Spatio-temporale Aspekte des Screening-Programms für Kolonkarzinome in Bayern

Master-Thesis

im Studiengang Biostatistik

Anna Rieger

Gauting, Januar 2011

Betreuung:

Dr. S. Greven
Prof. Dr. T. Hothorn

Institut für Statistik
Ludwig-Maximilians-Universität München

Zusammenfassung

Für die vorliegende Arbeit wurden Screening-Daten der Kassenärztlichen Vereinigung Bayerns ausgewertet. Es handelt sich um Daten zu Vorsorge-Untersuchungen gegen das Kolonkarzinom (Darmkrebs) aus den Jahren 2006 bis 2008. Das Ziel war es herauszufinden, wie die Nutzungsraten des Screenings in Bayern verteilt sind, d. h. ob regionale Unterschiede vorliegen. Dann sollten mögliche Einflussgrößen herausgefunden werden sowie der Zeitverlauf hinsichtlich der Stabilität der Nutzung solcher Vorsorge-Angebote modelliert werden.

Da eine regionale Verteilung untersucht werden sollte, wurden Modelle aus der räumlichen Statistik an die Daten angepasst. Zur Auswertung lagen zwei Arten von Kovariablen vor: Zum einen waren die individuellen Angaben zu den koloskopierten Personen wie Alter und Geschlecht im Datensatz enthalten. Des Weiteren lagen das Jahr der Untersuchung und der Wohnort der Klienten vor. Zum andern wurden verschiedene Sozialindizes der Beobachtungsregionen bzw. des Wohnortes zur räumlichen Analyse herangezogen. Dazu gehörte ein Deprivationsindex und die durchschnittliche Kaufkraft der Einwohner sowie die Religionsverteilung und die Gastroenterologendichte.

Die Erfassung der Wohnorte auf Basis der ersten drei Stellen der Postleitzahl stellte ein wesentliches Problem der Auswertung dar. Nachdem Nutzungsraten untersucht werden sollten, ist es nötig zu wissen, wie viele Personen das Angebot hätten nutzen können. Zu diesem Zweck müssen die zugrunde liegenden Bevölkerungsdaten bekannt sein. Allerdings gibt es für Postleitzahlbezirke keine Daten des statistischen Landesamtes. Die dort verfügbaren kostenlosen Daten der Gemeinden wurden zu 36 Gebieten zusammengefasst, so dass eine Zuordnung der Gemeinden und der Postleitzahlen zu einem Gebiet möglich war. Außerdem wurden Daten der Postleitzahlbezirke käuflich erworben und mit diesem Datensatz eine zweite Analyse durchgeführt.

Die Nutzungsraten selbst zeigten bei beiden Auswertungen eine leichte Schwankung über die Jahre. Abhängig vom Alter verhielten sich die Geschlechter in der Nutzung von Vorsorge-Untersuchungen unterschiedlich: Während die Nutzungsraten der Männer mit dem Alter stiegen, nahmen die anfangs hohen Raten der Frauen im Altersverlauf ab. Insgesamt zeigte sich die Nutzung der Darmkrebs-Vorsorge zwischen den Geschlechtern relativ ausgeglichen.
Im Osten und Südwesten wurden vergleichsweise niedrige Raten beobachtet, während sich vom Nordwesten nach Süden ein Band mit höheren Raten erstreckt.

Beide Datensätze der Amts- und der gekauften Daten wurden jeweils mit drei Modellarten ausgewertet: Neben einem Modell mit räumlich unstrukturiertem Effekt wurden zwei Modelle mit räumlich strukturiertem Effekt berechnet. Die Modelle mit unstrukturiertem Effekt wurden als generalisiertes gemischtes lineares Modell formuliert. Die Modelle mit räumlich strukturiertem Effekt enthielten ein Gauß-Markov-Zufallsfeld. Einmal wurden sie mit der Methode der penalisierten Quasi-Likelihood angepasst und einmal wurde die bayesianische Sichtweise zur Schätzung der Parameter herangezogen. Somit ergaben sich insgesamt sechs verschiedene Modelle, um die Screening-Daten auszuwerten. Die Ergebnisse dieser Modelle wurden bezüglich der Schätzer für die Einflüsse der Kovariablen sowie bezüglich der räumlichen Effekte verglichen. Sowohl die Regressionskoeffizienten als auch die räumlichen Effekte führten zu sehr ähnlichen Ergebnissen.

Insgesamt ließ sich anhand der berechneten Modelle Folgendes feststellen: Männer hatten eine geringere Chance, das Screening-Angebt zu nutzen und verhielten sich über das Alter und die Zeit hinweg anders als Frauen. Wenn die Personen älter waren, gingen Männer eher zum Screening als Frauen. Ebenfalls eine kleinere Chance, zum Screening zu gehen, wurde für ältere Personen beiderlei Geschlechts festgestellt. Eine Stabilität über die Jahre konnte nicht gezeigt werden. Insgesamt bestätigte sich somit der Eindruck, den man aus den Nutzungsraten gewonnen hatte, in den Modellen.
Die Anzahl Gastroenterologen und die Religionsverteilung hatten keinen signifikanten Einfluss auf die Nutzung der Vorsorge-Untersuchungen. Des Weiteren spielte überraschenderweise die soziale Situation keine Rolle, ob sich eine Person einem Darmkrebs-Screening unterzog oder nicht.

Das Ost-West-Gefälle der Nutzungsraten blieb in den räumlichen Effekten erhalten. Die in die Modelle aufgenommenen Einflussgrößen erklären die regionalen Unterschiede noch nicht ausreichend.

Ein Vergleich der Nutzungsraten mit zwei weiteren Auswertungen zeigte sehr ähnliche Ergebnisse.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1. Einführung

1.1. Einleitung

Kolonkarzinom ist der Fachausdruck für Darmkrebs. Dieser ist ein häufiger Krebs in Bayern. Laut Pritzkuleit et al. (2009) ist er die zweithäufigste Ursache für einen Krebstod sowohl bei Männern als auch bei Frauen. Er tritt meist erst in höherem Lebensalter auf (Nnoaham et al., 2010).
Ein Screening-Programm kann ein wirksames Mittel zur Früherkennung von Krebsen bieten. Im Falle des Kolonkarzinoms ist die zugehörige Screening-Untersuchung eine Vorsorge-Koloskopie, laienhaft gesprochen eine Darmspiegelung. Seit 2002 bietet die Gesetzliche Krankenversicherung in Bayern für ihre Mitglieder ab 55 Jahren zwei Früherkennungskoloskopien im Zehn-Jahres-Abstand an (Mansmann et al., 2008). In ca. 30% der Koloskopien wird eine Krebsvorstufe entdeckt (Mansmann et al., 2008), die sich innerhalb von ca. 10 Jahren zu einem Krebs auswachsen kann. Falls eine Krebsvorstufe während des Screenings gefunden wird, entfernt man es sogleich. Es wäre demzufolge wichtig, das Angebot der Gesetzlichen Krankenversicherung zu nutzen, um einer Krebserkrankung und damit einem möglichen Krebstod vorzubeugen – zumal die Chancen auf Heilung relativ gut sind, falls der Krebs frühzeitig entdeckt wird (Prof. Mansmann, München).

Für die Jahre 2006 bis 2008 liegt eine internetbasierte elektronische Dokumentation der durchgeführten Vorsorge-Untersuchungen vor. Die Daten aus dem Jahr 2006 wurde bereits von Mansmann et al. (2008) analysiert. U. a. ging es in dieser Arbeit darum, die Qualität der Screeningkoloskopien zu sichern.
Die Jahre 2006 bis 2008 werteten Pritzkuleit et al. (2009) jeweils einzeln aus. Das Ziel dieser Untersuchung war, Gründe für regionale Unterschiede in der Anfälligkeit für Darmkrebs herauszufinden.

Die vorliegende Arbeit befasst sich dagegen mit der Nutzung des Vorsorge-Angebots. Auch hier sollten regionale Unterschiede herausgearbeitet werden und mögliche Einflussgrößen auf die Nutzungsrate untersucht werden. Zudem wurde analysiert, ob sich über die Zeit eine Veränderung in der Häufigkeit der Nutzung ergeben hat, d. h. ob die Nutzung über

die Zeit stabil war oder ob sich Schwankungen ergaben.

Die per Internet erfassten Screening-Daten von der Kassenärztlichen Vereinigung Bayern (KVB) bestehen aus 182492 Einzelbeobachtungen der Jahre 2006 bis 2008. Für jeden Klienten, der am Darmkrebs-Screening teilgenommen hat, ist das Jahr der Untersuchung, sein Alter in Jahren, sein Geschlecht und die ersten drei Stellen der Postleitzahl seines Wohnortes im Datensatz enthalten. Die Basis für die Analyse bilden verschiedene Datensätze mit möglichen Einflussgrößen. Diese umfassen die zugrundeliegenden Bevölkerungsdaten, d. h. wie viele Personen des betreffenden Alters und Geschlechts das Screening-Angebot nutzen *könnten* sowie regionale Variablen wie z. B. Messgrößen für die soziale Situation in der Region.

An diese Daten wurden Modelle mit den Methoden der räumlichen Statistik angepasst und somit eine räumliche Analyse der Nutzungsraten in Bayern durchgeführt.

Große Schwierigkeiten bereitete der Umstand, dass das Bayerische Landesamt für Statistik auf Basis der Gemeinden oder Landkreise arbeitet. Somit waren die möglichen Einflussgrößen kostenfrei nur auf diesen Ebenen zu erhalten. Aus diesem Grund musste ein Kompromiss geschlossen werden zwischen den Gemeinden und den Postleitzahl-Bezirken, denn diese beiden räumlichen Einheiten stimmen weder überein noch ist eines eine Untereinheit des anderen. So wurde Bayern in 36 Gebiete aufgeteilt, zu denen sowohl die Gemeinden als auch die Postleitzahlbezirke eindeutig zugeordnet werden können. Kleinräumige regionale Unterschiede können mit dieser Arbeitsweise allerdings nicht mehr erkannt werden. Um die Stabilität der Ergebnisse auf Basis dieser 36 Gebiete zu überprüfen, wurde ein Datensatz hinzugekauft, welcher die Bevölkerungsdaten in den PLZ-Bezirken und die durchschnittliche Kaufkraft je Einwohner als Messgröße für die soziale Lage beinhaltet. Auch diese Daten wurden als Grundlage für die Berechnung von Nutzungsraten und zur Anpassung von Modellen der räumlichen Statistik verwendet und ihre Ergebnisse mit denen der 36 Gebiete verglichen.

Die Auswertung der Daten und die Berechnung der Ergebnisse erfolgte mit den statistischen Programmpaketen R (R Development Core Team, 2009) und BayesX (Belitz et al., 2009) (siehe hierzu Kap. A).

1.2. Übersicht

In Kapitel 2 werden zunächst die verschiedenen Datensätze sowie das Vorgehen in Bezug auf Datensatz-Bereinigung und -Bearbeitung vorgestellt. In Kapitel 2.2.1 wird der Algorithmus erklärt, nach dem die Gemeinden Bayerns und die dreistelligen PLZ-Bezirke zu

den 36 Gebieten zusammengefasst wurden. Des Weiteren wird in Kapitel 2.5 berichtet, wie die Karten erstellt wurden, damit die Ergebnisse der räumlichen Analyse visualisiert werden können.

In Kapitel 3 wird die der Analyse zugrunde liegende Theorie erläutert. Zunächst werden in Kapitel 3.1 die Möglichkeiten zur Verallgemeinerung von linearen Modellen aufgezeigt und genauer auf das Logit-Modell (Kap. 3.1.2) eingegangen. Diese Modellart ist notwendig, da die Responsevariable nicht normal-, sondern binomialverteilt ist. Eine Modellart, welches die Daten der benachbarten Regionen nicht berücksichtigt, ist das gemischte Modell. Die Theorie der gemischten Modelle wird in Kapitel 3.2 erklärt. Durch die Aufnahme eines strukturierten räumlichen Effekts ins Modell wird die Rolle der Nachbarschaftsstruktur geschätzt. Die entsprechende Theorie findet sich in Kapitel 3.3. Es gibt mehrere Methoden, einen strukturierten räumlichen Effekt anzupassen. In der vorliegenden Arbeit wird die Methode der penalisierten Quasi-Likelihood (Kap. 3.3.3.2) und die bayesianische Sichtweise (Kap. 3.3.3.3) verwendet.

In Kapitel 4 werden die Ergebnisse dargestellt und die Schätzer aus den Modellen interpretiert. Außerdem wird in Kapitel 5.1 ein Vergleich mit den oben erwähnten Auswertungen von Mansmann et al. (2008) und Pritzkuleit et al. (2009) gezogen. Auf spezielle Probleme der Datensätze und bei der Auswertung wird in Kapitel 5.2 näher eingegangen. Schließlich gibt Kapitel 5.3 eine Kurzfassung der gewonnenen Erkenntnisse wieder.

2. Datengrundlage

Die Datengrundlage für die Analyse der Nutzungsraten von Vorsorge-Koloskopien bilden verschiedene Datensätze aus mehreren Quellen. Diese Rohdaten wurden u. U. erst aggregiert und anschließend so zusammengefasst, dass die dabei neu entstandenen Datensätze eine Analyse der Screening-Daten ermöglicht haben.

Da bei der Erhebung der Screening-Daten nur die ersten drei Stellen der Postleitzahl erfragt wurden, ergaben sich einige Probleme bei der Auswertung. Denn diese stimmen weder mit den Gemeinden Bayerns überein noch bilden sie eine Übermenge derselben. Außerdem sind die Gemeinden in Bayern kleinräumiger als die dreistelligen Postleitzahlbezirke, so dass umgekehrt die Gemeinden auch keine Übermenge der dreistelligen Postleitzahlbezirke sind. Aus diesem Grund musste eine Zusammenfassung von Gemeinden und Postleitzahlbezirken zu eindeutigen Gebieten erfolgen, denn die Daten auf Basis der Gemeinden waren kostenlos verfügbar. Eine zweite Analyse zur Untersuchung der Sensitivität des Ergebnisses basierte auf kommerziell erworbenen Datensätzen, welche Informationen über die dreistelligen Postleitzahlbezirke lieferten.

2.1. Liste der Postleitzahlen mit Gemeindeschlüssel

Als Verbindung der Screening-Daten mit den auf Gemeinde-Basis erhobenen Daten hat die Bayerische Vermessungsverwaltung eine Liste zur Verfügung gestellt, welche die Gemeindeschlüssel in Bayern und deren zugehörige Postleitzahlen (PLZ) enthält.

Der amtliche Gemeindeschlüssel, auch “Gemeindekennziffer” genannt, identifiziert in acht Ziffern jede Gemeinde in Deutschland. Die ersten beiden Ziffern stehen dabei für das Bundesland. Bayern wird durch “09” gekennzeichnet. Die folgenden drei Ziffern bezeichnen den Landkreis, zu dem die Gemeinde gehört. Die letzten drei Ziffern identifizieren die Gemeinde selbst. Die Kennziffern von Landkreisen und kreisfreien Städten enden in der achtstelligen Variante des Gemeindeschlüssels deswegen auf “000”. Die Gemeinde 09671111 liegt also in Bayern und gehört zum Landkreis Nr. 671 (bzw. 09671000 in der achtstelligen Variante).

Dabei ist zu beachten, dass die 25 kreisfreien Städte als je eine einzige Gemeinde gelten. Landkreise dagegen sind eine Zusammenfassung mehrerer Gemeinden. Sie haben auch keine eigene Postleitzahl, im Gegensatz zu den kreisfreien Städten. Diese haben i. d. R. sogar so viele Postleitzahlen, dass sie sich bisweilen noch in der dritten Stelle unterscheiden.

Da die führende 09 allen Gemeinden dieser Liste gemeinsam ist, wurde der Gemeindeschlüssel für eine bessere Übersichtlichkeit auf sechs Stellen gekürzt.
Des Weiteren wurden die Postleitzahlen in der Liste der Bayerischen Vermessungsverwaltung auf die ersten drei Stellen verkürzt, da nur diese Information von Bedeutung für die folgende Analyse war. Doch dadurch kommen einige Zeilen mehrfach vor, da sich nun z. B. die Zeilen

PLZ5	Gemeindeschlüssel
63739	661000
63741	661000
63743	661000

nicht mehr unterscheiden, weil sie durch die Verkürzung der fünfstelligen Postleitzahl auf drei Stellen jeweils

PLZ3	Gemeindeschlüssel
637	661000

ergaben. Aus diesem Grund konnte man die Liste verkleinern, indem man die mehrfach vorhandenen Zeilen auf einen Eintrag reduziert.

2.2. Zusammensetzung der Beobachtungseinheiten

2.2.1. Die 36 Gebiete

Wie erwähnt, sind die dreistelligen PLZ-Bezirke weder eine Über- noch eine Untermenge der Gemeinden Bayerns. Diese problematische Eigenschaft gilt auch für die ersten zwei Stellen der Postleitzahlen sowie für vier oder fünf Stellen. Nur die erste Ziffer zu berücksichtigen, ist ebenfalls nicht zielführend. Aus diesem Grund ist für die weitere Analyse eine Zusammenfassung der Gemeinden und der dreistelligen PLZ-Bezirke notwendig, damit die Daten zu eindeutigen räumlichen Einheiten gehören.

Eine Zusammenfassung zu Gebieten sollte idealerweise die Eigenschaft haben, dass die Zuordnung sowohl der Gemeinden als auch der PLZ-Bezirke zu einem Gebiet eindeutig

ist. Dies wurde schrittweise erreicht mit der Liste, die von der Bayerischen Vermessungsverwaltung angefordert werden konnte (siehe Kap. 2.1). Zu Beginn wurde jeder dreistellige PLZ-Bezirk als ein Gebiet deklariert. Im nächsten Schritt wurden alle Gemeinden, die zu einem Gebiet gehören, in dieses Gebiet aufgenommen. Daraufhin wurden alle PLZ-Bezirke zu einem Gebiet zusammengefasst, welche zu den beteiligten Gemeinden gehören; dann wiederum wurden die angehörenden Gemeinden einem Gebiet zugeordnet. Es wurde also abwechselnd anhand der Postleitzahlen und der Gemeinden zusammengefasst. Diese Iterationen wurden so lange durchgeführt, bis sich keine Veränderung mehr in der Zuweisung ergeben hat. Auf diese Weise wurden aus den 2056 Gemeinden bzw. 117 Bezirken mit dreistelligen Postleitzahlen 36 Gebiete geformt.

Eine Liste der Gemeinden mit ihren dreistelligen Postleitzahlen und der zugehörigen Gebietsnummer findet sich in der Tabelle `gkz-plz-geb.txt` im Ordner "Daten" des elektronischen Anhangs (vgl. Kap. B).

Ein kleines Datenbeispiel soll die Vorgehensweise verdeutlichen: Angenommen, die Gemeinden in Tabelle 2.1 sind die einzigen Gemeinden in Bayern mit den Postleitzahlen, welche mit 832, 833 oder 834 beginnen. Im ersten Schritt wird jedem dieser *PLZ3-Bezirke* eine Gebietsnummer zugeordnet (Tab. 2.1, Iteration 1). In einem zweiten Schritt werden alle *Gemeinden*, die zum alten Gebiet Nr. 1 gehören, als neues Gebiet Nr. 1 deklariert. Die Gemeinden 189148, 189155, 189159 und 189160 gehören also zu Gebiet Nr. 1, denn sie alle haben die Postleitzahl 832xx. Jedoch gehört ein Teil der Gemeinde 189155 auch zum PLZ-Bezirk 833. Aus diesem Grund war in Zeile 9 zuvor Gebiet Nr. 2 eingetragen. Doch jetzt wird nach Gemeindeschlüsseln zusammengefasst und Gemeinde 189155 gehört somit komplett zu Gebiet 1. Das gleiche Verfahren führt dazu, dass die Gemeinden 172111, 172118, 172130, 189154 und 189157 dem Gebiet Nr. 2 zugeteilt werden. Somit ändert sich in Zeile 11 die Gebietszuordnung, denn die alte Zuordnung basierte auf den Postleitzahlen und nun wird anhand der Gemeindeschlüssel zusammengefasst.

Im nächsten Iterationschritt (Tab. 2.1, Iteration 2) werden alle an Gebiet 1 beteiligten PLZ-Bezirke einem Gebiet zugeordnet. Zum Gebiet 1 gehört neben der Postleitzahl 832xx auch die Postleitzahl 833xx, da Gemeinde 189155 beide Postleitzahlen führt. Somit umfasst das neue Gebiet 1 alle Zeilen, welche zu PLZ 832 und 833 gehören. Anschließend werden wieder alle beteiligten Gemeinden zusammengefasst und es ändert sich in Zeile 11 erneut die Gebietszuweisung. Gemeinde 172111 ist jetzt nämlich auch Gebiet 1 zuzuordnen.

Wird danach nochmals nach den dreistelligen PLZ-Bezirken zusammengefasst (Tab. 2.1, Iteration 3), so ergibt sich ein großes Gebiet, welches die PLZ-Bezirke 832, 833 und 834 umfasst sowie die Gemeinden 172111, 172112, 172118, 172130, 189148, 189154, 189155,

189157, 189159 und 189160. Diese Zuordnung ändert sich auch nicht mehr, wenn man die am Gebiet beteiligten Gemeindeschlüssel zusammen nimmt.

Tabelle 2.1.: Beispiel zur Zusammenfassung der 36 Gebiete. Veränderungen zur vorhergehenden Iteration sind fett gedruckt.

			Iteration 1		Iteration 2		Iteration 3	
Zeile	PLZ3	Gemeinde-schlüssel	(PLZ3) Gebiet	(Schl.) Gebiet	(PLZ3) Gebiet	(Schl.) Gebiet	(PLZ3) Gebiet	(Schl.) Gebiet
1	832	189148	1	1	1	1	1	1
2	832	189155	1	1	1	1	1	1
3	832	189159	1	1	1	1	1	1
4	832	189160	1	1	1	1	1	1
5	833	172111	2	2	**1**	1	1	1
6	833	172118	2	2	**1**	1	1	1
7	833	172130	2	2	**1**	1	1	1
8	833	189154	2	2	**1**	1	1	1
9	833	189155	2	**1**	1	1	1	1
10	833	189157	2	2	**1**	1	1	1
11	834	172111	3	**2**	2	**1**	1	1
12	834	172112	3	3	**2**	2	**1**	1

2.2.2. Bezirke der dreistelligen Postleitzahlen

Firmen vertreiben Daten auf der Ebene der dreistelligen oder oft auch der fünfstelligen Postleitzahlen. Erwirbt man einen solchen Datensatz, ergeben sich keine Zuordnungsprobleme wie bei Daten auf Gemeinde-Basis. Diese jedoch sind kostenlos bei den Bundes- und Landesämtern für Statistik verfügbar.

Bei der Analyse mit Hilfe der kommerziellen Daten war demnach keine besondere Zusammenfassung von Gemeinden o. Ä. notwendig. Die jeweiligen Beobachtungseinheiten bilden die Bezirke der dreistelligen Postleitzahlen.

2.3. Rohdaten

Nachdem im vorangegangenen Kapitel (Kap. 2.2) die Beobachtungseinheiten erklärt wurden, werden nun die verschiedenen Datensätze, ihre Quellen und ihre erste Bearbeitung erläutert. Zusätzlich werden sie kurz deskriptiv vorgestellt. Im nächsten Kapitel (Kap. 2.4) wird näher auf die Zusammensetzung der Datensätze zur letztendlich verwendeten Datengrundlage eingegangen.

2.3.1. Screening-Daten

Die "Screening-Daten" beinhalten die internetbasierte Dokumentation der Vorsorge-Koloskopien. Somit stellt dieser Datensatz die Informationen über die Zielgröße der vorliegenden Analyse.
Nachdem in der elektronischen Dokumentation auch andere Indikationen für eine Koloskopie (z. B. Krebsvorstufen-Nachsorge) erfasst wurden, musste der Datensatz verkleinert werden. Er enthielt dann statt knapp 800000 Beobachtungen nur mehr 185901 Zeilen. Für diese Klienten lag jeweils das Untersuchungsjahr, das Alter, das Geschlecht und der Wohnort in Form der ersten drei Stellen der Postleitzahl vor.
Mit Hilfe des angegebenen PLZ-Bezirks und der Liste über die Kombinationen aus Postleitzahlen und Gemeinden von der Vermessungsverwaltung (siehe Kap. 2.1) wurden die Screening-Daten um nicht-bayerische Klienten bereinigt, so dass am Ende 182492 Einzelbeobachtungen in die Analyse eingingen.

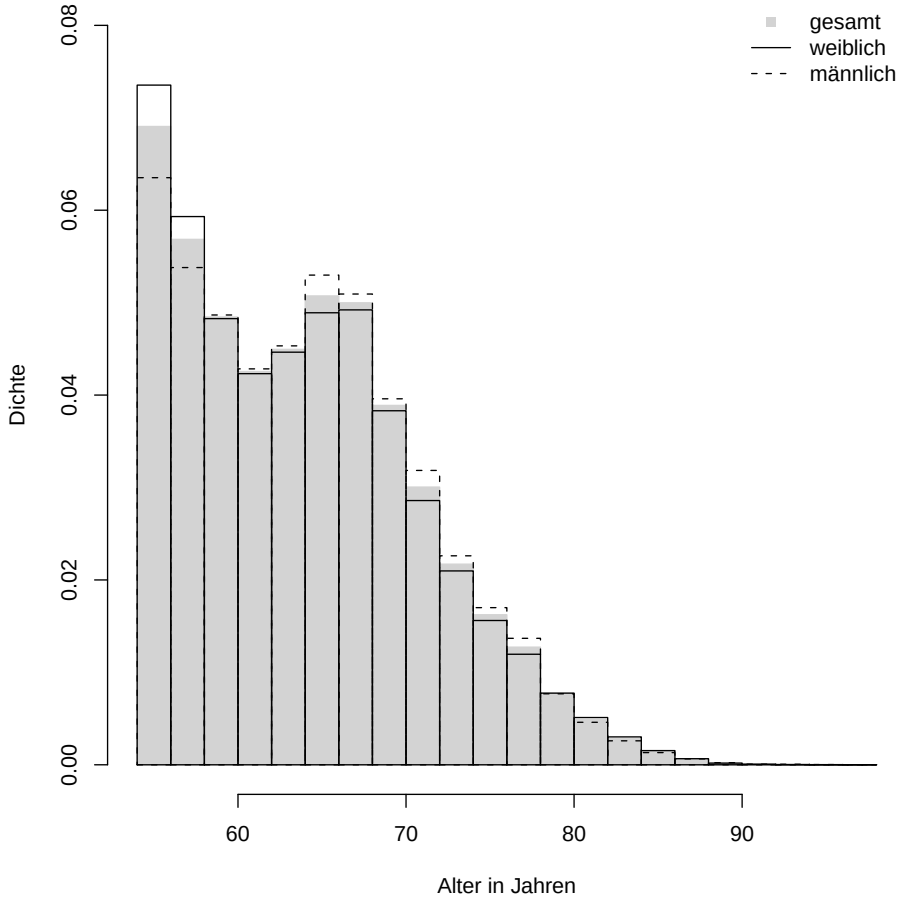

Abbildung 2.1.: Histogramm zur Verteilung des Alters in den Screening-Daten.

Betrachtet man nun die Verteilung der Nutzer im Datensatz über das Alter in einem Histogramm (Abb. 2.1), so zeigt sich ein erster Gipfel bei einem Alter von 55–56 Jahren. Ein zweiter, weniger ausgeprägter Gipfel ist im Alter von 64–65 Jahren zu beobachten. Danach fällt die Dichte der Nutzer mit dem Alter ab.
Bei getrennter Betrachtung von Männern und Frauen verschiebt sich die Dichte der Frauen mehr nach links. Bis zu einem Alter von etwa 59 Jahren liegt das Histogramm für die Frauen über demjenigen der Gesamtdaten. Die Verteilung der Männer neigt sich umgekehrt eher nach rechts und zeigt ab einem Alter von etwa 64 bis ca. 78 Jahren eine höhere Dichte als die Gesamtdaten.

2.3.1.1. Weitere Bearbeitung für die 36 Gebiete

Eine mögliche Einflussgröße auf die Nutzung des Koloskopie-Angebots ist das Alter der Klienten. Daher ist es für die Analyse notwendig, die Altersverteilung der Gesamtbevölkerung zu kennen. In den Daten über die Bevölkerung der Gemeinden Bayerns, welche beim Landesamt für Statistik und Datenverarbeitung zur Verfügung gestellt werden, ist das Alter in Klassen erhoben worden (siehe Kap. 2.3.2.1). Aus diesem Grund wurde die Altersvariable der Screening-Daten für die weiteren Analysen kategorisiert, so dass die Altersangaben in den Screening-Daten zu den Altersgruppen in den Daten der möglichen Einflussgrößen passten.
Dabei ist zu erwähnen, dass der Screening-Datensatz nur Beobachtungen mit einem Alter von mindestens 55 Jahren beinhaltet. Ab diesem Alter fördert die Gesetzliche Krankenversicherung die Krebsvorsorge durch zwei kostenlose Koloskopien im Zehn-Jahres-Abstand (Mansmann et al., 2008). In den Daten zur Altersverteilung in der Gesamtbevölkerung (auf Gemeinde-Basis) stehen aber nur die Altersgruppen “ab 50 bis unter 65”, “ab 65 bis unter 75” und “75 und älter” zur Verfügung. Aus diesem Grund musste dort ein Korrekturfaktor die Anzahlen in der Alterskategorie “ab 50 bis unter 65” an die Screening-Daten anpassen, da diese erst mit 55 Jahren beginnen. Die Vorgehensweise dazu wird an Ort und Stelle erklärt (siehe Kapitel 2.3.2.1). Pro Einzelbeobachtung der Screening-Daten wurde also festgehalten, ob der Klient zwischen 55 bis unter 65 Jahre, 65 bis unter 75 Jahre oder 75 Jahre und älter war. Somit passen die Altersangaben mit den u. U. korrigierten Anzahlen in den Daten der Gesamtbevölkerung zu den Altersklassen in den Screening-Daten.

Fasst man nun die Daten über die einzelnen Jahre hinweg zusammen, so zeigt sich folgende Verteilung von Nutzern (Tabelle 2.2): Über die Hälfte der betrachteten Personen ist jünger als 65 Jahre. Den geringsten Anteil (9.4%) stellt die Altersstufe der mindestens 75-Jährigen. Über die Altersgruppen hinweg sind mehr Frauen als Männer koloskopiert

worden, wobei der Frauenüberschuss mit zunehmendem Alter immer geringer wird. Knapp ein Drittel (29.7%) der Klienten sind Frauen bis 65 Jahre. Insgesamt sind etwa 55.3% der im Datensatz enthaltenen Personen weiblich und 44.7% männlich.

Tabelle 2.2.: Verteilung der Nutzer nach Altersklasse (dreistufig) und Geschlecht.

Geschlecht	< 65	< 75	≥ 75	Summe
weiblich	29.65%	20.57%	5.08%	55.31%
männlich	22.72%	17.70%	4.28%	44.69%
Summe	52.37%	38.27%	9.36%	100.00%

Wie in Kapitel 2.2.1 erläutert, müssen die Gemeinden und dreistelligen PLZ-Bezirke zu 36 Gebieten zusammengefasst werden. Auch die Screening-Daten wurden für die weitere Analyse entsprechend aggregiert. Für jeden PLZ-Bezirk ist die Gebietsnummer aufgrund der in Kapitel 2.2.1 beschriebenen Prozedur bekannt. Anhand dieser Information konnten die Anzahlen der Klienten für jedes Gebiet und jede Kombination aus Geschlecht, Altersklasse und Jahr zusammengefasst werden. In der vorliegenden Arbeit werden also die pro Gebiet aggregierten individuellen Daten analysiert.

Der Screening-Datensatz beinhaltet folgende Variablen:

- `gebiet`: die laufende Nummer des betrachteten Gebiets
- `maennlich`: 0 bezeichnet weibliche Klienten, 1 männliche.
- `altergrp`: eine Faktorvariable mit den Stufen
  - `5565`: Klienten ab 55 bis unter 65 Jahre
  - `6575`: Klienten ab 65 bis unter 75 Jahre
  - `7599`: Klienten ab 75 Jahre und älter
- `jahr`: das Untersuchungsjahr
- `Freq`: die Anzahl Klienten in der zugehörigen "Zelle" von Gebiet, Geschlecht, Altersstufe und Untersuchungsjahr

2.3.1.2. Weitere Bearbeitung für die Bezirke der dreistelligen Postleitzahlen

Der zweite Teil der Analyse basiert auf kommerziellen Daten, welche auf Ebene der dreistelligen PLZ erhoben wurden. Auch hier ist die Altersverteilung in der Gesamtbevölkerung nur in Kategorien verfügbar. Genauer liegen Daten in den Altersgruppen "55 bis unter 60", "60 bis unter 65", "65 bis unter 75" und "75 und älter" vor (siehe Kap. 2.3.2.1). Auch für diesen Teil der Auswertung wurde für jede Einzelbeobachtung der Screening-Daten

festgestellt, in welche Alterskategorie der Klient fällt. Durch diese Maßnahme passen die Altersangaben in den Screening-Daten zu den Daten der Gesamtbevölkerung (auf Basis der dreistelligen PLZ).

Diese Zusammenfassung ist in Tabelle 2.3 näher beschrieben. Die Anteile von Frauen und Männern (55.3% bzw. 44.7%) bleibt naturgemäß erhalten im Vergleich zu Tabelle 2.2. Der Bereich ab 65 Jahre hat sich durch diese Kategorisierung des Alters ebenfalls nicht geändert. In den neuen Altersgruppen "55 bis unter 60" und "60 bis unter 65" erkennt man, dass fast ein Drittel (30.3%) der koloskopierten Personen jünger als 60 Jahre ist. Auch in dieser Altersklasse sind – wie in allen anderen Klassen – mehr Frauen als Männer untersucht worden. Der Unterschied zwischen den Anzahlen von Frauen und Männern in dieser jüngsten Altersgruppe macht ca. 5% des gesamten Datensatzes aus. Nach einem Sinken der Anzahl der koloskopierten Personen zwischen 60 und 65 steigen die Zahlen sowohl bei Männern als auch bei Frauen in der nächsten Altersklasse wieder an. Bei beiden Geschlechtern liegt eine Steigerung um dem Faktor von ca. 1.7 vor.

Tabelle 2.3.: Verteilung der Nutzer nach Altersklasse (vierstufig) und Geschlecht.

Geschlecht	< 60	< 65	< 75	≥ 75	Summe
weiblich	17.52%	12.13%	20.57%	5.08%	55.31%
männlich	12.75%	9.97%	17.70%	4.28%	44.69%
Summe	30.26%	22.10%	38.27%	9.36%	100.00%

Da die Screening-Daten mit dem Altersklassen der Daten für die PLZ-Bezirke zum aktuellen Zeitpunkt der Bearbeitung noch Individualdaten sind, muss dieser Datensatz ebenfalls aggregiert werden: Für die Analyse mit den käuflich erworbenen Daten werden die Einzelbeobachtungen pro dreistelligem PLZ-Bezirk zu Anzahlen zusammengefasst.

Der Screening-Datensatz für die kommerziellen Daten beinhaltet folgende Variablen:

- `plz`: die ersten drei Stellen der Postleitzahl, welche zugleich die Identifikationsnummer der räumlichen Einheit ist
- `maennlich`: 0 bezeichnet weibliche Klienten, 1 männliche.
- `altergrp`: eine Faktorvariable mit den Stufen
  - `5559`: Klienten ab 55 bis unter 60 Jahre
  - `6064`: Klienten ab 60 bis unter 65 Jahre
  - `6574`: Klienten ab 65 bis unter 75 Jahre
  - `7599`: Klienten ab 75 Jahre und älter
- `jahr`: das Untersuchungsjahr

- `Freq`: die Anzahl Klienten in der zugehörigen "Zelle" von Gebiet, Geschlecht, Altersstufe und Untersuchungsjahr

2.3.2. Datensätze zu den Einflussgrößen

Im Kapitel 2.3.1 wurde die Zielgröße, die Screening-Daten, näher beschrieben. Auf das Verhalten der Menschen bei der Darmkrebs-Vorsorge haben verschiedene Größen Einfluss. Zum einen gibt es die individuellen Variablen wie Alter und Geschlecht der betreffenden Person. Zum andern liegt eine räumliche Information in der Zielgröße vor. Man kann also durch Aufnahme regionaler Variablen einen räumlichen Effekt untersuchen. Diese Variablen sind z. B. die Dichte an Gastroenterologen, also die Anzahl Ärzte je 100000 Einwohner am Wohnort der betreffenden Person. Hinzu kommt das durchschnittliche soziale Niveau bzw. die Kaufkraft der Einwohner im Gebiet. Diese Variablen sollen die Tatsache beschreiben, dass nur die gesetzlich Versicherten im Screening-Datensatz enthalten sind. Da man sich erst ab einem gewissen sozialen bzw. finanziellen Niveau privat versichert, dienen das durchschnittliche soziale Niveau und die Kaufkraft dazu, den Anteil privat versicherter und somit nicht erfasster Klienten im untersuchten Gebiet zu schätzen. Außerdem kann in den 36 Gebieten die Religionsverteilung (aus verschiedenen Gründen der Anteil christlicher Einwohner, siehe Kap. 2.3.2.2) ins Modell aufgenommen werden. In diesem Kapitel werden die Datensätze zu den möglichen Einflussgrößen vorgestellt.

2.3.2.1. Bevölkerungszahlen zu Alter und Geschlecht

Vom Bayerischen Landesamt für Statistik und Datenverarbeitung In der frei zugänglichen GENESIS-Datenbank des Bayerischen Landesamtes für Statistik und Datenverarbeitung ist der hier verwendete Datensatz nach einer Registrierung verfügbar. Er zeigt für jede der 2056 Gemeinden Bayerns die Bevölkerungszahlen in 13 Altersklassen. Das Alter ist dabei in Kategorien von "0 bis unter 3 Jahre" bis hin zu "50 bis unter 65 Jahre", "65 bis unter 75 Jahre" und "75 Jahre oder mehr" unterteilt. Der Stichtag für die Datenerhebung ist jeweils der 31. Dezember eines Jahres.
In jeder Altersklasse sind die gesamten Bevölkerungszahlen sowie die Anzahl von weiblichen Bewohnern verfügbar. Durch diese Angaben kann man auch die Anzahl der Männer in der Gemeinde und der gewünschten Altersklasse berechnen.
In der GENESIS-Datenbank kann auf alle drei für die Analyse benötigten Jahre, also 2006 bis 2008, zugegriffen werden. Diese drei Datensätze werden im Folgenden "*Gemeinde-Daten*" genannt.

Nachdem man die Datei für jedes Jahr im MS-Excel-Format herunter geladen hat, wurden die nicht benötigten Daten der Verwaltungsgemeinschaften, Regionen, Regierungsbezirke und Gesamtbayern entfernt. Des Weiteren wurde eine neue Variable `schl` eingefügt, welche den vorhandenen Regionalschlüssel `RSCHL` auf sechs Stellen erweitert: Die Schlüssel der Landkreise und kreisfreien Städte werden von rechts mit Nullen aufgefüllt. Mit Hilfe dieser Angaben können die Gemeinden eindeutig identifiziert werden. Die Variable `schl` dient somit als eine Art ID. Die auf diese Weise bearbeiteten Datensätze wurden jeweils als Tabstop-getrennte Text-Datei gespeichert.

Diese `txt`-Dateien konnten nun in R eingelesen werden. Allerdings musste weitere Datenbereinigung betrieben werden. Zum einen gab es im Datensatz von 2006 eine leere Spalte, zum andern waren in jedem Jahres-Datensatz leere Zeilen vorhanden. Diese wurden gelöscht. Anschließend wurden alle Spalten, deren Altersklassen nicht näher von Interesse waren, entfernt, um die Datensätze schlanker zu gestalten.
Außerdem sind in den ursprünglichen Gemeinde-Daten aus der GENESIS-Datenbank noch Informationen über die Landkreise enthalten. Anhand des Regionalschlüssels oder der Variable `schl` lassen sich die Landkreise nicht von den kreisfreien Städten unterscheiden. Dies ist jedoch wichtig, da die kreisfreien Städte als Gemeinde gelten, die Landkreise jedoch Zusammenfassungen der zugehörigen Gemeinden sind. Somit ist es etwas aufwendig, die Zeilen der Landkreise bereits im Excel-Format zu löschen: Entweder muss man den gesamten Datensatz manuell durchgehen und alle Zeilen mit nur dreistelligem `RSCHL` löschen, denen die Zeilen von zugehörigen Gemeinden vorstehen. Dies kennzeichnet nämlich die Landkreise. Den Zeilen der kreisfreien Städte dagegen geht keine Auflistung von Gemeinden voraus. Anhand des Regionalschlüssels `RSCHL` zu sortieren, ist keine praktikable Alternative. Denn in diesem Fall ist wieder nicht erkennbar, ob es sich um aggregierte Landkreis-Daten oder um eine kreisfreie Stadt handelt. Einfacher ist es deswegen, diese Zeilen in R zu löschen: Die Zeilen, welche zusammengefasste Daten eines Landkreises darstellen, haben einen (ursprünglich nur dreistelligen) Regionalschlüssel, der nicht in der Liste über die Kombinationen aus Postleitzahlen und Gemeindeschlüssel von der Vermessungsverwaltung (Kap. 2.1) enthalten ist. Denn Landkreise haben keine eigene Postleitzahl wie die kreisfreien Städte, sondern die dem Landkreis untergeordneten Gemeinden sind mit ihren Postleitzahlen in der Liste enthalten. Durch die Überprüfung, ob der betrachtete Schlüssel in der Liste steht, kann man die Landkreise klar von den kreisfreien Städten trennen. Nachdem die Landkreis-Zeilen ebenfalls gelöscht wurden, hat der Datensatz noch 2056 Zeilen, was den 2056 politisch selbstständigen Gemeinden Bayerns entspricht.

In den so weit bearbeiteten Gemeinde-Daten sind die Kategorien der Zielgröße voll abgedeckt, bis auf die Grenze in den Altersklassen. Die Screening-Daten beginnen erst ab einem

Alter von 55 Jahren. Aus diesem Grund mussten die Anteile der 50- bis unter 55-Jährigen geschätzt werden.

