and P. Nijkamp, eds.

2014. *Handbook of Regional Science*, 1 edition. Berlin Heidelberg: Springer-Verlag.

Gani, J., ed.

1986. *The Craft of Probabilistic Modelling: A Collection of Personal Accounts*, volume 1 of *Applied Probability*, 1 edition. Springer-Verlag New York.

Geary, R. C.

1954. The Contiguity Ratio and Statistical Mapping. *The Incorporated Statistician*, 5(3):115–146. Publisher: [Royal Statistical Society, Wiley].

Gelfand, A. E., P. J. Diggle, M. Fuentes, and P. Guttorp, eds.

2010. *Handbook of Spatial Statistics*, Handbook of Spatial Statistics. CRC Press.

Goodchild, M. F.

1986. *Spatial autocorrelation*, volume 47. Geo Books.

Goodchild, M. F.

2004. The Validity and Usefulness of Laws in Geographic Information Science and Geography. *Annals of the Association of American Geographers*, 94(2):300–303. Publisher: Routledge.

Haining, R.

2003. *Spatial Data Analysis: Theory and Practice*. Cambridge: Cambridge University Press.

Monmonier, M.

2013. *Eins zu einer Million: die Tricks und Lügen der Kartographen*. Birkhäuser Basel.

Moran, P. A. P.

1950. Notes on Continuous Stochastic Phenomena. *Biometrika*, 37(1/2):17–23. Publisher: [Oxford University Press, Biometrika Trust].

Openshaw, S., P. J. Taylor, and N. Wrigley

1979. Statistical applications in the spatial sciences. *London: Edited by N. Wrigley. Pion*, Pp. 127–144.

Schabenberger, O. and C. A. Gotway

2005. *Statistical Methods for Spatial Data Analysis*. Chapman & Hall/CRC. Google-Books-ID: I20NDgAAQBAJ.

Sui, D. Z.

2004. Tobler's First Law of Geography: A Big Idea for a Small World? *Annals of the Association of American Geographers*, 94(2):269–277. Publisher: Routledge.

Tobler, W. R.

1970. A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, 46:234–240. Publisher: [Clark University, Wiley].

Tobler, W. R.

2004. On the First Law of Geography: A Reply. *Annals of the Association of American Geographers*, 94(2):304–310. Publisher: Routledge.

Waller, L. A. and C. A. Gotway

2004. *Applied Spatial Statistics for Public Health Data*, Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, New Jersey.

Whittle, P.

1954. On stationary processes in the plane. *Biometrika*, 41(3-4):434–449.