Diese Schätzung wurde anhand eines Datensatzes durchgeführt, der eine feinere Altersstruktur aufweist. Dies geht jedoch zu Lasten der Anzahl an Erhebungsgebieten, denn nur auf Ebene der Landkreise sind in der GENESIS-Datenbank Daten mit 17 Altersklassen von "0 bis unter 3 Jahre" bis "18 bis unter 20 Jahre" und ab einem Alter von 20 Jahren in Fünf-Jahres-Schritten bis zu "50 bis unter 55 Jahre", "55 bis unter 60 Jahre", "60 bis unter 65 Jahre", "65 bis unter 75 Jahre" und '75 Jahre oder mehr" verfügbar. Pro Landkreis sind in jeder Alterskategorie die Anzahlen der männlichen und der weiblichen Bewohner sowie eine Gesamtzahl aufgelistet. Ebenso wie bei den Gemeinde-Daten wurde dieser Datensatz, im Folgenden "*Landkreis-Daten*" genannt, zum Stichtag 31. Dezember des jeweiligen Jahres erhoben. Auch für die Landkreis-Daten kann man in der GENESIS-Datenbank wie für die Gemeinde-Daten alle drei benötigten Jahre anfordern.

Die Landkreis-Daten mussten ebenfalls etwas bereinigt werden, nachdem sie aus MS-Excel in eine Text-Datei exportiert und in R eingelesen wurden. Auch diese Datensätze besaßen nämlich leere Zeilen. Außerdem waren die Daten in einer unpraktischen Struktur abgespeichert: Für jeden Landkreis waren im Datensatz drei Zeilen enthalten – je eine Zeile für Männer, für Frauen und für die Gesamtzahlen. Diese Struktur wurde dahingehend geändert, dass nur noch eine Zeile vorlag, die diese Informationen enthielt. Außerdem wurden nicht benötigte Informationen über die Regierungsbezirke und Gesamtbayern entfernt.

Nach einer Verkleinerung des Datensatzes auf die relevanten Altersklassen wurden pro Landkreis die Anzahlen der 50 bis unter 65-jährigen Personen berechnet. Im ursprünglichen Datensatz liegen nämlich die Altersgruppen "50 bis unter 55", "55 bis unter 60" und "60 bis unter 65" einzeln vor und mussten addiert werden.

In den Landkreis-Daten wurde nun pro Landkreis und für jedes Jahr getrennt der Anteil der 50- bis unter 55-Jährigen in der neu geschaffenen Altersklasse der "50- bis unter 65-Jährigen" ermittelt. Die Berechnung dieser Prozentzahlen erfolgte nach Geschlecht getrennt sowie für die Gesamtzahl der "50- bis unter 65-Jährigen".

Die Gemeinden sind durch ihre Gemeindekennziffer eindeutig ihrem Landkreis zugeordnet. Infolgedessen ist es möglich, die Gemeinde durch den übergeordneten Landkreis zu approximieren. Die Anzahl der "55- bis unter 65-Jährigen" wurde nun erstellt, indem man den Anteil der 50- bis unter 55-Jährigen aus den Landkreis-Daten mit der Anzahl der "50- bis unter 65-Jährigen" in den Gemeinde-Daten multipliziert hat und diese Zahl von der Gesamtzahl der "50- bis unter 65-Jährigen" der Gemeinde-Daten abgezogen hat. Die Berechnung wurde für Frauen und für die Gesamtzahl durchgeführt, da die Anzahlen der

Männer in den Gemeinde-Daten nicht erwähnt werden. Außerdem wurden die Anzahlen in der Alterskategorie der "50- bis unter 65-Jährigen" für die drei Untersuchungsjahre einzeln korrigiert. Auf diese Weise wurde eine neue Altersklasse "55 bis unter 65 Jahre" in den Gemeinde-Daten gebildet. Die Bevölkerungszahlen in den Alterskategorien sind durch diese Maßnahme besser vergleichbar mit den Altersangaben im Datensatz der Zielgröße, den Screening-Daten. Aus diesem Grund können im Folgenden die Gemeinde-Daten zur Analyse herangezogen werden.

Tabelle 2.4.: Verteilung der Einwohner nach Altersklasse (dreistufig) und Geschlecht.

Geschlecht	< 65	< 75	≥ 75	Summe
weiblich	18.78%	18.96%	17.32%	55.05%
männlich	18.40%	16.89%	9.65%	44.95%
Summe	37.18%	35.85%	26.97%	100.00%

Betrachtet man die Verteilung der Einwohner über die Jahre und Gemeinden hinweg (siehe Tab. 2.4), bietet sich bei einem Vergleich mit den Screening-Daten (siehe Tab. 2.2) folgendes Bild: Die Geschlechter sind insgesamt gut in den Screening-Daten repräsentiert. Dort sind die Klienten zu ca. 44.7% männlich. In den Gemeinde-Daten befinden sich zu ca. 45.0% Männer. Vergleicht man allerdings anhand der Altersklassen, so zeigt sich eine Verschiebung zu jüngeren Personen in den Screening-Daten. Denn während in den Gemeinde-Daten jede Altersklasse grob ein Drittel der Personen ausmacht (37.2%, 35.9% bzw. 27.0%), sind in den Screening-Daten über die Hälfte der Klienten (52.4%) unter 65 Jahre und fallen somit in die jüngste Altersgruppe. Dadurch ist die ältere Bevölkerung in den Screening-Daten unterrepräsentiert.
Ein weiterer Unterschied fällt in der Geschlechterverteilung in den Altersklassen auf. Es wurden mehr Frauen als Männer koloskopiert, doch dieser Unterschied zwischen den Geschlechtern wird mit fortschreitendem Alter immer geringer. In der zu Grunde liegenden Bevölkerung werden dagegen die Unterschiede zwischen Frauen und Männern mit zunehmendem Alter immer größer: In der jüngsten Altersklasse bis 65 Jahre sind die Geschlechter nahezu gleich stark vertreten (18.8% bzw. 18.4%). Das Verhältnis von Frauen zu Männern nimmt dann zu, bis es in der Gruppe der über 75-Jährigen bei fast 1.8 liegt. In dieser Altersklasse sind also ca. 1.8-mal mehr Frauen als Männer vertreten.
Außerdem fällt auf, dass sich die Frauen nahezu gleichmäßig auf die drei Altersklassen verteilen, während die Verteilung der Männer deutlich rechtsschief ist und es mit höherem Alter immer weniger Männer gibt. Über 75-jährige Männer machen nur mehr ca. 9.7% der Bevölkerung über 55 aus.

Wegen der Erhebung der Screening-Daten auf Basis der dreistelligen Postleitzahlbezir-

ke mussten auch die Gemeinde-Daten zu den 36 Gebieten zusammengefasst werden. Zu diesem Zweck wurden wieder die Gebietsnummern der Gemeinden aus Kapitel 2.2.1 benötigt. Anhand dieser Nummern wurden die Anzahlen der Personen für jedes Gebiet und die jeweilige Kombination aus Geschlecht und Altersklasse pro Jahr zusammengefasst.

Die Gemeinde-Daten beinhalten nach der Korrektur und anschließender Aggregation folgende Variablen:

- `gebiet`: die laufende Nummer des betrachteten Gebiets
- `maennlich`: 0 bezeichnet weibliche Klienten, 1 männliche.
- `altergrp`: eine Faktorvariable mit den Stufen
  - `5565`: Klienten ab 55 bis unter 65 Jahre
  - `6575`: Klienten ab 65 bis unter 75 Jahre
  - `7599`: Klienten ab 75 Jahre und älter
- `jahr`: das Untersuchungsjahr
- `einw`: die Anzahl Einwohner in der zugehörigen "Zelle" von Gebiet, Geschlecht, Altersstufe und Untersuchungsjahr

Sie entsprechen damit dem Aufbau der Screening-Daten für die Gemeinden.

Von der Schober Information Group Deutschland GmbH Falls die Zielgröße und die möglichen Einflussgrößen in verschieden großen räumlichen Einheiten erhoben wurden, liegt ein so genanntes "spatial misalignment" vor. Bisher hat sich keine statistische Methode gefunden, um dieses Problem zu lösen. Aus diesem Grund mussten die Daten, welche auf Gemeinde-Basis erhoben wurden, zu 36 Gebieten aggregiert werden. Nur so ist eine Analyse der Zielgröße, nämlich der auf Ebene der ersten drei Stellen der Postleitzahlen erhobenen Screening-Daten, möglich. Eine andere Möglichkeit, die Screening-Daten zu analysieren, ist der Erwerb von Bevölkerungsdaten auf Postleitzahl-Basis. Von der Schober Information Group wurde freundlicherweise der benötigte Datensatz gegen eine Aufwandsentschädigung zur Verfügung gestellt. Dieser Datensatz beinhaltet u. a. die Einwohnerzahlen für jeden dreistelligen PLZ-Bezirk in Bayern. Neben der Auftrennung der Zahlen nach Geschlecht liegt auch die Verteilung über 17 Altersklassen von "0 bis unter 3" bis hin zu "55 bis unter 60", "60 bis unter 65", "65 bis unter 75" und "75 und älter" vor. Für jede Alterskategorie ist außerdem enthalten, wie viele männliche und weibliche Einwohner im jeweiligen PLZ-Bezirk wohnen.

Nachdem die Daten der einzelnen Jahre in der MS-Excel-Datei untereinander gesetzt wor-

den waren, wurden die Überschriften der Spalten angepasst, so dass sie problemlos in R erkannt werden können. Anschließend wurde die MS-Excel-Datei als Tabstop-getrennte Text-Datei (Endung `.txt`) gespeichert und in R eingelesen. Leere Zeilen sowie die Spalten der nicht benötigten Altersklassen wurden gelöscht. Danach musste der Datensatz umstrukturiert werden, da pro dreistelliger Postleitzahl eine Zeile für jedes Jahr vorhanden war. Für die Analyse ist es jedoch nötig, dass für jede *Kombination* aus Jahr, Altersklasse und Geschlecht eine Zeile vorhanden ist.

Tabelle 2.5.: Verteilung der Einwohner nach Altersklasse (vierstufig) und Geschlecht.

Geschlecht	< 60	< 65	< 75	≥ 75	Summe
weiblich	10.15%	8.76%	18.88%	17.43%	55.21%
männlich	10.02%	8.59%	16.72%	9.46%	44.79%
Summe	20.16%	17.36%	35.59%	26.89%	100.00%

Vergleicht man die kommerziellen Daten der Schober Information Group (siehe Tab. 2.5) mit den Gemeinde-Daten aus der GENESIS-Datenbank (siehe Tab. 2.4) in Bezug auf die Alters- und Geschlechtsverteilung, so kann man erkennen: In der Auftretenshäufigkeit von Männern und Frauen ist der Unterschied gering. In den kommerziellen Daten sind ca. 55.2% der Personen weiblich, in den Gemeinde-Daten sind zu ca. 55.1% Frauen enthalten. Betrachtet man zudem die Altersverteilung, werden die Unterschiede nicht größer: In der Gruppe der ältesten Personen (75 Jahre und älter) befinden im kostenpflichtigen Datensatz ca. 26.9% aller Personen. In den Gemeinde-Daten sind 27.0% der einfließenden Beobachtungen dieser Altersklasse zuzurechnen. Ein ebenso kleiner Unterschied ist in der Altersgruppe der 65- bis 75-Jährigen zu sehen. In den Gemeinde-Daten sind dort 35.9% der Personen anzusiedeln, während es in den kommerziellen Daten 35.6% sind. In den kommerziellen Daten sind ca. 17.4% zwischen 60 und 65 Jahre alt und 20.2% zwischen 55 und 60 Jahre. Insgesamt sind in diesem Datensatz also 37.6% jünger als 65 Jahre. In den Gemeinde-Daten stellen die Personen ab 55 bis unter 65 Jahre insgesamt 37.2%.
Auch die Verteilung der Altersklassen in den Geschlechtern verhält sich in den Daten der Schober Information Group sehr ähnlich zu den Gemeinde-Daten. Der größte Unterschied beträgt ca. 0.21%. Es handelt sich um die Männer, welche jünger als 65 Jahre sind. Dieser geringe Unterschied zwischen den Zahlen in der jüngsten Alterskategorie zeigt, dass die Korrektur der Altersklassen in den Gemeinde-Daten gut funktioniert hat.

Bei einem Vergleich der kommerziellen Daten (siehe Tab. 2.5) mit den Screening-Daten (siehe Tab. 2.3) zeigt sich, wie bei den Gemeinde-Daten auch, dass die Geschlechterverteilung der Screening-Daten insgesamt derjenigen der Bevölkerung nahe ist (55.3% bzw. 55.2% Frauen). In der Altersverteilung zeigt sich eine häufigere Nutzung der Vorsorge-

Untersuchung durch jüngere Personen (30.3%) als der Anteil jüngerer Personen in der Bevölkerung (20.2%) nahelegt. Der Anteil von koloskopierten Personen zwischen 60 und 75 Jahren ist nahe am Anteil dieser Alterklassen in der Bevölkerung. Eine deutliche Unterrepräsentation lässt sich bei den Personen der ältesten Kategorie erkennen: Während nur 9.4% der Personen, die zum Screening gegangen sind, dieser Alterskategorie zuzuordnen sind, entfallen 26.9% der Bevölkerung ab 55 auf diese Altersklasse. Betrachtet man die Altersklassen noch differenziert nach Geschlecht, so zeigt sich, dass entsprechend der Aufteilung in der Grundbevölkerung mehr Frauen als Männer zur Darmkrebs-Vorsorge gingen. Allerdings ist der Unterschied in den Anzahlen der Männer und der Frauen mit zunehmendem Alter immer geringer. In der Bevölkerung dagegen zeigt sich eine immer größere Differenz zwischen den Anteilen männlicher und weiblicher Einwohner, je höher die Alterkategorie ist. Ähnlich wie bei den Gemeinde-Daten (Tab. 2.4) sind etwa gleich viele Frauen wie Männer zwischen 55 und 65 Jahre alt (10.2% bzw. 10.0%). In der ältesten Kategorie gibt es jedoch ca. 1.8-mal mehr Frauen als Männer (17.4% bzw. 9.5%).

Die Bevölkerungsdaten der Schober Information Group beinhalten nach einer ersten Bearbeitung folgende Variablen:

- `plz`: die ersten drei Stellen der Postleitzahl, welche zugleich die Identifikationsnummer der räumlichen Einheit ist
- `maennlich`: 0 bezeichnet weibliche Klienten, 1 männliche.
- `altergrp`: eine Faktorvariable mit den Stufen
  - `5559`: Klienten ab 55 bis unter 60 Jahre
  - `6064`: Klienten ab 60 bis unter 65 Jahre
  - `6574`: Klienten ab 65 bis unter 75 Jahre
  - `7599`: Klienten ab 75 Jahre und älter
- `jahr`: das Untersuchungsjahr
- `einw`: die Anzahl Einwohner in der zugehörigen "Zelle" von Gebiet, Geschlecht, Altersstufe und Untersuchungsjahr

2.3.2.2. Religionsverteilung

Wie die Religionen in den einzelnen Gemeinden verteilt sind, kann man ebenfalls in der GENESIS-Datenbank erfahren. Leider beruhen die einzigen verfügbaren Daten auf der letzten Volkszählung aus dem Jahr 1987. Die urspünglichen Daten liegen auf Basis der Gemeinden vor, so dass sie zu den 36 Gebieten aggregiert werden konnten. Beruhend auf

den damaligen Bevölkerungszahlen, welche ebenfalls im Datensatz enthalten sind, wurden Prozentzahlen ausgerechnet. Diese Prozentzahlen kann man als Variable in das Modell aufnehmen. Die Berechnung geschah folgendermaßen:

Der Datensatz wurde als csv-Datei abgespeichert. Nachdem nicht benötigte Zeilen wie die Tabellenüberschrift gelöscht wurden, konnten die Daten nach dem Regionalschlüssel sortiert werden und die Datenzeilen von Bayern (gesamt) und den Regierungsbezirken gelöscht werden. Die Landkreise konnten, wie bei den Gemeinde-Daten zuvor auch, nicht pauschal gelöscht werden, da sich unter den Einträgen mit dreistelligem Regionalschlüssel auch kreisfreie Städte befinden, deren Daten noch benötigt werden (siehe Kap. 2.3.2.1).

Anschließend wurde der Regionalschlüssel, welcher unterschiedlich viele Stellen hat, auf sechs Stellen verallgemeinert. Die Anfangs-9 wurde gelöscht und die Landkreise und kreisfreien Städte durch Multiplikation mit 1000 auf sechs Stellen erweitert.

Des Weiteren wurde der Tabellenkopf vereinfacht, da dann das Einlesen in R leichter ist. Statt der Überschriften “Römisch-katholische Kirche”, “Evangelische Kirche (einschl. Freikirchen)” etc. wurden kurze Bezeichnungen gewählt wie “rk”, “ev” etc.

Der soweit bearbeitete Datensatz wurde nun in R eingelesen und weiter bearbeitet. Der ursprüngliche Regionalschlüssel wurde entfernt. Eine Besonderheit in diesem Datensatz ist, dass “nicht vorhanden” mit “–” statt mit 0 gekennzeichnet ist. Dies wurde umgewandelt, indem die Variablen als `integer` definiert wurden. Dadurch entstehen `NA`s an den Stellen, an denen vorher “–” stand. Die `NA`s konnten dann in Nullen verwandelt werden.

Im Anschluss wurden die gemeindefreien Gebiete und Gebiete, deren Postleitzahl nicht in der Liste der Postleitzahlen mit Gemeindeschlüssel (siehe Kap. 2.1) steht, aus dem Datensatz entfernt. Nach dieser Maßnahme hat der Religionsdatensatz noch 2056 Zeilen, was genau den 2056 Gemeinden in Bayern entspricht.

Danach konnte ein neuer Datensatz erstellt werden, der die pro Gebiet aggregierten Daten enthält. Für jedes Gebiet wurde zunächst herausgefunden, welche dreistelligen Postleitzahlen darin enthalten sind. Anschließend wurden die zugehörigen Gemeindeschlüssel ermittelt. Anhand dieser konnten die interessierenden Zeilen im Religionsdatensatz extrahiert werden, welche genau die Gemeinden in betrachteten Gebiet widerspiegelten. In diesem Subdatensatz wurden die Spaltensummen der einzelnen Anzahlen berechnet und danach zusammen mit der Gebietsnummer abgespeichert.

Während der ganzen Prozedur wurde mit den Bevölkerungszahlen des Jahres 1987, welche neben den Anzahlen der Anhänger der verschiedenen Religionsgemeinschaften im ursprünglichen Datensatz enthalten waren, genauso verfahren wie mit den Religionszuge-

hörigkeiten. Dadurch war es nun möglich, pro Gebiet Prozentzahlen zu errechnen. Der endgültige Datensatz enthält folgende Variablen:

- `gebiet`: die laufende Nummer des betrachteten Gebiets
- `rk`: die Prozentzahl der in diesem Gebiet wohnenden Anhänger der römisch-katholischen Kirche, gemessen an der Gesamtbevölkerung 1987
- `ev`: die Prozentzahl der in diesem Gebiet wohnenden Anhänger der evangelischen Kirche (einschl. Freikirchen), gemessen an der Gesamtbevölkerung 1987
- `jd`: die Prozentzahl der in diesem Gebiet wohnenden Anhänger der jüdischen Religionsgesellschaft, gemessen an der Gesamtbevölkerung 1987
- `sonst`: die Prozentzahl der in diesem Gebiet wohnenden Anhänger sonstiger religiöser und christlicher Gemeinschaften, gemessen an der Gesamtbevölkerung 1987
- `kein`: die Prozentzahl der in diesem Gebiet wohnenden Personen, die keiner Religionsgesellschaft rechtlich zugehörig sind, gemessen an der Gesamtbevölkerung 1987

In Abbildung 2.2 ist die Verteilung der Religionszugehörigkeiten mittels Boxplots dargestellt. Wie erwartet, sind die Anteile der christlichen Religionen gegengleich verteilt. Denn der Norden Bayerns ist traditionell evangelisch geprägt, während im Süden hauptsächlich der Katholizismus praktiziert wird. Tatsächlich liegt die Korrelation bei -0.9897, es liegt also fast perfekte negative Korrelation vor. Aus diesem Grund kann man nicht beide Variablen ins Modell aufnehmen. Deswegen wurde die Variable `chr` geschaffen, welche beide Anteile aufaddiert. Sie repäsentiert also den Anteil der Christen unter den Einwohnern eines Gebiets. Allerdings ist diese Variable mit den Variablen `sonst` und `kein` fast ebenso hoch korreliert. Der Korrelationskoeffizient beträgt ca. -0.89 in beiden Fällen. Aus diesem und aus Gründen der Relevanz wurde in der weiteren Analyse nur die Variable `chr` in die Modelle aufgenommen.

2.3.2.3. Deprivationsscore

Wie weiter oben in Kapitel 2.3.2 erläutert, soll das durchschnittliche soziale Niveau bzw. die Kaufkraft der Einwohner dazu dienen, den Anteil privat versicherter und somit im Screening-Datensatz nicht erfasster Klienten zu schätzen. Aus Datenschutzgründen ist ein Datensatz zum verfügbaren Einkommen auf Gemeinde-Ebene nicht in GENESIS verfügbar. Deswegen griff man für die Analyse der 36 Gebiete auf einen so genannten “Index multipler Deprivation” (IMD) zurück. Er wurde von Werner Maier vom Helmholtz-

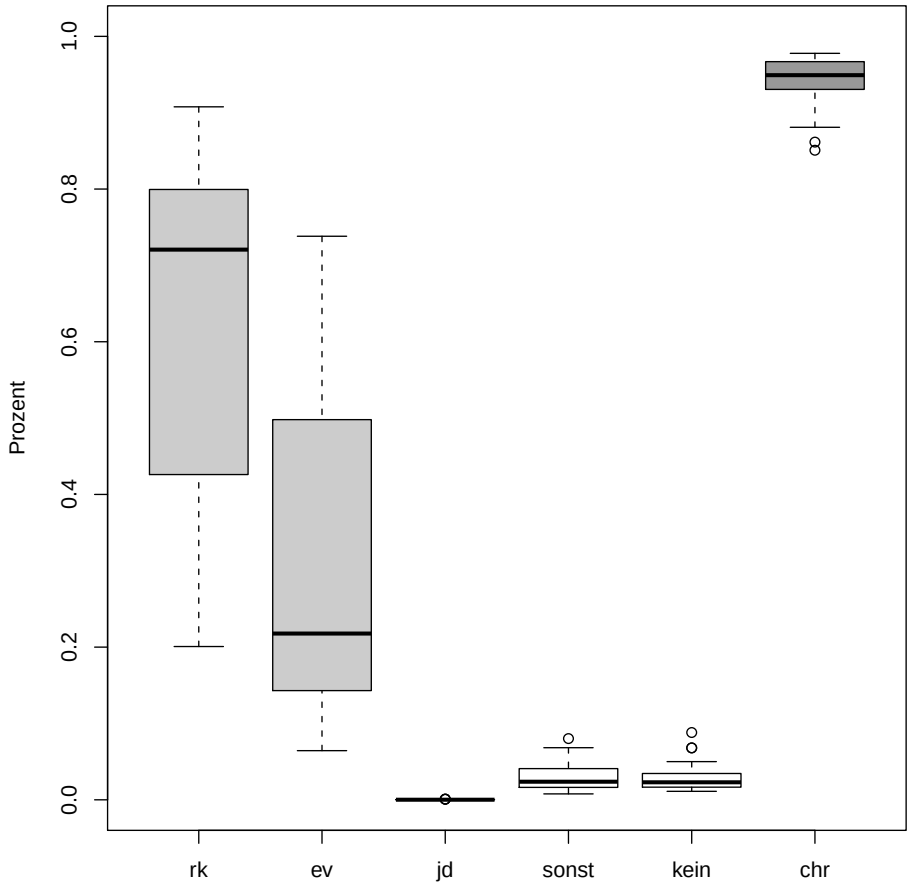

Abbildung 2.2.: Boxplots zur Verteilung der Religionszugehörigkeit in den 36 Gebieten. Die Abkürzungen entsprechen den Variablennamen im Datensatz der Religionsverteilung. Der Boxplot `chr` repräsentiert die Anteile römisch-katholischer und evangelischer Einwohner, also der Christen insgesamt.

Zentrum München[1] zur Verfügung gestellt. Herr Maier hat diesen nach dem Vorbild des IMD gebildet, welcher im Vereinigten Königreich (engl. United Kingdom, Abkürzung: UK) verwendet wird. Der Begriff Deprivation beschreibt einen Mangelzustand und die daraus folgenden Benachteiligungen. Deprivation kann sich auf einzelne Personen beziehen, welche einen Mangel an Nahrung, Kleidung, Unterkunft, Bildung, Arbeitssituation, etc. haben. Ebenso kann sich Deprivation auf ganze Populationsgruppen erstrecken, die sich in diesen Bereichen oder Teilen davon unterhalb des üblichen Niveaus befinden (Townsend, 1979).

Der hier verwendete Index (Maier et al., in Bearbeitung) misst die materielle und soziale Deprivation in den 2056 Gemeinden Bayerns. Die einfließenden Variablen stammen laut Werner Maier vom Bayerischen Landesamt für Statistik und geben den Stand des

[1] http://www.helmholtz-muenchen.de

Jahres 2006 wieder. Sie wurden zu sieben so genannten Domänen zusammengefasst. Diese sind: Einkommensdeprivation, Beschäftigungsdeprivation, Bildungsdeprivation, kommunale Einkommensdeprivation, Sozialkapitaldeprivation, Umweltdeprivation und Sicherheitsdeprivation (Maier et al., in Bearbeitung). In jeder Domäne wurde ein Score aus den Variablen berechnet. Dabei kam es in einigen Gemeinden vor, dass in manchen Variablen Werte fehlten und der ein oder andere Domänen-Score für die Gemeinde auf einer einzigen Messung beruhte.
Anschließend wurden Ränge über die Scores der Domänen bebildet, in Zahlen zwischen 0 und 1 umgerechnet und dann exponentialtransformiert. Aus diesen Zahlen wurde der Deprivationsindex als gewichteter Mittelwert berechnet. Das größte Gewicht erhält dabei die Domäne "Einkommensdeprivation". Die Werte sind durch die Exponentialtransformation positiv.
Bei der Interpretation ist darauf zu achten, dass ein hoher Score für hohe Deprivation steht. Je höher der Index ist, desto mehr Mangel herrscht also in der betreffenden Gemeinde. Der Vorteil des IMD ist, dass man entweder den Gesamt-Score verwenden kann oder nur einzelne Domänen in die Analyse einfließen lässt.

Nachdem der Datensatz aus MS-Excel als Tabstop-getrennte Text-Datei exportiert und in R importiert wurde, wurden erst nicht benötigte Spalten gelöscht. Anschließend wurde auch beim Deprivationsscore der Gemeindeschlüssel auf acht Stellen verringert und die führende 9 gestrichen. Danach konnten die Gemeinde-Daten von 2006 (vgl. Kap. 2.3.2.1) eingelesen werden und die Gesamtbevölkerung jeder Gemeinde in den IMD-Datensatz aufgenommen werden. Diese Zahl war für die weitere Bearbeitung von Bedeutung. Denn der Index eines jeden der 36 Gebiete wurde als gewichteter Mittelwert aus den beteiligten Gemeinden berechnet. Als Gewichte dienten dabei die jeweiligen Einwohnerzahlen.

Der IMD-Datensatz enthält nach der Aggregation durch einen gewichteten Mittelwert folgende Variablen:

- `gebiet`: die laufende Nummer des betrachteten Gebiets
- `bev.ges`: die Anzahl Einwohner im Gebiet
- `imd`: der gewichtete Mittelwert aus den Deprivationsindizes der beteiligten Gemeinden, gewichtet mit der Einwohnerzahl der Gemeinden

2.3.2.4. Kaufkraft je Einwohner

Die Schober Information Group stellte neben den Einwohnern je Postleitzahl-Bezirk auch die Kaufkraft auf der dreistelligen Ebene der Postleitzahlen zur Verfügung. Die von der

Schober Information Group verwendete Kaufkraft entspricht dem verfügbaren Einkommen der Wohnbevölkerung. Das verfügbare Einkommen ist laut dem Statistischen Bundesamt[2] das Einkommen, das den privaten Haushalten zufließt und die für Konsum- und Sparzwecke verwendet werden können. Die Zahlen basieren v. a. auf dem Nettoeinkommen. Diese Variable kann also, ähnlich dem Deprivationsscore aus Kapitel 2.3.2.3, als Schätzer für den nicht erfassten Anteil privat versicherter Personen im PLZ-Bezirk verwendet werden. Denn die Versicherung bei einer privaten Krankenkasse ist erst ab einem gewissen Einkommensniveau möglich.
Des Weiteren ist zu beachten, dass die Kaufkraft nie rückwirkend erhoben wird, sondern als Prognose für das kommende Jahr bereit steht. Somit gehört die Kaufkraft 2007 zu den Bevölkerungsdaten aus dem Jahr 2006. Denn die Kaufkraft, welche für das Jahr 2007 prognostiziert wird, basiert auf den Zahlen von 2006. Die Daten wurden von der Schober Information Group u. a. als "Kaufkraft je Einwohner in Euro" angegeben.

Ebenso wie die Einwohnerzahlen war die Kaufkraft für die verschiedenen Jahre in der MS-Excel-Tabelle nebeneinander angeordnet. Diese Spalten wurden untereinander gesetzt und die Überschriften R-freundlich gestaltet. Als Text-Datei abgespeichert und in R eingelesen, wurde der Datensatz zusammen mit den Bevölkerungszahlen in das für die Analyse benötigte Format gebracht (siehe Kap. 2.3.2.1). Der Einfachheit halber wurde dabei die zum Jahr 2006 gehörende Kaufkraft für 2007 unter dem Jahr 2006 abgespeichert. Mit den Daten zu den beiden anderen Jahren wurde analog verfahren.

Wie Abbildung 2.3 zeigt, ist zwischen den Jahren kaum ein Unterschied in der Verteilung der Kaufkraft zu erkennen. Die Boxplots zeigen eine sehr leichte Tendenz nach oben.

Die Kaufkraft ist im Bevölkerungsdatensatz für die dreistelligen PLZ-Bezirke der Schober Information Group enthalten. Die für die Kaufkraft relevanten Variablen sind:

- `plz`: die ersten drei Stellen der Postleitzahl, welche zugleich die Identifikationsnummer der räumlichen Einheit ist
- `jahr`: das Jahr, auf dem die Daten der prognostizierten Kaufkraft basieren
- `kk.einw`: die Kaufkraft je Einwohner in Euro

[2] http://www.destatis.de/jetspeed/portal/cms/Sites/destatis/Internet/DE/Content/Statistiken/VolkswirtschaftlicheGesamtrechnungen/Begriffserlaeuterungen/VerfuegbaresEinkommenHaushalte.psml

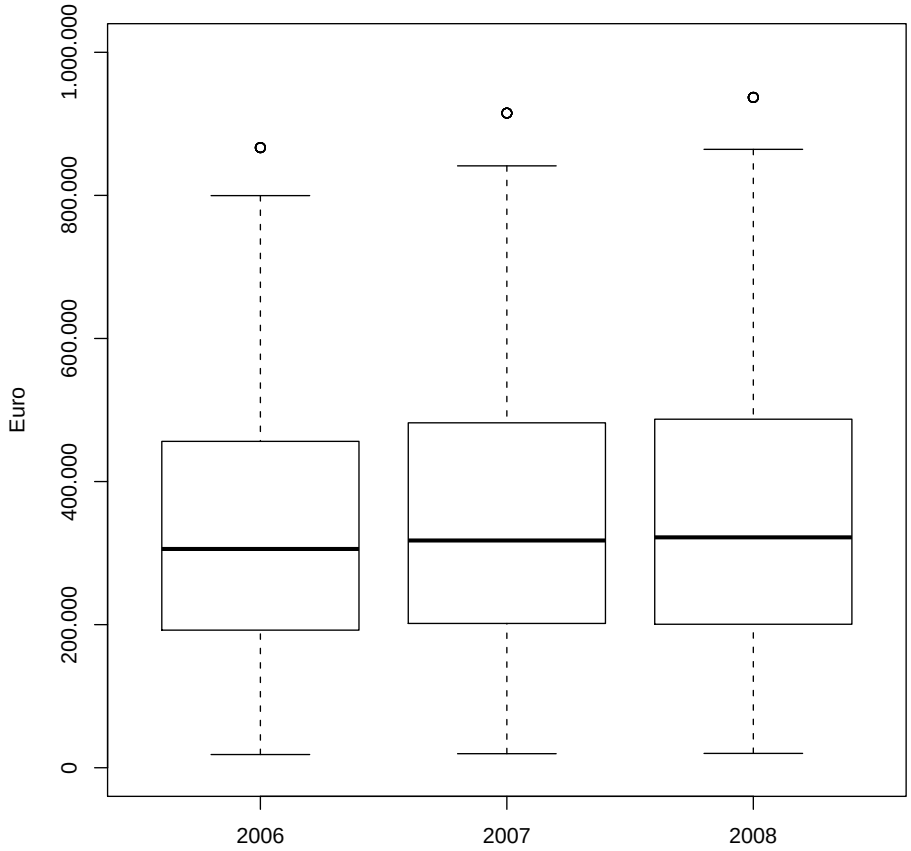

Abbildung 2.3.: Boxplots zur Verteilung der durchschnittlichen Kaufkraft je Einwohner in den Bezirken der dreistelligen Postleitzahlen.

2.3.2.5. Anzahl der Gastroenterologen

Von Prof. Mansmann wurde ein `csv`-File bereit gestellt, welches die Gastroenterologen enthält, die für die internetbasierte Dokumentation der Darmkrebs-Screenings registriert sind. Die Arztpraxen sind mit allen fünf Stellen der Postleitzahl erfasst.

Die Datei wurde in R eingelesen und aus den fünfstelligen dreistellige Postleitzahlen hergestellt. Wiederum wurde die in Kapitel 2.2.1 hergestellte Gebietsnummer für jeden dreistelligen PLZ-Bezirk verwendet, um die Ärzte den 36 Gebieten zuzuordnen. Für jedes der 36 Gebiete wurde dann die Anzahl an Gastroenterologen festgehalten. Dabei hat sich ergeben, dass in drei Gebieten kein Gastroenterologe ansässig ist – zumindest keiner, der die Vorsorge-Koloskopien im Internet dokumentiert. Dies betrifft die Gebiete 2 (PLZ 844 und 845) und 15 (PLZ 951 bis 955 und 963) sowie Gebiet 22 (PLZ 915). Um einen vollständigen Datensatz zu erhalten, wurde in diesen Gebieten als Anzahl der Gastroenterologen

"0" festgehalten.

Da eine Arztdichte meist pro 100000 Einwohner angegeben wird, wurden anschließend die Gemeinde-Daten für das Jahr 2008 in R geladen und pro Gebiet die Anzahl Gastroenterologen pro 100000 Einwohner berechnet. Die Gemeinde-Daten des Jahres 2008 wurden verwendet, da es sich bei der Auflistung der Gastroenterologen um den aktuellen Stand (2010) handelt.

Die größte Arztdichte von ca. 18 Gastroenterologen pro 100000 Einwohner wird in Gebiet 33 erreicht (Raum Ochsenfurt). Der Median liegt bei 3.4 Ärzten, der Mittelwert bei 4.0 Ärzten je 100000 Einwohner.

Der Gastroenterologen-Datensatz für die Gebiete enthält die Variablen:

- `gebiet`: die laufende Nummer des betrachteten Gebiets
- `Freq`: die absolute Anzahl an Gastroenterologen
- `ge.pro.100000.einw`: die relative Anzahl an Gastroenterologen je 100000 Einwohner

Ein ähnliches Vorgehen wurde bei der Bearbeitung des `csv`-Files für die Bezirke der dreistelligen Postleitzahlen gewählt: Nach dem Einlesen in R wurde eine Variable geschaffen, welche die dreistelligen Postleitzahlen jedes Arztes enthält. Anschließend konnten die Daten pro PLZ-Bezirk aggregiert werden.

Analog zu den Gebieten, in welchen es keine Gastroenterologen gibt, die mit der internetbasierten Dokumentation arbeiten, wurden die PLZ-Bezirke ohne Gastroenterologen festgestellt: Es handelt sich um die Bezirke 745xx, 747xx, 808xx, 817xx, 831xx, 851xx, 884xx, 890xx, 906xx, 916xx, 918xx, 927xx, 941xx, 945xx, 955xx, 957xx, 975xx und 979xx. Auch hier wurde der Datensatz vervollständigt, indem die PLZ-Bezirke mit "0" als Anzahl an Gastroenterologen aufgenommen wurden.

Die Arztdichte pro 100000 Einwohner wurde für die PLZ-Bezirke ebenfalls berechnet. Dazu wurde der Bevölkerungsdatensatz für die dreistelligen PLZ-Bezirke von der Schober Information Group eingelesen und die Daten für das Jahr 2008 extrahiert. Außerdem waren nur die Gesamtbevölkerungszahlen von Interesse. Die übrigen Einträge im Datensatz wurden gelöscht.
Im weiteren Verlauf zeigte sich, dass die Postleitzahlen 745xx, 747xx, 884xx und 890xx im Bevölkerungsdatensatz der Schober Information Group gar nicht enthalten sind. Für die Berechnung der Arztdichte wurden die betreffenden Postleitzahlen aus dem Gastroenterologen-Datensatz genommen. Genaueres zu diesem Problem wird im Kapitel 2.4.2 erläutert.

Nun konnte problemlos die Anzahl Ärzte je 100000 Einwohner berechnet werden. Die größte Arztdichte ergab sich für einen Bezirk in München: Der PLZ-Bezirk 805 verfügt über ca. 24 Gastroenterologen pro 100000 Einwohner. Der Median des gesamten Datensatzes liegt bei 3.1 Ärzten, der Mittelwert bei 3.6 Ärzten je 100000 Einwohner. Diese Zahlen sind geringfügig kleiner als die Zahlen für die gebietsweise Zusammenfassung der Einwohner.

Der Gastroenterologen-Datensatz für die Bezirke der dreistelligen Postleitzahlen enthält die Variablen:

- `PLZ3`: die ersten drei Stellen der Postleitzahl, welche zugleich die Identifikationsnummer der räumlichen Einheit ist
- `Freq`: die absolute Anzahl an Gastroenterologen
- `ge.pro.100000.einw`: die relative Anzahl an Gastroenterologen je 100000 Einwohner

2.4. Zusammengeführte Datensätze

2.4.1. Für die 36 Gebiete

Vor der Auswertung der Screening-Daten wurden die Gemeinde-Daten (Kap. 2.3.2.1), die Religionszugehörigkeit (Kap. 2.3.2.2), die Anzahl der Gastroenterologen (Kap. 2.3.2.5) und der Deprivationsindex (Kap. 2.3.2.3) anhand der Gebietsnummern zu einem Datensatz zusammengeführt. Die Jahreszahlen sollten als Faktor in das jeweilige Modell aufgenommen werden, so dass sie in Faktorstufen von 0 (Jahr 2006) bis 2 (Jahr 2008) überführt worden sind.

Der zur Analyse verwendete Datensatz beinhaltet folgende Kovariablen:

- `gebiet`: die laufende Nummer des betrachteten Gebiets
- `jahr`: das Untersuchungsjahr als Faktor
  Diese Variable wurde aufgenommen, um zu überprüfen, ob die Nutzung des Screening-Angebotes über die Zeit stabil ist.
- `maennlich`: 0 bezeichnet weibliche Klienten, 1 männliche.
  Das Geschlecht wurde erfasst, um zu überprüfen, ob Männer und Frauen ein unterschiedliches Nutzungsverhalten aufweisen.
- `altergrp`: eine Faktorvariable mit den Stufen
  - `5565`: Klienten ab 55 bis unter 65 Jahre
  - `6575`: Klienten ab 65 bis unter 75 Jahre

  - `7599`: Klienten ab 75 Jahre und älter

  Das Alter könnte ebenfalls bei der Nutzung der Darmkrebs-Vorsorge eine Rolle spielen.

- `anz.ge`: die relative Anzahl an Gastroenterologen je 100000 Einwohner (Stand 2008)
  Diese Variable sollte modellieren, ob die Verfügbarkeit des Screening-Angebots einen Einfluss auf die Nutzung hat.

- `imd`: der gewichtete Mittelwert aus den Indizes multipler Deprivation (Werner Maier) der beteiligten Gemeinden, gewichtet mit der Einwohnerzahl der Gemeinden (Stand 2006)
  Der Deprivationsindex wurde ins Modell genommen, um zu überprüfen, ob der Anteil der nicht-erfassten privat versicherten Personen in einer Region Auswirkungen auf die Anzahl registrierter Koloskopien hat. Der Index ist dafür geeignet, da man sich erst ab einem gewissen sozialen Niveau privat versichert. Würde sich ergeben, dass in einem Gebiet mit einem hohen Deprivationsindex viele Koloskopien erfasst wurden, würde der Anteil der nicht-erfassten Privatpatienten eine Rolle spielen. Andernfalls kann man diese Personen für die Auswertung mehr oder weniger ignorieren.

- `christlich`: die Prozentzahl der in diesem Gebiet wohnenden Anhänger der römisch-katholischen oder evangelischen Kirche (einschl. Freikirchen), gemessen an der Gesamtbevölkerung 1987
  Die Religionszugehörigkeit hat manchmal durchaus Auswirkungen auf die Entscheidung, einen medizinischen Eingriff vornehmen zu lassen oder nicht. Aus Kollinearitätsgründen konnte hier nur untersucht werden, ob die Zugehörigkeit zu einer christlichen Gemeinschaft einen Einfluss auf den Gang zur Darmkrebs-Vorsorge hat oder nicht.

Die Zielgröße `scree` sowie die Bevölkerungsdaten wurde für jede Modellart an die verwendete Software angepasst. Die Beschreibung der betreffenden Datensatz-Spalten findet sich an der jeweiligen Stelle von Kapitel 3.5.

Ein Beispiel für den Datensatz-Aufbau findet sich in Tabelle 2.6. Pro Kovariablen-Kombination k wurde hier festgehalten, wie viele Personen in der betreffenden Region mit dieser Kombination aus Geschlecht, Alter und Untersuchungsjahr das Screening nutzten (`scree` $= 1$) und wieviele Personen das Screening hätten nutzen können (`scree` $= 0$). Dies entspricht den Einwohnern ab 55 mit der entsprechenden Kovariablen-Kombination. Außerdem wurden pro Gebiet die regionalen Variablen aufgenommen (Index multipler Deprivation, Anteil christlicher Einwohner und Anzahl Gastroenterologen). Diese verändern

sich nicht, sondern bleiben für alle Beobachtungen eines Gebietes gleich.

2.4.2. Für die Bezirke der dreistelligen Postleitzahlen

In den Screening-Daten waren Postleitzahlen vorhanden, die nicht von der Schober Information Group geliefert wurden. Es handelt sich jeweils um einen Gemeindeteil, der als einziger in Bayern diese Zahlenkombination auf den ersten drei Stellen seiner Postleitzahl aufweist. Aus diesem Grund wurden diese Postleitzahlen in den Screening-Daten für die Auswertung der PLZ-Bezirke gelöscht, da ja zu diesen Regionen keine Bevölkerungsdaten vorhanden waren.

Zur Berechnung der Modelle wurden die Kaufkraft-Zahlen (Kap. 2.3.2.4), die Bevölkerungszahlen (Kap. 2.3.2.1) und die Anzahl der Gastroenterologen (Kap. 2.3.2.5) zu einem Datensatz zusammengefasst und ebenfalls wie bei den 36 Gebieten die Jahreszahlen in Faktorstufen umgewandelt.

Der zur Analyse verwendete Datensatz beinhaltet folgende Kovariablen:

- `plz`: die ersten drei Stellen der Postleitzahl, welche zugleich die Identifikationsnummer der räumlichen Einheit ist
- `jahr`: das Untersuchungsjahr als Faktor
  Die Variable `jahr` wurde ins Modell aufgenommen, um die Stabilität der Nutzung über die Zeit zu untersuchen.
- `maennlich`: 0 bezeichnet weibliche Klienten, 1 männliche.
  Ein Effekt des Geschlechts ist zu beobachten, wenn Männer und Frauen sich bei der Nutzung des Screening-Angebots unterschiedlich verhalten.
- `altergrp`: eine Faktorvariable mit den Stufen
  - `5559`: Klienten ab 55 bis unter 60 Jahre
  - `6064`: Klienten ab 60 bis unter 65 Jahre
  - `6574`: Klienten ab 65 bis unter 75 Jahre
  - `7599`: Klienten ab 75 Jahre und älter

  Das Alter spielt in der Nutzung von Krebsvorsorge-Angeboten eventuell eine Rolle.
- `anz.ge`: die relative Anzahl an Gastroenterologen je 100000 Einwohner (Stand 2008)
  Womöglich hat die Verfügbarkeit von Gastroenterologen eine Auswirkung, ob in einer Region häufig eine Koloskopie zur Krebsvorsorge durchgeführt wird oder nicht.
- `kk.einw`: die Kaufkraft je Einwohner in Euro (Schober Information Group)

Tabelle 2.6.: Der Beginn des Datensatzes, der im Modell mit unstrukturiertem räumlichen Effekt auf Basis der 36 Gebiete verwendet wurde, soll als Beispiel für den Datensatz-Aufbau dienen. Das Jahr 0 entspricht dem Jahr 2006, in der Spalte Geschlecht bezeichnet 1 männliche Personen. Die Gastroenterologen wurden in Ärzte je 100000 Einwohner angegeben. Die Spalte `scree` gibt hier an, ob die Anzahl Personen der betreffenden Zeile zum Screening ging (`scree` $= 1$) oder ob es sich um die screeningberechtigte Bevölkerung handelt (`scree` $= 0$). Die Tabelle ist so zu lesen, dass z. B. in Gebiet 1 23'066 Personen mit Kombination $k = 1$ (Frauen im Jahr 2006 zwischen 55 und 65 Jahre) wohnten und somit screeningberechtigt waren. In Gebiet 1 ist außerdem 92.91% der Bevölkerung christlichen Glaubens, es gibt dort 3.42 Gastroenterologen pro 100000 Einwohner und der Index multipler Deprivation liegt bei 22.74. (Je höher der Index ist, desto mehr Mangel herrscht.)

Gebiet (s)	k	Jahr	Geschlecht	Alter	christlich	Gastroent.	IMD	`scree`	Anzahl
1	1	0	0	< 65	0.9291	3.42	22.74	0	23066
1	2	0	0	< 75	0.9291	3.42	22.74	0	24824
1	3	0	0	≥ 75	0.9291	3.42	22.74	0	22479
1	4	0	1	< 65	0.9291	3.42	22.74	0	24059
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	7	1	0	< 65	0.9291	3.42	22.74	0	23806
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	18	2	1	≥ 75	0.9291	3.42	22.74	0	13859
1	1	0	0	< 65	0.9291	3.42	22.74	1	648
1	2	0	0	< 75	0.9291	3.42	22.74	1	483
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2	1	0	0	< 65	0.9778	0.00	25.16	0	2354
2	2	0	0	< 75	0.9778	0.00	25.16	0	2682
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Die Kaufkraft sollte modellieren, ob der Anteil der nicht-erfassten privat versicherten Personen eine Auswirkung auf die Nutzung des Darmkrebs-Screenings hat. Nachdem man sich erst ab einem gewissen Einkommen privat versichern kann, erscheint die Kaufkraft sinnvoll, um diesen Umstand angemessen zu erfassen.

Die Zielgröße sowie die Bevölkerungsstruktur wurde für jede Modellart aufgrund der verwendeten Software verändert. Deren Beschreibung finden sich an der jeweiligen Stelle von Kapitel 3.5. Die Datensätze für die Postleitzahlbezirke sind ähnlich aufgebaut wie diejenigen der 36 Gebiete (vgl. hierzu Tab. 2.6).

2.5. Karten von Bayern

Um die Nutzungsraten und auch die Ergebnisse der Untersuchungen visualisieren zu können, sind Karten notwendig. Wie bei den Daten über die Bevölkerung ist es schwierig, an Material zu kommen. Die Internetseite `http://www.gadm.org` bietet sowohl `ESRI`- als auch `RData`-Objekte zum kostenfreien Download an. Aus diesen `RData`-Objekten wurden die zu Bayern gehörenden Kartenteile entnommen. Dies geschah durch Überprüfung des Namens der Bundesländer. Des Weiteren stimmte die Reihenfolge, in der die zu zeichnenden Daten den Polygonzügen zugewiesen werden ("plotOrder"), nicht mit den hier verwendeten Gemeinde- bzw. Landkreis-Daten überein. Diese Tatsache hat sich im weiteren Verlauf als unwichtig erwiesen, da dieser Kartentyp nur zur Überprüfung der "äußeren Form" von Bayern verwendet werden konnte. Dennoch wurde die Zeichen-Reihenfolge schon früh geändert und die Karten so abgespeichert. Im weiteren Verlauf wurde nur noch auf die gespeicherten Objekte zugegriffen, darum ist dieser Umstand hier kurz erwähnt. Ein weiterer Arbeitsschritt dieser Art ist die Umbenennung einiger Polygone in die offiziell verwendeten Landkreis-Namen. Diese wurden ebenfalls mit Hilfe der Gemeinde- bzw. Landkreis-Daten angepasst.

Die `RData`-Objekte von `http://www.gadm.org` enthalten nur Polygonzüge auf Landkreis-Ebene. Die Landkreise und Gemeinden sind jedoch, genauso wie bei den Bevölkerungsdaten (siehe Kap. 2.2.1), für eine Visualisierung der auf Postleitzahlebene erhobenen Screening-Daten ungeeignet. Aus diesem Grund kann dieser Kartentyp nicht weiter verwendet werden und die Änderung der Namen und der Zeichen-Reihenfolge wurde mit Feststellung der Inkompatibilität von PLZ-Bezirken und Gemeinden bzw. Landkreisen hinfällig.

Deswegen wurde neben dem Bevölkerungsdatensatz für die dreistelligen PLZ-Bezirke von der Schober Information Group auch ein Shapefile der Bezirke der dreistelligen Postleit-

zahlen in Deutschland erworben. Das `shp`-file konnte mit Hilfe des Paketes `maptools` (Lewin-Koh et al., 2009) in R eingelesen werden und in einen `SpatialPolygonsDataFrame` umgewandelt werden. Dieser R-Objekttyp ist typisch für die Darstellung von räumlichen Daten in Karten. Zur weiteren Bearbeitung wird das Paket `sp` (Pebesma und Bivand, 2005) benötigt. Anschließend wurden über die Postleitzahlen, welche auch in der Liste des Vermessungsamtes (Kap 2.1) enthalten sind, die Bezirke Bayerns extrahiert.

Allerdings sind die Bezirke der dreistelligen Postleitzahlen nicht nur nicht mit Gemeinde- oder Landkreisgrenzen konform, sondern reichen sogar über die Bundeslandgrenzen hinaus. Deswegen wurde zur genauen Abgrenzung von Bayern das `RData`-Objekt von `http://www.gadm.org` verwendet. Dies führte zu einem weiteren Problem: Das `shp`-file enthielt die Koordinaten im Gauß-Krüger-System, während die Koordinaten im `RData`-Objekt geographische Koordinaten waren. Die Umwandlung der Gauß-Krüger- in geographische Koordinaten war der nächste Arbeitsschritt. Dazu wurde eine R-Funktion geschrieben, welche nach den Formeln von Schödlbauer (1982, S. 79ff.) die Umrechnung durchführt. Diese wird innerhalb einer weiteren Funktion aufgerufen, die die Polygone in einem `SpatialPolygonsDataFrame`-Objekt nach und nach durchgeht und die Koordinaten umwandelt. Die Zeichnung beider Karten in das gleiche Koordinatensystem zeigt eine leichte, aber durchaus akzeptable Verschiebung in Ost-West-Richtung (siehe Abb. 2.4).

2.5.1. Karte der Bezirke der dreistelligen Postleitzahlen

Nachdem die Koordinaten vom Gauß-Krüger- ins geographische Koordinatensystem umgewandelt waren, konnten die Grenzbereinigungen vorgenommen werden. Die Postleitzahlbezirke, welche mit 745, 747, 884, 890 beginnen, wurden komplett entfernt. Der Postleitzahlbezirk 979xx besteht aus zwei Regionen, von denen nur eine in Bayern liegt. Die Region außerhalb Bayerns wurde ebenfalls gelöscht. Der PLZ-Bezirk 978 wird von der Landesgrenze durchschnitten. Hier war etwas mehr Aufwand nötig, um die genauen Treffpunkte der beiden Polygone von der Schober Information Group und von `http://www.gadm.org` zu bestimmen. Anschließend wurde der Teil aus dem `gadm`-Objekt herausgefiltert, der zwischen diesen beiden Schnittpunkten liegt, und in den `SpatialPolygonsDataFrame` an entsprechender Stelle eingefügt. Im PLZ-Bezirk 875 ist etwas zuviel Fläche eingeschlossen. Bei dieser Verkleinerung wurde genauso wie bei der Region 978 verfahren.

Somit war die Bereinigung abgeschlossen und es lag eine Karte der dreistelligen Postleitzahlbezirke Bayerns vor.

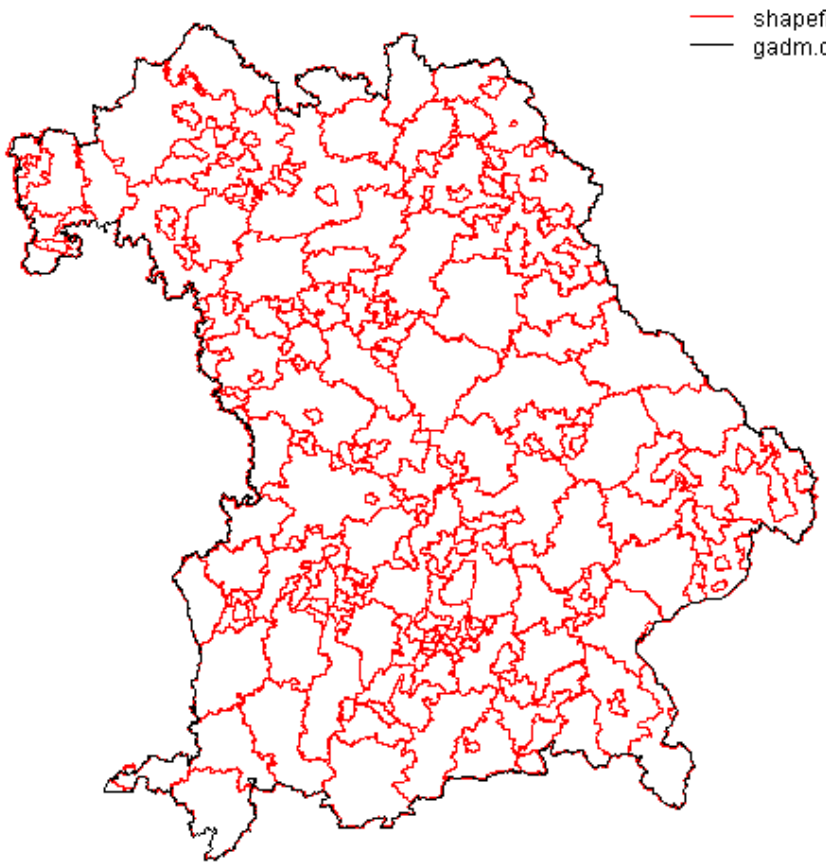

Abbildung 2.4.: Karten von Schober Information Group und http://www.gadm.org. Die leichte Abweichung zwischen den beiden Karten in Ost-West-Richtung ist am besten im Nordwesten zu erkennen.

2.5.2. Karte der 36 Gebiete

Aus der Karte der dreistelligen PLZ-Bezirke konnte nun eine Karte der 36 Gebiete geformt werden. Wiederum wurde eine R-Funktion geschrieben, welche dies bewerkstelligt. Sie durchläuft die Liste der PLZ mit zugehöriger Gebietsnummer aus Kapitel 2.2.1 (siehe auch Tabelle `gkz-plz-geb.txt` im Ordner "Daten" des elektronischen Anhangs, Kap. B) und bestimmt pro Gebiet die zugehörigen Polygonzüge des `SpatialPolygonsDataFrame` und speichert diese als ein `list()`-Objekt. Die Liste wird dann als *eine* räumliche Einheit definiert, ähnlich wie mehrere Inseln o. Ä. ein (politisches) Land formen. Aus der Gesamtheit dieser Listen konnte wiederum ein `SpatialPolygons`-Objekt erzeugt werden. Zusammen mit den IDs von 1 bis 36 bildet es einen `SpatialPolygonsDataFrame`, mit dem im Weiteren die Karten der 36 Gebiete gezeichnet werden konnten.

2.6. Die Nutzungsraten

In Kapitel 5.1 werden die in Kapitel 4 beschriebenen Ergebnisse dieser Arbeit mit vorangegangenen Analysen verglichen. Dort wurde allerdings kein statistisches Modell gerechnet (Pritzkuleit et al., 2009) bzw. dessen Schätzer nicht veröffentlicht (Mansmann et al., 2008), so dass ein Vergleich nur aufgrund der Nutzungsraten möglich ist. In beiden Vergleichsartikeln geben Prozentzahlen an, welcher Anteil der berechtigten Bevölkerung das Screening in Anspruch genommen hat. Die Nutzungsraten des Screening-Programms für Kolonkarzinome aus dem vorliegenden Datensatz sind in Kapitel 2.6.1 zu finden.

Diese variieren allerdings räumlich stark, wie bei einer Beschreibung der Daten anhand von Karten zu erkennen ist.
Für die Karten in den Kapiteln 2.6.2 und 2.6.3 gilt: Die Daten wurden pro Region über die Jahre gemittelt. Je mehr die Farbe des jeweiligen Gebiets ins Rötliche tendiert, desto niedriger ist dort die Nutzungsrate. Die Farbgebung basiert auf den Mittelwerten der Gebiete und der PLZ-Bezirke, so dass gleiche Farbe gleiches Werte-Intervall bedeutet und die Karten der Gebiete mit den Karten der PLZ-Bezirke farblich direkt vergleichbar sind.

2.6.1. Die Nutzungsraten im Überblick

Über die drei Jahre gemittelt ist die Nutzungsrate gemessen an der Screening-berechtigten Bevölkerung für die 36 Gebiete 1.597%. In den Bezirken der dreistelligen PLZ gingen 1.603% der berechtigten Personen zum Screening. Diese leichte Unstimmigkeit ist wohl auf die unterschiedlichen zugrunde liegenden Bevölkerungsdatensätze zurückzuführen. Gleiches gilt natürlich auch für die Tabellen 2.7 und 2.8.

Tabelle 2.7.: Nutzungsraten nach Altersklasse (dreistufig) und Geschlecht für die Jahre 2006 bis 2008.

	2006			2007			2008		
Alter	Männer	Frauen	gesamt	Männer	Frauen	gesamt	Männer	Frauen	gesamt
< 65	1.738%	2.341%	2.042%	2.122%	2.685%	2.406%	2.048%	2.535%	2.295%
< 75	1.554%	1.603%	1.580%	1.819%	1.890%	1.856%	1.643%	1.703%	1.675%
≥ 75	0.655%	0.413%	0.498%	0.761%	0.513%	0.602%	0.708%	0.478%	0.562%
gesamt	1.440%	1.478%	1.461%	1.716%	1.728%	1.722%	1.604%	1.604%	1.604%

In den 36 Gebieten wird das Screeningangebot mit dem Alter seltener in Anspruch genommen (Tab. 2.7). Außerdem lässt sich feststellen, dass in den jüngeren Altersklassen Frauen eher zur Koloskopie gehen als Männer. Dieser Effekt des Geschlechts dreht sich mit zunehmendem Alter um: Bei den über 75-jährigen nutzen – relativ gesehen – mehr Männer als

Frauen die Vorsorgeuntersuchung. Durch diese Umkehr des Geschlechter-Effekts bleiben die Gesamt-Raten für Männer und Frauen recht ausgeglichen. Man kann erkennen, dass die Teilnahmeraten vom Jahr 2006 auf das Jahr 2007 ansteigen. Von 2007 auf das Jahr 2008 werden die Raten wieder geringer, jedoch bleiben sie in allen Altersklassen und bei beiden Geschlechtern über dem Niveau von 2006. Das Teilnahme-Muster in Abhängigkeit von Alter und Geschlecht bleibt über die Jahre bestehen.

Tabelle 2.8.: Nutzungsraten nach Altersklasse (vierstufig) und Geschlecht für die Jahre 2006 bis 2008.

	2006			2007			2008		
Alter	Männer	Frauen	gesamt	Männer	Frauen	gesamt	Männer	Frauen	gesamt
< 60	1.733%	2.495%	2.115%	2.178%	2.944%	2.563%	2.195%	2.844%	2.523%
< 65	1.661%	2.072%	1.869%	2.064%	2.411%	2.239%	1.861%	2.185%	2.024%
< 75	1.603%	1.634%	1.619%	1.835%	1.889%	1.864%	1.655%	1.707%	1.683%
≥ 75	0.675%	0.414%	0.505%	0.780%	0.514%	0.608%	0.725%	0.476%	0.565%
gesamt	1.451%	1.476%	1.465%	1.732%	1.731%	1.731%	1.614%	1.605%	1.609%

Auch bei einer Betrachtung der Bezirke der dreistelligen PLZs (Tab. 2.8) nimmt die Nutzung der Vorsorge-Koloskopie mit zunehmendem Alter ab. Frauen nehmen bis zu einem Alter von ca. 65 Jahren eher an einem Screening teil als Männer der gleichen Alterskategorien. In der Klasse der 65- bis 75-Jährigen ist die Teilnahme bei beiden Geschlechtern etwa gleich. Im hohen Alter (ab 75 Jahre und älter) geht ein größerer Anteil Männer zur Koloskopie als es bei den Frauen der Fall ist. Insgesamt sind die Nutzungsraten von Männern und Frauen etwa gleich. Auch bei den PLZ-Bezirken ist eine Zunahme der Nutzung von 2006 auf das folgende Jahr zu beobachten. Die Raten sinken im Jahr 2008 wieder, allerdings nicht mehr auf das Niveau von 2006. Die Variation bzgl. Alter und Geschlecht wiederholt sich jedes Jahr nach dem gleichen Muster.

2.6.2. Karten für die 36 Gebiete

Die Abbildung 2.5 zeigt eine Karte der Nutzungsraten in den 36 Gebieten. Die Daten wurden über die drei Jahre gemittelt. Je roter die Farbe einer Region ist, desto niedriger ist dort die Nutzungsrate. Die kleinste mittlere Nutzungsrate lag bei 0.52 Screenings pro 100 berechtigte Einwohner (Gebiet 24: Postleitzahlen 926xx rund um Weiden i.d.OPf.), die größte lag beim Vierfachen, nämlich bei 2.02 Screenings je 100 screeningberechtigte Einwohner (Gebiet 2: Postleitzahlen 745xx, 916xx in Mittelfranken an der Grenze zu Baden-Württemberg). Die Medianbeobachtung teilen sich die Gebiete 7 (Postleitzahlen 844xx und 845xx in Oberbayern an der Grenze zu Oberösterreich) und 32 (Postleitzahlen 970xx und 972xx rund um Würzburg) mit mittleren Raten von 1.65 bzw. 1.58 Screenings

je 100 berechtigte Einwohner. Im Mittel wurden 1.609 Screenings erfasst. Die genaue Lage der Gebiete kann Abbildung D.1 entnommen werden.
Durch die Färbung ist ein klares Ost-West-Gefälle erkennbar – mit Ausnahme des Allgäus im Südwesten, welches niedrigere Nutzungsraten als der restliche Westteil Bayerns aufweist. Die höchsten Raten wurden in den Ballungsräumen um Nürnberg bzw. Fürth und Ansbach sowie in der Nähe von Neu-Ulm beobachtet. Die Raten von München gehen in die Raten des zugehörigen Gebietes Nr. 3 ein. Dieses ist jedoch so groß, dass man den Effekt des Ballungsraums nicht mehr erkennt.

Weil die 36 Gebiete z. T. sehr groß sind, wurde auf eine Glättung der Nutzungsraten verzichtet.

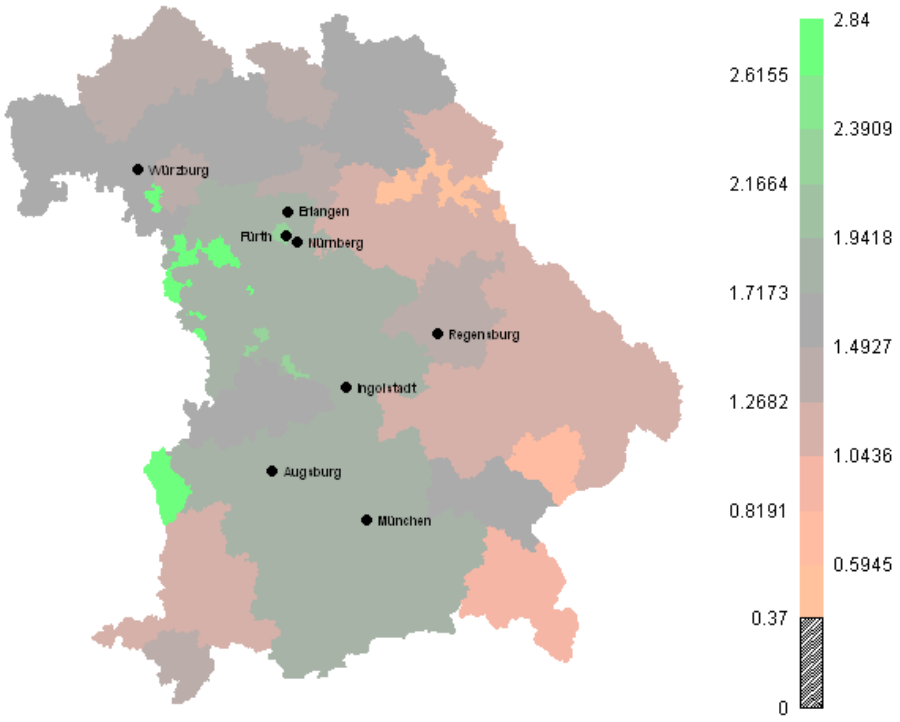

Abbildung 2.5.: Karte der 36 Gebiete zur Verteilung der Nutzungsraten. Gezeigt sind die Mittelwerte der Nutzungsraten über den gesamten Zeitraum im jeweiligen Gebiet. Angaben in Screenings je 100 screeningberechtigte Einwohner.

2.6.3. Karten für die Bezirke der dreistelligen Postleitzahlen

Nachdem es 113 dreistellige Postleitzahlen in Bayern gibt, sind die Nutzungsraten auf den Karten dieser PLZ-Bezirke naturgemäß kleinräumiger dargestellt, als es bei den 36 Gebieten möglich ist. Die Abbildung 2.6 zeigt aus diesem Grund eine differenziertere Farbgebung

als Karte 2.5.

Wie oben erwähnt, wurde die Farbgebung so gewählt, dass die Karten der 36 Gebiete direkt mit den Karten der dreistelligen PLZ-Bezirke vergleichbar sind. Bei der Karte der 36 Gebiete (Abb. 2.5) zeigt sich eine starke Häufung bei mittleren Nutzungsraten. Die Verteilung der Nutzungsraten in den Bezirken der dreistelligen Postleitzahlen ist flacher und deswegen treten die "extremeren" Farbtöne, also stark rötlich und grünlich gefärbte Regionen, vermehrt auf (Abb. 2.6).

Auch bei den Nutzungsraten der PLZ-Bezirke wurden die Daten pro Bezirk über die Jahre gemittelt. Die niedrigste Rate beträgt 0.37 Screenings pro 100 berechtigte Einwohner (PLZ 924xx rund um Schwandorf), die höchste 2.84 Screenings (PLZ 822xx rund um Fürstenfeldbruck). Im Mittel gingen 1.605 von 100 screeningberechtigten Personen zur Darmkrebs-Vorsorge. Der Median von 1.64 Screenings pro 100 berechtigte Einwohner wurde in Postleitzahl-Bezirk 930xx (rund um Regensburg) beobachtet. Die Lage der Postleitzahlbezirke ist in Kapitel D.2 zu entnehmen.
Wie bei den 36 Gebieten zeigen die Nutzungsraten einen Ost-West-Trend. Ebenfalls ist hier zu erkennen, dass der Südwesten im Verhältnis niedrige Raten aufweist. Die höchsten Raten konnten die Gebiete nordwestlich und südlich von Würzburg, rund um Kulmbach, in den Ballungsräumen Nürnberg/Fürth und München, zwischen Nürnberg und Regensburg und in der Gegend von Nördlingen erreichen.

Räumlich geglättete Nutzungsraten erhält man, indem man alle Nachbarn einer Region bestimmt und aus deren Raten einen gewichteten Mittelwert berechnet. Das Gewicht dabei sind die Einwohner der benachbarten Regionen – in diesem Fall die screeningberechtigten Einwohner. Aufgrund der Definition, dass man sich nicht selbst Nachbar sein kann, ist die eigentliche Beobachtung bei der Mittelung ausgeschlossen. Man leiht sich also die Beobachtungen der umliegenden PLZ-Bezirke, um Aussagen über eine Region machen zu können. Da die meisten Regionen mehrere Nachbarn haben und ihre Daten somit in mehrere Berechnungen einfließen, werden sich die Raten benachbarter Bezirke durch die Glättung ähnlicher. Diese geglättete Version der Mittelwerte je PLZ-Bezirk ist in Abbildung 2.7 gezeigt. Der Wertebereich der Raten ist durch die Glättung enger geworden – es werden mind. 0.7344 (in PLZ-Bezirk 925xx; roh: 0.370 in PLZ-Bezirk 924xx) und max. 2.1940 (in PLZ-Bezirk 823xx; roh: 2.840 in PLZ-Bezirk 822xx) Screenings je 100 berechtigte Einwohner beobachtet. Der Mittelwert ändert sich kaum (1.601 statt vorher 1.605). Die geglätteten Nutzungsraten der PLZ-Bezirke ähneln stark der Raten der 36 Gebiete (vgl. Abb. 2.7 mit 2.5).

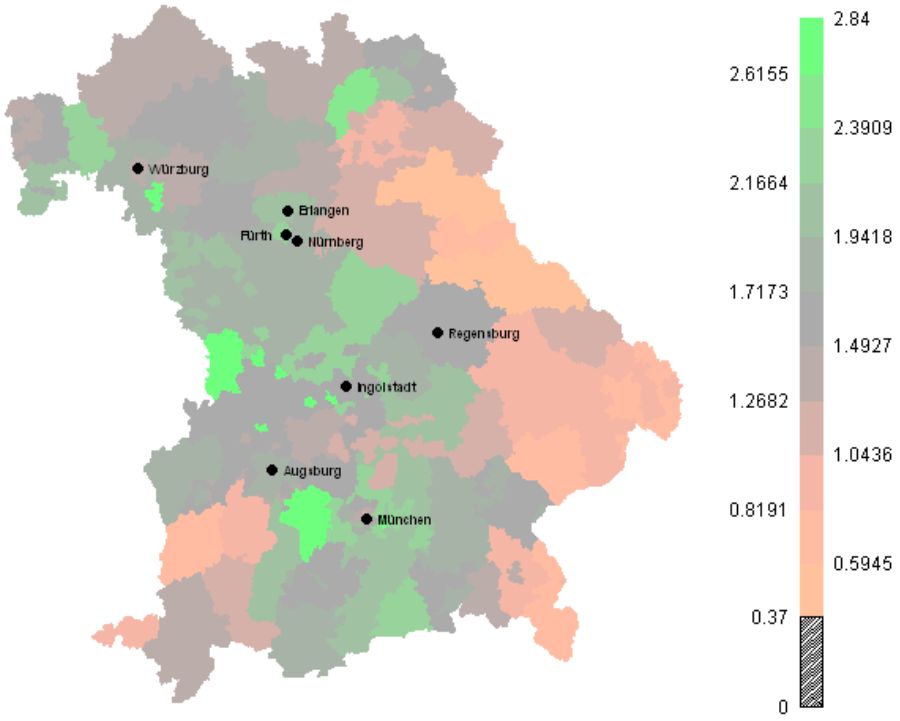

Abbildung 2.6.: Karte der Bezirke der dreistelligen Postleitzahlen zur Verteilung der Nutzungsraten. Gezeigt sind die Mittelwerte der Nutzungsraten über den gesamten Zeitraum im jeweiligen Bezirk. Angaben in Screenings je 100 screeningberechtigte Einwohner.

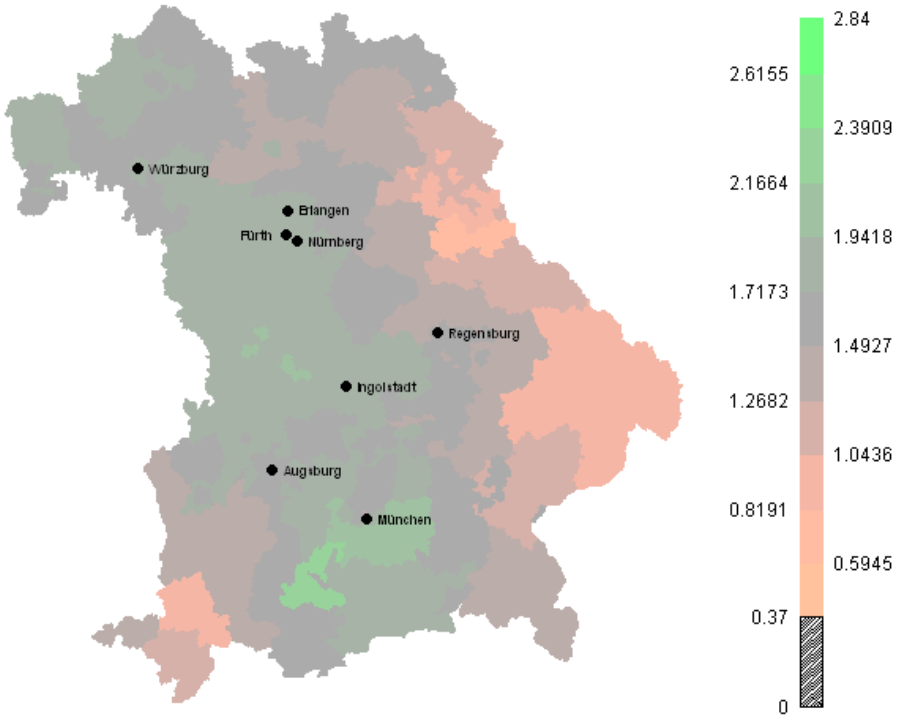

Abbildung 2.7.: Karte der Bezirke der dreistelligen Postleitzahlen zur Verteilung der Nutzungsraten. Gezeigt sind die über die Nachbarbezirke geglätteten Mittelwerte der Nutzungsraten über den gesamten Zeitraum im jeweiligen Bezirk. Angaben in Screenings je 100 screeningberechtigte Einwohner.

3. Methodik

Die Auswertung der Screening-Daten mit einem einfachen linearen Modell ist der Datenstruktur nicht angemessen. In Kapitel 3.1 wird deswegen die Verallgemeinerung des linearen Modells auf andere Verteilungsannahmen erklärt, denn die Screening-Daten sind nicht normalverteilt. Die Zielgröße lautet "Screening ja/nein" und ist somit binomialverteilt.
Da die Daten auf Basis räumlicher Einheiten wie z. B. Gemeinden oder Postleitzahlbezirke vorliegen, wurde ein räumlicher Effekt in die Modellgleichung aufgenommen. In Kapitel 3.3.1.2 werden die Hauptunterschiede zwischen einem unstrukturierten und einem strukturierten räumlichen Effekt erläutert.
Die Auswertung unter der Annahme eines unstrukturierten räumlichen Effekts erfolgte mit Hilfe eines generalisierten linearen gemischten Modells. Die Theorie hierzu findet sich in Kapitel 3.2.
Der strukturierte räumliche Effekt kann auf mehrere Arten geschätzt werden. In der vorliegenden Arbeit wurde der Ansatz der penalisierten Quasi-Likelihood (PQL) (Kap. 3.3.3.2) mit dem bayesianischen Ansatz (Kap. 3.3.3.3) verglichen.

3.1. Generalisierte lineare Modelle

3.1.1. Verallgemeinerung des linearen Modells

Im linearen Modell liegt die Annahme einer Normalverteilung zugrunde (siehe (3.1)).

$$\begin{aligned} y_i &= \boldsymbol{x}_i{}^T\boldsymbol{\beta} + \epsilon_i \\ \text{mit } \epsilon_i &\overset{\text{iid}}{\sim} N(0,\ \sigma^2) \\ \iff Y_i \,|\, \boldsymbol{x}_i &\sim N(\boldsymbol{x}_i{}^T\boldsymbol{\beta},\ \sigma^2) \end{aligned} \tag{3.1}$$

Der Index i läuft über die einzelnen Beobachtungen ($i = 1, \ldots, n$). Der Vektor $\boldsymbol{x}_i$ hat $P + 1$ Elemente und beinhaltet eine Eins für den Intercept und die Werte der P ins Modell aufgenommenen Kovariablen für Beobachtung i. Der Vektor $\boldsymbol{\beta}$ hat ebenfalls $P +$

1 Komponenten und enthält neben dem Intercept β_0 die jeweils zum p-ten Parameter gehörenden Regressionskoeffizienten ($p = 1, \ldots, P$).

Nicht immer jedoch sind die Zielgrößen bzw. der Messfehler (approximativ) normalverteilt. Insbesondere wird diese Annahme bei binären Response-Variablen verletzt. In diesem Fall möchte man der Modell-Anpassung eher eine Binomialverteilung zugrunde legen. Somit muss man die Annahmen des linearen Modells verallgemeinern und zu einem generalisierten linearen Modell (Abkürzung: GLM) übergehen. Die Annahmen in dieser Modellklasse beruhen auf einer Verteilung der *Exponentialfamilie*. Sie ist also u. a. auch für Binomialverteilungen definiert.

Der Erwartungswert des Response wird mit Hilfe einer *Link-Funktion* mit den Daten verbunden. Im linearen Modell ist die Link-Funktion die Identität (siehe (3.2)).

$$\begin{aligned} g^{-1}(\boldsymbol{x}_i{}^T\boldsymbol{\beta}) = \mathbb{E}(Y_i) = \mu_i = \boldsymbol{x}_i{}^T\boldsymbol{\beta} \\ g(\mu_i) = \boldsymbol{x}_i{}^T\boldsymbol{\beta} = \mu_i \end{aligned} \tag{3.2}$$

3.1.2. Das Logit-Modell

Bei der anschließenden Analyse der vorliegenden Daten wurde modelliert, ob eine gewisse Gruppe von Menschen das Screening-Angebot nutzt ($\texttt{scree}_i = 1$) oder nicht. Das Logit-Modell ist für solche binäre Response-Variablen geeignet. Statt der Normalverteilung wird nun aufgrund der binären Kodierung von Y_i eine *Binomialverteilung* angenommen: $Y_i \,|\, \boldsymbol{x}_i \sim Bin(1,\ p(\boldsymbol{x}_i))$. Der Erwartungswert des Response $\mathbb{E}(Y_i)$ wird durch den Parameter $p(\boldsymbol{x}_i) = \mathbb{P}(Y_i = 1)$ bestimmt (siehe auch (3.6)).

Zur Herleitung des Logit-Modells kann man eine latente Beobachtung $\tilde{y}_i$ betrachten, welche einem linearen Modell folgt:

$$\tilde{y}_i = \gamma_0 + \boldsymbol{z}_i{}^T\boldsymbol{\beta} + \epsilon_i, \quad \boldsymbol{\beta} = (\beta_1, \ldots, \beta_p) \tag{3.3}$$

Das eigentlich beobachtete y_i gibt einen Status wieder, welcher erreicht wird, falls $\tilde{y}_i$ eine gewisse Schranke ϑ überschreitet, und ist somit binär kodiert:

$$y_i = 1 \iff \tilde{y}_i \geq \vartheta$$

Als Beispiel kann man sich hier als $\tilde{y}_i$ den Gesundheitszustand einer Person i vorstellen, welcher ab einem gewissen Maß an Beschwerden ϑ zum Arztbesuch ($y_i = 1$) führt. Bei einer Vorsorge-Situation wie im vorliegenden Datensatz ist die Schranke durch ein gewisses

Gesundheitsbewusstsein gegeben.

Die Wahrscheinlichkeit $p(\boldsymbol{z}_i)$, dass bei gegebenen Kovariablen, welche die Gesundheit beeinflussen, ein Arzt aufgesucht wird, kann folgendermaßen hergeleitet werden (Fahrmeir et al., 2007):

$$\begin{aligned}
p(\boldsymbol{z}_i) &= \mathbb{P}(Y_i = 1 \,|\, \boldsymbol{z}_i) \\
&= \mathbb{P}(\tilde{Y}_i \geq \vartheta \,|\, \boldsymbol{z}_i) \\
&= \mathbb{P}(\gamma_0 + \boldsymbol{z}_i{}^T\boldsymbol{\beta} + \epsilon_i \geq \vartheta) \\
&= \mathbb{P}(\epsilon_i \geq \underbrace{\vartheta - \gamma_0}_{=-\beta_0=-1^T\beta_0} - \boldsymbol{z}_i{}^T\boldsymbol{\beta}) \\
&= 1 - \mathbb{P}(\epsilon_i < -\boldsymbol{x}_i{}^T\boldsymbol{\beta}), && \boldsymbol{\beta} = (\beta_0, \ldots, \beta_p),\ \boldsymbol{x}_i = (1, \boldsymbol{z}_i) \\
&= 1 - F_{\epsilon_i}(-\boldsymbol{x}_i{}^T\boldsymbol{\beta}), && \epsilon_i \text{ stetig verteilt} \\
&= F_{\epsilon_i}(\boldsymbol{x}_i{}^T\boldsymbol{\beta}), && \epsilon_i \text{ symmetrisch um 0 verteilt} \\
&= p(\boldsymbol{x}_i)
\end{aligned} \tag{3.4}$$

Dadurch ist der Parameter $p(\boldsymbol{x}_i)$ schon durch $F_{\epsilon_i}(\boldsymbol{x}_i{}^T\boldsymbol{\beta})$ auf eine gewisse Art und Weise mit den Daten verbunden. Die Zufallsgröße ϵ_i wird im Logit-Modell als logistisch verteilt angenommen. Somit lautet die Verteilungsfunktion

$$F_{\epsilon_i}(e) = \frac{1}{1 + \exp(-e)} = \frac{\exp(e)}{1 + \exp(e)}. \tag{3.5}$$

Durch diese Annahme werden die Voraussetzungen für obige Umformungen in (3.4), eine stetige und um 0 symmetrische Verteilung, erfüllt.

Der Erwartungswert des Response kann also folgendermaßen geschrieben werden:

$$\begin{aligned}
\mathbb{E}(Y_i) &= p(\boldsymbol{x}_i) \\
&= F_{\epsilon_i}(\boldsymbol{x}_i{}^T\boldsymbol{\beta}) && \text{mit (3.4)} \\
&= \frac{\exp(\boldsymbol{x}_i{}^T\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i{}^T\boldsymbol{\beta})} && \text{mit (3.5)}
\end{aligned} \tag{3.6}$$

Also ist

$$Y_i \,|\, \boldsymbol{x}_i \sim Bin\left(1,\ p(\boldsymbol{x}_i) = \frac{\exp(\boldsymbol{x}_i{}^T\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i{}^T\boldsymbol{\beta})}\right)$$

Da die Link-Funktion den Erwartungswert mit den Daten verbinden soll, muss (3.6) ent-

sprechend umgeformt werden:

$$\begin{aligned} p(\boldsymbol{x}_i) &= \frac{\exp({\boldsymbol{x}_i}^T\boldsymbol{\beta})}{1+\exp({\boldsymbol{x}_i}^T\boldsymbol{\beta})} \\ p(\boldsymbol{x}_i) + p(\boldsymbol{x}_i)\exp({\boldsymbol{x}_i}^T\boldsymbol{\beta}) &= \exp({\boldsymbol{x}_i}^T\boldsymbol{\beta}) \\ p(\boldsymbol{x}_i) &= [1-p(\boldsymbol{x}_i)]\exp({\boldsymbol{x}_i}^T\boldsymbol{\beta}) \\ \frac{p(\boldsymbol{x}_i)}{1-p(\boldsymbol{x}_i)} &= \exp({\boldsymbol{x}_i}^T\boldsymbol{\beta}) \\ \log\left(\frac{p(\boldsymbol{x}_i)}{1-p(\boldsymbol{x}_i)}\right) &= {\boldsymbol{x}_i}^T\boldsymbol{\beta} \\ \Rightarrow g(p(\boldsymbol{x}_i)) = {\boldsymbol{x}_i}^T\boldsymbol{\beta} &= \log\left(\frac{p(\boldsymbol{x}_i)}{1-p(\boldsymbol{x}_i)}\right) \end{aligned} \tag{3.7}$$

Die *Link-Funktion* der Binomialverteilung ist also die Logit-Funktion. Die Voraussetzung dafür ist die Verteilungsannahme der logistischen Verteilung für den Messfehler ϵ_i. Die Modellgleichung des Logit-Modells lautet

$$\log\left(\frac{p(\boldsymbol{x}_i)}{1-p(\boldsymbol{x}_i)}\right) = {\boldsymbol{x}_i}^T\boldsymbol{\beta} \tag{3.8}$$

3.1.3. Schätzen im Logit-Modell

Die Schätzung im Logit-Modell erfolgt z. B. über das Prinzip der kleinsten Quadrate (KQ). Dazu wird die Quadratsumme der Differenzen aus wahrem Wert y_i und Vorhersage $\widehat{y}_i$ bezüglich der Koeffizienten $\boldsymbol{\beta}$ minimiert. Die zweite Potenz der Differenzen verhindert, dass sich Abweichungen nach unten und nach oben gegenseitig aufheben können.

$$\begin{aligned} KQ(\beta) &= \sum_{i=1}^{n}(y_i-\widehat{y}_i)^2 \\ &= \sum_{i=1}^{n}[y_i-\mathbb{E}(Y_i)]^2 \\ &= \sum_{i=1}^{n}[y_i-p(\boldsymbol{x}_i)]^2 \\ &= \sum_{i=1}^{n}\left(y_i-\frac{\exp({\boldsymbol{x}_i}^T\boldsymbol{\beta})}{1+\exp({\boldsymbol{x}_i}^T\boldsymbol{\beta})}\right)^2 \longrightarrow \min_{\boldsymbol{\beta}} \end{aligned} \tag{3.9}$$

Diese Art von Schätzung berücksichtigt allerdings nicht die Varianzheterogenität des Logit-Modells. Diese kommt zustande, da $\mathbb{V}ar(Y_i) = p(\boldsymbol{x}_i)\,[1-p(\boldsymbol{x}_i)]$ für jede Beobachtung i verschieden sein kann. Dieses Problem umgeht die Schätzung nach dem Maximum-Likelihood-

Prinzip (Fahrmeir et al., 2007).

Die Likelihood lautet

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} L_i(\boldsymbol{\beta}) = \prod_{i=1}^{n} p(\boldsymbol{x}_i)^{y_i} \left(1 - p(\boldsymbol{x}_i)\right)^{1-y_i}$$

Die log-Likelihood ist wegen (3.8) und $1 - p(\boldsymbol{x}_i) = \frac{1+\exp(\boldsymbol{x}_i{}^T\boldsymbol{\beta})-\exp(\boldsymbol{x}_i{}^T\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i{}^T\boldsymbol{\beta})} = \frac{1}{1+\exp(\boldsymbol{x}_i{}^T\boldsymbol{\beta})}$

$$\begin{aligned} l(\boldsymbol{\beta}) = \sum_{i=1}^{n} l_i(\boldsymbol{\beta}) &= \sum_{i=1}^{n} y_i \log\left(p(\boldsymbol{x}_i)\right) + (1 - y_i) \log\left(1 - p(\boldsymbol{x}_i)\right) \\ &= \sum_{i=1}^{n} y_i \log\left(\frac{p(\boldsymbol{x}_i)}{1 - p(\boldsymbol{x}_i)}\right) + \log\left(1 - p(\boldsymbol{x}_i)\right) \\ &= \sum_{i=1}^{n} y_i \boldsymbol{x}_i{}^T\boldsymbol{\beta} + \log\left(\frac{1}{1 + \exp(\boldsymbol{x}_i{}^T\boldsymbol{\beta})}\right) \\ &= \sum_{i=1}^{n} y_i \boldsymbol{x}_i{}^T\boldsymbol{\beta} - \log\left(1 + \exp(\boldsymbol{x}_i{}^T\boldsymbol{\beta})\right) \end{aligned}$$

Die Scorefunktion ist die Ableitung der log-Likelihood nach $\boldsymbol{\beta}$:

$$s(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \boldsymbol{x}_i - \frac{1}{1 + \exp(\boldsymbol{x}_i{}^T\boldsymbol{\beta})} \cdot \exp(\boldsymbol{x}_i{}^T\boldsymbol{\beta}) \boldsymbol{x}_i = \sum_{i=1}^{n} \boldsymbol{x}_i \left(y_i - p(\boldsymbol{x}_i)\right)$$

Die ML-Gleichung ist

$$s(\widehat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} \boldsymbol{x}_i \left(y_i - \frac{\exp(\boldsymbol{x}_i{}^T\widehat{\boldsymbol{\beta}})}{1 + \exp(\boldsymbol{x}_i{}^T\widehat{\boldsymbol{\beta}})} \right) = \mathbf{0}$$

und hängt somit nicht-linear von $\boldsymbol{\beta}$ ab. Die P Gleichungen der P Parameter werden i. A. iterativ mit Fisher-Scoring gelöst. Ausgehend von der Taylor-Entwicklung um den Schätzer $\widehat{\boldsymbol{\beta}}$ erhält man

$$\begin{aligned} s\left(\widehat{\boldsymbol{\beta}}\right) &\approx s\left(\boldsymbol{\beta}^{(k)}\right) + \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(k)}\right) \cdot \boldsymbol{F}\left(\boldsymbol{\beta}^{(k)}\right) \stackrel{!}{=} 0 \\ -\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(k)}\right) &= s\left(\boldsymbol{\beta}^{(k)}\right) \cdot \boldsymbol{F}\left(\boldsymbol{\beta}^{(k)}\right)^{-1} \\ \widehat{\boldsymbol{\beta}} &= \boldsymbol{\beta}^{(k)} - s\left(\boldsymbol{\beta}^{(k)}\right) \cdot \boldsymbol{F}\left(\boldsymbol{\beta}^{(k)}\right)^{-1} \end{aligned}$$

mit

$$\begin{aligned}
\boldsymbol{F}(\boldsymbol{\beta}) &= -\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} l(\boldsymbol{\beta}) = -\frac{\partial}{\partial\boldsymbol{\beta}} \sum_{i=1}^{n} \boldsymbol{x}_i \left(y_i - \frac{\exp(\boldsymbol{x}_i{}^T\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i{}^T\boldsymbol{\beta})} \right) \\
&= -\sum_{i=1}^{n} -\boldsymbol{x}_i \frac{\left[1+\exp(\boldsymbol{x}_i{}^T\boldsymbol{\beta})\right]\exp(\boldsymbol{x}_i{}^T\boldsymbol{\beta})\boldsymbol{x}_i{}^T - \exp(\boldsymbol{x}_i{}^T\boldsymbol{\beta})\exp(\boldsymbol{x}_i{}^T\boldsymbol{\beta})\boldsymbol{x}_i{}^T}{\left[1+\exp(\boldsymbol{x}_i{}^T\boldsymbol{\beta})\right]^2} \\
&= \sum_{i=1}^{n} \boldsymbol{x}_i \frac{\exp(\boldsymbol{x}_i{}^T\boldsymbol{\beta})\boldsymbol{x}_i{}^T}{\left[1+\exp(\boldsymbol{x}_i{}^T\boldsymbol{\beta})\right]^2} \\
&= \sum_{i=1}^{n} \boldsymbol{x}_i p(\boldsymbol{x}_i) \frac{1}{1+\exp(\boldsymbol{x}_i{}^T\boldsymbol{\beta})} \boldsymbol{x}_i{}^T \\
&= \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i{}^T p(\boldsymbol{x}_i)\left(1-p(\boldsymbol{x}_i)\right)
\end{aligned}$$

Mit einem Startwert $\widehat{\boldsymbol{\beta}}^{(0)}$, z. B. dem KQ-Schätzer, wird also

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} - s\left(\boldsymbol{\beta}^{(k)}\right) \cdot \boldsymbol{F}\left(\boldsymbol{\beta}^{(k)}\right)^{-1} \qquad k = 0, 1, 2, \ldots$$

so lange iteriert, bis ein Abbruchkriterium erfüllt ist, z. B. die Veränderung von einem Schritt zum nächsten nur noch sehr klein ist.

3.1.4. Interpretation der Schätzer im Logit-Modell

Ausgehend von einem einfachen Logit-Modell mit der Modellgleichung

$$\log\left(\frac{p(\boldsymbol{x})}{1-p(\boldsymbol{x})}\right) = \boldsymbol{x}^T\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 \tag{3.10}$$

mit

$$x_1 = \begin{cases} 0 & \text{Referenzpopulation, z. B. Frauen} \\ 1 & \text{z. B. Männer} \end{cases}$$

ist die Interpretation der Schätzer folgendermaßen (Tutz, 2000):

Die Gleichung (3.10) kann nach dem Parameter β_0 aufgelöst werden, falls $x_1 = 0$ gilt.

$$\log\left(\frac{p(x_1=0)}{1-p(x_1=0)}\right) = \beta_0 \tag{3.11}$$

Der Parameter β_0 gibt dann die *Logits*, also die logarithmierten Chancen der Referenzpo-

pulation an bzw.

$$\exp(\beta_0) = \frac{p(x_1 = 0)}{1 - p(x_1 = 0)} = \frac{p(x_1 = 0)}{p(x_1 = 1)}$$

ergibt die Chancen der Referenzpopulation, $y = 1$ als Response zu erhalten. Eine Chance ist das Verhältnis der Trefferwahrscheinlichkeit zur Wahrscheinlichkeit, dass kein Treffer eintritt. Ist die Chance $\exp(\beta_0)$ größer als 1 bzw. das Logit β_0 größer als 0, so ist es wahrscheinlicher, einen Treffer zu erzielen, also bei der Beobachtung von $x_1 = 0$ als Response $y = 1$ zu haben, als keinen Treffer zu erzielen. Umgekehrt ist es wahrscheinlicher, $y = 0$ zu erhalten, falls $\exp(\beta_0) < 1$ ist bzw. β_0 einen negativen Wert hat. Ist die Chance etwa 1, so beläuft sich die Trefferwahrscheinlichkeit und somit auch ihre Gegenwahrscheinlichkeit auf ca. 0.5.

Falls $x_1 = 1$ ist, so ist β_1 in der Modellgleichung (3.10) enthalten. Löst man die Gleichung dann nach β_1 auf, so erhält man

$$\log\left(\frac{p(x_1 = 1)}{1 - p(x_1 = 1)}\right) = \beta_0 + \beta_1 x_1$$

$$\log\left(\frac{p(x_1 = 1)}{1 - p(x_1 = 1)}\right) - \beta_0 = \beta_1$$

$$\log\left(\frac{p(x_1 = 1)}{1 - p(x_1 = 1)}\right) - \log\left(\frac{p(x_1 = 0)}{1 - p(x_1 = 0)}\right) = \beta_1 \qquad \text{mit (3.11)}$$

Der Parameter β_1 gibt eine Differenz von Logits an. Wendet man die Exponentialfunktion darauf an, ergibt sich eine so genannte *Odds Ratio*, also ein Verhältnis von Chancen. Deswegen wird die Odds Ratio auch "relative Chance" genannt. Diese kann man als Assoziationsmaß interpretieren, ähnlich dem χ^2-Koeffizienten.

$$\beta_1 = \log\left(\frac{\frac{p(x_1=1)}{1-p(x_1=1)}}{\frac{p(x_1=0)}{1-p(x_1=0)}}\right)$$

$$\exp(\beta_1) = \frac{\frac{p(x_1=1)}{1-p(x_1=1)}}{\frac{p(x_1=0)}{1-p(x_1=0)}}$$

Ist nun $\beta_1 \approx 0$, so ist die Chance etwa 1 und es liegt kein Zusammenhang zwischen x_1 und dem Response y vor. Falls $\beta_1 > 0$ ist, ergibt sich eine Odds Ratio > 1, d. h. die Chance der Population mit $x_1 = 1$, im Beispiel die Männer, ist größer als für die Referenzpopulation (also für die Frauen). Umgekehrt gilt: $\beta_1 < 0$ ergibt eine Odds Ratio kleiner als 1, also ist die Chance auf den Response $y = 1$ der Referenzpopulation größer als diejenige der anderen.

In der in Kapitel 4 vorgestellten Analyse werden auch Interaktionen zwischen kategoria-

len Variablen in das Modell aufgenommen. Ein einfaches Logit-Modell mit einer solchen Interaktion ist folgendes:

$$\log\left(\frac{p(\boldsymbol{x})}{1-p(\boldsymbol{x})}\right) = \boldsymbol{x}^T\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 \tag{3.12}$$

mit

$$x_1 = \begin{cases} 0 & \text{z. B. Frauen} \\ 1 & \text{z. B. Männer} \end{cases}$$

und

$$x_2 = \begin{cases} 0 & \text{z. B. jung} \\ 1 & \text{z. B. alt} \end{cases}$$

Die Referenzpopulation ist dann $x_1 = 0 \cap x_2 = 0$, im Beispiel also junge Frauen.

Der Parameter β_0 gibt wieder die Logits der Referenzpopulation an, da für dessen Berechnung $x_1 = 0$ und $x_2 = 0$ erfüllt sein muss.

$$\log\left(\frac{p(x_1 = 0, x_2 = 0)}{1 - p(x_1 = 0, x_2 = 0)}\right) = \beta_0 \tag{3.13}$$

Um β_1 berechnen zu können, muss $x_1 = 1$ und $x_2 = 0$ sein. Dann erhält man β_1 auf die gleiche Weise wie oben in Modell (3.10) als

$$\begin{aligned} \log\left(\frac{p(x_1 = 1, x_2 = 0)}{1 - p(x_1 = 1, x_2 = 0)}\right) &= \beta_0 + \beta_1 \\ \log\left(\frac{p(x_1 = 1, x_2 = 0)}{1 - p(x_1 = 1, x_2 = 0)}\right) - \beta_0 &= \beta_1 \\ \log\left(\frac{\frac{p(x_1=1,x_2=0)}{1-p(x_1=1,x_2=0)}}{\frac{p(x_1=0,x_2=0)}{1-p(x_1=0,x_2=0)}}\right) &= \beta_1 \qquad \text{mit (3.13)} \end{aligned} \tag{3.14}$$

Der Parameter β_1 ergibt also wieder eine logarithmierte Odds Ratio wie im einfachen Modell (3.10) ohne Interaktion. Die Größe $\exp(\beta_1)$ kann wieder als Assoziationsmaß interpretiert werden. Allerdings kann sie nur Aussagen über die Subpopulation mit $x_2 = 0$ machen. Doch in dieser Untergruppe von Beobachtungen ist die Interpretation die gleiche wie im einfachen Modell (3.10): Ein Wert größer als 1 besagt eine größere Chance auf $y = 1$ für die Subpopulation mit $x_1 = 1$ (und $x_2 = 0$) im Vergleich zur Subpopulation mit $x_1 = 0$ (und $x_2 = 0$).

Genau umgekehrt ergibt sich β_2, wenn $x_1 = 0$ und $x_2 = 1$ ist:

$$\begin{aligned} \log\left(\frac{p(x_1=0,x_2=1)}{1-p(x_1=0,x_2=1)}\right) - \beta_0 &= \beta_2 \\ \log\left(\frac{\frac{p(x_1=0,x_2=1)}{1-p(x_1=0,x_2=1)}}{\frac{p(x_1=0,x_2=0)}{1-p(x_1=0,x_2=0)}}\right) &= \beta_2 \qquad \text{mit (3.13)} \end{aligned} \tag{3.15}$$

Nach Anwenden der Exponentialfunktion ist $\exp(\beta_2)$ die Odds Ratio zwischen den Populationen mit $x_2 = 1$ bzw. mit $x_2 = 0$. Auch hier ist die Interpretation analog zum einfachen Modell ohne Interaktion mit der Gleichung (3.10) – lediglich bezogen auf die Subpopulationen, die entstehen, wenn $x_1 = 0$ festgehalten wird und x_2 variiert.

Die Interaktion wird folgendermaßen hergeleitet (als verkürzende Schreibweise wird $p_{ij} = p(x_1 = i, x_2 = j)$ verwendet):

$$\begin{aligned} \log\left(\frac{p_{11}}{1-p_{11}}\right) - (\beta_0 + \beta_1 + \beta_2) &= \beta_3 \\ \log\left(\frac{p_{11}}{1-p_{11}}\right) - \left[\log\left(\frac{p_{00}}{1-p_{00}}\right) + \log\left(\frac{\frac{p_{10}}{1-p_{10}}}{\frac{p_{00}}{1-p_{00}}}\right)\right] - \beta_2 &= \beta_3 \\ \log\left(\frac{p_{11}}{1-p_{11}}\right) - \log\left(\frac{p_{00}}{1-p_{00}} \cdot \frac{\frac{p_{10}}{1-p_{10}}}{\frac{p_{00}}{1-p_{00}}}\right) - \beta_2 &= \beta_3 \\ \log\left(\frac{p_{11}}{1-p_{11}}\right) - \log\left(\frac{p_{10}}{1-p_{10}}\right) - \beta_2 &= \beta_3 \\ \log\left(\frac{\frac{p_{11}}{1-p_{11}}}{\frac{p_{10}}{1-p_{10}}}\right) - \beta_2 &= \beta_3 \\ \log\left(\frac{\frac{p_{11}}{1-p_{11}}}{\frac{p_{10}}{1-p_{10}}}\right) - \log\left(\frac{\frac{p_{01}}{1-p_{01}}}{\frac{p_{00}}{1-p_{00}}}\right) &= \beta_3 \qquad \text{mit (3.15)} \end{aligned}$$

Der exponentialtransformierte Regressionskoeffizient $\exp(\beta_3)$ der Interaktion ergibt ein Verhältnis von Odds Ratios, also ein Verhältnis von Verhältnissen von Chancen. Die Odds Ratio der ersten Population mit $x_1 = 1$ ist folglich größer ($\exp(\beta_3) > 1$) oder kleiner ($\exp(\beta_3) < 1$) als die Odds Ratio in der zweiten Population mit $x_1 = 0$. Es variiert also jeweils x_2 in den Odds Ratios, während $x_1 = 1$ bzw. $x_1 = 0$ die Subpopulation bildet, für die die Odds Ratio in Bezug auf x_2 eine Aussage machen kann. Falls $\beta_3 = 0$ bzw. $\exp(\beta_3) = 1$ geschätzt wird, hat die Interaktion keinen Einfluss auf den Response, denn die Odds Ratios sind gleich groß. Im Beispiel würde es bedeuten, dass der Zusammenhang zwischen dem Alter x_2 und dem Response nicht geschlechtsspezifisch (x_1) ist. Die Chance auf $y = 1$ ändert sich zwar mit dem Alter, aber es kommt zu keiner Änderung, wenn man statt Männern Frauen betrachtet.

Genauso gilt dann: Der Zusammenhang zwischen dem Geschlecht und dem Response ist nicht altersspezifisch, denn

$$\begin{aligned}
\log\left(\frac{p_{11}}{1-p_{11}}\right) - (\beta_0+\beta_2+\beta_1) &= \beta_3 \\
\log\left(\frac{p_{11}}{1-p_{11}}\right) - \log\left(\frac{p_{00}}{1-p_{00}} \cdot \frac{\frac{p_{01}}{1-p_{01}}}{\frac{p_{00}}{1-p_{00}}}\right) - \beta_1 &= \beta_3 \\
\log\left(\frac{\frac{p_{11}}{1-p_{11}}}{\frac{p_{01}}{1-p_{01}}}\right) - \log\left(\frac{\frac{p_{10}}{1-p_{10}}}{\frac{p_{00}}{1-p_{00}}}\right) &= \beta_3 \qquad \text{mit (3.14)}
\end{aligned}$$

Wenn $\exp(\beta_3) \neq 1$ ist, unterscheiden sich die Odds Ratios, welche den Zusammenhang zwischen x_1 und dem Response y angeben – je nachdem, ob $x_2 = 1$ oder $x_2 = 0$ beobachtet wird. Falls $\exp(\beta_3)$ größer als 1 ist, ist die Odds Ratio der Subpopulation mit $x_2 = 1$ größer als die Odds Ratio der Population mit $x_2 = 0$ in Bezug auf x_1. Falls $\exp(\beta_3) < 1$ ist, hat die Population mit $x_2 = 1$ eine größere relative Chance auf einen Treffer ($y = 1$) als die andere Subpopulation mit $x_2 = 0$.
Das Verhältnis von Odds Ratios ist demzufolge symmetrisch.

In der Analyse aus Kapitel 4 kommen auch stetige Kovariablen vor. Das einfache Logit-Modell (3.16) soll zur Erklärung der Interpretation dienen.

$$\log\left(\frac{p(\boldsymbol{x})}{1-p(\boldsymbol{x})}\right) = \boldsymbol{x}^T\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 \tag{3.16}$$

mit

$$x_1 = \text{z. B. Alter in Jahren}$$

Die Referenzpopulation ist hier ebenso $x_1 = 0$, im Beispiel also Personen mit 0 Jahren.

Wie bei kategorialen Einflussgrößen gibt β_0 die Logits der Referenzpopulation an. Nimmt die stetige Variable x_1 einen bestimmten Wert ${x_1}^*$ an, so ergibt sich

$$\begin{aligned}
\log\left(\frac{p(x_1 = {x_1}^*)}{1-p(x_1 = {x_1}^*)}\right) &= \beta_0 + \beta_1 {x_1}^* \\
\log\left(\frac{p(x_1 = {x_1}^*)}{1-p(x_1 = {x_1}^*)}\right) - \beta_0 &= \beta_1 {x_1}^* \\
\log\left(\frac{p(x_1 = {x_1}^*)}{1-p(x_1 = {x_1}^*)}\right) - \log\left(\frac{p(x_1 = 0)}{1-p(x_1 = 0)}\right) &= \beta_1 {x_1}^*
\end{aligned}$$

Die Odds Ratio kann durch Exponentialtransformation gewonnen werden:

$$\frac{\frac{p(x_1=x_1{}^*)}{1-p(x_1=x_1{}^*)}}{\frac{p(x_1=0)}{1-p(x_1=0)}} = \exp(\beta_1 x_1{}^*) = \exp(\beta_1)^{x_1{}^*}$$

Falls die Variable x_1 um eine Einheit zunimmt, also den Wert $x_1{}^* + 1$ hat, so ändert sich die Odds Ratio um einen multiplikativen Term, nämlich $\exp(\beta_1)$:

$$\begin{aligned}
\log\left(\frac{p(x_1 = x_1{}^* + 1)}{1 - p(x_1 = x_1{}^* + 1)}\right) &= \beta_0 + \beta_1 \cdot (x_1{}^* + 1) \\
\log\left(\frac{p(x_1 = x_1{}^* + 1)}{1 - p(x_1 = x_1{}^* + 1)}\right) - \beta_0 &= \beta_1 x_1{}^* + \beta_1 \\
\frac{\frac{p(x_1=x_1{}^*+1)}{1-p(x_1=x_1{}^*+1)}}{\frac{p(x_1=0)}{1-p(x_1=0)}} &= \exp(\beta_1 x_1{}^* + \beta_1) = \exp(\beta_1 x_1{}^*)\exp(\beta_1) \\
&= \exp(\beta_1)^{x_1{}^*} \cdot \exp(\beta_1)
\end{aligned}$$

Ist $\exp(\beta_1)$ kleiner als 1, so verringert sich die Odds Ratio mit zunehmendem Alter jedes Jahr um den Faktor $\exp(\beta_1)$. Ist $\exp(\beta_1)$ größer als 1, nimmt die Odds Ratio mit steigendem Alter jedes Jahr um den Faktor $\exp(\beta_1)$ zu.

Setzt man die beiden Odds Ratios zueinander ins Verhältnis, erhält man

$$\begin{aligned}
\frac{\frac{\frac{p(x_1=x_1{}^*+1)}{1-p(x_1=x_1{}^*+1)}}{\frac{p(x_1=0)}{1-p(x_1=0)}}}{\frac{\frac{p(x_1=x_1{}^*)}{1-p(x_1=x_1{}^*)}}{\frac{p(x_1=0)}{1-p(x_1=0)}}} &= \frac{\exp(\beta_1)^{x_1{}^*} \cdot \exp(\beta_1)}{\exp(\beta_1)^{x_1{}^*}} \\
\frac{\frac{p(x_1=x_1{}^*+1)}{1-p(x_1=x_1{}^*+1)}}{\frac{p(x_1=x_1{}^*)}{1-p(x_1=x_1{}^*)}} &= \exp(\beta_1)
\end{aligned}$$

Die Größe $\exp(\beta_1)$ gibt also – neben der multiplikativen Veränderung der Odds Ratio – statt wie üblich die Odds Ratio bezüglich der Referenzpopulation die Odds Ratio zwischen den Populationen mit $x_1 = x_1{}^* + 1$ und $x_1 = x_1{}^*$ an.

3.2. Gemischte Modelle

Gemischte Modelle werden in der Regel für longitudinale Daten verwendet. Sie sind also eine Modellklasse für Zeitreihen und wiederholte Messungen.
Ein typisches Beispiel dafür ist die Messung des Blutdrucks im Tagesverlauf. Die Messungen einer Person sind untereinander tendenziell ähnlicher als die Messungen von ver-

schiedenen Personen. Jede Person hat ein eigenes Level, um das ihre Blutdruck-Messungen schwanken. Dieser subjektspezifische Effekt kann neben den festen Effekten der möglichen Einflussgrößen ins Modell aufgenommen werden. Man geht davon aus, dass er zufällig ist und nimmt deswegen eine Verteilung für ihn an.

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + b_i + \boldsymbol{\epsilon}_i \tag{3.17}$$

Für eine einzelne Beobachtung lautet das Modell:

$$y_{it} = \boldsymbol{x}_{it}{}^T\boldsymbol{\beta} + b_i + \epsilon_{it}$$

In der Modellgleichung (3.17) bezeichnet $\boldsymbol{X}_i$ die Designmatrix der (festen) Effekte und $\boldsymbol{\beta}$ deren Koeffizientenvektor wie im (generalisierten) linearen Modell. Der Vektor $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iT})^T$ beinhaltet die Messungen zu den Zeitpunkten 1 bis T für Subjekt i. Der zufällige Effekt ist durch b_i dargestellt. I. d. R. nimmt man an, dass b_i normalverteilt ist mit Erwartungswert 0 $\left(b_i \overset{\text{iid}}{\sim} N(0,\ d^2)\right)$. Der Messfehler $\boldsymbol{\epsilon}_i$ ist unabhängig von den zufälligen Effekten b_i, jedoch genau wie dieser als normalverteilt angenommen $\left(\boldsymbol{\epsilon}_i \overset{\text{iid}}{\sim} N(0,\ \sigma^2\boldsymbol{I})\right)$.
Der Response ist somit auch normalverteilt mit

$$\boldsymbol{Y}_i \,|\, \boldsymbol{X}_i, b_i \sim N\left(\boldsymbol{X}_i\boldsymbol{\beta} + b_i,\ (d^2 + \sigma^2)\boldsymbol{I}\right)$$

Da in diesem Modell sowohl feste als auch zufällige Effekte vorkommen, wird es als "gemischtes Modell" bezeichnet. Auf die Schätzung der Parameter $\boldsymbol{\beta}$ und $\boldsymbol{b}$ wird in Kapitel 3.3.3.2 näher eingegangen (vgl. dort Schätzung von $\boldsymbol{\beta}$ und $\boldsymbol{r}$).

Durch die Aufnahme eines zufälligen Effektes kann man modellieren, dass jedes Subjekt einen eigenen Effekt auf den Response $\boldsymbol{y}_i$ hat. Gemischte Modelle können durch Definition einer geeigneten Designmatrix für die zufälligen Effekte erweitert werden. Dann ist es auch möglich, subjektspezifische Wachstumsraten o. Ä. aufzunehmen.

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i b_i + \boldsymbol{\epsilon}_i$$

Des Weiteren ist es möglich, wie im linearen Modell, auch im gemischten Modell andere Verteilungen als die Normalverteilung zugrunde zu legen. Man spricht dann von einem "generalisierten linearen gemischten Modell".
Nimmt man wiederum eine Binomialverteilung des Response an, so lautet nach denselben

Umformungen wie in Kap. 3.1.2 die Modellgleichung

$$\log\left(\frac{p(\boldsymbol{x}_i)}{1-p(\boldsymbol{x}_i)}\right) = {\boldsymbol{x}_i}^T\boldsymbol{\beta} + b_i$$

3.3. Räumliche Statistik

3.3.1. Einführung in die räumliche Statistik

Sowohl die Screening-Daten als auch die Datensätze zu den möglichen Einflussgrößen liegen auf Basis räumlicher Einheiten wie z. B. Gemeinden oder PLZ-Bezirke vor. Somit handelt es sich um ein unregelmäßiges Gitter von Daten. Regelmäßige Gitterdaten wären z. B. die Grauwerte der Pixel eines digitalen Bildes (Schmid, 2010), da die Pixel regelmäßig auf die Länge und Breite des Bildes verteilt sind.

In der räumlichen Statistik werden die Daten als Realisierung eines räumlichen stochastischen Prozesses angesehen (Schmid, 2010).

3.3.1.1. Räumliche stochastische Prozesse

Die für Gitterdaten üblicherweise verwendete Klasse von stochastischen Prozessen besitzen die so genannte Markov-Eigenschaft. Sie ist eine Erweiterung von Markov-Ketten in höhere Dimensionen (Schmid, 2010).

Markov-Ketten sind der einfachste Fall eines stochastischen Prozesses. Sie sind jeweils eine Folge von Zufallsvariablen Y_n, deren Verteilungen bedingt unabhängig sind. Die Wahrscheinlichkeit, dass im n-ten Schritt der Zustand y_n eintritt, gegeben alle vorherigen Realisationen 1 bis $n-1$, ist gleich der Wahrscheinlichkeit, dass im n-ten Schritt der Zustand y_n eintritt, gegeben die unmittelbar vorhergehende Realisation $n-1$:

$$\mathbb{P}(Y_n = y_n \,|\, y_0, \ldots, y_{n-1}) = \mathbb{P}(Y_n = y_n \,|\, y_{n-1})$$

Durch die Annahme, dass diese Eigenschaft nicht von n abhängig ist, wird der Markov-Prozess stationär. (Meintrup und Schäffler, 2005)

Die Erweiterung auf höher-, insbesondere zweidimensionale Räume besagt, dass die Verteilung der aktuell betrachteten Zufallsvariable, auf eine Nachbarschaft bedingt, unabhängig von den Verteilungen der anderen Zufallsvariablen ist. Man bezieht sich nicht mehr nur auf den "Vorgänger" wie bei der eindimensionalen Markov-Kette, sondern betrachtet die umliegenden Y_i. Dies kann auf verschiedene Art und Weise geschehen. Beispiele für

verschiedene Nachbarschaftsstrukturen in einem regelmäßigen Gitter sind in Abb. 3.1 zu sehen. Bei einem unregelmäßigen Gitter wird die Nachbarschaft in der Regel so definiert, dass zwei Einheiten mit mindestens einem gemeinsamen Grenzpunkt Nachbarn sind. Sich seöbst kann eine Einheit in beiden Gitterarten kein Nachbar sein.

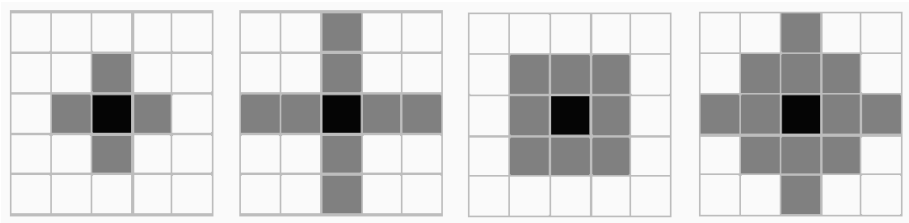

Abbildung 3.1.: Beispiele für Nachbarschaftsstrukturen. Der schwarze Punkt ist der betrachtete Gitterpunkt, die grauen Punkte stellen die dazu gehörige Nachbarschaft dar. Entnommen aus Schmid (2010).

Bei Verwendung von d Dimensionen bezeichnet $s \in \mathbb{R}^d$ einen Ort im d-dimensionalen Raum. Dann ist Y_s die dort lokalisierte Zufallsvariable. Die Nachbarschaft von s wird mit $N(s)$ bezeichnet. Somit ergibt sich als Markov-Eigenschaft im d-dimensionalen Raum:

$$\mathbb{P}[Y_s = y_s \,|\, y_{s'}, s \neq s' \in \mathbb{R}^d] = \mathbb{P}[Y_s = y_s \,|\, y_{s'}, s \neq s' \in N(s)] \tag{3.18}$$

Die Wahrscheinlichkeit, dass die Zufallsvariable Y an der Stelle s die Realisation y_s ergibt, ist, gegeben die Realisationen im übrigen Raum, gleich der Wahrscheinlichkeit, dass die Zufallsvariable Y an der Stelle s die Realisation y_s ergibt, gegeben die Realisationen in der Nachbarschaft um s.

3.3.1.2. Unterschied zwischen unstrukturierten und strukturierten räumlichen Effekten

Bei der Analyse räumlicher Daten kann man die Nachbarschaft mit einbeziehen oder nicht. Falls der Modellierung die Annahme zugrunde gelegt werden soll, dass benachbarte Regionen einen ähnlichen Effekt haben, nimmt man einen strukturierten räumlichen Effekt in das Modell auf. Das geschieht meist durch Gauß-Markov-Zufallsfelder (siehe Kap. 3.3.3.1). In diesem Fall ist die Nachbarschaftsstruktur von Belang.
Nimmt man dagegen an, alle Regionen sind voneinander unabhängig, modelliert man einen unstrukturierten räumlichen Effekt. Dies ist v. a. dann sinnvoll, wenn man davon ausgeht, dass z. B. die Nähe zu einer Großstadt keinen Einfluss auf die Werte der umliegenden Regionen hat.

In der Regel erhält man durch einen strukturierten räumlichen Effekt eine geglättete Variante des unstrukturierten räumlichen Effekts. Denn dann wird angenommen, dass z. B. die Nähe zur Großstadt eine Auswirkung auf die umgebenden Regionen hat und diese ähnliche Werte aufweisen.

3.3.2. Unstrukturierter räumlicher Effekt

Die Modellierung eines unstrukturierten räumlichen Effekts erfolgt durch Aufnahme eines zufälligen Effekts pro räumlicher Einheit. Dadurch erhält man ein "gemischtes Modell". Die Nachbarschaftsstruktur hat in diesem Fall keinen Einfluss.

Überträgt man das gemischte Modell aus Kapitel 3.2 auf räumliche Daten, kann man als "Wiederholungen" mehrere Beobachtungen aus der gleichen räumlichen Einheit statt mehrerer Beobachtungen über die Zeit haben. Durch die Aufnahme eines zufälligen Effekts werden die u. U. verschiedenen Levels der räumlichen Einheiten im Modell abgebildet. Der zufällige Effekt entspricht dem geschätzten räumlichen Effekt der zugehörigen Beobachtungseinheit und gleichzeitig der Abweichung dieser Beobachtung vom globalen Mittelwert. Die Modellgleichung (3.17) ändert sich dadurch leicht. Das Modell lautet nun für eine einzelne, nämlich die i-te Beobachtung aus Region s ($i = 1, \ldots, n_s$)

$$y_{is} = \boldsymbol{x}_{is}{}^T \boldsymbol{\beta} + b_s + \epsilon_{is}$$

Der zufällige Effekt ist nun nicht mehr auf die Beobachtung gezogen, sondern auf die räumliche Einheit. An der Schätzung etc. ändert sich nichts.

3.3.3. Strukturierter räumlicher Effekt

3.3.3.1. Gauß-Markov-Zufallsfelder

Bei Aufnahme eines unstrukturierten räumlichen Effekts hat die Nachbarschaftsstruktur der Markov-Eigenschaft (3.18) keine Auswirkung. Will man den räumlichen Effekt jedoch mit einer gewissen Struktur ins Modell aufnehmen, geschieht dies über die so genannte Nachbarschaftsmatrix. Sie bildet die angenommene Nachbarschaftsstruktur ab. Dies wird in der Praxis meist durch ein Gauß-Markov-Zufallsfeld realisiert. Im Folgenden wird die Aufnahme eines strukturierten räumlichen Effekts in ein lineares Modell erklärt. In Kapitel 3.3.3.2 und 3.3.3.3 wird dann die Schätzung eines GMRF in einem generalisierten linearen Modell erläutert.

Für ein Gauß-Markov-Zufallsfeld (engl. Gaussian Markov Random Field, Abkürzung:

GMRF) wird die zuvor erwähnte Markov-Eigenschaft (3.18) sowie eine Normalverteilung angenommen. Der räumliche Effekt $\boldsymbol{R}$ ist also normalverteilt mit einem gewissen Erwartungswert $\mathbb{E}(\boldsymbol{R})$ und einer Kovarianzmatrix, deren Inverse Σ^{-1} die Nachbarschafts- oder auch Präzisionsmatrix $\boldsymbol{P}$ darstellt. Die Präzisionsmatrix $\boldsymbol{P}$ ist dabei als symmetrisch und positiv definit angenommen.

$$\boldsymbol{R} \sim N\left(\mathbb{E}(\boldsymbol{R}),\ \boldsymbol{\Sigma} = \boldsymbol{P}^{-1}\right)$$

Für die Einträge der Präzisionsmatrix gilt: Die Einträge sind 0, wenn die betreffenden räumlichen Einheiten bedingt unabhängig sind ($p_{ss'} = 0 \Leftrightarrow r_s \perp r_{s'} \,|\, \boldsymbol{r}_{-ss'}$ bzw. $p_{ss'} \neq 0 \Leftrightarrow s' \in N(s)$). Die Schreibweise $\boldsymbol{r}_{-ss'}$ bedeutet, dass der komplette Vektor $\boldsymbol{r}$ gemeint ist, außer das s-te und s'-te Element. Der Index s läuft über $1, \ldots, S$ und bezeichnet die räumlichen Einheiten.
Durch diese Einträge enthält die Präzisionsmatrix indirekt über die Markov-Eigenschaft die Nachbarschaftsstruktur der Beobachtungseinheiten.

Durch Angabe der bedingten Dichten ist ein Gauß-Markov-Zufallsfeld ebenfalls vollständig spezifiziert:

$$R_s \,|\, \boldsymbol{R}_{-s} \sim N\left(\mathbb{E}(R_s \,|\, \boldsymbol{R}_{-s}),\ \mathbb{V}ar(R_s \,|\, \boldsymbol{R}_{-s})\right),$$

wobei $\mathbb{E}(R_s \,|\, \boldsymbol{R}_{-s}) = \mathbb{E}(R_s) - \frac{1}{p_{ss}} \sum_{s' \in N(s)} p_{ss'}[r_{s'} - \mathbb{E}(R_{s'})]$ und $\mathbb{V}ar(R_s \,|\, \boldsymbol{R}_{-s}) = {\tau_s}^2 = \frac{1}{p_{ss}}$.
Der Erwartungswert ist durch den unbedingten Erwartungswert und einem gewichteten Mittelwert aus den Abweichungen der benachbarten Responsewerte von ihrem jeweiligen Erwartungswert beschrieben. Die Gewichte sind dabei durch die Nachbarschaftsbeziehung, also die Einträge in der Präzisionsmatrix, gegeben.

Üblicherweise wird nun als Erwartungswert-Vektor der Nullvektor angenommen. Somit reduziert sich der bedingte Erwartungswert etwas. Aus dieser Beziehung lässt sich Folgendes herleiten:

$$\begin{aligned}
\mathbb{E}(R_s \,|\, \boldsymbol{R}_{-s}) &= \sum_{s' \in N(s)} -\frac{1}{p_{ss}} p_{ss'} r_{s'} \\
\Longleftrightarrow \mathbb{E}(R_s \,|\, \boldsymbol{R}_{-s}) &= \sum_{s' \in N(s)} \underbrace{-{\tau_s}^2 p_{ss'}}_{=c_{ss'}} r_{s'} \\
\Longrightarrow p_{ss'} &= -\frac{c_{ss'}}{{\tau_s}^2}
\end{aligned}$$

Der Diagonaleintrag p_{ss} entsteht aus der bedingten Varianz:

$$\mathbb{V}ar(R_s \mid \boldsymbol{R}_{-s}) = {\tau_s}^2 = \frac{1}{p_{ss}}$$
$$\Longrightarrow p_{ss} = \frac{1}{{\tau_s}^2}$$

Laut Kneib (2009/2010) werden die Parameter $c_{ss'}$ und ${\tau_s}^2$ üblicherweise auf Gewichten basierend gewählt. So kann man $c_{ss'}$ durch einen Art Anteil am Gesamtgewicht für die Beobachtungseinheit s darstellen:

$$c_{ss'} = \lambda \frac{w_{ss'}}{w_{s+}}$$

mit $w_{s+} = \sum_{s' \in N(s)} w_{ss'}$ und $\lambda \in [0, 1]$. Die Varianz besteht aus einem konstanten Wert, der durch die Gesamtzahl der Gewichte aller Nachbarn geteilt wird:

$${\tau_s}^2 = \frac{\tau^2}{w_{s+}}$$

Die bedingte Varianz wird kleiner, je mehr Nachbarn die Beobachtungseinheit s hat. Dafür muss die Nachbarschaftsstruktur bekannt sein.
Durch die Annahme, dass sich die Parameter durch Gewichte ausdrücken lassen, vereinfachen sich die Einträge der Präzisionsmatrix weiter:

$$p_{ss'} = -\frac{c_{ss'}}{{\tau_s}^2} = -\frac{\lambda \frac{w_{ss'}}{w_{s+}}}{\frac{\tau^2}{w_{s+}}} = -\lambda \frac{w_{ss'}}{\tau^2}$$
$$p_{ss} = \frac{1}{{\tau_s}^2} = \frac{w_{s+}}{\tau^2}$$

Somit kann man zusammenfassend schreiben: $p_{ss'} = \frac{1}{\tau^2} k_{ss'}$ mit

$$k_{ss'} = \begin{cases} -\lambda w_{ss'} & s \neq s' \\ w_{s+} & s = s' \end{cases} \tag{3.19}$$

Ein weitere übliche Wahl ist $w_{ss'} = 1$. Daraus resultiert $w_{s+} = |N(s)|$.
Die Zusammenfassung (3.19) lässt sich auch auf die Matrix-Schreibweise übertragen, da $\frac{1}{\tau^2}$ ein konstanter Faktor ist. Also ist

$$\boldsymbol{P} = \frac{1}{\tau^2} \boldsymbol{K} \quad \text{bzw.} \quad \boldsymbol{\Sigma} = \boldsymbol{P}^{-1} = \tau^2 \boldsymbol{K}^{-1}$$

Wählt man $\lambda = 1$, so erhält man eine uneigentliche Normalverteilung, das GMRF ist "intrinsisch". Der Rang der Präzisionsmatrix ist $\text{rang}(\boldsymbol{P}) = S - 1 = \text{rang}(\boldsymbol{K})$. Somit lässt

sie sich nicht mehr invertieren, um die Kovarianzmatrix $\boldsymbol{\Sigma}$ zu erhalten. Deswegen wird in diesem Fall die Determinante von $\boldsymbol{P}$ als Produkt der Eigenwerte, welche nicht 0 sind, definiert (Schmid, 2010). Man erhält:

$$|\boldsymbol{\Sigma}^{-1}| = |\boldsymbol{P}| = \left|\frac{1}{\tau^2}\boldsymbol{K}\right| = \left(\tau^2\right)^{-\text{rang}(\boldsymbol{K})} \tag{3.20}$$

Durch die vorangegangenen Vereinfachungen und mit den gewählten Parametern $\lambda = 1$ und $w_{ss'} = 1$ lautet die bedingt unabhängige Verteilung der räumlichen Effekte:

$$R_s \,|\, \boldsymbol{R}_{-s} \sim N\left(\frac{1}{|N(s)|}\sum_{s'\in N(s)} r_{s'},\ \frac{\tau^2}{|N(s)|}\right) \tag{3.21}$$

Die gemeinsame Verteilungsfunktion im intrinsischen Gauß-Markov-Zufallsfeld lautet dann wegen Brook's Lemma (siehe Kap. E.1)

$$f_{\boldsymbol{R}}(\boldsymbol{r}) \propto \left(\tau^2\right)^{-\frac{1}{2}\text{rang}(\boldsymbol{K})} \exp\left(-\frac{1}{2\tau^2}\boldsymbol{r}^T\boldsymbol{K}\boldsymbol{r}\right) \tag{3.22}$$

d. h. $\boldsymbol{R} \sim N(0,\ \tau^2\boldsymbol{K}^{-1})$.

Die Erstellung der Präzisionsmatrix, welche die Nachbarschaftsstruktur widerspiegelt, wird am Beispiel der sieben Regierungsbezirke Bayerns demonstriert: In Abb. 3.2 ist eine Karte Bayerns dargestellt. Betrachtet man z. B. den Bezirk Niederbayern, so ergeben sich zwei Nachbarn: Oberbayern und Oberpfalz. Somit erhalten die Zellen (1, 2) bzw. (2, 1) für die Nachbarschaftsbeziehung Niederbayern – Oberbayern und (1, 4) bzw. (4, 1) für die Nachbarschaftsbeziehung Niederbayern – Oberpfalz der Matrix $\boldsymbol{K}$ den Eintrag $-\lambda w_{ss'}$ (siehe (3.19)). Im intrinsischen GMRF ist $\lambda = 1$, außerdem wurde $w_{ss'} = 1$ gewählt. Somit lauten die Einträge -1.
Die Diagonale wird mit w_{s+} gefüllt, was bei der vorangegangenen Wahl von $w_{ss'} = 1$ der Anzahl der Nachbarn entspricht. Der Eintrag in Zelle (1, 1) ist also 2 für die Nachbarn Oberbayern und Oberpfalz.
Das Resultat für die gesamte Karte ist in Tabelle 3.1 gezeigt.

Das GMRFwird als räumlicher Effekt neben den anderen Kovariablen ins Modell aufgenommen:

$$y_{is} = \boldsymbol{x}_{is}{}^T\boldsymbol{\beta} + r_s + \epsilon_{is}$$

mit $\boldsymbol{R} \sim N(0,\ \tau^2\boldsymbol{K}^{-1})$. Die Beobachtung is ist die i-te Beobachtung aus der Region s. Da die Nachbarschaftsstruktur $\boldsymbol{K}$ gegeben ist, muss für den räumlichen Effekt nur der Varianzparameter τ^2 geschätzt werden.

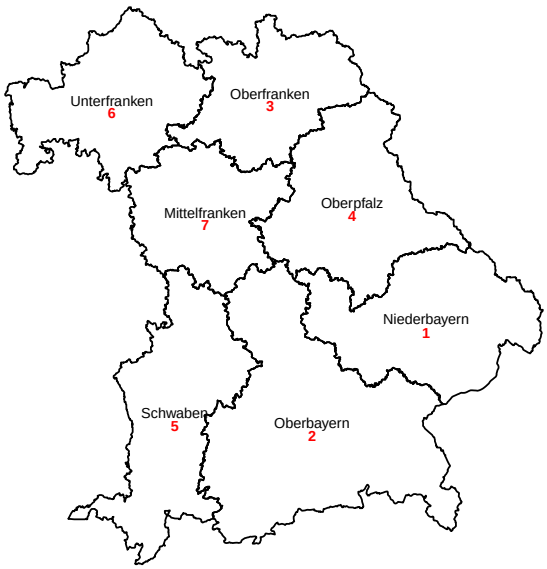

Abbildung 3.2.: Karte der Regierungsbezirke Bayerns für das Beispiel zu den Einträgen in der Präzisionsmatrix.

Tabelle 3.1.: $\boldsymbol{K}$-Matrix zur Karte der Regierungsbezirke Bayerns bei Annahme eines intrinsischen Gauß-Markov-Zufallsfeldes mit $w_{ss'} = 1$.

Bezirk	1	2	3	4	5	6	7
1	2	-1	0	-1	0	0	0
2	-1	4	0	-1	-1	0	-1
3	0	0	3	-1	0	-1	-1
4	-1	-1	-1	4	0	0	-1
5	0	-1	0	0	2	0	-1
6	0	0	-1	0	0	2	-1
7	0	-1	-1	-1	-1	-1	5

Natürlich ist es auch in generalisierten linearen Modellen möglich, räumlich strukturierte Effekte ins Modell aufzunehmen. Allerdings ist dann die Schätzung der Parameter erschwert.

3.3.3.2. Schätzung durch den Ansatz der penalisierten Quasi-Likelihood

Die Schätzung der räumlichen Effekte in einem generalisierten linearen Modell kann z. B. durch eine penalisierte Quasi-Likelihood (Abkürzung: PQL) geschehen. Für diese Schätzmethode werden die Daten so approximiert, dass man sie als einfaches lineares gemischtes Modell auffassen kann. Die Parameter werden durch iterative Verfahren geschätzt.

Aufstellen der Likelihood: Allgemein kann man den Erwartungswert und die Varianz eines generalisierten linearen Modells durch

$$\mathbb{E}(Y_i) = \mu_i = b'(\vartheta_i) \quad \text{und} \quad \mathbb{V}ar(Y_i) = \frac{\varphi}{\omega_i} b''(\vartheta_i) = \frac{\varphi}{\omega_i} v(\mu_i) \tag{3.23}$$

angeben (Fahrmeir et al., 2007). Je nach angenommener Verteilung ist die Varianzfunktion $v(\mu_i)$ anders spezifiziert. Der Varianzparameter φ ist u. U. unbekannt, bei einer Binomialverteilung jedoch konstant 1. Die Gewichte ω_i sind bekannt. Den Erwartungswert und die Varianz in einem Logit-Modell mit binomialverteilten Zielgrößen kann man durch

$$\mathbb{E}(Y_{is} \,|\, \boldsymbol{r}) = p(\boldsymbol{x}_{is}) \quad \text{und} \quad \mathbb{V}ar(Y_{is} \,|\, \boldsymbol{r}) = \frac{1}{\omega_{is}} v(p(\boldsymbol{x}_{is}))$$

ausdrücken.

Durch die Link-Funktion ist der Erwartungswert mit den Daten verbunden:

$$\begin{aligned} g(p(\boldsymbol{x}_{is})) &= \log\left(\frac{p(\boldsymbol{x}_{is})}{1 - p(\boldsymbol{x}_{is})}\right) \\ &= {\boldsymbol{x}_{is}}^T \boldsymbol{\beta} + r_s \\ g(\boldsymbol{p}) &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{r} \qquad\qquad \text{mit } g(\boldsymbol{p}) = [g(p(\boldsymbol{x}_{11})), \ldots, g(p(\boldsymbol{x}_{n_S S}))]^T \\ g^{-1}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{r}) &= \boldsymbol{p} \\ &= \mathbb{E}(\boldsymbol{Y} \,|\, \boldsymbol{r}) \end{aligned}$$

Allgemein wird angenommen, dass $\boldsymbol{R}$ einer Normalverteilung mit $\boldsymbol{R} \sim N(\boldsymbol{0},\ \boldsymbol{D}(\boldsymbol{\vartheta}))$ folgt (Breslow und Clayton, 1993). Der Parameter $\boldsymbol{\vartheta}$ ist unbekannt und muss geschätzt werden. In einem Logit-Modell mit räumlich strukturierten Effekten lautet die Verteilungsannahme der regionalen Effekte

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

Der unbekannte Parametervektor $\boldsymbol{\vartheta}$ reduziert sich auf den Skalar τ^2, da die Nachbarschaftsmatrix $\boldsymbol{K}$ gegeben ist.

Die Likelihood zum vorliegenden Problem lautet

$$L(\boldsymbol{\beta}, \tau^2) = \int f(\boldsymbol{y} \,|\, \boldsymbol{\beta}, \boldsymbol{r}) \cdot f(\boldsymbol{r} \,|\, \tau^2)\, \mathrm{d}\boldsymbol{r} \tag{3.24}$$

wobei $f(\boldsymbol{y} \,|\, \boldsymbol{\beta}, \boldsymbol{r}) = \prod_{i=1}^{n_S} f(y_{is} \,|\, \boldsymbol{\beta}, \boldsymbol{r})$ mit $f(y_{is} \,|\, \boldsymbol{\beta}, \boldsymbol{r})$ die Dichte der Binomialverteilung mit $p(\boldsymbol{x}_{is}) = g^{-1}({\boldsymbol{x}_{is}}^T \boldsymbol{\beta} + r_s)$ bezeichnet, und $f(\boldsymbol{r} \,|\, \tau^2)$ die Dichte der gemeinsamen Verteilung von $\boldsymbol{R}$, also wie in (3.22) ist. Diese integrierte Form zur Schätzung der Parameter $\boldsymbol{\beta}$ und

τ^2 entspricht der Likelihood eines generalisierten linearen gemischten Modells (Fahrmeir et al., 2007).

Konstruktion der Quasi-Likelihood: Um die Likelihood $L(\boldsymbol{\beta}, \tau^2)$ zu vereinfachen, wird eine Quasi-Likelihood definiert. Aufgrund ihrer Eigenschaften kann sie wie eine Likelihood zum Schätzen der Parameter verwendet werden. Z. B. muss die Ableitung der Quasi-Likelihood einen Erwartungswert von 0 haben, denn die Ableitung der Quasi-Likelihood entspricht der Scorefunktion.

Für die Konstruktion der Quasi-Likelihood geht man deswegen auch von der Scorefunktion eines generalisierten linearen Modells (GLM) aus. Ausgehend von der Definition einer Exponentialfamilie als

$$f(\boldsymbol{y} \mid \vartheta_i, \varphi, \omega_i) = \prod_{i=1}^{n} \exp\left(\frac{\omega_i}{\varphi}(y_i\vartheta_i - b(\vartheta_i)) + c(y_i, \varphi, \omega_i)\right)$$

nach Fahrmeir et al. (2007) mit (3.23) lautet die

$$\begin{aligned}
\text{Likelihood} \qquad & L(\boldsymbol{\beta}) = \prod_{i=1}^{n} L_i(\boldsymbol{\beta}) = \prod_{i=1}^{n} \exp\left(\frac{\omega_i}{\varphi}(y_i\vartheta_i - b(\vartheta_i)) + c(y_i, \varphi, \omega_i)\right) \\
\text{log-Likelihood} \qquad & l(\boldsymbol{\beta}) = \sum_{i=1}^{n} l_i(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{\omega_i}{\varphi}(y_i\vartheta_i - b(\vartheta_i)) + c(y_i, \varphi, \omega_i) \\
\text{Scorefunktion} \qquad & s(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\vartheta}} l_i(\boldsymbol{\beta}) \cdot \frac{\partial}{\partial \boldsymbol{\mu}} \vartheta_i \cdot \frac{\partial}{\partial \boldsymbol{\eta}} \mu_i \cdot \frac{\partial}{\partial \boldsymbol{\beta}} \eta_i
\end{aligned}$$

Die einzelnen Ableitungen lauten

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\vartheta}} l_i(\boldsymbol{\beta}) &= \frac{\omega_i}{\varphi}(y_i - b'(\vartheta_i)) &&= \frac{\omega_i}{\varphi}(y_i - \mu_i) \\
\frac{\partial}{\partial \boldsymbol{\eta}} \mu_i &= \frac{\partial}{\partial \boldsymbol{\eta}} g^{-1}(\eta_i) = \frac{\partial}{\partial \boldsymbol{\eta}} g^{-1}(\boldsymbol{x}_i{}^T \boldsymbol{\beta}) &&:= d_i \\
\frac{\partial}{\partial \boldsymbol{\beta}} \eta_i &= \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{x}_i{}^T \boldsymbol{\beta} &&= \boldsymbol{x}_i
\end{aligned}$$

Die Ableitung von ϑ_i gestaltet sich etwas komplizierter:

$$\begin{aligned}
\mu_i &= b'(\vartheta_i) \\
\iff \vartheta_i &= (b')^{-1}(\mu_i) \\
\frac{\partial}{\partial \boldsymbol{\mu}}\vartheta_i &= \frac{\partial}{\partial \boldsymbol{\mu}}(b')^{-1}(\mu_i) := \frac{\partial}{\partial \boldsymbol{\mu}} f^{-1}(\mu_i) \\
&= \frac{1}{f'[f^{-1}(\mu_i)]} && \text{durch den Umkehrsatz} \\
&= \frac{1}{b''[\underbrace{(b')^{-1}(\mu_i)}_{\vartheta_i}]} \\
&= \frac{1}{b''(\vartheta_i)} \\
&= \frac{1}{\mathbb{V}ar(Y_i)}\frac{\varphi}{\omega_i} && \text{mit (3.23)}
\end{aligned}$$

Insgesamt ergibt sich die Scorefunktion als

$$s(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{\omega_i}{\varphi}(y_i - \mu_i)\frac{1}{\mathbb{V}ar(Y_i)}\frac{\varphi}{\omega_i} d_i \boldsymbol{x}_i = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\mathbb{V}ar(Y_i)} d_i \boldsymbol{x}_i \overset{\text{mit (3.23)}}{=} \sum_{i=1}^{n} \frac{y_i - \mu_i}{\frac{\varphi}{\omega_i} v(\mu_i)} d_i \boldsymbol{x}_i$$

Geht man nun die log-Transformation und Differentiation duch Integration und Anwenden der Exponentialfunktion rückwärts, so lautet die Quasi-Likelihood

$$\exp\left(\int_{y_i}^{\mu_i} \sum_{i=1}^{n} \frac{y_i - u}{\frac{\varphi}{\omega_i} v(u)} d_i \boldsymbol{x}_i \, \mathrm{d}u\right) \propto \exp\left(\int_{y_i}^{\mu_i} \sum_{i=1}^{n} \frac{y_i - u}{\frac{\varphi}{\omega_i} v(u)} \, \mathrm{d}u\right)$$

Dies entspricht $f(\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{r})$ aus (3.24). Zusammen mit der Normalverteilung der räumlichen Effekte $\boldsymbol{R}$ aus (3.22) ergibt sich die integrierte Quasi-Likelihood

$$\begin{aligned}
QL(\boldsymbol{\beta}, \tau^2) &= \int \exp\left(\int_{y_i}^{\mu_i} \sum_{i=1}^{n} \frac{y_i - u}{\frac{\varphi}{\omega_i} v(u)} \, \mathrm{d}u\right) \cdot (\tau^2)^{-\frac{1}{2}\mathrm{rang}(\boldsymbol{K})} \exp\left(-\frac{1}{2\tau^2} \boldsymbol{r}^T \boldsymbol{K} \boldsymbol{r}\right) \, \mathrm{d}\boldsymbol{r} \\
&= (\tau^2)^{-\frac{1}{2}\mathrm{rang}(\boldsymbol{K})} \int \exp\left(\sum_{i=1}^{n} \int_{y_i}^{\mu_i} \frac{y_i - u}{\frac{\varphi}{\omega_i} v(u)} \, \mathrm{d}u - \frac{1}{2\tau^2} \boldsymbol{r}^T \boldsymbol{K} \boldsymbol{r}\right) \, \mathrm{d}\boldsymbol{r}
\end{aligned}$$

wie es auch in Breslow und Clayton (1993) und Dean et al. (2004) beschrieben ist. Diese Funktion ist das Pendant zur Likelihood im "regulären" Likelihood-Ansatz.

Penalisierung: Der Bestrafungsterm ist dabei $\boldsymbol{r}^T \boldsymbol{K} \boldsymbol{r}$. Durch ihn wird berücksichtigt, dass $\boldsymbol{R}$ aus einer Verteilung stammt und kein fester Effekt ist (Fahrmeir et al., 2007). Durch die

Gestalt der Nachbarschaftsmatrix $\boldsymbol{K}$ werden Abweichungen der räumlichen Effekte von denen der benachbarten Regionen bestraft, denn es gilt (Kneib, 2009/2010):

$$\boldsymbol{r}^T\boldsymbol{K}\boldsymbol{r} = \sum_s \sum_{\substack{s' \in N(s) \\ s' < s}} (r_s - r_{s'})^2$$

Der Glättungsparameter ist zugleich die Varianz τ^2 des räumlichen Effekts.

Laplace-Approximation: Somit ist die penalisierte Quasi-Likelihood (PQL) aufgestellt und kann bezüglich $\boldsymbol{\beta}$ und $\boldsymbol{r}$ maximiert werden. Dazu müsste man die ersten Ableitungen der integrierten Quasi-Likelihood nullsetzen. Allerdings kann man dieses Integral nicht in geschlossener Form berechnen. Aus diesem Grund ist eine Approximation notwendig. Im PQL-Verfahren wird eine Laplace-Approximation verwendet. Der Integrand wird durch eine Taylor-Entwicklung um $\widehat{\boldsymbol{r}}$ angenähert:
Setze $k(\boldsymbol{r}) = \sum_{i=1}^n \int_{y_i}^{\mu_i} \frac{y_i - u}{\frac{\varphi}{\omega_i} v(u)}\, \mathrm{d}u - \frac{1}{2\tau^2}\boldsymbol{r}^T\boldsymbol{K}\boldsymbol{r}$. Dann bezeichnet $k'(\boldsymbol{r})$ den Vektor $\frac{\partial}{\partial \boldsymbol{r}} k(\boldsymbol{r})$ und $k''(\boldsymbol{r})$ die Matrix $\frac{\partial^2}{\partial \boldsymbol{r}^2} k(\boldsymbol{r})$. Der Vektor $\widehat{\boldsymbol{r}}$ ist die Lösung von $k'(\boldsymbol{r}) = 0$.

$$\begin{aligned}
QL(\boldsymbol{\beta}, \tau^2) &= \left(\tau^2\right)^{-\frac{1}{2}\mathrm{rang}(\boldsymbol{K})} \int \exp\left(k(\boldsymbol{r})\right)\, \mathrm{d}\boldsymbol{r} = \left|\tau^2\boldsymbol{K}^{-1}\right|^{-\frac{1}{2}} \int \exp\left(k(\boldsymbol{r})\right)\, \mathrm{d}\boldsymbol{r} && \text{mit (3.20)} \\
\log\left(QL(\boldsymbol{\beta}, \tau^2)\right) &= -\frac{1}{2}\log\left(\left|\tau^2\boldsymbol{K}^{-1}\right|\right) + \log\left(\int \exp\left(k(\boldsymbol{r})\right)\, \mathrm{d}\boldsymbol{r}\right) \\
&\approx -\frac{1}{2}\log\left(\left|\tau^2\boldsymbol{K}^{-1}\right|\right) + \log\left[\int \exp\left(k(\widehat{\boldsymbol{r}}) + \frac{1}{2}(\boldsymbol{r} - \widehat{\boldsymbol{r}})^T k''(\widehat{\boldsymbol{r}})(\boldsymbol{r} - \widehat{\boldsymbol{r}})\right)\, \mathrm{d}\boldsymbol{r}\right] \\
&= -\frac{1}{2}\log\left(\left|\tau^2\boldsymbol{K}^{-1}\right|\right) + k(\widehat{\boldsymbol{r}}) + \log\left[\int \exp\left(\frac{1}{2}(\boldsymbol{r} - \widehat{\boldsymbol{r}})^T k''(\widehat{\boldsymbol{r}})(\boldsymbol{r} - \widehat{\boldsymbol{r}})\right)\, \mathrm{d}\boldsymbol{r}\right] \\
&\propto -\frac{1}{2}\log\left(\left|\tau^2\boldsymbol{K}^{-1}\right|\right) + k(\widehat{\boldsymbol{r}}) + \log\left(|k''(\widehat{\boldsymbol{r}})|^{-\frac{1}{2}}\right) && \text{mit (E.1)} \\
&= -\frac{1}{2}\log\left(\left|\tau^2\boldsymbol{K}^{-1}\right|\right) + k(\widehat{\boldsymbol{r}}) - \frac{1}{2}\log\left(|k''(\widehat{\boldsymbol{r}})|\right)
\end{aligned}$$

Schätzung der Parameter: Die Matrix $k''(\boldsymbol{r})$ kann man durch Definition einer Matrix $\boldsymbol{W} = diag\left(\left[\frac{\varphi}{\omega_i} v(\mu_i)[g'(\mu_i)]^2\right]^{-1}\right)$ (Breslow und Clayton, 1993) als $k''(\boldsymbol{r}) = \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z} - \frac{1}{\tau^2}\boldsymbol{K}$

schreiben. Damit ergibt sich die Quasi-log-Likelihood

$$\begin{aligned}\log\left(QL(\boldsymbol{\beta},\tau^2)\right) &= -\frac{1}{2}\log\left(\left|\tau^2\boldsymbol{K}^{-1}\right|\right) + k(\widehat{\boldsymbol{r}}) - \frac{1}{2}\log\left(\left|\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z} - \frac{1}{\tau^2}\boldsymbol{K}\right|\right)\\ &= k(\widehat{\boldsymbol{r}}) - \frac{1}{2}\left[\log\left(\left|\tau^2\boldsymbol{K}^{-1}\right|\right) + \log\left(\left|\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z} - \frac{1}{\tau^2}\boldsymbol{K}\right|\right)\right]\\ &= k(\widehat{\boldsymbol{r}}) + \frac{1}{2}\log\left(\left|\tau^2\boldsymbol{K}^{-1}\right| \cdot \left|\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z} - \frac{1}{\tau^2}\boldsymbol{K}\right|\right)\\ &= k(\widehat{\boldsymbol{r}}) + \frac{1}{2}\log\left(\left|\tau^2\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z}\boldsymbol{K}^{-1} - \boldsymbol{I}\right|\right)\end{aligned}$$

Der neu definierte Arbeitsresponsevektor $c_i = \boldsymbol{x}_i{}^T\boldsymbol{\beta} + (y_i - \mu_i)g'(\mu_i)$ folgt einem linearen gemischten Modell (siehe Kap. 3.2): $\boldsymbol{C} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{r} + \boldsymbol{\epsilon}$, wobei $\boldsymbol{\epsilon} \sim N(\boldsymbol{0},\ \boldsymbol{W}^{-1})$ und $\boldsymbol{R}$ wie gewohnt normalverteilt mit $N(0,\ \tau^2\boldsymbol{K}^{-1})$. Um $\boldsymbol{\beta}$, $\boldsymbol{r}$ und τ^2 zu schätzen, wird ein iterativer Algorithmus angewendet:

0. Mit einem geeigneten Startwert für τ^2 werden
1. erst $\boldsymbol{\beta}$ und $\boldsymbol{r}$ geschätzt.
2. Anschließend wird die Schätzung für τ^2 durchgeführt.
3. Am Ende wird der Arbeitsresponse $\boldsymbol{C}$ aktualisiert und erneut alle Schritte 1. – 3. durchlaufen, bis sich keine Veränderung mehr ergibt.

Der Schätzer für $\boldsymbol{\beta}$ kann einfach gefunden werden, indem man das marginale Modell $\boldsymbol{C} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}^*$ mit $\boldsymbol{\epsilon}^* = \boldsymbol{r} + \boldsymbol{\epsilon}$ betrachtet. Damit ist die Verteilung von $\boldsymbol{C} \sim N(\boldsymbol{X}\boldsymbol{\beta},\ \boldsymbol{W}^{-1} + \tau^2\boldsymbol{K}^{-1})$, da $\boldsymbol{r}$ und $\boldsymbol{\epsilon}$ voneinander unabhängig sind. Dies entspricht einem verallgemeinerten linearen Modell. Unter der Annahme, dass die Kovarianz $\boldsymbol{V} := \boldsymbol{W}^{-1} + \tau^2\boldsymbol{K}^{-1}$ bekannt ist, ist der Schätzer der festen Effekte ein Aitken-Schätzer (Toutenburg, 2003):

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{C}$$

Der beste Schätzer für $\boldsymbol{r}$ ist laut Fahrmeir et al. (2007) der bedingte Erwartungswert $\mathbb{E}(\boldsymbol{r} \mid \boldsymbol{C})$. Diesen erhält man am einfachsten über die gemeinsame Verteilung von $\boldsymbol{C}$ und $\boldsymbol{r}$. Für zwei multivariat normalverteilte Zufallsvariablen $\boldsymbol{A}$ und $\boldsymbol{B}$ lautet der bedingte Erwartungswert $\boldsymbol{\mu}_{\boldsymbol{B}|\boldsymbol{A}} = \boldsymbol{\mu}_{\boldsymbol{B}} + \boldsymbol{\Sigma}_{\boldsymbol{A}\boldsymbol{B}}\boldsymbol{\Sigma}_{\boldsymbol{A}}{}^{-1}(\boldsymbol{A} - \boldsymbol{\mu}_{\boldsymbol{A}})$. Übertragen auf das vorliegende Problem lautet der Schätzer für $\boldsymbol{r}$

$$\widehat{\boldsymbol{r}} = \boldsymbol{0} + \tau^2\boldsymbol{K}^{-1}\boldsymbol{V}^{-1}(\boldsymbol{C} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$$

Der Schätzer für τ^2 wird ebenfalls aus dem marginalen Modell $\boldsymbol{C} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}^*$ hergeleitet.

Nachdem $\boldsymbol{C}$ einer Normalverteilung folgt, lautet die Likelihood

$$l(\boldsymbol{\beta}, \tau^2) = -\frac{1}{2}\log(|\boldsymbol{V}|) - \frac{1}{2}(\boldsymbol{C} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{V}^{-1}(\boldsymbol{C} - \boldsymbol{X}\boldsymbol{\beta})$$

Setzt man $\widehat{\boldsymbol{\beta}}$ in $l(\boldsymbol{\beta}, \tau^2)$ ein, so erhält man die Profil-log-Likelihood $l(\widehat{\boldsymbol{\beta}}, \tau^2)$. Da jedoch der ML-Schätzer verzerrt ist, verwendet man in der Regel den REML-Schätzer $\widehat{\tau^2_R}$, welcher $l_R(\tau^2)$ maximiert (Fahrmeir et al., 2007).

$$\begin{aligned}
l_R(\tau^2) &= \log\left(\int L(\boldsymbol{\beta}, \tau^2)\, \mathrm{d}\boldsymbol{\beta}\right) \\
&= \log\left(\int |V|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{C} - \boldsymbol{X}\boldsymbol{\beta})^T V^{-1}(\boldsymbol{C} - \boldsymbol{X}\boldsymbol{\beta})\right)\, \mathrm{d}\boldsymbol{\beta}\right) \\
&= \log \int |V|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{C} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^T V^{-1}(\boldsymbol{C} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})\right) \qquad \text{mit (E.3)} \\
&\qquad \cdot \exp\left(-\frac{1}{2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \boldsymbol{X}^T V^{-1} \boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right)\, \mathrm{d}\boldsymbol{\beta} \\
&= \log\left[|V|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{C} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^T V^{-1}(\boldsymbol{C} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})\right)\right] \\
&\quad + \log\left[\int \exp\left(-\frac{1}{2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \boldsymbol{X}^T V^{-1} \boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right)\, \mathrm{d}\boldsymbol{\beta}\right]
\end{aligned}$$

Der Integrand hat die Form einer multivariaten Normalverteilung von $\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta},\ (\boldsymbol{X}^T V^{-1} \boldsymbol{X})^{-1})$, es fehlt nur die Komponente außerhalb der Exponentialfunktion. Diese wird mit einem 1-Trick eingefügt, so dass das Integral den Wert 1 ergibt.

$$\begin{aligned}
l_R(\tau^2) &= -\frac{1}{2}\log(|\boldsymbol{V}|) - \frac{1}{2}(\boldsymbol{C} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^T \boldsymbol{V}^{-1}(\boldsymbol{C} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) \\
&\quad + \log\left[\int \frac{\left|(\boldsymbol{X}^T V^{-1} \boldsymbol{X})^{-1}\right|^{\frac{1}{2}}}{\left|(\boldsymbol{X}^T V^{-1} \boldsymbol{X})^{-1}\right|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \boldsymbol{X}^T V^{-1} \boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right)\, \mathrm{d}\boldsymbol{\beta}\right] \\
&= -\frac{1}{2}\log(|\boldsymbol{V}|) - \frac{1}{2}(\boldsymbol{C} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^T \boldsymbol{V}^{-1}(\boldsymbol{C} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) \\
&\quad + \log\left[\left|(\boldsymbol{X}^T V^{-1} \boldsymbol{X})^{-1}\right|^{\frac{1}{2}} \int \left|(\boldsymbol{X}^T V^{-1} \boldsymbol{X})^{-1}\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \boldsymbol{X}^T V^{-1} \boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right)\, \mathrm{d}\boldsymbol{\beta}\right] \\
&= -\frac{1}{2}\log(|\boldsymbol{V}|) - \frac{1}{2}(\boldsymbol{C} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^T \boldsymbol{V}^{-1}(\boldsymbol{C} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) + \log\left(\left|(\boldsymbol{X}^T V^{-1} \boldsymbol{X})\right|^{-\frac{1}{2}}\right) \\
&= -\frac{1}{2}\log(|\boldsymbol{V}|) - \frac{1}{2}(\boldsymbol{C} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^T \boldsymbol{V}^{-1}(\boldsymbol{C} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) - \frac{1}{2}\log\left(\left|\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X}\right|\right)
\end{aligned}$$

Der Schätzer $\widehat{\tau^2_R}$ muss jedoch numerisch durch Fisher-Scoring o. Ä. bestimmt werden: Nachdem man $l_R(\tau^2)$ maximieren möchte, setzt man die Ableitung gleich 0. Durch eine Taylor-Entwicklung der Ableitung, also der Score-Funktion, erhält man die Formel, welche

zur Aktualisierung verwendet wird:

$$\begin{aligned} s_R\left(\widehat{\tau_R^2}\right) &\approx s_R\left(\tau^{2(k)}\right) + \left(\widehat{\tau_R^2} - \tau^{2(k)}\right) \cdot F_R\left(\tau^{2(k)}\right) \stackrel{!}{=} 0 \\ -\left(\widehat{\tau_R^2} - \tau^{2(k)}\right) &= s_R\left(\tau^{2(k)}\right) \cdot F_R\left(\tau^{2(k)}\right)^{-1} \\ \widehat{\tau_R^2} &= \tau^{2(k)} - s_R\left(\tau^{2(k)}\right) \cdot F_R\left(\tau^{2(k)}\right)^{-1} \end{aligned}$$

mit $s_R(\tau^2) = \frac{\mathrm{d}}{\mathrm{d}\tau^2} l_R(\tau^2)$ und $F_R(\tau^2) = -\frac{\mathrm{d}^2}{\mathrm{d}(\tau^2)^2} l_R(\tau^2)$. Die Formel zur Aktualisierung lautet also

$$\tau^{2(k+1)} = \tau^{2(k)} - s_R\left(\tau^{2(k)}\right) \cdot F_R\left(\tau^{2(k)}\right)^{-1}$$

für $k = 0, 1, 2, \ldots$.

3.3.3.3. Bayesianischer Ansatz für Gauß-Markov-Zufallsfelder

Eine andere Möglichkeit, die räumlichen Effekte in einem generalisierten linearen Modell zu schätzen, ist die bayesianische Sichtweise. Dazu wird eine Priori für alle Parameter angenommen. Mit Hilfe der Priori kann man Vorwissen in die Modellierung einbringen. In einem räumlich strukturierten Modell wird für die Regressionskoeffizienten der festen Effekte i. d. R. eine nicht-informative Priori gewählt ($f(\boldsymbol{\beta}) \propto const$), für die räumlichen Effekte wird die Normalverteilung $N(0,\ \tau^2 \boldsymbol{K}^{-1})$ als Priori aufgefasst. Für den Parameter τ^2 benötigt man eine Hyperpriori – meist eine inverse Γ-Verteilung, da diese konjugiert ist zur Normalverteilung des GMRF.

Die Posteriori ist durch das Bayes-Theorem proportional zum Produkt aus Priori und Likelihood.

$$\begin{aligned} f(\boldsymbol{\beta}, \boldsymbol{r}, \tau^2 \,|\, \boldsymbol{y}) &\propto f(\boldsymbol{y} \,|\, \boldsymbol{\beta}, \boldsymbol{r}) \cdot f(\boldsymbol{\beta}) f(\boldsymbol{r} \,|\, \tau^2) f(\tau^2) \\ &= \prod_{i,s} f(y_{is} \,|\, \boldsymbol{\beta}, \boldsymbol{r}) \cdot f(\boldsymbol{\beta}) f(\boldsymbol{r} \,|\, \tau^2) f(\tau^2) \end{aligned}$$

Für die Auswertung der Screening-Daten liegen folgende Annahmen zugrunde:

$$\begin{aligned} Y_{is} &\sim Bin\left(1,\ p(\boldsymbol{x}_{is}) = g^{-1}(\boldsymbol{x}_{is}^T \boldsymbol{\beta} + r_j)\right) \\ \boldsymbol{\beta} &\sim N(\boldsymbol{m},\ \boldsymbol{S}) \qquad \text{mit } \boldsymbol{S} \to \infty \\ \boldsymbol{R} &\sim N(0,\ \tau^2 \boldsymbol{K}^{-1}) \\ \tau^2 &\sim IG(a = 0.001,\ b = 0.001) \end{aligned} \tag{3.25}$$

Der Posteriori-Erwartungswert ist eine Möglichkeit, Punktschätzer für die Parameter zu

erhalten. So ergibt sich beispielsweise $\widehat{\beta}_p$ als $\mathbb{E}(\beta_p \,|\, \cdot) = \int \beta_p f(\beta_p \,|\, \cdot)\, \mathrm{d}\beta_p$. Dafür werden für jeden Parameter die so genannten "vollständig bedingten Dichten" benötigt. Sie lauten:

$$
\begin{aligned}
f(\boldsymbol{\beta} \,|\, \cdot) &\propto \prod_{i,s} f(y_{is} \,|\, \boldsymbol{\beta}, \boldsymbol{r}) f(\boldsymbol{\beta}) \\
f(\boldsymbol{r} \,|\, \cdot) &\propto \prod_{i} f(y_{is} \,|\, \boldsymbol{\beta}, \boldsymbol{r}) f(\boldsymbol{r} \,|\, \tau^2) \\
f(\tau^2 \,|\, \cdot) &\propto f(\boldsymbol{r} \,|\, \tau^2) f(\tau^2)
\end{aligned}
$$

Meist wird der Posteriori-Erwartungswert zur Schätzung nicht analytisch berechnet, sondern ein Monte-Carlo-Verfahren angewendet. Dazu werden aus den vollständig bedingten Dichten K Zufallszahlen erzeugt und dann der Erwartungswert durch das arithmetische Mittel geschätzt (Gesetz der großen Zahlen).

$$
\widehat{\beta}_p = \widehat{\mathbb{E}}(\beta_p \,|\, \cdot) = \frac{1}{K} \sum_{k=1}^{K} {\beta_p}^{(k)}
$$

Da aber y_{is} keiner Normalverteilung folgt sondern binomialverteilt ist, ist die jeweilige vollständig bedingte Dichte keine bekannte Verteilungsfunktion. Deswegen muss man z. B. mit Hilfe eines Markov-Ketten-Monte-Carlo-Verfahren (engl. Markov chain Monte Carlo, Abkürzung: MCMC-Verfahren) Zufallszahlen simulieren. Denn eine Markov-Kette konvergiert unter bestimmten Voraussetzungen gegen ihre stationäre Verteilung. Als stationäre Verteilung wird die Posteriori bzw. die vollständig bedingte Dichte verwendet. Bis die Stationarität erreicht ist, muss die Markov-Kette ein so genanntes *Burn-in* durchlaufen. Die Ziehungen des Burn-in werden verworfen. Um die naturgemäße Autokorrelation der Ziehungen zu reduzieren, kann man eine Ausdünnung durchführen und nur jede t-te Zufallsvariable speichern.

Möchte man eine gewisse Anzahl Zufallsvariablen erzeugen, muss man K Schritte durchführen. Die Anzahl der Schritte muss dabei das Burn-in und das Ausdünnen berücksichtigen. Ist man z. B. bei einem Burn-in von 200 und einer Ausdünnung von $t = 5$ an einer Ziehung von 1000 Zufallszahlen interessiert, so ist $K = 200 + 5 \cdot 1000 = 5200$.

Ein bekanntes MCMC-Verfahren ist der Metropolis-Hastings-Algorithmus. Dazu werden die Zufallszahlen aus einer Vorschlagsdichte h generiert und mit einer gewissen Wahrscheinlichkeit als Ziehung aus f akzeptiert. Die Vorschlagsdichte kann dabei jeweils von der vorangegangenen Ziehung abhängen (Markov-Eigenschaft; vgl. Kap. 3.3.1.1).

In der hier verwendeten Software BayesX ist es jedoch etwas anders gelöst. Basierend auf der Theorie von Holmes und Held (2002) wird dort von einem latenten linearen Modell

ausgegangen, wie es auch schon in der Herleitung des Logit-Modells Verwendung fand (siehe Formel (3.3) in Kap. 3.1.2):

$$\tilde{y}_{is} = \boldsymbol{x}_{is}{}^T\boldsymbol{\beta} + r_s + \epsilon_{is}$$

Der Fehler ϵ_{is} ist normalverteilt mit $\epsilon_{is} \sim N(0,\ \lambda_{is})$. Der Parameter λ_{is} benötigt eine Priori. Holmes und Held (2002) wählten die folgende Darstellung:

$$\begin{aligned}
&\tilde{Y} \sim N\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{R},\ \boldsymbol{\Lambda}\right) && \text{mit } \boldsymbol{\Lambda} = diag\left(\lambda_{is}\right) = diag\left(\boldsymbol{\lambda}\right) \\
&\boldsymbol{\beta} \sim N(\boldsymbol{m},\ \boldsymbol{S}) && \text{mit } \boldsymbol{S} \to \infty \\
&\boldsymbol{R} \sim N(0,\ \tau^2\boldsymbol{K}^{-1}) \\
&\tau^2 \sim IG(a = 0.001,\ b = 0.001) \\
&\lambda_{is} = (2\psi_{is})^2 \\
&\psi_{is} \sim \text{Kolmogorov-Smirnov}
\end{aligned} \tag{3.26}$$

Die vollständig bedingte Dichte bzw. die Likelihood ist eine bei 0 trunkierte Normalverteilung (Holmes und Held, 2002):

$$f(\tilde{y}_{is}\,|\,\cdot) \propto \begin{cases} \exp\left(-\frac{1}{2\lambda_{is}}\left(\tilde{y}_{is} - \boldsymbol{x}_{is}{}^T\boldsymbol{\beta} - r_j\right)^2\right)\mathbb{I}(\tilde{y}_{is} > 0) & \text{falls } y_{is} = 1 \\ \exp\left(-\frac{1}{2\lambda_{is}}\left(\tilde{y}_{is} - \boldsymbol{x}_{is}{}^T\boldsymbol{\beta} - r_j\right)^2\right)\mathbb{I}(\tilde{y}_{is} \leq 0) & \text{falls } y_{is} = 0 \end{cases}$$

Mit den Annahmen aus (3.26) folgen die vollständig bedingten Dichten nun fast alle einer bekannten Verteilung, so dass eine Simulation von Zufallszahlen nicht mehr nötig ist. Für die festen Effekte lautet die vollständig bedingte Dichte

$$\begin{aligned}
f(\boldsymbol{\beta}\,|\,\cdot) &\propto f(\tilde{\boldsymbol{y}}\,|\,\boldsymbol{\beta}, \boldsymbol{r}, \boldsymbol{\lambda}) f(\boldsymbol{\beta}) \\
&\propto \exp\left(-\frac{1}{2}\left(\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{r}\right)^T\boldsymbol{\Lambda}^{-1}\left(\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{r}\right)\right)\exp\left(-\frac{1}{2}\left(\boldsymbol{\beta} - \boldsymbol{m}\right)^T\boldsymbol{S}^{-1}\left(\boldsymbol{\beta} - \boldsymbol{m}\right)\right) \\
&\propto \exp\left(-\frac{1}{2}\left(\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{\Lambda}^{-1}\tilde{\boldsymbol{y}} + \boldsymbol{\beta}^T\boldsymbol{S}^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\boldsymbol{S}^{-1}\boldsymbol{m}\right)\right) \\
&= \exp\left(-\frac{1}{2}\boldsymbol{\beta}^T\left(\boldsymbol{X}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{X} + \boldsymbol{S}^{-1}\right)\boldsymbol{\beta} + \boldsymbol{\beta}^T\left(\boldsymbol{X}^T\boldsymbol{\Lambda}^{-1}\tilde{\boldsymbol{y}} + \boldsymbol{S}^{-1}\boldsymbol{m}\right)\right) \\
\Rightarrow \boldsymbol{\beta}\,|\,\cdot &\sim N\left(\left(\boldsymbol{X}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{X} + \boldsymbol{S}^{-1}\right)^{-1}\left(\boldsymbol{X}^T\boldsymbol{\Lambda}^{-1}\tilde{\boldsymbol{y}} + \boldsymbol{S}^{-1}\boldsymbol{m}\right),\ \left(\boldsymbol{X}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{X} + \boldsymbol{S}^{-1}\right)^{-1}\right)
\end{aligned}$$

wegen (E.2) und

$$\begin{aligned}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} &= \left(\boldsymbol{X}^T\boldsymbol{\Lambda}^{-1}\tilde{\boldsymbol{y}} + \boldsymbol{S}^{-1}\boldsymbol{m}\right) = \left(\boldsymbol{X}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{X} + \boldsymbol{S}^{-1}\right)\boldsymbol{\mu} \\ \Rightarrow \boldsymbol{\mu} &= \left(\boldsymbol{X}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{X} + \boldsymbol{S}^{-1}\right)^{-1}\left(\boldsymbol{X}^T\boldsymbol{\Lambda}^{-1}\tilde{\boldsymbol{y}} + \boldsymbol{S}^{-1}\boldsymbol{m}\right)\end{aligned}$$

Die vollständig bedingte Dichte der räumlichen Effekte lautet mit denselben Umformungen wie für die festen Effekte

$$\begin{aligned}f(\boldsymbol{r}\,|\,\cdot) &\propto f(\tilde{\boldsymbol{y}}\,|\,\boldsymbol{\beta},\boldsymbol{r},\boldsymbol{\lambda})f(\boldsymbol{r}\,|\,\tau^2) \\ &\propto \exp\left(-\frac{1}{2}\left(\tilde{\boldsymbol{y}}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{r}\right)^T\boldsymbol{\Lambda}^{-1}\left(\tilde{\boldsymbol{y}}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{r}\right)\right)\exp\left(-\frac{1}{2\tau^2}\boldsymbol{r}^T\boldsymbol{K}\boldsymbol{r}\right) \\ &\propto \exp\left(-\frac{1}{2}\left(-2\boldsymbol{r}^T\boldsymbol{\Lambda}^{-1}\tilde{\boldsymbol{y}} + 2\boldsymbol{r}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{r}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{r} + \frac{1}{\tau^2}\boldsymbol{r}^T\boldsymbol{K}\boldsymbol{r}\right)\right) \\ &= \exp\left[-\frac{1}{2}\boldsymbol{r}^T\left(\boldsymbol{\Lambda}^{-1}+\frac{1}{\tau^2}\boldsymbol{K}\right)\boldsymbol{r} + \boldsymbol{r}^T\left(\boldsymbol{\Lambda}^{-1}\tilde{\boldsymbol{y}} - \boldsymbol{\Lambda}^{-1}\boldsymbol{X}\boldsymbol{\beta}\right)\right] \\ \Rightarrow \boldsymbol{r}\,|\,\cdot &\sim N\left(\left(\boldsymbol{\Lambda}^{-1}+\frac{1}{\tau^2}\boldsymbol{K}\right)^{-1}\left(\boldsymbol{\Lambda}^{-1}\tilde{\boldsymbol{y}} - \boldsymbol{\Lambda}^{-1}\boldsymbol{X}\boldsymbol{\beta}\right),\ \left(\boldsymbol{\Lambda}^{-1}+\frac{1}{\tau^2}\boldsymbol{K}\right)^{-1}\right)\end{aligned}$$

Der Varianzparameter τ^2 kann ebenfalls aus seiner vollständig bedingten Dichte gezogen werden, da es sich um eine inverse Γ-Verteilung handelt:

$$\begin{aligned}f(\tau^2\,|\,\cdot) &\propto f(\boldsymbol{r}\,|\,\tau^2)f(\tau^2) \\ &\propto \left(\tau^2\right)^{-\frac{1}{2}\mathrm{rang}(\boldsymbol{K})}\exp\left(-\frac{1}{2\tau^2}\boldsymbol{r}^T\boldsymbol{K}\boldsymbol{r}\right)(\tau^2)^{-a-1}\exp\left(-\frac{b}{\tau^2}\right) \\ &= \left(\tau^2\right)^{-\left(a+\frac{1}{2}\mathrm{rang}(\boldsymbol{K})\right)-1}\exp\left[-\frac{1}{\tau^2}\left(\frac{1}{2}\boldsymbol{r}^T\boldsymbol{K}\boldsymbol{r}+b\right)\right] \\ \Rightarrow \tau^2\,|\,\cdot &\sim IG\left(a+\frac{1}{2}\mathrm{rang}(\boldsymbol{K}),\ \frac{1}{2}\boldsymbol{r}^T\boldsymbol{K}\boldsymbol{r}+b\right)\end{aligned}$$

Nur die Dichte des Fehlervarianzparameters $\boldsymbol{\lambda}$ ist unbekannt:

$$f(\boldsymbol{\lambda}\,|\,\cdot) \propto f(\tilde{\boldsymbol{y}}\,|\,\boldsymbol{\beta},\boldsymbol{r},\boldsymbol{\lambda})f(\boldsymbol{\lambda})$$

Diese wird durch *Rejection Sampling* simuliert. Der Algorithmus hat generell folgende Form, wenn man eine Zufallszahl aus f_Λ ziehen möchte:

- Ziehe eine Zufallszahl κ aus $f_K(\kappa)$. Dabei ist $f_K(\kappa)$ möglichst ähnlich zu $f_\Lambda(\lambda)$, folgt aber einer bekannten Verteilung.
- Ziehe eine Zufallszahl u aus einer Gleichverteilung auf $[0,\ 1]$.

- Berechne
$$\alpha(K) = \frac{1}{M}\frac{f_\Lambda(K)}{f_K(K)}$$
wobei $M \geq 1$ so, dass $\alpha(\kappa) \leq 1 \, \forall \kappa$
- Akzeptiere κ als Zufallszahl aus $f_\Lambda(\lambda)$, falls $u \leq \alpha(\kappa)$

Dass dieser Algorithmus die Dichte $f_\Lambda(\lambda)$ simuliert, kann leicht gezeigt werden (Leisch, 2008):

$$\begin{aligned}\mathbb{P}(\Lambda \leq \lambda) &= \mathbb{P}(K \leq \lambda \,|\, U \leq \alpha(K)) \\ &= \frac{\mathbb{P}(K \leq \lambda,\ U \leq \alpha(K))}{\mathbb{P}(U \leq \alpha(\lambda))}\end{aligned}$$

Wegen

$$\begin{aligned}\mathbb{P}(K \leq \lambda,\ U \leq \alpha(K)) &= \int_{-\infty}^{\lambda}\int_{0}^{\alpha(\kappa)} f_K(\kappa) \cdot 1 \,\mathrm{d}u \,\mathrm{d}\kappa \\ &= \int_{-\infty}^{\lambda}\int_{0}^{\alpha(\kappa)} 1 \,\mathrm{d}u f_K(\kappa) \,\mathrm{d}\kappa \\ &= \int_{-\infty}^{\lambda} \alpha(\kappa) f_K(\kappa) \,\mathrm{d}\kappa \\ &= \int_{-\infty}^{\lambda} \frac{1}{M}\frac{f_\Lambda(\kappa)}{f_K(\kappa)} f_K(\kappa) \,\mathrm{d}\kappa \\ &= \frac{1}{M}\int_{-\infty}^{\lambda} f_\Lambda(\kappa) \,\mathrm{d}\kappa\end{aligned}$$

und

$$\begin{aligned}\mathbb{P}(U \leq \alpha(\lambda)) &= \mathbb{P}(K \leq \infty,\ U \leq \alpha(\lambda)) \\ &= \int_{-\infty}^{\infty}\int_{0}^{\alpha(\kappa)} 1 \,\mathrm{d}u f_K(\kappa) \,\mathrm{d}\kappa \\ &= \int_{-\infty}^{\infty} \alpha(\kappa) f_K(\kappa) \,\mathrm{d}\kappa \\ &= \frac{1}{M}\int_{-\infty}^{\infty} f_\Lambda(\kappa) \,\mathrm{d}\kappa \\ &= \frac{1}{M}\end{aligned}$$

ist

$$\begin{aligned}\mathbb{P}(\Lambda \leq \lambda) &= \frac{\frac{1}{M}\int_{-\infty}^{\lambda} f_\Lambda(\kappa)\,\mathrm{d}\kappa}{\frac{1}{M}} \\ &= \int_{-\infty}^{\lambda} f_\Lambda(\kappa)\,\mathrm{d}\kappa\end{aligned}$$

Bei Holmes und Held (2002) und somit auch in BayesX wird als "Vorschlagsdichte" $f_K(\kappa)$ eine generalisierte inverse Normalverteilung verwendet. Die Zufallszahlen daraus werden mittels Inversionsmethode aus einer Standard-Normalverteilung gewonnen. Der Algorithmus für ein λ_{is} findet sich in Holmes und Held (2002, S. 164). Dieser wird für jedes λ_{is} wiederholt. Das ist möglich, da die λ_{is} unabhängig voneinander sind.

Insgesamt wird hiermit ein MCMC-Verfahren angewendet, da die vollständig bedingten Dichten von den Ziehungen der anderen Parameter aus dem vorherigen Sampling-Schritt abhängen. Nach einer Initialisierung

0. Berechne den Posterior-Modus der normalverteilten vollständig bedingten Dichten von $\boldsymbol{\beta}$ und $\boldsymbol{r}$ bei gegebenem $\frac{\lambda_{is}}{\tau^2} = 0.1$ (Default von BayesX).

können die K Schritte durchgeführt werden. Im Schritt $k = 1, \ldots, K$ wird Folgendes berechnet:

1. Aktualisiere die festen Effekte $\boldsymbol{\beta}^{(k)}$ aus der normalverteilten vollständig bedingten Dichte $f\left(\boldsymbol{\beta} \,|\, \boldsymbol{r}^{(k-1)}, \left(\tau^2\right)^{(k-1)}, \boldsymbol{\lambda}^{(k-1)}, \tilde{\boldsymbol{y}}^{(k-1)}\right)$.
2. Aktualisiere die festen Effekte $\boldsymbol{r}^{(k)}$ aus der normalverteilten vollständig bedingten Dichte $f\left(\boldsymbol{r} \,|\, \boldsymbol{\beta}^{(k)}, \left(\tau^2\right)^{(k-1)}, \boldsymbol{\lambda}^{(k-1)}, \tilde{\boldsymbol{y}}^{(k-1)}\right)$.
3. Aktualisiere die Varianzparameter $\boldsymbol{\lambda}^{(k)}$ mittels Rejection Sampling und $\left(\tau^2\right)^{(k)}$ aus der invers-Γ-verteilten vollständig bedingten Dichte $f\left(\tau^2 \,|\, \boldsymbol{\beta}^{(k)}, \boldsymbol{r}^{(k)}, \boldsymbol{\lambda}^{(k)}, \tilde{\boldsymbol{y}}^{(k-1)}\right)$.
4. Aktualisiere die latenten Beobachtungen $\tilde{\boldsymbol{y}}^{(k)}$ durch $f\left(\tilde{y}_{is} \,|\, \boldsymbol{\beta}^{(k)}, \boldsymbol{r}^{(k)}, \left(\tau^2\right)^{(k)}, \boldsymbol{\lambda}^{(k)}\right)$.

3.4. Generalisierter linearer Hypothesentest

Um p-Werte zu den verwendeten Kovariablen zu erhalten, werden Signifikanztests durchgeführt. Mit den p-Werten kann man z. B. durch Rückwärts-Selektion Modellwahl betreiben oder eine Aussage über die Signifikanz des Einflusses der Kovariablen treffen. Ist der p-Wert kleiner als ein vorgegebenes Signifikanzniveau α, so wird die Null-Hypothese verworfen. Für jede Kovariable wird eine eigene Null-Hypothese auf Signifikanz aufgestellt, diese je-

doch anhand der gleichen Daten geprüft. Dadurch liegt ein multiples Testproblem vor. Dies kann man durch die so genannte Bonferroni-Korrektur umgehen. Allerdings ist dieses Verfahren konservativ, d. h. das Signifikanzniveau wird nicht voll ausgeschöpft. Durch ein multiples Test-Verfahren wird diesem Problem Rechnung getragen.

Bei einem linearen Hypothesentest, wie er im R-Paket `multcomp` (Hothorn et al., 2008) implementiert ist, wird angenommen, dass der Schätzer $\widehat{\boldsymbol{\vartheta}}$ für den Parametervektor $\boldsymbol{\vartheta}$ asymptotisch normalverteilt ist: $\widehat{\boldsymbol{\vartheta}} \overset{a}{\sim} N(\boldsymbol{\vartheta},\ \boldsymbol{S})$, wobei $\boldsymbol{S}$ ein Schätzer für $Cov\left(\widehat{\boldsymbol{\vartheta}}\right)$ ist. Eine lineare Transformation des Schätzers folgt auch einer asymptotischen Normalverteilung: $\boldsymbol{C}\widehat{\boldsymbol{\vartheta}} \overset{a}{\sim} N\left(\boldsymbol{C\vartheta},\ \boldsymbol{CSC}^T\right)$. Eine Standardisierung der Transformation ändert ebenfalls nichts an der Verteilungsfamilie: $\boldsymbol{D}^{-\frac{1}{2}}\left(\boldsymbol{C}\widehat{\boldsymbol{\vartheta}} - \boldsymbol{C\vartheta}\right) \overset{a}{\sim} N(0,\ \boldsymbol{R})$ mit $\boldsymbol{R} = \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{CSC}^T\boldsymbol{D}^{-\frac{1}{2}}$ und $\boldsymbol{D} = diag(\boldsymbol{CSC}^T)$ (Hothorn et al., 2008). Die Schätzer $\widehat{\boldsymbol{\vartheta}}$ und $\boldsymbol{S}$ erhält man in R aus dem jeweiligen Modell.

Die zu testende Null-Hypothese hat die Form

$$H_0 : \boldsymbol{C\vartheta} = \boldsymbol{d} \quad \text{gegen} \quad H_1 : \boldsymbol{C\vartheta} \neq \boldsymbol{d} \tag{3.27}$$

Unter H_0 ist dann $\boldsymbol{T} = \boldsymbol{D}^{-\frac{1}{2}}\left(\boldsymbol{C}\widehat{\boldsymbol{\vartheta}} - \boldsymbol{d}\right) \overset{H_0}{\sim} N(0,\ \boldsymbol{R})$. Die individuellen Null-Hypothesen lauten

$$H_0{}^p : \boldsymbol{c}_p\vartheta_p = d_p \quad \text{gegen} \quad H_1{}^p : \boldsymbol{c}_p\vartheta_p \neq d_p$$

Im Spezialfall $d_p = 0$ testet man die Signifikanz der zu ϑ_p gehörenden Effekte.

Die globale Null-Hypothese aus (3.27) kann mit einem χ^2- oder F-Test überprüft werden. Eine andere Möglichkeit besteht darin, das Maximum von $\boldsymbol{T}$ zu verwenden. Diese Teststatistik ist unter H_0 t-verteilt bzw. approximativ normalverteilt (Hothorn et al., 2008). Somit lässt sich die Verteilungsfunktion der Teststatistik folgendermaßen berechnen:

$$\begin{aligned}\mathbb{P}\left(\max(|\boldsymbol{T}|) \leq t\right) &\approx \int_{-t}^{t} \dots \int_{-t}^{t} \frac{1}{2\pi^{\frac{P}{2}}|\boldsymbol{R}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{R}^{-1}\boldsymbol{x}\right)\ \mathrm{d}x_1 \dots\ \mathrm{d}x_P \\ &:= g(\boldsymbol{R}, t)\end{aligned}$$

Mit Hilfe dieser Form kann genau gesagt werden, welche der individuellen Null-Hypothesen $H_0{}^p$ verworfen werden muss. Denn der individuelle p-Wert des p-ten Parameters kann mittels $p_p = 1 - g(\boldsymbol{R}, t_p)$ bestimmt werden, falls $t = (t_1, \dots, t_P)^T$ als Teststatistik beobachtet wurde (Hothorn et al., 2008).

Die Null-Hypothese $H_0{}^p$ wird verworfen, falls der p-Wert höchstens so groß ist wie das zuvor festgelegte Signifikanzlevel α ($p_p \leq \alpha$).

3.5. Zusammenfassung und Anwendung

3.5.1. Das unstrukturierte Modell

Die unstrukturierten Modelle wurden als generalisiertes lineares gemischtes Modell mit zufälligem Effekt pro Region formuliert (vgl. Kap. 3.3.2). Die Berechnung erfolgte in R.

Der Response ist binomialverteilt mit Logit-Link g:

$$Y_{is} \overset{\text{iid}}{\sim} Bin\left(1,\ p(\boldsymbol{x}_{is}) = g^{-1}(\boldsymbol{x}_{is}{}^T\boldsymbol{\beta})\right)$$

wobei Y_{is} angibt, ob Person i aus Region s zum Screening gegangen ist oder nicht und der Index s das Gebiet ($s = 1, \ldots, 36$) bzw. den Postleitzahlbezirk ($s = 1, \ldots, 113$) bezeichnet, in der die Person lebt.
Der zufällige Effekt pro Region ist normalverteilt:

$$b_s \overset{\text{iid}}{\sim} N(0,\ d^2)$$

Dies sind die Annahmen für die Individualdaten. Insbesondere die Kovariablen konnten aber nicht auf Individualebene erfasst werden, sodass eine aggregierte Auswertung erfolgte. Jede Kombination k der Kovariablen aus 2.4 (vgl. insb. Tab. 2.6) erhielt ein Gewicht w_k, welches aussagt, wie oft diese Kombination beobachtet worden ist. Die Schätzung erfolgt im R-Paket `lme4` (Bates und Maechler, 2009) durch gewichtete KQ-Schätzung. Anstatt der üblicherweise verwendeten Summe von Quadraten $\sum_i (y_i - \widehat{y}_i)^2$ wird die gewichtete Summe $\sum_k w_k (y_k - \widehat{y}_k)^2$ minimiert (vgl. (3.9)). Das Gewicht wird durch die Variable `anz` im Datensatz ausgedrückt. Die Zielvariable heißt in diesem Datensatz `scree`.

- `scree`: 0 bezeichnet die Personen ab 55, welche nicht am Screening teilgenommen haben, 1 bezeichnet Screening-Klienten.
- `anz`: die Anzahl Personen, auf welche die Kovariablen-Kombination dieser Zeile zutrifft.

Die Modellgleichung lautet

$$\log\left(\frac{p(\boldsymbol{x}_{is})}{1 - p(\boldsymbol{x}_{is})}\right) = \boldsymbol{x}_{is}{}^T\beta + b_s$$

Für die 36 Gebiete lautet der Aufruf in R

```
R> library(lme4)
R> modell1a <- lmer(scree ~ jahr * altergrp * maennlich + christlich +
```

```
+     anz.ge + imd + (1 | gebiet), family = binomial(), data = daten4,
+     weights = daten4$anz)
```

Für die Bezirke der dreistelligen PLZs lautet der Aufruf in R

```
R> library(lme4)
R> modell1b <- lmer(scree ~ jahr * altergrp * maennlich + kk.einw +
+     anz.ge + (1 | plz), family = binomial(), data = daten2,
+     weights = daten2$anz)
```

3.5.2. Das strukturierte Modell in R

Die strukturierten Modelle in R wurden für das Paket `mgcv` als generalisiertes additives gemischtes Modell formuliert, wobei der Smoothing-Term ein GMRF für den strukturierten räumlichen Effekt ist. So ist die Anpassung mittels einer penalisierten Quasi-Likelihood (PQL, vgl. Kap. 3.3.3.2) möglich. Die Software, um ein GMRF im R-Paket `mgcv` (Wood, 2008) anzupassen, stammt von Prof. Kneib.

Der Response ist skaliert binomialverteilt mit Logit-Link g:

$$Y_{is} \overset{\text{iid}}{\sim} Bin\left(w_{is},\ p(\boldsymbol{x}_{is}) = g^{-1}({\boldsymbol{x}_{is}}^T\boldsymbol{\beta})\right)/w_{is}$$

wobei Y_{is} angibt, welcher Anteil der w_{is} Personen mit Kovariablen-Kombination i zum Screening gegangen ist. Die skalierte Binomialverteilung hat den Vorteil, dass der Erwartungswert $\mathbb{E}(Y_{is}) = w_{is}p(\boldsymbol{x}_{is})/w_{is} = p(\boldsymbol{x}_{is})$ ist, sodass man ein gewöhnliches Logit-Modell anpassen kann.

Die Datensatz-Spalte

- `prop` zeigt Y_{is}.
- `anz` zeigt die Anzahl Personen, auf welche die Kovariablen-Kombination dieser Zeile zutrifft, repräsentiert also w_{is}.

Die anderen Kovariablen sind entsprechend den in Kapitel 2.4 vorgestellten Beschreibungen. Der Index s bezeichnet das Gebiet ($s = 1, \ldots, 36$) bzw. den PLZ-Bezirk ($s = 1, \ldots, 113$), in der die Kovariablen-Kombination i beobachtet wurde.

Der räumlich strukturierte Effekt für die Regionen ist multivariat normalverteilt:

$$\boldsymbol{R}\,|\,\tau^2 \sim N\left(\boldsymbol{0},\ \tau^2\boldsymbol{K}^{-1}\right)$$

(GMRF). Dabei ist $\boldsymbol{K}$ die Nachbarschaftsmatrix der Regionen. Sie beinhaltet für jede Region s auf der Diagonale die Anzahl der Nachbarn, und bei jedem Nachbarn $s' \in N(s)$ lautet der Eintrag "-1". Es wird also ein intrinsisches GMRF mit Gewichten $w_{ss'} = 1$ angenommen. Der Varianzparameter τ^2 wird per penalisierter Quasi-Likelihood geschätzt.

Die Modellgleichung lautet

$$\log\left(\frac{p(\boldsymbol{x}_{is})}{1-p(\boldsymbol{x}_{is})}\right) = {\boldsymbol{x}_{is}}^T\beta + r_s$$

Für die 36 Gebiete lautet der Aufruf in R

```
R> library("mgcv")
R> source("von Thomas Kneib/spatialdesign.R")
R> modell2a <- gamm(prop ~ jahr * altergrp * maennlich + christlich +
+     anzge + imd + s(gebiet, bs = "mrf", xt = list(geb.adja2,
+     geb.id2)), family = binomial(), data = daten3, weights = daten3$einw)
```

Für die Bezirke der dreistelligen PLZs lautet der Aufruf in R

```
R> library("mgcv")
R> source("von Thomas Kneib/spatialdesign.R")
R> modell2b <- gamm(prop ~ jahr * altergrp * maennlich + kk.einw +
+     anz.ge + s(plz, bs = "mrf", xt = list(plz.adja2, plz.id2)),
+     family = binomial(), data = daten1, weights = daten1$einw)
```

3.5.3. Das strukturierte Modell in BayesX

In BayesX muss man die Dummy-Variablen selbst erzeugen. Dadurch hat sich die Zusammensetzung des Datensatzes erweitert. Außerdem ist dort statt `scree` und der Gewichtsvariable `anz` eine andere Darstellung notwendig. Die Variable

- `einw`: bezeichnet die Anzahl Personen der allgemeinen Bevölkerung ab 55, auf welche die Kovariablen-Kombination dieser Zeile zutrifft, und
- `scree`: bezeichnet die Anzahl Screening-Klienten, auf welche die Kovariablen-Kombination dieser Zeile zutrifft

Die strukturierten Modelle in BayesX wurden bayesianisch formuliert (vgl. Kap. 3.3.3.3). Für den strukturierten räumlichen Effekt wurde als Priori ein GMRF angenommen. Die anderen Kovariablen wurden als feste Effekte mit einer nicht-informativen Priori ($p(\beta_p) \propto const$) ins Modell aufgenommen. Die Anpassung erfolgt per MCMC-Simulation. Der Burn-in be-

trägt 2000 Ziehungen, jede 25. Ziehung wird gespeichert. Damit wird bei einer Gesamtzahl von 27000 Iterationen ein Sample von 1000 Werten pro Parameter erzeugt.

Der Response ist binomialverteilt mit Logit-Link g:

$$Y_{is} \overset{\text{iid}}{\sim} Bin\left(w_{is},\ p(\boldsymbol{x}_{is}) = g^{-1}(\boldsymbol{x}_{is}{}^T\boldsymbol{\beta})\right)$$

wobei Y_{is} angibt, wie viele der w_{is} Personen mit Kovariablen-Kombination i zum Screening gegangen ist. Die Variable `einw` repräsentiert also w_{is}, die Spalte `scree` zeigt die Anzahl der "Treffer" Y_{is}. Der Index s bezeichnet das Gebiet ($s = 1, \ldots, 36$) bzw. den PLZ-Bezirk ($s = 1, \ldots, 113$), in der die Kovariablen-Kombination i beobachtet wurde.

Der räumlich strukturierte Effekt für die Regionen ist multivariat normalverteilt:

$$\boldsymbol{R} \,|\, \tau^2 \sim N\left(\mathbf{0},\ \tau^2 \boldsymbol{K}^{-1}\right)$$

wobei $\boldsymbol{K}$ die Nachbarschaftsmatrix der Regionen ist (GMRF; wie in 3.5.2). Die Hyperpriori für τ^2 ist eine inverse Gammaverteilung:

$$\tau^2 \sim IG(a = 0.001,\ b = 0.001)$$

Die Modellgleichung lautet

$$\log\left(\frac{p(\boldsymbol{x}_{is})}{1 - p(\boldsymbol{x}_{is})}\right) = \boldsymbol{x}_{is}{}^T\beta + r_s$$

Der räumlich strukturierte Effekt $\boldsymbol{R}$ folgt einer Normalverteilung mit Varianzparameter τ^2, wie in den Prioris spezifiziert.

Für die 36 Gebiete lautet der Aufruf in BayesX

```
BayesX> bayesreg modell3a
BayesX> modell3a.regress scree = jahr1 + jahr2 + altergrp6575 +
     +     altergrp7599 + maennlich + jahr1altergrp6575 +
     +     jahr1altergrp7599 + jahr2altergrp6575 + jahr2altergrp7599 +
     +     jahr1maennlich1 + jahr2maennlich1 + altergrp6575maennlich1 +
     +     altergrp7599maennlich1 + jahr1altergrp6575maennlich1 +
     +     jahr2altergrp6575maennlich1 + jahr1altergrp7599maennlich1 +
     +     jahr2altergrp7599maennlich1 + christlich + anzge + imd +
     +     gebiet(spatial, map = m) weight einw, burnin=2000
```

```
     +     iterations=27000 step=25 family=binomial predict using d
```

Für die Bezirke der dreistelligen PLZs lautet der Aufruf in BayesX

```
BayesX> bayesreg modell3b
BayesX> modell3b.regress scree = jahr1 + jahr2 + altergrp6064 +
     +     altergrp 6574 + altergrp7599 + maennlich + jahr1altergrp6064 +
     +     jahr1altergrp6574 + jahr1altergrp7599 + jahr2altergrp6064 +
     +     jahr2altergrp6574 + jahr2altergrp7599 + jahr1maennlich1 +
     +     jahr2maennlich1 + altergrp6064maennlich1 +
     +     altergrp6574maennlich1 + altergrp7599maennlich1 +
     +     jahr1altergrp6064maennlich1 + jahr1altergrp6574maennlich1 +
     +     jahr1altergrp7599maennlich1 + jahr2altergrp6064maennlich1 +
     +     jahr2altergrp6574maennlich1 + jahr2altergrp7599maennlich1 +
     +     anzge + kkeinw + plz(spatial, map = m) weight einw,
     +     burnin=2000 iterations=27000 step=25 family=binomial
     +     predict using d
```

4. Ergebnisse

4.1. Die festen Effekte

In den folgenden Tabellen der Ergebnisübersicht gilt: Falls der Schätzer bei einem generalisierten linearen Hypothesentest (siehe Kap. 3.4) auf einem Niveau von 5% von 0 verschieden war, wurde der Schätzer fett gedruckt. Die "Signifikanz" der bayesianischen Modelle wurde dadurch ermittelt, ob das 95% Credibility Interval die Null enthält oder nicht.

Das 95% Credibility Interval ist das Intervall, das vom 2.5%- und vom 97.5%-Quantil der Posteriori eingeschlossen wird. Somit liegt der wahre Wert mit 95%-iger Wahrscheinlichkeit zwischen den beiden Intervallgrenzen.

4.1.1. Der Effekt des Geschlechts

Der Effekt des Geschlechts ist in den verschiedenen Modellen sehr ähnlich geschätzt worden. Er war in allen Modellen auf dem 5%-Niveau signifikant (siehe Tab. 4.1).

Tabelle 4.1.: Geschätzter Effekt des Geschlechts.

Modell	Gebiete	PLZ3
unstrukturiert	**-0.3015**	**-0.3719**
strukturiert – PQL	**-0.3015**	**-0.3687**
strukturiert – Bayes	**-0.3017**	**-0.3713**
Referenzkategorie	weiblich ($x = 0$)	

Der Effekt des Geschlechts ist sowohl in den 36 Gebieten als auch in den Bezirken der dreistelligen PLZ negativ. In den 36 Gebieten hat der Effekt einen Betrag von rund 0.30. Die Odds Ratio (vgl. Kap. 3.1.4) beträgt $\frac{\frac{p(x=1)}{1-p(x=1)}}{\frac{p(x=0)}{1-p(x=0)}} = \exp(0.30) = 0.74$. Sie ist somit kleiner als 1. Nachdem das Geschlecht so kodiert war, dass 1 männlich und 0 weiblich bedeutet, ist die Chance, dass Männer zu Screening gehen, kleiner als diejenige der Frauen, falls alle anderen Kovariablen gleich ausgeprägt sind.

In den Postleitzahlbezirken beträgt der Schätzer in den drei Modellarten je ca. -0.37. Die

Odds Ratio ergibt sich zu 0.69. Damit hat das Geschlecht in der Analyse auf Basis der PLZ-Bezirke einen etwas größeren Effekt als in den 36 Gebieten. Aber auch hier bedeutet das Schätzergebnis, dass die Chance, dass Männer das Screening-Angebot nutzen, kleiner ist als die Chance der Frauen mit gleicher Kovariablen-Kombination.

In den Nutzungsraten aus Kapitel 2.6.1 zeigte sich insgesamt keine Abhängigkeit vom Geschlecht. Die Teilnahmeraten von Frauen und Männern waren nahezu gleich. Dort spielte nur die Interaktion des Geschlechts mit dem Alter eine Rolle.

4.1.2. Der Effekt des Zeitverlaufs

Ob der Zeitverlauf einen Effekt auf die Nutzung der Vorsorge-Koloskopie hat, sollte durch eine Faktorvariable modelliert werden. Es wurde je ein Schätzer für den Unterschied zwischen den Jahren 2006 und 2007 bzw. zwischen 2006 und 2008 berechnet. Die verschiedenen Modellanpassungen kommen jeweils zu sehr ähnlichen Ergebnissen. Auch diese Regressionskoeffizienten waren alle auf dem 5%-Niveau signifikant (siehe Tab. 4.2).

Tabelle 4.2.: Geschätzter Effekt des Zeitverlaufs.

	Jahr 2007		Jahr 2008	
Modell	Gebiete	PLZ3	Gebiete	PLZ3
unstrukturiert	**0.1419**	**0.1716**	**0.0842**	**0.1364**
strukturiert – PQL	**0.1419**	**0.1752**	**0.0842**	**0.1415**
strukturiert – Bayes	**0.1417**	**0.1774**	**0.0841**	**0.1439**
Referenzkategorie	Jahr 2006			

Der Effekt des Zeitverlaufs war in allen Modellen positiv, aber abnehmend. Der Effekt vom Basisjahr 2006 auf das Jahr 2007 wurde in den 36 Gebieten mit rund 0.14 geschätzt. Die Odds Ratio vom Jahr 2006 auf das Jahr 2007 beträgt somit 1.15 und ist größer als 1. Die Chance, dass jemand 2007 zum Screening geht, ist folglich größer als die Chance, dass sich jemand mit der gleichen Kovariablen-Ausprägung im Jahr 2006 dem Screening unterzieht.
Der Effekt vom Jahr 2006 auf das Jahr 2008 wurde als rund 0.08 geschätzt. Die Odds Ratio von 2006 auf das Jahr 2008 beträgt somit 1.09 und ist größer als 1. Das bedeutet, die Chance, dass jemand in 2008 zum Screening geht, ist größer als die Chance, dass sich jemand im Jahr 2006 dem Screening unterzieht, aber nicht so groß wie die Chance, dass man 2007 am Screening teilnimmt (je mit gleichen Werten in den anderen Kovariablen).

In den Bezirken der dreistelligen PLZ ist die Situation genau gleich, nur die Zahlen variieren etwas. Der Effekt vom Basisjahr auf das Jahr 2007 wurde dort auf rund 0.17 geschätzt. Die Odds Ratio vom Jahr 2006 auf das Jahr 2007 ist somit 1.19, also größer als 1. Die Chance,

dass jemand im Jahr 2007 zum Screening geht, ist folglich größer als die Chance, dass sich jemand mit den gleichen Kovariablen-Werten im Jahr 2006 dem Screening unterzieht.
Der Effekt vom Jahr 2006 auf das Jahr 2008 wurde als rund 0.14 geschätzt. Die Odds Ratio von 2006 auf das Jahr 2008 beträgt 1.15 und ist somit größer als 1. Das bedeutet, die Chance, dass jemand 2008 zum Screening geht, ist größer als die Chance, dass sich jemand im Jahr 2006 dem Screening unterzieht, aber nicht so groß wie die Chance, dass jemand mit der gleichen Kovariablen-Kombination im Jahr 2007 am Screening teilnimmt.

Dass die Nutzung von 2006 auf 2007 stark zunimmt und von 2006 auf 2008 ebenfalls eine – wenn auch geringere – Steigerung zu beobachten ist, zeichnete sich bereits bei den Nutzungsraten in Kapitel 2.6.1 ab.

4.1.3. Die Interaktion zwischen Geschlecht und Zeitverlauf

Die Ergebnisse zur Interaktion zwischen Geschlecht und Zeitverlauf sind in den 36 Gebieten denen der dreistelligen Postleitzahlen sehr ähnlich. Der Haupteffekt des Zeitverlaufs zeigte dagegen unterschiedliche Niveaus für die verschiedenen räumlichen Einheiten. Hier unterscheidet sich lediglich die Signifikanz des Schätzers von Modell zu Modell (siehe Tab. 4.3).

Tabelle 4.3.: Geschätzter Effekt der Interaktion zwischen Geschlecht und Zeitverlauf.

	Geschlecht ∗ Jahr 2007		Geschlecht ∗ Jahr 2008	
Modell	Gebiete	PLZ3	Gebiete	PLZ3
unstrukturiert	**0.0634**	0.0630	**0.0864**	**0.1067**
strukturiert – PQL	0.0634	0.0636	0.0864	**0.1079**
strukturiert – Bayes	**0.0636**	**0.0662**	**0.0863**	**0.1118**
Referenzkategorie	Frauen im Jahr 2006			

Die Interaktion zwischen Geschlecht und Zeitverlauf war in den Nutzungsraten (Kap. 2.6.1) nicht so stark ausgeprägt, dass er dort aufgefallen wäre. Der Effekt ist in den Modellen auch dementsprechend gering im Betrag: Die Interaktion zwischen Geschlecht und Zeitverlauf ist für das Jahr 2007 sowohl in den 36 Gebieten als auch in den PLZ-Bezirken auf rund 0.06 geschätzt worden.
Eine Interaktion ist schwieriger zu interpretieren als ein Haupteffekt. Denn hier ergeben sich keine Odds Ratios, sondern *Verhältnisse* von Odds Ratios. Das Verhältnis beträgt hier 1.07, ist also knapp größer als 1. Die Odds Ratio der Frauen vom Jahr 2006 auf das Jahr 2007 ist etwas größer als die entsprechende Odds Ratio der Männer. Die Odds Ratios der Jahre werden folglich vom Geschlecht beeinflusst. Ebenso wirkt der Zeitverlauf auf die Odds Ratios der Geschlechter, da die Odds Ratio und auch das Verhältnis von Odds

Ratios symmetrisch ist (vgl. Kap. 3.1.4).

Für den Übergang von 2006 auf das Jahr 2008 ergaben sich leicht unterschiedliche Schätzer in den Gebieten und den Bezirken der dreistelligen PLZs: 0.09 bzw. 0.11. Die Verhältnisse der Odds Ratios ergeben 1.09 und 1.11, beide wieder knapp größer als 1. In den 36 Gebieten zeigt sich also – verglichen mit dem Jahr 2007 – kaum ein Unterschied in der Interaktion. Auch hier gilt: Der Zeitverlauf beeinflusst die Odds Ratios der Geschlechter. Gleiches gilt in den Bezirken der dreistelligen Postleitzahlen, wobei dort die Interaktion etwas stärker ausgeprägt ist und somit der Effekt vom Geschlecht auf den Unterschied vom Basisjahr 2006 auf das Jahr 2008 und umgekehrt etwas deutlicher ausfällt.

4.1.4. Der Effekt des Alters

Der Effekt des Alters ergibt in den verschiedenen Modellen für die 36 Gebiete und für die dreistelligen Postleitzahlbezirke jeweils sehr ähnliche Schätzer pro Übergang von der jüngsten zur betrachteten Altersklasse. Auch diese Schätzer sind – wie schon die Haupteffekte des Geschlechts und des Zeitverlaufs zuvor – bei einem generalisierten linearen Hypothesentest auf dem 5%-Niveau signifikant von 0 verschieden (siehe Tab. 4.4).

Tabelle 4.4.: Geschätzter Effekt der Altersgruppen.

	Gebiete	
Modell	65 – 75	≥ 75
unstrukturiert	**-0.3793**	**-1.7421**
strukturiert – PQL	**-0.3792**	**-1.7420**
strukturiert – Bayes	**-0.3792**	**-1.7417**
Referenzkategorie	55 bis unter 65 Jahre	

	PLZ3		
Modell	60 – 64	65 – 74	≥ 75
unstrukturiert	**-0.1901**	**-0.4316**	**-1.8160**
strukturiert – PQL	**-0.1970**	**-0.4259**	**-1.8065**
strukturiert – Bayes	**-0.1979**	**-0.4271**	**-1.8078**
Referenzkategorie	55 bis unter 60 Jahre		

Der Effekt des Alters war in allen Modellen negativ und mit steigendem Alter weiter abnehmend. Für die Gebiete ergaben sich Schätzer von -0.38 bzw. -1.74 für die Übergänge von der jüngsten Altergruppe bis 65 Jahre zur mittleren (bis 75 Jahre) bzw. zu den ältesten Personen über 75 Jahre. Die Odds Ratios ergeben 0.68 und 0.16. Die Chance, dass jemand aus der mittleren Altersgruppe zum Screening geht ist also kleiner als jene, dass eine "junge" Person zum Screening geht. Die Chance, dass jemand aus der ältesten Kategorie, aber sonst gleichen Ausprägungen der anderen Kovariablen eine Vorsorgeuntersuchung

besucht ist noch einmal um ein Vielfaches geringer.

Für die Bezirke der dreistelligen PLZ ergibt sich aufgrund der anderen Alterstruktur in den zugrunde liegenden Bevölkerungsdaten von der Schober Information Group eine weitere Kategorie. Die Gesamtsituation bleibt allerdings gleich: Mit zunehmendem Alter wird die Chance, dass jemand zum Screening geht, immer kleiner im Vergleich zur jüngsten Altersgruppe, welche hier die 55- bis 60-jährigen Personen umfasst. Der Schätzer für die 60- bis 65-Jährigen beläuft sich auf -0.19, -0.43 ergibt sich für den Vergleich zwischen 55- bis 60-Jährigen und 65- bis 75-Jährigen und -1.82 wurde für den Unterschied zwischen der jüngsten und der ältesten Altersstufe geschätzt. Damit werden Odds Ratios von 0.83 und 0.65 für die beiden mittleren Altersklassen erreicht. Bis zu einem Alter von 75 Jahren nimmt die Chance, dass eine Person zur Vorsorge-Untersuchung geht, im Vergleich zur Chance von 55- bis 60-jährigen Personen mit gleichen Werten in den anderen Kovariablen, ab. Ein weiterer großer Abfall in der Chance ergibt sich für die älteste Personengruppe: Dort beträgt die Odds Ratio nur noch 0.16 (wie bei den Modellen der 36 Gebiete), alte Menschen gehen also nur noch selten zum Screening.

Das Alter zeigte in den Nutzungsraten von Kapitel 2.6.1 einen deutlich negativen Effekt. Pro Altersklasse nahmen die Teilnahmeraten merklich ab. Dieser Effekt ist sowohl in den 36 Gebieten mit drei Altersstufen als auch in den dreistelligen PLZ-Bezirken mit vier Altersstufen erkennbar und zieht sich durch alle Beobachtungsjahre. Bei den Postleitzahlbezirken ist der Unterschied zwischen der Gruppe der 60- bis 65-Jährigen und der Gruppe der 65- bis 75-Jährigen nicht so groß, wie auch die Schätzer und Odds Ratios in den Modellen bestätigen.

4.1.5. Die Interaktion zwischen Geschlecht und Alter

Die Interaktion zwischen Geschlecht und Alter ist in den verschiedenen Modellen pro Faktorstufe ebenfalls sehr ähnlich geschätzt worden. Diese Interaktion ist im Gegensatz zu jener zwischen Geschlecht und Zeitverlauf (Kap. 4.1.3) in allen Modellen auf dem 5%-Niveau signifikant. Die Schätzer haben auch einen größeren Betrag (siehe Tab. 4.5).

Die Interaktion zwischen Geschlecht und Alter zeigt durchweg positive Vorzeichen. In der Analyse der 36 Gebiete ergibt sich ein Schätzer in Höhe von 0.27 für den Übergang von Frauen bis 65 Jahre zu Männern zwischen 65 und 75 Jahren. Für den Sprung von weiblichen Personen zwischen 55 und 65 hin zu Männern der mittleren Altersgruppe wurde ein Effekt von 0.76 geschätzt.
Die Odds Ratio der Frauen von der jüngsten zur mittleren Altersklasse ist größer als die entsprechende Odds Ratio der Männer – das Verhältnis der Odds Ratios beträgt 1.31.

Tabelle 4.5.: Geschätzter Effekt der Interaktion zwischen Geschlecht und Alter.

	Gebiete	
Modell	Geschl. $*$ 65 – 75	Geschl. $*$ $\geq$ 75
unstrukturiert	**0.2689**	**0.7632**
strukturiert – PQL	**0.2689**	**0.7631**
strukturiert – Bayes	**0.2685**	**0.7642**
Referenzkategorie	Frauen ab 55 bis unter 65 Jahre	

	PLZ3		
Modell	Geschl. $*$ 60 – 64	Geschl. $*$ 65 – 74	Geschl. $*$ $\geq$ 75
unstrukturiert	**0.1466**	**0.3522**	**0.8633**
strukturiert – PQL	**0.1438**	**0.3464**	**0.8578**
strukturiert – Bayes	**0.1468**	**0.3498**	**0.8606**
Referenzkategorie	Frauen ab 55 bis unter 60 Jahre		

Also werden die Odds Ratios der Altersklassen durch das Geschlecht negativ beeinflusst. Einen deutlicheren Effekt hat das Geschlecht auf die Odds Ratios des Übergangs von der jüngsten zur ältesten Personengruppe, da dieses Verhältnis der Odds Ratios mit einem Betrag von 2.15 deutlich von 1 verschieden ist.

In der Analyse der PLZ-Bezirke kommt eine weitere Altersstufe hinzu, was zu einem weiteren Schätzer und Odds-Ratio-Verhältnis führt. Der Schätzer für den Übergang zur Gruppe der 60- bis 65-jährigen Männer ist ca. 0.15. Für die 65- bis 75-Jährigen ergab sich ein Schätzer von 0.35 und für den Unterschied zwischen den Frauen der jüngsten und den Männern der ältesten Altersklasse ergab sich 0.86. Das Verhältnis der Odds Ratios zwischen den Geschlechtern ist 1.16 für die 60- bis 65-Jährigen im Vergleich zu den 55- bis 65-Jährigen, wobei sich eine größere Odds Ratio für die Frauen ergibt. Die Odds Ratio der Frauen von der jüngsten zur Altersgruppe der 65- bis 75-Jährigen ist 1.42-mal größer als die entsprechende Odds Ratio der Männer. Am deutlichsten ist der Unterschied zwischen den Geschlechtern beim Übergang der unter 65-Jährigen zu den über 75-Jährigen: Die Odds Ratios haben ein Verhältnis von 2.37 zu 1 für die Frauen. Der Einfluss des Geschlechts auf die Odds Ratios wird mit zunehmendem Alter immer größer.

D.h. wenn die Personen älter sind, gehen Männer eher zum Screening als Frauen. Insgesamt mit den negativen Hautpeffekten von Alter und Geschlecht nimmt die Nutzung in beiden Geschlechtern mit zunehmendem Alter aber ab. Dieses Verhaltensmuster war bereits in den Nutzungsraten von Kapitel 2.6.1 zu erkennen.

4.1.6. Die Interaktion zwischen Geschlecht, Alter und Zeitverlauf

Der Vollständigkeit halber sind in Tabelle 4.6 die Schätzer der Interaktion zwischen Geschlecht, Alter und Zeitverlauf aufgeführt, welche ebenfalls in alle Modelle aufgenommen war. Allerdings ergeben solche Interaktionen von drei kategorialen Variablen ein Verhältnis von Verhältnissen von Odds Ratios und sind praktisch nicht mehr interpretierbar.
Die Signifikanz ist sehr variabel. Es lässt sich keine Regelmäßigkeit dabei erkennen.
Der Effekt ist meist negativ mit geringem Betrag. Eine Ausnahme davon ist die Kategorie der Männer zwischen 60 und 65 Jahren im Jahr 2007 (Analyse mit den dreistelligen PLZ-Bezirken), diese Faktorstufe hat als einzige einen positiven, aber sehr kleinen Effekt.

Tabelle 4.6.: Geschätzter Effekt der Interaktion zwischen Geschlecht, Alter und Zeitverlauf.

	Gebiete	
Modell	Geschl. $*$ 65 – 75 $*$ Jahr 2007	Geschl. $*$ 65 – 75 $*$ Jahr 2008
unstrukturiert	-0.0705	**-0.0905**
strukturiert – PQL	-0.0705	-0.0905
strukturiert – Bayes	**-0.0698**	**-0.0900**
Modell	Geschl. $*$ $\geq$ 75 $*$ Jahr 2007	Geschl. $*$ $\geq$ 75 $*$ Jahr 2008
unstrukturiert	**-0.1305**	**-0.1564**
strukturiert – PQL	-0.1305	-0.1558
strukturiert – Bayes	**-0.1314**	**-0.1569**
Referenzkategorie	Frauen ab 55 bis unter 65 Jahre im Jahr 2006	

	PLZ3	
Modell	Geschl. $*$ 60 – 64 $*$ Jahr 2007	Geschl. $*$ 60 – 64 $*$ Jahr 2008
unstrukturiert	0.0037	-0.0450
strukturiert – PQL	0.0037	-0.0453
strukturiert – Bayes	0.0016	**-0.0488**
Modell	Geschl. $*$ 65 – 74 $*$ Jahr 2007	Geschl. $*$ 65 – 74 $*$ Jahr 2008
unstrukturiert	-0.0726	**-0.1182**
strukturiert – PQL	-0.0725	-0.1183
strukturiert – Bayes	**-0.0773**	-0.1241
Modell	Geschl. $*$ $\geq$ 75 $*$ Jahr 2007	Geschl. $*$ $\geq$ 75 $*$ Jahr 2008
unstrukturiert	**-0.1352**	**-0.1752**
strukturiert – PQL	-0.1360	-0.1772
strukturiert – Bayes	**-0.1396**	**-0.1818**
Referenzkategorie	Frauen ab 55 bis unter 60 Jahre im Jahr 2006	

4.1.7. Weitere Einflussgrößen

Sowohl Geschlecht und Alter als auch der Zeitverlauf wurden als kategoriale Variablen in die Modelle aufgenommen. Die regionalen Einflussgrößen, welche sich nur zwischen den Regionen unterscheiden, sind jedoch stetige Größen. Die Interpretation ändert sich dadurch geringfügig (vgl. Kap. 3.1.4).

Die Anzahl Gastroenterologen hat in den drei Modellarten folgenden geschätzten Einfluss (siehe Tab. 4.7):
Die Schätzer zeigen alle einen negativen Einfluss mit geringem bis sehr geringem Betrag. In den dreistelligen PLZ-Bezirken ist dieser noch kleiner als in den 36 Gebieten, falls man einen räumlich strukturierten Effekt im Modell aufgenommen hat. Der Effekt der Gastroenterologenanzahl, also die Verfügbarkeit der Vorsorge-Untersuchung, ergibt in allen Modellen eine multiplikative Änderung der Odds Ratio von etwa 0.99, falls die Anzahl um eine Einheit steigt. Somit hat die Zahl in der Region ansässiger Gastroenterologen nicht ganz erwartungsgemäß keinen Einfluss auf die Entscheidung, eine Koloskopie durchführen zu lassen oder nicht. Überraschenderweise scheint eine große Anzahl Gastroenterologen sogar eher einen negativen Effekt zu haben. Allerdings waren die Schätzer in fünf der sechs Modelle nicht signifikant von 0 verschieden. Die Signifikanz der bayesianischen Modelle wurde – wie erwähnt – über das 95% Credibility Interval bestimmt. Im Fall der PLZ-Bezirke ist die einzige Signifikanz dieses Regressionskoeffizienten sehr knapp erreicht worden.

Tabelle 4.7.: Geschätzter Effekt der Anzahl Gastroenterologen.

Modell	Gebiete	PLZ3
unstrukturiert	-0.0005	-0.0050
strukturiert – PQL	-0.0158	-0.0043
strukturiert – Bayes	**-0.0341**	-0.0050

Die Einkommenssituation wurde in den 36 Gebieten über den "Index multipler Deprivation" (Werner Maier) modelliert. Dieser ist eine Messgröße für die generelle soziale Situation in den Gemeinden Bayerns und beinhaltet u. a. das durchschnittliche Einkommen der Einwohner. In den Bezirken der dreistelligen Postleitzahlen wurde die Variable "Kaufkraft pro Einwohner" (Schober Information Group) ins Modell aufgenommen. Dadurch sind die Schätzer nicht absolut vergleichbar.
Beide Variablen sollten modellieren, ob der Anteil nicht-gesetzlich Versicherter eine Rolle spielt. Diese Personen sind im Screening-Datensatz der KVB nämlich nicht erfasst.

Tabelle 4.8.: Geschätzter Effekt der sozialen Situation.

Modell	Gebiete	PLZ3
unstrukturiert	-0.0038	0.0000
strukturiert – PQL	0.0033	0.0000
strukturiert – Bayes	0.0060	**0.0000**

Die Schätzer haben allesamt einen Wert sehr nahe an der Null (siehe Tab. 4.8). Ein Problem beim Index multipler Deprivation könnte sein, dass dieser ja pro Gebiet als gewichteter Mittelwert aus den Gesamtindizes der Gemeinden hergestellt wird und somit durch die Mittelung keine extremen Werte mehr vorkommen. Da aber diese gemittelten Werte einen so geringen Effekt haben und auch die reine Kaufkraft bei den kleinräumigeren Postleitzahlbezirken keine größeren Effekt hat, werden die "rohen" Daten des Deprivationsindexes pro Gemeinde auch keinen großen Einfluss ergeben.
Die multiplikative Änderung der Odds Ratios ist ca. 1. Sie ändern sich also quasi nicht. Nur einer der Schätzer ist signifikant von 0 verschieden. Dieser ist vom Betrag auch der größte aus allen Modellen. Seine Signifikanz wurde nur sehr knapp erreicht. Die soziale Situation einer Region bzw. seiner Bewohner spielt demnach keine Rolle, ob viele Personen oder eher wenige die Darmkrebs-Vorsorge in Anspruch nehmen. Aus den Ergebnissen für diesen Schätzer lässt sich vermuten, dass entweder keine Region in Bayern im Vergleich zu den anderen Regionen einen hohen Anteil privat versicherter Bewohner hat oder dass diese Anteile in allen Regionen sehr gering sind.

Der Anteil christlicher Einwohner war nur in den Gemeinden verfügbar. Diese Kovariable konnte deswegen nur in die Modelle, welche in den 36 Gebieten angepasst wurden, aufgenommen werden. Die Schätzer unterscheiden sich zwischen den Modellen stark im Betrag (siehe Tab. 4.9). Allerdings wurde in jedem Modell ein negativer Effekt des Anteils christlicher Bevölkerung geschätzt. Die multiplikative Änderung der Odds Ratio ist in allen Fällen kleiner als 1: Nimmt der Anteil christlicher Einwohner um eine Einheit zu, so wird die Odds Ratio sinken. Somit ist die Chance, dass viele Personen eine Krebsvorsorge durchführen lassen, in Regionen mit einem hohen Anteil christlicher Bevölkerung geringer als in Regionen mit niedrigerem Anteil. Der Schätzer dieses Effekts war jedoch nur im bayesianischen Modell signifikant von 0 verschieden. Die Religionszugehörigkeit scheint demnach keine große Rolle zu spielen, wenn eine Person vor der Entscheidung steht, eine Darmkrebs-Vorsorgeuntersuchung durchführen zu lassen oder nicht.

Tabelle 4.9.: Geschätzter Effekt des Anteil christlicher Einwohner.

Modell	Gebiete
unstrukturiert	-3.8333
strukturiert – PQL	-1.9768
strukturiert – Bayes	**-3.0674**

4.2. Die räumlichen Effekte

Die Farbgebung aller folgenden Abbildungen ist so gewählt, dass die Werte-Skala überall gleich ist. Dadurch sind die Abbildungen direkt miteinander vergleichbar, da gleiche Farben gleiches Werte-Intervall bedeuten.

Der Vergleich der Modelle, welche aufgrund der Daten aus den 36 Gebieten angepasst wurden, zeigt Folgendes:

Im Vergleich des unstrukturierten mit dem strukturierten Modell, das mit PQL angepasst wurde, (Abb. 4.1) zeigt sich kaum ein Unterschied in den räumlichen Effekten. Der Werte der einzelnen Gebiete sind im strukturierten Modell nicht so extrem, dafür insgesamt etwas ins Positive verlagert. Dies betrifft hauptsächlich die Gegend von Nürnberg/Fürth bis nach München. Somit ist der erwartete Smoothing-Effekt eingetreten.

In einem Vergleich zwischen dem unstrukturierten und dem strukturierten Modell, das bayesianisch angepasst wurde, (Abb. 4.2) ist zu erkennen, dass auch bei dieser Art, die Nachbarschaftsstruktur zu berücksichtigen, die Werte weniger extrem sind und dass hauptsächlich negative Werte näher an die Null rücken.

Betrachtet man die räumlichen Effekte der Modelle, die die Daten der Bezirke der dreistelligen Postleitzahlen analysieren, kann man eine Umkehrung der Wirkung des regionalen Effekts erkennen:

Vergleicht man das unstrukturierte mit dem strukturierten, PQL-angepassten Modell der PLZ-Bezirke, zeigt sich, dass die Werte eher ins Negative verlagert sind statt in Positive wie bei den 36 Gebieten. Hier zeigt sich kein Smoothing-Effekt beim Übergang von der unstrukturierten zur strukturierten Analyse. Auffällig ist, dass die Umgebungen der Großstädte nicht von den Verlagerungen betroffen sind.

Ebenso ergibt sich eine umgekehrte Wirkung bei der Gegenüberstellung des Modells mit unstrukturiertem räumlichen Effekt und des bayesianischen Modells mit Berücksichtigung der Nachbarschaftsstruktur. Auch bei einer bayesianischen Anpassung fallen die räumlichen Effekte eher negativ aus als beim unstrukturierten Modell. Die Änderungen konzentrieren sich auf die Mitte Bayerns in Höhe von Nürnberg/Fürth und Regensburg. Der Rest

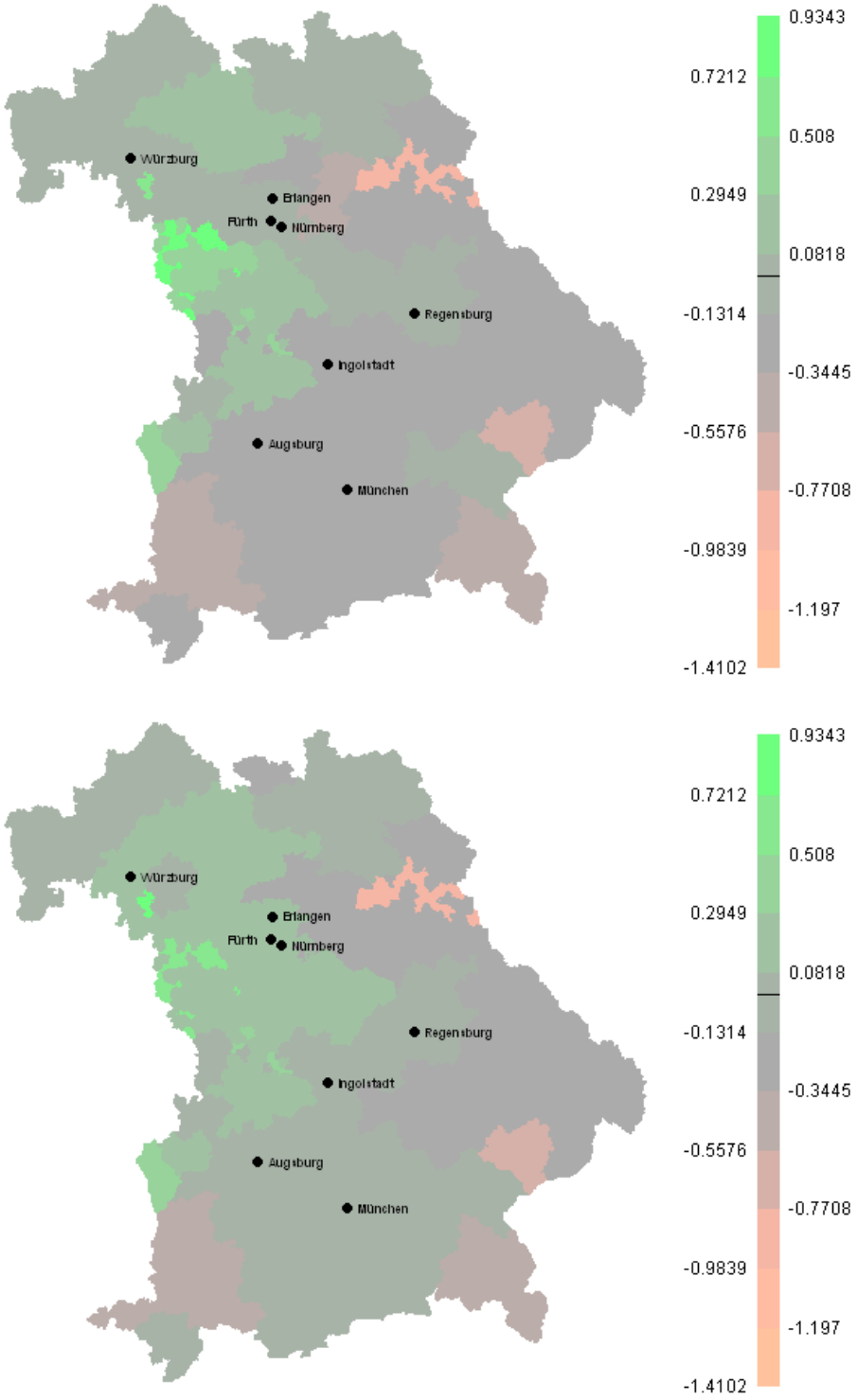

Abbildung 4.1.: *Oben:* Regionaler Effekt im unstrukturierten Modell der Gebiete (1a). *Unten:* Regionaler Effekt im strukturierten Modell der Gebiete, ausgewertet mit R per PQL (2a).

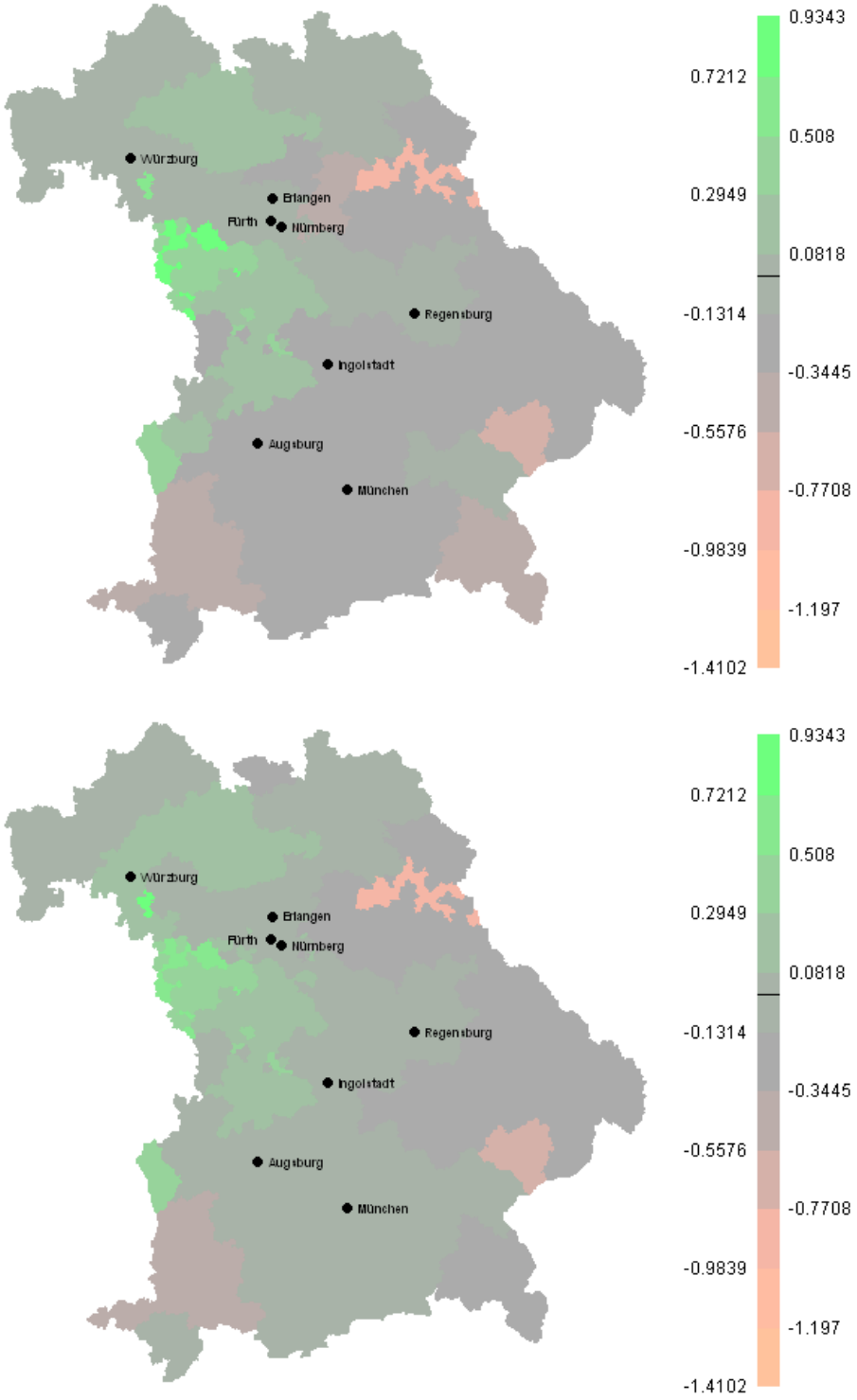

Abbildung 4.2.: *Oben:* Regionaler Effekt im unstrukturierten Modell der Gebiete (1a). *Unten:* Regionaler Effekt im strukturierten Modell der Gebiete, ausgewertet mit BayesX (3a).

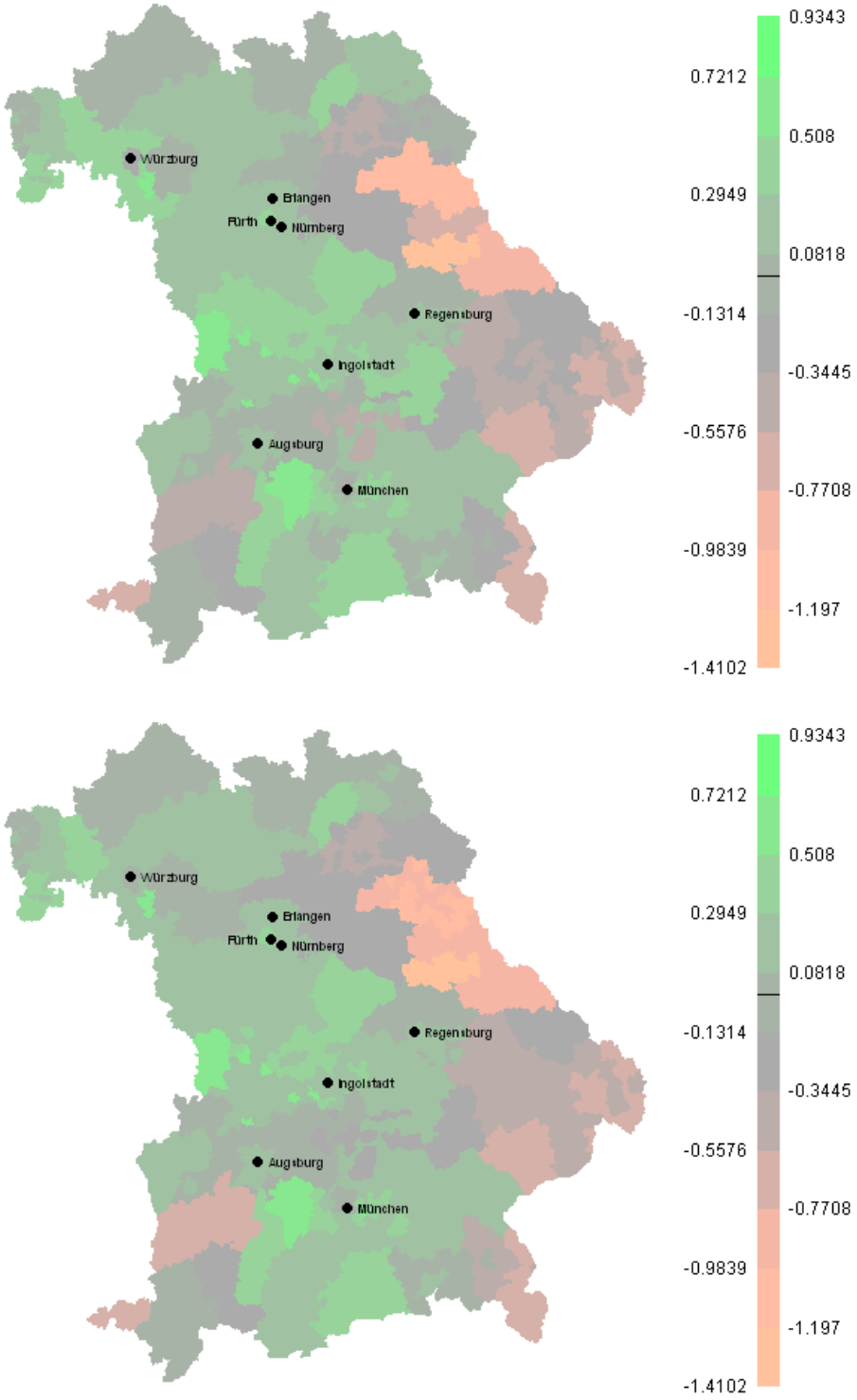

Abbildung 4.3.: *Oben:* Regionaler Effekt im unstrukturierten Modell der Bezirke der dreistelligen PLZs (1b). *Unten:* Regionaler Effekt im strukturierten Modell der Bezirke der dreistelligen PLZs, ausgewertet mit R per PQL (2b).

Bayerns hat Werte im gleichen Intervall wie schon im unstrukturierten Modell. Somit ist auch hier kein Smoothing-Effekt sichtbar.

Die Fehlerstruktur bzw. die räumlichen Effekte zeigen in allen Modellen einen deutlichen regionalen Unterschied. Der Osten und Südwesten weisen negative Vorzeichen auf. In den Modellen, welche auf den PLZ-Bezirken basieren, zieht sich vom Nordwesten zum Süden ein Streifen mit positivem Einfluss durch. Diese "Streifenbildung" ist in der Analyse der 36 Gebiete nicht zu erkennen, da die Gebiete sehr groß sind und ein kleinräumiger Effekt untergeht.
Das erwartete Smoothing in der strukturierten Analyse ist nur bedingt sichtbar.

Der ausgeprägte räumliche Effekt deutet daraufhin, dass nicht alle zur Erklärung des Screening-Verhaltens relevanten Einflussgrößen in den Modellen enthalten sind. Ein weiterer Hinweis darauf ist die Signifikanz fast aller untersuchter Variablen.

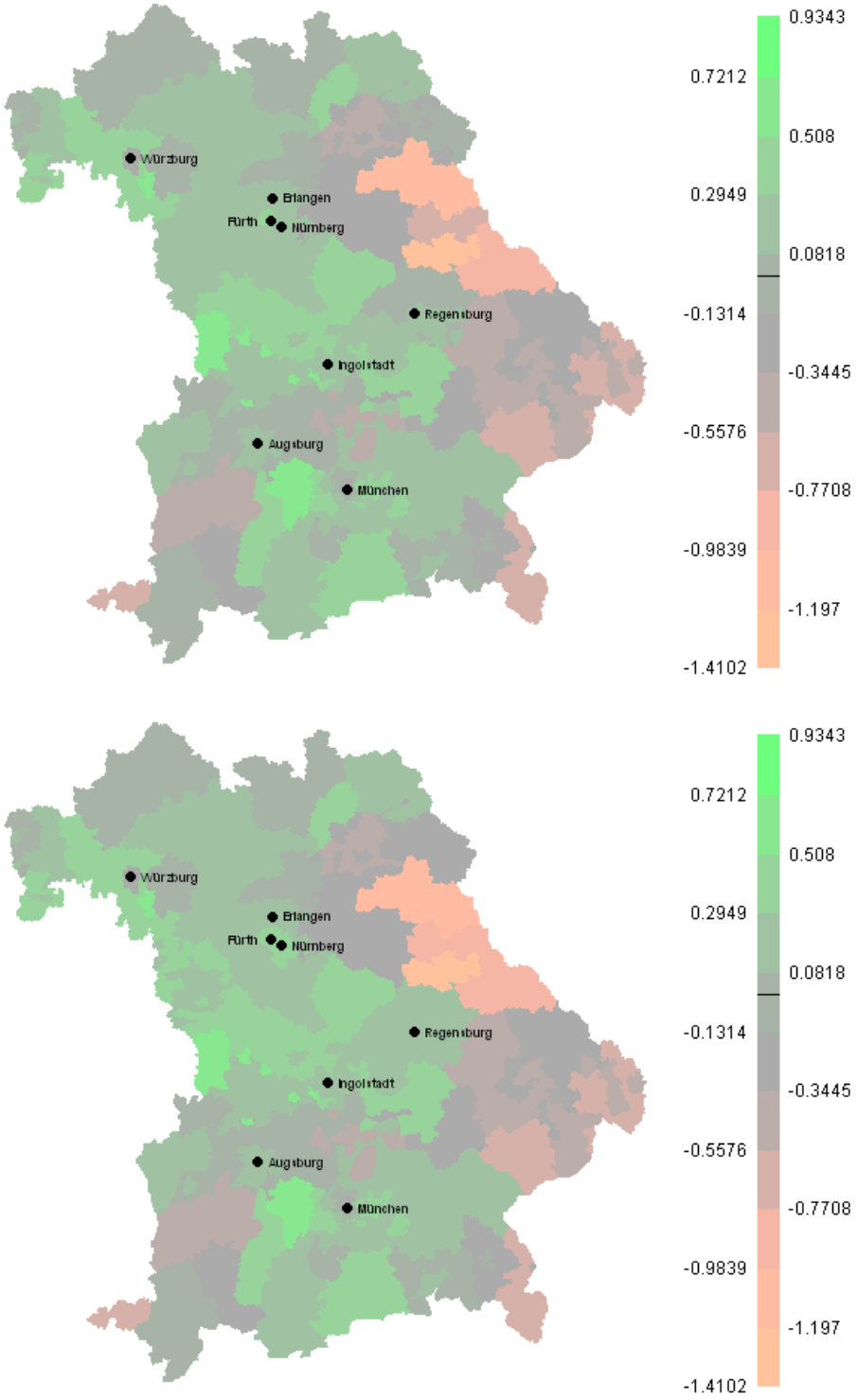

Abbildung 4.4.: *Oben:* Regionaler Effekt im unstrukturierten Modell der Bezirke der dreistelligen PLZs (1b). *Unten:* Regionaler Effekt im strukturierten Modell der Bezirke der dreistelligen PLZs, ausgewertet mit BayesX (3b).

5. Diskussion

5.1. Vergleich mit der bisherigen Auswertung

Die Auswertung von Mansmann et al. (2008) bezieht sich auf den Teil des Datensatzes, der im Jahr 2006 erhoben wurde. Er wurde auf die Qualität der Koloskopien hin untersucht. Weil nur die Daten von 2006 analysiert wurden, lässt sich der Zeitverlauf nicht vergleichen. Bei Mansmann et al. (2008) wurden außerdem die Schätzer des dort berechneten Poisson-Modells nicht veröffentlicht, jedoch detaillierte Nutzungsraten beschrieben. Diese sollen im Folgenden einem groben Vergleich der Ergebnisse dienen.

Eine weitere Auswertung der Daten stammt von Pritzkuleit et al. (2009). Die Jahre 2006 bis 2008 wurden jeweils einzeln ausgewertet. Das Ziel dieser Untersuchung war, Gründe für regionale Unterschiede in der Anfälligkeit für Darmkrebs herauszufinden.
Von Pritzkuleit et al. (2009) wurde kein statistisches Modell an die Daten angepasst, sondern lediglich verschiedene Raten (z. B. Teilnahme- oder Inzidenzraten) und Korrelationen berechnet. Diese Raten sollen ebenfalls zu einem groben Vergleich herangezogen werden.

Insgesamt sind die Nutzungsraten in beiden Artikeln denjenigen aus Kapitel 2.6.1 sehr ähnlich. Der zugrunde liegende Screening-Datensatz wurde jeweils von der KVB bezogen. Sowohl Pritzkuleit et al. (2009) als auch Mansmann et al. (2008) verwenden ebenfalls Bevölkerungsdaten vom Statistischen Landesamt, wie es hier bei den Modellen für die 36 Gebiete geschah, allerdings unterscheiden sich die Raten trotzdem leicht untereinander und zusätzlich von den hier berechneten. Die Altersgruppen in den Artikeln von Mansmann et al. (2008) und Pritzkuleit et al. (2009) entsprechen aufgrund der Datenlage denen, welche hier zur Analyse der 36 Gebiete verwendet wurden.

Der Effekt des Geschlechts wirkt sich in der Untersuchung von Mansmann et al. (2008) kaum aus. Die Nutzungsraten von Frauen und Männern sind nahezu identisch. Auch bei Pritzkuleit et al. (2009) ist die Teilnahme zwischen den Geschlechtern etwa gleich verteilt. In der vorliegenden Arbeit dagegen war der Effekt des Geschlechts signifikant von 0 verschieden, so dass die Chance, dass Männer zum Screening gehen, geringer ist als die der

Frauen (vgl. Kap. 4.1.1).

Der Effekt des Alters zeigt bei Mansmann et al. (2008) einen negativen Effekt auf die Nutzungsraten. Die jüngste Altersklasse "55 – 64 Jahre" hat die höchsten Raten aufzuweisen, die ältesten Teilnehmer (über 75 Jahre) haben die niedrigste Nutzungsrate. Die gleiche Wirkung des Älterwerdens ist in den Nutzungsraten von Pritzkuleit et al. (2009) zu sehen. Auch dort nimmt die Nutzung mit dem Alter deutlich ab. In der Tendenz sind die Ergebnisse mit demjenigen der vorliegenden Arbeit vergleichbar. Die Odds Ratios sind hier für alle Altersgruppen kleiner als 1. Der Einfluss des Alters ist signifikant (vgl. Kap. 4.1.4).

Die Interaktion zwischen Geschlecht und Alter zeigt bei Mansmann et al. (2008) ein komplexes Muster. In der jüngsten Altersklasse gehen mehr Frauen als Männer zum Screening. In der mittleren Kategorie sind die Raten fast identisch. In der höchsten Altersgruppe war die Nutzungsrate der Frauen niedriger als die der Männer. Die Nutzungsraten bei Pritzkuleit et al. (2009) zeigen das Gleiche.

Dieses komplexe Muster zeigt sich im vorliegenden Datensatz bei den Nutzungsraten (Kap. 2.6.1) ebenfalls, allerdings bestätigt kein berechnetes Modell die Ergebnisse: Der negative Effekt des Geschlechts überwiegt die an sich positive Interaktion. Denn ein Mann der mittleren Altersklasse hat eine niedrigere Chance zur Koloskopie zu gehen als eine Frau der jüngsten Altersklasse (Odds Ratio von 0.66, Analyse der 36 Gebiete). Gleiches gilt für Männer aus der höchsten Altersklasse im Vergleich zu den Frauen aus der niedrigsten Altersstufe in stärkerem Maße (Odds Ratio von 0.28, Analyse der 36 Gebiete). Bei der Auswertung auf Basis der PLZ-Bezirke zeigt sich das gleiche Bild. Der Schätzer dieser Interaktion unterschied sich in allen berechneten Modellen signifikant von 0 (vgl. Kap. 4.1.5).

Der Effekt des Zeitverlaufs kann – wie oben erwähnt – bei Mansmann et al. (2008) nicht verglichen werden. Bei Pritzkuleit et al. (2009) ist ein Anstieg in den Raten von 2006 auf 2007 zu erkennen sowie ein Sinken von 2007 auf das Jahr 2008. Diese Beobachtung wurde von allen hier berechneten und in Kapitel 4 vorgestellten Modellen bestätigt. Der Effekt des Zeitverlaufs ist positiv, aber abnehmend und in allen Modellen signifikant (vgl. Kap. 4.1.2).

Eine Interaktion zwischen Geschlecht und Zeitverlauf ist bei Pritzkuleit et al. (2009) nicht offensichtlich. Die hier gerechneten Modelle ergaben einen leicht positiven Effekt der

Interaktion. Frauen scheinen demzufolge von 2007 auf 2008 nicht so stark in der Nutzung nachgelassen zu haben bzw. zeigten einen stärkeren Anstieg von 2006 auf 2007. Fast alle Schätzer waren trotz ihres geringen Betrags signifikant von 0 verschieden (vgl. Kap. 4.1.3).

Die Fehlerstruktur des von Mansmann et al. (2008) berechneten Poisson-Modells zeigt für den Osten Bayerns niedrige Raten bzw. einen negativen regionalen Effekt. Während der Südwesten ebenfalls niedrige Nutzungsraten zeigt, wurden die höchsten Teilnahmeraten im Umkreis von Großstädten beobachtet. Die Karten in Pritzkuleit et al. (2009) beschreiben die gleiche Verteilung im Raum: Im Osten und Südwesten Bayerns sind die Teilnahmeraten am geringsten. Dies konnte man auch bei der vorliegenden Arbeit in den regionalen Effekten erkennen. Besonders bei der Analyse, welche auf den Daten der dreistelligen PLZ-Bezirke basiert, fällt die Ähnlichkeit zu den Arbeiten von Mansmann et al. (2008) und Pritzkuleit et al. (2009) auf (vgl. z. B. Abb. 4.3).

5.2. Probleme bei der Auswertung

Es ergaben sich einige Probleme bei der Auswertung der Screening-Daten:

Ein sehr großes Problem stellte die Erfassung der Wohnorte durch die KVB dar. Im Moment werden die ersten drei Ziffern der Postleitzahl des Wohnortes der Klienten erfasst. Dies ist problematisch, da für die Bezirke der dreistelligen Postleitzahlen kaum Daten zur Bevölkerung verfügbar sind. Viele Einflussgrößen sind nur auf Gemeinde- oder Landkreisebene erhoben. Dies führt dazu, dass man Gebiete erstellen muss, welche aber z. T. sehr flächendeckend sind und somit kleinräumige regionale Unterschiede, z. B. in der Nähe von Großstädten, nicht mehr erkennbar sind (vgl. Kap. 2.2.1). Dies ist hier z. B. bei Gebiet 3 der Fall, welches München, Ingolstadt, Rosenheim und weitere Städte umfasst. Gebiet 5 ist ebenfalls flächenmäßig sehr groß.
Wesentlich einfacher wäre eine Auswertung der Daten, wenn die KVB den Wohnort nicht postalisch erfassen würde, sondern nach dem Landkreis oder der Gemeinde fragen würde, in welcher der Klient wohnt. Dann könnte man auf die amtlich erfassten Daten der GENESIS-Datenbank zurückgreifen, was einfacher wäre und auch eine Einsparung von Kosten bedeuten würde, da man diese Datensätze nicht käuflich erwerben muss wie Daten auf Basis der dreistelligen PLZ-Bezirke.

Die Ergebnisse der Analyse könnten durch folgende Umstände verzerrt sein:

Zum einen ist eine falsche Abrechnung der Ärzte eine Fehlerquelle des Datensatzes. Eventuell wurde eine eigentlich präventive Koloskopie – welche hier ausgewertet wurden – bei

einem Fund von Krebs oder Krebsvorstufen als kurative Koloskopie abgerechnet und somit im Datensatz falsch erfasst.

Zum andern könnte die internetbasierte Dokumentation der Daten problematisch sein. Eventuell sind die Ärzte in den ländlichen Regionen Bayerns aufgrund des noch immer lückenhaften Breitband-Ausbaus nicht gewillt, sich mit der Dokumentation per Internet zu befassen oder aber mit der Nutzung des Internets nicht vertraut.

Ein weiteres Problem ist, dass Privatpatienten in den Statistiken nicht erfasst werden. Aus diesem Grund muss eine Einflussgröße in das Modell aufgenommen werden, welche den Anteil der nicht-erfassten Koloskopien modelliert. Ein Ausweg wäre, die Anzahl gesetzlich versicherter Personen je räumlicher Einheit in den Datensatz der KVB aufzunehmen. Im Idealfall basiert der gesamte Datensatz auf Landkreisen oder Gemeinden.

Weiterhin ist bemerkenswert, dass Pritzkuleit et al. (2009) die Gemeinden und PLZ-Bezirke Bayerns zu 77 Gebieten zusammenfassen konnten, während hier nur 36 Gebiete eine eindeutige Zuordnung gewährleisten. Dies liegt wohl an unterschiedlichen, der Zuordnung zugrunde liegenden Listen der Gemeinden mit zugehörigen Postleitzahlen.

5.3. Schluss

Die in dieser Arbeit vorgestellte Auswertung von Screening-Daten der Kassenärztlichen Vereinigung Bayerns (KVB) erfolgte flächendeckend

1. in 36 Gebieten, in denen die Zuordnung von Gemeinden und PLZ-Bezirken eindeutig ist (vgl. Kap. 2.2.1)
2. in den Bezirken der dreistelligen Postleitzahlen mit Hilfe der gekauften Daten von der Schober Information Group Deutschland GmbH

Das Ziel der Analyse war es, herauszufinden,

- ob in Bayern regionale Unterschiede in der Nutzung des Screenings vorhanden sind,
- welche Größen darauf Einfluss haben und
- ob die Nutzung über die Zeit stabil ist.

Zu diesem Zweck wurden räumlich unstrukturierte und räumlich strukturierte Modelle berechnet. Die Auswertung erfolgte mit den statistischen Programmpaketen R (R Development Core Team, 2009) und BayesX (Belitz et al., 2009).

Man konnte schon bei der deskriptiven Analyse der Daten erkennen, dass die Nutzungsraten räumlich stark variieren (Kap. 2.6.2 und 2.6.3). Nach der Anpassung der Modelle

bleibt das bereits festgestellte Ost-West-Gefälle in den räumlichen Effekten bestehen (Kap. 4.2). Dies deutet darauf hin, dass eventuell noch nicht alle Größen gefunden wurden, die Einfluss auf das Nutzungsverhalten von Krebsvorsorge-Angeboten haben.

Was sich ebenfalls schon bei der deskriptiven Betrachtung der Daten zeigte, war der Effekt des Geschlechts, des Alters und deren Interaktion (Kap. 2.6.1). Diese Einflussgrößen waren in allen vorgestellten Modellen auf dem 5%-Niveau signifikant (Kap. 4.1.1, 4.1.4 und 4.1.5). Allerdings bestätigten die Schätzer dieser festen Effekte das Muster in den Nutzungsraten nicht. Es zeigte sich in den Modellen weniger kompliziert.

Überraschenderweise hat die Anzahl an Gastroenterologen einen leicht negativen Einfluss auf die Nutzung. Der Effekt ist jedoch sehr gering und nur in einem Modell signifikant von 0 verschieden.
Eine weitere Überraschung zeigte sich beim Sozialstatus bzw. beim Einkommen. Diese Variablen waren ebenfalls nicht signifikant, ihre Schätzer waren nahezu 0. In Auswertungen anderer Datensätze, z. B. bei Rückinger et al. (2008) (Brustkrebs-Screening in Bayern) oder bei Nnoaham et al. (2010) (Darmkrebs-Screening in Südengland), spielte die soziale Situation durchaus eine Rolle, ob Screening-Angebote genutzt werden oder nicht.
Der Anteil christlicher Einwohner war nur für die 36 Gebiete verfügbar und zeigte einen negativen Einfluss. Allerdings war auch der Schätzer dieses Effekts nur in einem Modell signifikant.

Deskriptiv ergab sich eine leichte Schwankung in den Nutzungsraten (Kap. 2.6.1) und auch in allen Modellen war der Faktor "Zeit" signifikant (Kap. 4.1.2). Somit konnte eine Stabilität über die Zeit nicht gezeigt werden.
Zwischen den Geschlechtern konnte ein kleiner Unterschied in der Nutzung über die Zeit festgestellt werden. Trotz seines geringen Betrags war dieser feste Effekt mitunter signifikant (Kap. 4.1.3).

Falls weiterführende Analysen der Nutzung des Screening-Programms für Kolonkarzinome durchgeführt werden sollen, wird eine Bearbeitung des Formulars zur Erfassung der Daten dringend empfohlen. Statt wie bisher nach der Postleitzahl des Wohnorts des Klienten zu fragen, sollten der Landkreis oder die Gemeinde aufgenommen werden. Denn auf Basis der amtlichen Einheiten sind viel mehr Variablen verfügbar als auf Basis der postalischen Regionen. Man könnte als weitere Einflussgrößen z. B. die Einwohnerdichte als Messgröße für eher städtische oder eher ländliche Regionen in die Modelle aufnehmen. Ebenso wäre das verfügbare Einkommen in den Landkreisen kostenfrei in der GENESIS-Datenbank enthalten. Eventuell spielt der kulturelle Hintergrund eine Rolle. Die Ausländerstatistik des Landesamtes könnte auch in dieser Frage weiterhelfen. Weitere

soziale Faktoren wie Arbeitslosigkeit oder die Ausbildung der Einwohner sind ebenfalls in der GENESIS-Datenbank verzeichnet und könnten – falls der Landkreis und nicht der PLZ-Bezirk bekannt ist – relativ einfach in weitere Analysen einbezogen werden.

Literaturverzeichnis

Bates, D. und Maechler, M. (2009): *lme4: Linear mixed-effects models using S4 classes*. URL http://CRAN.R-project.org/package=lme4. R package version 0.999375-32.

Belitz, C., Brezger, A., Kneib, T. und Lang, S. (2009): *BayesX – Software for Bayesian inference in structured additive regression models*. URL http://www.stat.uni-muenchen.de/~bayesx.

Bivand, R., with contributions by Hisaji Ono und Dunlap, R. (2009a): *classInt: Choose univariate class intervals*. URL http://CRAN.R-project.org/package=classInt. R package version 0.1-14.

Bivand, R., with contributions by Luc Anselin, Assunção, R., Berke, O., Bernat, A., Carvalho, M., Chun, Y., Christensen, B., Dormann, C., Dray, S., Halbersma, R., Krainski, E., Lewin-Koh, N., Li, H., Ma, J., Millo, G., Mueller, W., Ono, H., Peres-Neto, P., Piras, G., Reder, M., Tiefelsdorf, M. und Yu., D. (2009b): *spdep: Spatial dependence: weighting schemes, statistics and models*. URL http://CRAN.R-project.org/package=spdep. R package version 0.4-56.

Breslow, N. E. und Clayton, D. G. (1993): *Approximate inference in generalized linear mixed models*. Journal of the American Statistical Association, **88**(421), 9–25.

Dean, C. B., Ugarte, M. D. und Militino, A. F. (2004): *Penalized quasi-likelihood with spatially correlated data*. Computational Statistics & Data Analysis, **45**(2), 235–248.

Fahrmeir, L., Kneib, T. und Lang, S. (2007): *Regression – Modelle, Methoden und Anwendungen*. Springer-Verlag, Berlin, Heidelberg.

Holmes, C. C. und Held, L. (2002): *Bayesian auxiliary variable models for binary and multinomial regression*. Bayesian Analysis, **1**(1), 145–168.

Hothorn, T., Bretz, F. und Westfall, P. (2008): *Simultaneous inference in general parametric models*. Biometrical Journal, **50**(3), 346–363.

Kneib, T. (2009/2010): *Skript zur Vorlesung "Räumliche Statistik"*. Carl-von-Ossietzky-Universität, Oldenburg.

Kneib, T., Heinzl, F., Brezger, A. und Bove, D. S. (2009): *BayesX: R Utilities Accompanying the Software Package BayesX*. URL http://CRAN.R-project.org/package=BayesX. R package version 0.2-4.

Leisch, F. (2002): *Sweave: Dynamic generation of statistical reports using literate data analysis*. In: Härdle, W. und Rönz, B. (Hg.), *Compstat 2002 — Proceedings in Computational Statistics*. Physica Verlag, Heidelberg, 575–580. URL http://www.stat.uni-muenchen.de/~leisch/Sweave. ISBN 3-7908-1517-9.

Leisch, F. (2008): *Skript zur Vorlesung "Computerintensive Methoden"*. Ludwig-Maximilians-Universität, München.

Lewin-Koh, N. J., Bivand, R., contributions by Edzer J. Pebesma, Archer, E., Baddeley, A., Bibiko, H.-J., Dray, S., Forrest, D., Giraudoux, P., Golicher, D., Gómez-Rubio, V., Hausmann, P., Jagger, T., Luque, S. P., MacQueen, D., Niccolai, A. und Short, T. (2009): *maptools: Tools for reading and handling spatial objects*. URL http://CRAN.R-project.org/package=maptools. R package version 0.7-26.

Maier, W., Fairburn, J. und Mielck, A. (in Bearbeitung): *Deprivation und Mortalität in Bayern: Entwicklung eines "Index Multipler Deprivation" auf Gemeindebasis*. Zeitschrift noch unbekannt.

Mansmann, U., Crispin, A., Henschel, V., Adrion, C., Augustin, V., Birkner, B. und Munte, A. (2008): *Bilanz der Qualitätssicherung ambulanter Koloskopien nach 245000 Untersuchungen*. Deutsches Ärzteblatt, **105**(24), 434–440.

Meintrup, D. und Schäffler, S. (2005): *Stochastik – Theorie und Anwendungen*. Springer-Verlag, Berlin, Heidelberg.

Nnoaham, K. E., Frater, A., Roderick, P., Moon, G. und Halloran, S. (2010): *Do geodemographic typologies explain variations in uptake in colorectal cancer screening? an assessment using routine screening data in the south of england*. Journal of Public Health, **32**(4), 572–581.

Pebesma, E. J. und Bivand, R. S. (2005): *Classes and methods for spatial data in R*. R News, **5**(2), 9–13. URL http://CRAN.R-project.org/doc/Rnews/.

Pritzkuleit, R., Katalinic, A., Polenz, K. und Kretschmer, W. (2009): *Räumliche Karzinomanalyse: regionaler Vergleich der Inzidenz kolorektaler Karzinome in Bayern im Zeitraum 2006–2008*. Abschlussbericht.

R Development Core Team (2009): *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org. ISBN 3-900051-07-0.

Rückinger, S., von Kries, R., Pauli, S., a. Munte und Mielck, A. (2008): *Die krebsfrüherkennungsuntersuchung für frauen wird in regionen mit niedrigerem haushaltseinkommen seltener in anspruch genommen – analyse von daten der kassenärztlichen vereinigung bayerns*. Das Gesundheitswesen, **70**, 393–397.

Schödlbauer, A. (1982): *Rechenformeln und Rechenbeispiele zur Landesvermessung. Teil 2: Geodätische Berechnungen im System der Gaußschen konformen Abbildung eines Bezugsellipsoids unter besonderer Berücksichtigung des Gauß-Krüger- und des UTM-Koordinatensystems im Bereich der Bundesrepublik Deutschland*, Bd. 2 von *Wichmann-Skripten*. Herbert-Wichmann-Verlag, Karlsruhe.

Schmid, V. (2010): *Skript zur Vorlesung "Räumliche Statistik"*. Ludwig-Maximilians-Universität, München.

Toutenburg, H. (2003): *Lineare Modelle – Theorie und Anwendungen*. Physika-Verlag, Heidelberg, 2. Aufl.

Townsend, P. (1979): *Poverty in the United Kingdom: a Survey of Household Resources and Standards of Living.* University of California Press, Berkeley und Los Angeles.

Tutz, G. (2000): *Die Analyse kategorialer Daten.* Oldenbourg Wissenschaftsverlag.

Wood, S. N. (2008): *Fast stable direct fitting and smoothness selection for generalized additive models.* Journal of the Royal Statistical Society (B), **70**(3), 495–518.

Zeileis, A., Hornik, K. und Murrell, P. (2009): *Escaping RGBland: Selecting colors for statistical graphics.* Computational Statistics & Data Analysis, **53**, 3259–3270.

Verzeichnis der Datenquellen

Bayerische Vermessungsverwaltung
Bayerisches Landesamt für Vermessung und Geoinformation
München
http://vermessung.bayern.de

GENESIS-Datenbank
Bayerisches Landesamt für Statistik und Datenverarbeitung
München
http://www.statistikdaten.bayern.de

Prof. Thomas Kneib
Institut für Mathematik
Fakultät für Mathematik und Naturwissenschaften
Carl-von-Ossietzky-Universität, Oldenburg
http://www.staff.uni-oldenburg.de/thomas.kneib

Werner Maier
Helmholtz Zentrum München
Institut für Gesundheitsökonomie und Management im Gesundheitswesen
Neuherberg
http://www.helmholtz-muenchen.de/igm/institut/team/mitarbeiter/wernermaier

Prof. Ulrich Mansmann
Institut für medizinische Informationsverarbeitung, Biometrie und Epidemiologie
Medizinische Fakultät
Ludwig-Maximilians-Universität, München
http://www.ibe.med.uni-muenchen.de/organisation/mitarbeiter/010_direktorat/mansmann

Schober Information Group Deutschland GmbH
Standort München
http://www.schober.de

A. Technische Daten

Die meisten Analysen und das Daten-Management wurden mit R-Version 2.9.2 durchgeführt (R Development Core Team, 2009). Zur Bearbeitung räumlicher Daten sind die Pakete `sp` (Version 0.9-57, Pebesma und Bivand (2005)) und `spdep` (Version 0.4-56, Bivand et al. (2009b)) notwendig. Für das Einlesen der Karten von `http://www.gadm.org` wurde das Paket `maptools` (Version 0.7-26, Lewin-Koh et al. (2009)) gebraucht. Die Ergebnisse der unstrukturierten Modelle wurden mit Hilfe des Pakets `lme4` (Version 0.999375-32, Bates und Maechler (2009)) erhalten. Die strukturierten Modelle konnte durch die Software des Paketes `mgcv` (Version 1.5-6, Wood (2008)) und den Code von Prof. Kneib angepasst werden. Die generalisierten Hypothesentests wurden jeweils mit dem Paket `multcomp` (Version 1.1-4, Hothorn et al. (2008)) durchgeführt. Für die Vorbereitung der bayesianischen Modelle und zur Verarbeitung ihrer Ergebnisse wurde das Paket `BayesX` (Version 0.2-4, Kneib et al. (2009)) verwendet. Von der Software BayesX (Belitz et al., 2009) selbst wurde die Version 2.0.1 verwendet. Die Grafiken der Nutzungsraten und der regionalen Effekte wurden jeweils so gestaltet, dass sie untereinander vergleichbar sind. Dazu wurden die Intervalle mit dem Paket `classInt` (Version 0.1-14, Bivand et al. (2009a)) bestimmt. Die Farben kommen aus dem Paket `colorspace` (Version 1.0-1, Zeileis et al. (2009)). Sie entsprechen den Grundeinstellungen aus BayesX.

Diese Arbeit wurde mit den Programmen LaTeX (MiKTeX Version 2.7) und z. T. mit Sweave (Leisch, 2002) erstellt. Zur Kompilierung des Sweave-Codes ist das R-Paket `tools` (Version 2.9.2, R Development Core Team (2009)) notwendig.

B. Elektronischer Anhang

Der elektronische Anhang enthält den R- und BayesX-Code sowie drei Ordner. Der Ordner `Daten` enthält sämtliche Datensätze in unbearbeitetem Zustand sowie diverse Zwischenschritte bis hin zu den endgültig für die Modellierung verwendeten Daten. Der Ordner `Modelle` beinhaltet die Ergebnisse der Modell-Berechnungen. Im Ordner `Grafiken mit Code` befinden sich die Grafiken, welche für diesen Text verwendet wurden, sowie der R-Code, um diese und Tabellen des Textes zu erstellen.

Im Einzelnen enthalten die folgenden Dateien Code für das Daten-Management:

- `01 Liste der PLZ mit Gemeindeschluessel.R`: Code für Kapitel 2.1
- `02 Zsmfassung zu Gebieten.R`: Code für Kapitel 2.2.1
- `03a Screening-Daten.R` und `03b Screening-Daten der Gebiete.R`: Daten-Management für die Screening-Daten (Kap. 2.3.1)
- `04 Gemeinde-Daten 2006.R`: Daten-Management für die Gemeinde-Daten (Kap. 2.3.2.1) des Jahres 2006 inkl. Korrektur der Altersklasse von "50 bis 65" auf "55 bis 65" mit Hilfe der Landkreis-Daten des Jahres 2006
- `05 Gemeinde-Daten 2007.R`: Daten-Management für die Gemeinde-Daten (Kap. 2.3.2.1) des Jahres 2007 inkl. Korrektur der Altersklasse von "50 bis 65" auf "55 bis 65" mit Hilfe der Landkreis-Daten des Jahres 2007
- `06 Gemeinde-Daten 2008.R`: Daten-Management für die Gemeinde-Daten (Kap. 2.3.2.1) des Jahres 2008 inkl. Korrektur der Altersklasse von "50 bis 65" auf "55 bis 65" mit Hilfe der Landkreis-Daten des Jahres 2008
- `07 PLZ3-Daten und Kaufkraft je Einwohner.R`: Daten-Management für die Bevölkerungsdaten (Kap. 2.3.2.1) von der Schober Information Group inkl. Daten-Management für die Kaufkraft je Einwohner aus Kapitel 2.3.2.4
- `08 Religion.R`: Daten-Management für die Religionsverteilung aus Kapitel 2.3.2.2
- `09 Deprivationsindex.R`: Daten-Management für den Index multipler Deprivation von Werner Maier (Kap. 2.3.2.3)
- `10a Gastroenterologen der Gebiete.R` und `10b Gastroenterologen der PLZ3.R`: Daten-Management für den Gastroenterologen-Datensatz (Kap. 2.3.2.5)

Die folgenden Dateien erstellen und bearbeiten die Karten von Bayern (Kap. 2.5):

- `11a Bayernkarte von gadm-org.R`: Code, um das Kartenobjekt von `http://www.gadm.org` für die weitere Verwendung aufzubereiten
- `11b Karte geografisch.R`: Umwandlung der Koordinaten des Shapefiles von der Schober Information Group vom Gauß-Krüger- ins geografische Koordinatensystem
- `11c Bayernkarten mit shapefile.R`: Erstellung der Karten der dreistelligen PLZ-

Bezirke und der 36 Gebiete (Kap. 2.5.1 und 2.5.2)

Der Code in diesen Dateien führt die einzelnen Datensätze jeweils geeignet für das folgende Modell zusammen und passt das Modell an:

- `12a Modell unstrukturiert Gebiete.R`: Code für die 36 Gebiete aus Kapitel 3.5.1
- `12b Modell unstrukturiert PLZ3.R`: Code für die Bezirke der dreistelligen Postleitzahlen aus Kapitel 3.5.1
- `13 Nachbarschaftsstruktur.R`: Erstellung der Nachbarschaftsstrukturen beider Kartentypen (36 Gebiete und PLZ)
- `14a Modell strukturiert PQL Gebiete.R`: Code für die 36 Gebiete aus Kapitel 3.5.2
- `14b Modell strukturiert PQL PLZ3.R`: Code für die Bezirke der dreistelligen Postleitzahlen aus Kapitel 3.5.2
- `15a BayesX-Datensatz für Gebiete.R`: Erstellung der Dummyvariablen für die 36 Gebiete zur Verwendung in BayesX (Kap. 3.5.3)
- `15b BayesX-Datensatz für PLZ3.R`: Erstellung der Dummyvariablen für die dreistelligen PLZ-Bezirke zur Verwendung in BayesX (Kap. 3.5.3)
- `16aGebiete.txt`: BayesX-Code zu Kapitel 3.5.3 für die 36 Gebiete
- `16bPLZ3.txt`: BayesX-Code zu Kapitel 3.5.3 für die dreistelligen PLZ-Bezirke

Die Grafiken und Tabellen dieser Arbeit wurden mit den folgenden R-Dateien erstellt:

- `01 Histogramm Screening-Daten.R`: Code für Abbildung 2.1
- `02 Boxplot Religionszugehoerigkeit.R`: Code für Abbildung 2.2
- `03 Boxplot Kaufkraft.R`: Code für Abbildung 2.3
- `04 Verschiebung.R`: Code für Abbildung 2.4
- `05a Nutzungsraten.R`: Code zur Berechnung der Nutzungsraten (Kap. 2.6)
- `05b Grossstaedte.R`: Code, um die Großstädte in die Karten einzuzeichnen (Ergebnis wird in `05c Plotfunktionen.R` verwendet)
- `05c Plotfunktionen.R`: Code zum einfacheren Zeichnen der Karten (wird von den betreffenden Dateien jeweils intern verwendet)
- `05d Plots der Nutzungsraten.R`: Code für die Abbildungen 2.5, 2.6 und 2.7
- `06 Praezisionsmatrix Beispiel Karte + TAB.R`: Code für Abbildung 3.2 und Tabelle 3.1
- `07a vektor.R`: wird in `07b Regionale Effekte.R` verwendet
- `07b Regionale Effekte.R`: Code für die Abbildungen 4.1, 4.2, 4.3 und 4.4
- `08a Gebietekarte.r`: Code für Abbildung D.1
- `08b PLZ3karte.r`: Code für die Abbildungen D.2, D.3 und D.4
- `TAB 01 Zsmfassung zu Gebieten.R`: Code für Tabelle 2.1
- `TAB 02 Verteilung der Screening-Nutzer.R`: Code für die Tabellen 2.2 und 2.3

- `TAB 03 Verteilung der Einwohner.R`: Code für die Tabellen 2.4 und 2.5
- `TAB 04 Verteilung der Nutzungsraten.R`: Code für die Tabellen 2.7 und 2.8
- `TAB 05 Gemeindeschluessel mit PLZ und Gebietsnr.R`: Code für die Liste der Gemeindeschlüssel mit Postleitzahlen und Gebietsnummer. Diese Tabelle findet sich aufgrund ihres Umfangs nur im elektronischen Anhang. Sie ist dort als `txt`-Datei im Order "Daten" unter dem Namen `gkz-plz-geb.txt` abgespeichert.

Die Sweave-Dateien wurden mit `SweaveZuTex.R` in LaTeX-Code umgewandelt und konnten dann in den laufenden Text eingearbeitet werden.

- `Modell-Code.rnw`: zur Erstellung der Code-Abschnitte in Kapitel 3.5
- `TAB feste Effekte.rnw`: Code für die Tabellen 4.1, 4.2, 4.3, 4.4, 4.5, 4.6, 4.7, 4.8 und 4.9

C. Postleitzahl-Bezirke pro Gebiet

Tabelle C.1 zeigt, welche Gemeinden mit mind. 25000 Einwohnern welchem Gebiet zugeordnet wurden. Diese Tabelle soll, ähnlich wie Abbildung D.2 für die Postleitzahlbezirke, einer groben Orientierung dienen. Genauere Angaben über die Zugehörigkeit der Gemeinden zu den Gebieten findet sich in der Tabelle `gkz-plz-geb.txt` im Ordner "Daten" des elektronischen Anhangs. Die Lage der Gebiete in Bayern kann Abbildung D.1 entnommen werden.

Tabelle C.1.: Tabelle der Gebiete mit zugehörigen Postleitzahlen und Gemeinden mit mind. 25000 Einwohner (Stand 2006).

Gebiet	Postleitzahlen	enthaltene Gemeinden ($\geq$ 25000 Einwohner, Stand 2006)
1	637–639, 747, 978, 979	Aschaffenburg
2	745, 916	
3	803–809, 812–825, 830, 831, 835–837, 850–857, 861, 863–865, 867–869, 933	Ingolstadt, München, Rosenheim, Dachau, Erding, Freising, Fürstenfeldbruck, Germering, Garmisch-Partenkirchen, Landsberg am Lech, Unterschleißheim, Augsburg, Friedberg, Königsbrunn
4	832–834	
5	840, 841, 922, 925, 934, 940–945	Landshut, Passau, Straubing, Deggendorf, Amberg
6	843	
7	844, 845	
8	866	Neuburg a.d.Donau
9	874, 876, 877, 881	Kaufbeuren, Kempten (Allgäu), Memmingen
10	875	
11	884, 890, 892	Neu–Ulm
12	893	
13	894	
14	904, 905, 910, 914, 923	Neumarkt i.d.OPf., Erlangen, Nürnberg, Zirndorf
15	906	
16	907	Fürth
17	911	Schwabach
18	912	Lauf a.d.Pegnitz
19	913	Forchheim
20	915	Ansbach
21	917	
22	918	
23	924, 930, 931	Regensburg, Schwandorf
24	926	Weiden i.d.OPf.
25	927, 956, 957	
26	950	Hof
27	951–955, 963	Bayreuth, Kulmbach
28	960	Bamberg
29	961, 974, 975	Schweinfurt
30	962	
31	964	Coburg
32	970, 972	Würzburg
33	971	
34	973	
35	976	
36	977	

D. Kartenteil

Die folgenden Karten sollen dazu dienen, ein bestimmtes Gebiet oder einen gewissen Postleitzahlbezirk zu identifizieren.

D.1. Karte der 36 Gebiete

Die 36 Gebiete sind dergestalt, dass sie sich u. U. aus mehreren Teilen zusammensetzen, Enklaven in anderen Gebieten haben oder auch wechselseitig ineinander liegen. Aus diesem Grund ist eine einfache Darstellung der Grenzen und das Zeichnen der zugehörigen Gebietsnummer innerhalb dieser nicht möglich. Die Darstellung in Abbildung D.1 ist eine platzsparende Möglichkeit, die Gebiete so darzustellen, dass man ein gesuchtes Gebiet relativ einfach identifizieren kann.
Die Gebietsnummern steigen mit den Postleitzahlen: Niedrige Postleitzahlen wie z. B. 637xx bilden i. d. R. ein Gebiet mit niedriger Gebietsnummer. Im Beispiel bildet 637xx zusammen mit anderen Postleitzahlen das Gebiet 1. Die genauen Zuordnungen sind in Tabelle C.1 zu sehen.

D.2. Karte der Bezirke der dreistelligen Postleitzahlen

Ebenfalls wie die 36 Gebiete sind die Postleitzahlbezirke keine zusammenhängenden Flächen, sondern bestehen mitunter aus mehreren kleineren Flächen, die innerhalb anderer Bezirke liegen können. Aufgrund der größeren Zahl von Regionen wurde die Bayernkarte dazu zweigeteilt und in Nord- und Südbayern getrennt. Für eine erste Orientierung dient die Karte der zweistelligen PLZ-Bezirke (Abb. D.2). Die Postleitzahlbezirke Nordbayerns bis etwa zur Donau sind in Abbildung D.3 dargestellt. Mithilfe der Codierung aus Farbe und Musterung sollte es einigermaßen möglich sein, einen gesuchten PLZ-Bezirk ausfindig zu machen. Gleiches gilt für die entsprechend aufbereitete Karte von Südbayern (Abb. D.4). Die Landeshauptstadt München besteht aus mehreren Postleitzahlbezirken. Diese sind daher sehr kleinräumig und deswegen schlecht voneinander zu unterscheiden. Es betrifft die Postleitzahlen von 803xx bis 820xx.

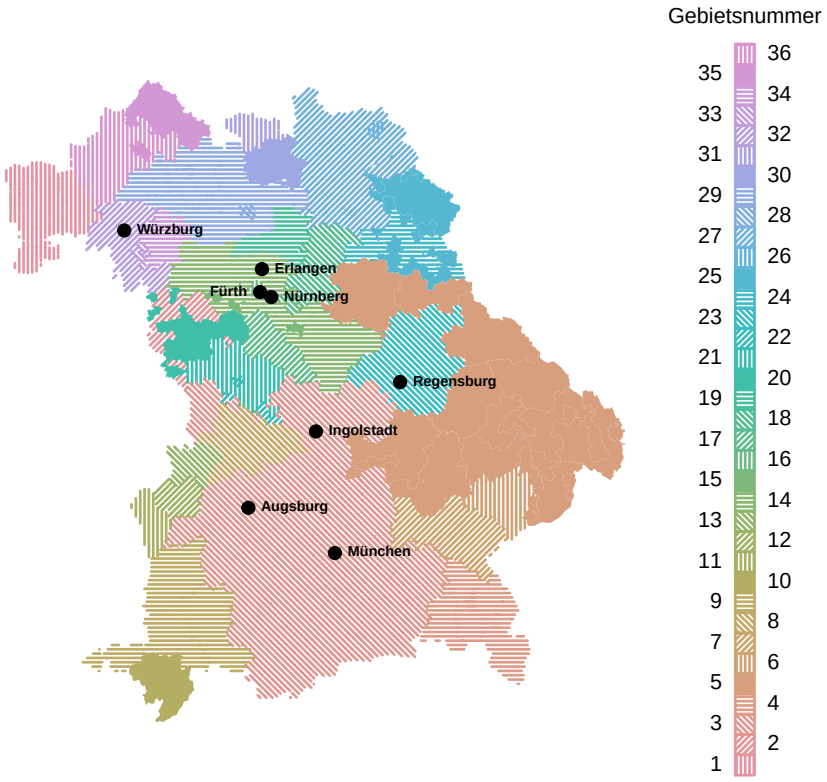

Abbildung D.1.: Karte der 36 Gebiete.

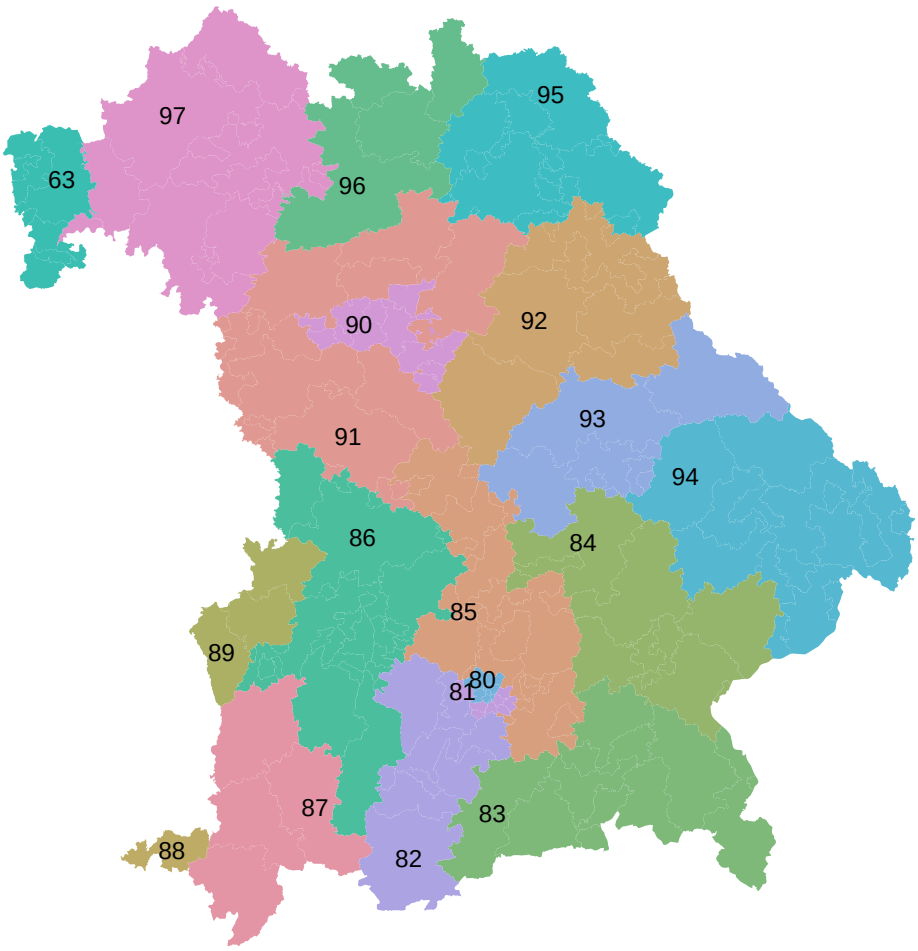

Abbildung D.2.: Karte der Bezirke der ersten zwei Stellen der Postleitzahlen in Bayern.

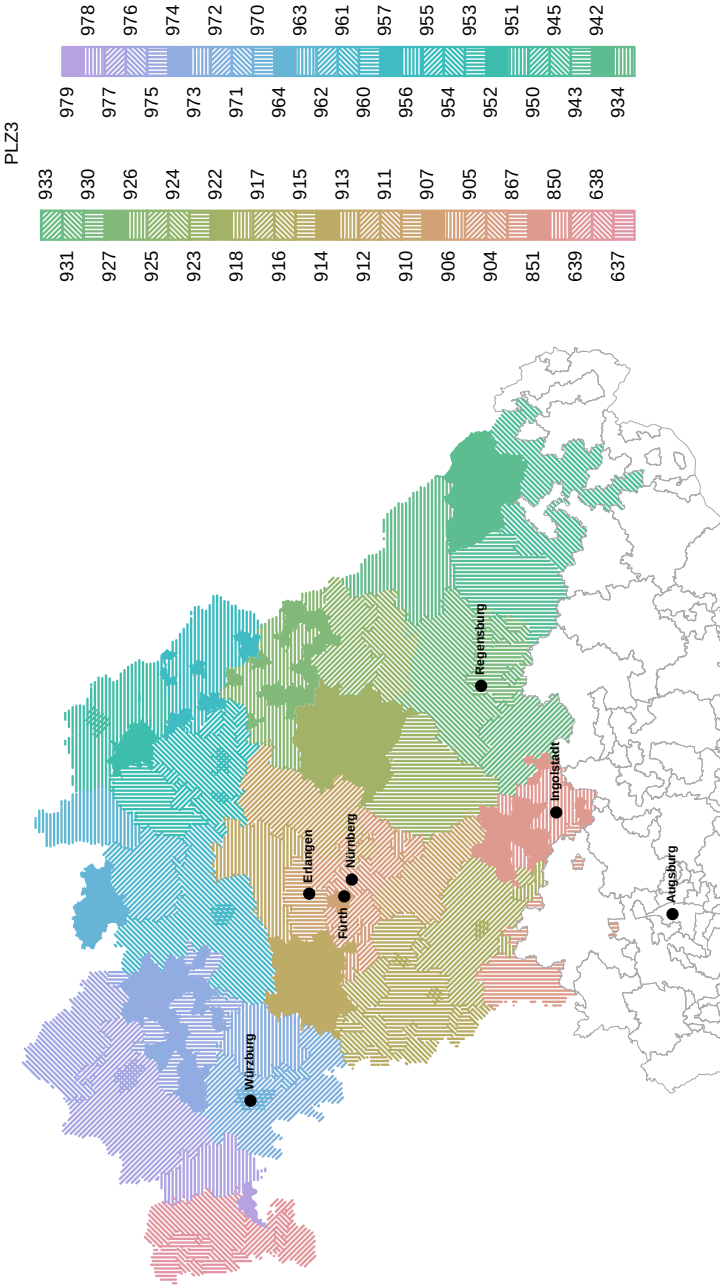

Abbildung D.3.: Karte der Bezirke der ersten drei Stellen der Postleitzahlen in Nordbayern.

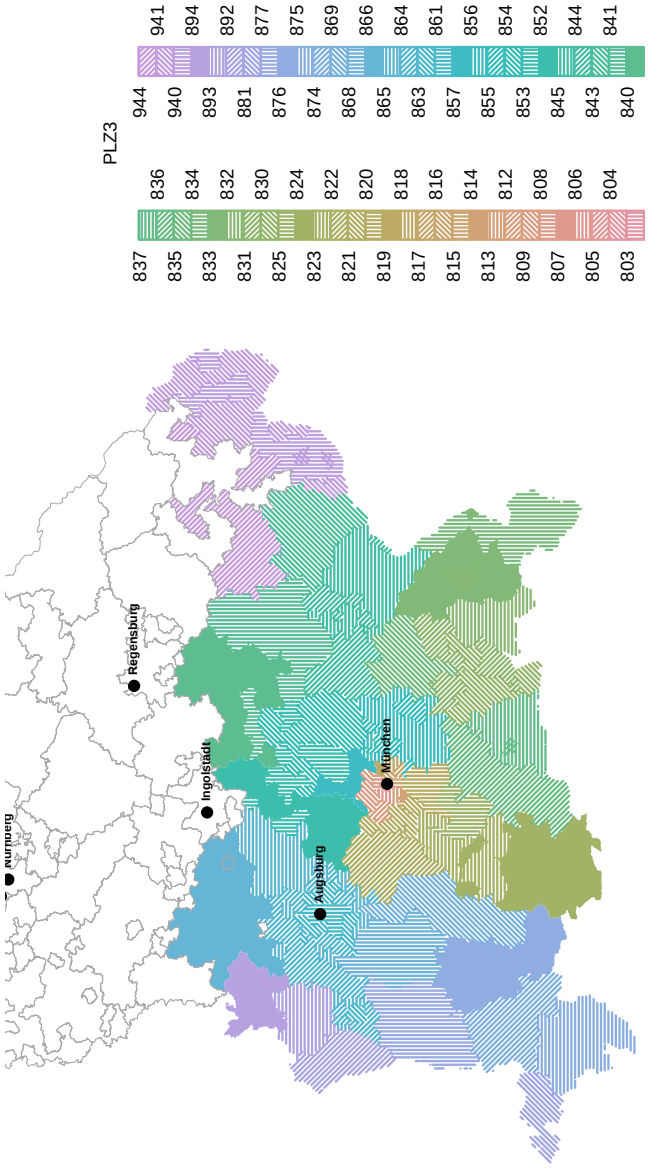

Abbildung D.4.: Karte der Bezirke der ersten drei Stellen der Postleitzahlen in Südbayern.

E. Benutzte Formeln

E.1. Brook's Lemma

Die Voraussetzung für die Anwendung dieses Lemmas ist die Positivität: Wenn T_s als Träger von r_s definiert ist

$$T_s = \{r_s \,|\, f_{R_s} > 0\}\,,\ s = 1, \ldots, S$$

und der Träger T des Vektors $\boldsymbol{r}$ das kartesische Produkt der einzelnen Träger ist

$$\{\boldsymbol{r} \,|\, f_{\boldsymbol{R}} > 0\} = T = T_1 \times \ldots \times T_S$$

dann ist für die gemeinsame Verteilung $f_{\boldsymbol{R}}(\boldsymbol{r})$ die Positivität erfüllt (Schmid, 2010).

Falls Positivität gilt, wird ein beliebiger, aber fester Punkt $\boldsymbol{r}^{(0)} = \left(r_1^{(0)}, \ldots, r_S^{(0)}\right)^T \in T$ definiert. Brook's Lemma besagt dann

$$f(\boldsymbol{r}) = \frac{f\left(r_1 \,|\, r_2, \ldots, r_S\right)}{f\left(r_1^{(0)} \,|\, r_2, \ldots, r_S\right)} \cdot \frac{f\left(r_2 \,|\, r_1^{(0)}, r_3, \ldots, r_S\right)}{f\left(r_2^{(0)} \,|\, r_1^{(0)}, r_3, \ldots, r_S\right)} \cdot \frac{f\left(r_3 \,|\, r_1^{(0)}, r_2^{(0)}, r_4, \ldots, r_S\right)}{f\left(r_3^{(0)} \,|\, r_1^{(0)}, r_2^{(0)}, r_4, \ldots, r_S\right)} \cdot \ldots \frac{f\left(r_S \,|\, r_1^{(0)}, \ldots, r_{S-1}^{(0)}\right)}{f\left(r_S^{(0)} \,|\, r_1^{(0)}, \ldots, r_{S-1}^{(0)}\right)} \cdot f\left(r_1^{(0)}, \ldots, r_S^{(0)}\right)$$

Die gemeinsame Verteilung ist bis auf die Normierungskonstante durch die vollständig bedingten Dichten gegeben (Schmid, 2010).

Verwendet man nun für $\boldsymbol{r}^{(0)} = \mathbb{E}(\boldsymbol{R}) = \boldsymbol{0}$, so ergibt sich in der logarithmierten Form

(Schmid, 2010)

$$\begin{aligned}
\log\left(\frac{f(\boldsymbol{r})}{f(\boldsymbol{0})}\right) &= \log\left(\prod_{s=1}^{S}\frac{f\left(r_s \,|\, 0_1,\ldots,0_{s-1},r_{s+1},\ldots,r_S\right)}{f\left(0_s \,|\, 0_1,\ldots,0_{s-1},r_{s+1},\ldots,r_S\right)}\right) \\
&\propto -\frac{1}{2}\sum_{s=1}^{S}\frac{|N(s)|}{\tau^2}\left(r_s - \frac{1}{|N(s)|}\sum_{s'=s+1}^{S} r_{s'}\right)^2 \\
&\quad + \frac{1}{2}\sum_{s=1}^{S}\frac{|N(s)|}{\tau^2}\left(0_s - \frac{1}{|N(s)|}\sum_{s'=s+1}^{S} r_{s'}\right)^2 && \text{mit (3.21)} \\
&= -\frac{1}{2}\sum_{s=1}^{S}\frac{|N(s)|}{\tau^2}\left[(r_s)^2 - 2r_s\frac{1}{|N(s)|}\sum_{s'=s+1}^{S} r_{s'} + \left(\frac{1}{|N(s)|}\sum_{s'=s+1}^{S} r_{s'}\right)^2\right] \\
&\quad + \frac{1}{2}\sum_{s=1}^{S}\frac{|N(s)|}{\tau^2}\left(\frac{1}{|N(s)|}\sum_{s'=s+1}^{S} r_{s'}\right)^2 \\
&= -\frac{1}{2}\sum_{s=1}^{S}\frac{|N(s)|}{\tau^2}(r_s)^2 + \sum_{s=1}^{S} r_s\frac{1}{\tau^2}\sum_{s'=s+1}^{S} r_{s'} \\
&= -\frac{1}{2}\boldsymbol{r}^T\frac{1}{\tau^2}\boldsymbol{K}\boldsymbol{r}
\end{aligned}$$

Also ist

$$f(\boldsymbol{r}) \propto f(\boldsymbol{0})\cdot\exp\left(-\frac{1}{2\tau^2}\boldsymbol{r}^T\boldsymbol{K}\boldsymbol{r}\right)$$

Nachdem $f(\boldsymbol{0})$ aber nur eine Normalisierungskonstante ist, kann man schreiben:

$$f_{\boldsymbol{R}}(\boldsymbol{r}) \propto \exp\left(-\frac{1}{2\tau^2}\boldsymbol{r}^T\boldsymbol{K}\boldsymbol{r}\right)$$

E.2. Umformungen der multivariaten Normalverteilung

Sei $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Dann gilt

$$\begin{aligned}
\int f_{\boldsymbol{X}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} &= 1 && \text{da Dichte} \\
\int \frac{1}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)\,\mathrm{d}\boldsymbol{x} &= 1 \\
\frac{1}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}}\int \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)\,\mathrm{d}\boldsymbol{x} &= 1 \\
\int \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)\,\mathrm{d}\boldsymbol{x} &= (2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}} \propto |\boldsymbol{\Sigma}|^{\frac{1}{2}} && \text{(E.1)}
\end{aligned}$$

$$
\begin{aligned}
f_{\boldsymbol{X}}(\boldsymbol{x}) &= \frac{1}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \\
&= \frac{1}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \\
&= \frac{1}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}\left(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - 2\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\right] \\
&\propto \exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)
\end{aligned} \tag{E.2}
$$

E.3. Umformungen bei der Herleitung des REML-Schätzers für den Varianzparameter des GMRF

$$
\begin{aligned}
(\boldsymbol{C}-\boldsymbol{X}\boldsymbol{\beta})^T V^{-1}(\boldsymbol{C}-\boldsymbol{X}\boldsymbol{\beta}) &= (\boldsymbol{C}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}+\boldsymbol{X}\widehat{\boldsymbol{\beta}}-\boldsymbol{X}\boldsymbol{\beta})^T V^{-1}(\boldsymbol{C}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}+\boldsymbol{X}\widehat{\boldsymbol{\beta}}-\boldsymbol{X}\boldsymbol{\beta}) \\
&= (\boldsymbol{C}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})^T V^{-1}(\boldsymbol{C}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}) + (\boldsymbol{X}\widehat{\boldsymbol{\beta}}-\boldsymbol{X}\boldsymbol{\beta})^T V^{-1}(\boldsymbol{X}\widehat{\boldsymbol{\beta}}-\boldsymbol{X}\boldsymbol{\beta}) \\
&\quad + 2\underbrace{(\boldsymbol{C}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})^T V^{-1}(\boldsymbol{X}\widehat{\boldsymbol{\beta}}-\boldsymbol{X}\boldsymbol{\beta})}_{=0,\text{ siehe (E.4)}} \\
&= (\boldsymbol{C}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})^T V^{-1}(\boldsymbol{C}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}) + (\boldsymbol{X}\widehat{\boldsymbol{\beta}}-\boldsymbol{X}\boldsymbol{\beta})^T V^{-1}(\boldsymbol{X}\widehat{\boldsymbol{\beta}}-\boldsymbol{X}\boldsymbol{\beta}) \\
&= (\boldsymbol{C}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})^T V^{-1}(\boldsymbol{C}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}) + (\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T\boldsymbol{X}^T V^{-1}\boldsymbol{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})
\end{aligned} \tag{E.3}
$$

$$
\begin{aligned}
(\boldsymbol{C}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})^T V^{-1}(\boldsymbol{X}\widehat{\boldsymbol{\beta}}-\boldsymbol{X}\boldsymbol{\beta}) &= \boldsymbol{C}^T V^{-1}\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T V^{-1}\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{C}^T V^{-1}\boldsymbol{X}\boldsymbol{\beta} + \widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T V^{-1}\boldsymbol{X}\boldsymbol{\beta} \\
&= \boldsymbol{C}^T V^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T V^{-1}\boldsymbol{X}\widehat{\boldsymbol{\beta}} \\
&\quad - \boldsymbol{C}^T V^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})\boldsymbol{\beta} + \widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T V^{-1}\boldsymbol{X}\boldsymbol{\beta} \\
&= \widehat{\boldsymbol{\beta}}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T V^{-1}\boldsymbol{X}\widehat{\boldsymbol{\beta}} \\
&\quad - \widehat{\boldsymbol{\beta}}^T(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})\boldsymbol{\beta} + \widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T V^{-1}\boldsymbol{X}\boldsymbol{\beta} \\
&= 0
\end{aligned} \tag{E.4}
$$

Hiermit versichere ich, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel verfasst habe.

Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

Gauting, den 18. Januar 2011

Anna Rieger