Analyse potenzieller Einflussgrößen für das Auftreten pneumologischer Erkrankungen und Entwicklung eines zeitlich-räumlichen Prognosemodells

Diplomarbeit
im Studiengang Statistik

Fakultät für Mathematik, Informatik und Statistik
Ludwig-Maximilians-Universität München

angefertigt von
Andreas Bayerstadler

betreuende Hochschullehrer
PD Dr. Christian Heumann
Prof. Dr. Ludwig Fahrmeir

München, den 15. September 2010

Selbstständigkeitserklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe.

.....
Andreas Bayerstadler

Danksagung

Vorab möchte ich folgenden Personen ganz herzlich für Ihre Hilfe und Unterstützung bei dieser Diplomarbeit danken:

- Herrn PD Dr. Christian Heumann für die Vermittlung und äußerst engagierte Betreuung dieser Diplomarbeit, insbesondere für die vielen langen Beratungsgespräche und zahlreichen guten Ideen
- Herrn Prof. Dr. Ludwig Fahrmeir für seine ständige Bereitschaft mir meine Fragen (vor allem zur bayesianischen Modellierung) zu beantworten und für seine richtungsweisenden Ratschläge
- Frau Dr. Eva Wanka und Frau Dr. Uta Ferrari für die ausführliche Beratung im medizinischen Bereich und das gewissenhafte Korrekturlesen
- Herrn Dr. Christoph Bergemann und Herrn Dr. Julian Mayer-Arnek für die Bereitstellung der Meteorologie- und Luftqualitätsdaten und den gesamten damit verbundenen Aufwand
- Herrn Dr. Werner Maier für die Bereitstellung des multiplen Deprivationsindex und die diesbezügliche Beratung
- meinen Eltern für ihre Geduld und den Verzicht auf sämtliche Computer, die mit diversen Modellen beschäftigt waren
- Lena für ihr Verständnis und ihre seelische Unterstützung

Inhaltsverzeichnis

1 Einleitung

Atemwegserkrankungen wie Asthma bronchiale und chronisch obstruktive Lungenerkrankung (COPD) haben in den letzten Jahrzehnten dramatisch zugenommen und zählen inzwischen zu den so genannten Volkskrankheiten. Asthma bronchiale ist eine chronisch entzündliche Erkrankung der Atemwege, die durch eine bronchiale Überempfindlichkeit (Hyperreaktivität) und eine variable Atemwegsobstruktion charakterisiert ist. Repräsentative Erhebungen zeigen in Deutschland eine Prävalenz von 9% bis 14% im Kindesalter und 4% bis 5% bei Erwachsenen. Damit ist Asthma eine der am häufigsten auftretenden chronischen Erkrankungen bei Kindern und Jugendlichen. Der Begriff der COPD (chronic obstructive pulmonary disease) erfasst die chronisch obstruktive Bronchitis, das Lungenemphysem und deren Kombinationen, schließt das Asthma hingegen aus. Nach der WHO Definition liegt eine chronische Bronchitis vor, wenn Husten und Auswurf über wenigstens 3 Monate in mindestens 2 aufeinanderfolgenen Jahren bestehen. Weltweit ist COPD gegenwärtig die vierthäufigste Todesursache. Für die nächsten Jahrzehnte ist ein weiterer Anstieg von Prävalenz, Morbidität und Mortalität zu erwarten, so dass die COPD bis zum Jahr 2020 unter den häufigsten Todesursachen auf den 3. Platz vorrücken wird. Aktuell wird die Prävalenz von COPD in Deutschland auf 10% bis 15% geschätzt. Obwohl bei der Diagnose eine Differenzierung zwischen den Krankheitsbildern Asthma und COPD anzustreben ist, ist dies in der Praxis, vor allem bei der ärztlichen Routineversorgung, nicht immer möglich (für voranstehende sowie ausführlichere Informationen vgl. Lingner et al. (2007, Kap. I)).

Als Risikofaktoren für beide Erkrankungen gilt neben verschiedenen genetischen Faktoren die Exposition gegenüber diversen Luftschadstoffen, wie z.B. Feinstaub, Schwefeldioxid, Stickstoffdioxid und Ozon (vgl. z. B. D'Amato et al. (2002), Atkinson et al. (2001) und Harré et al. (1997)). Das Risiko an COPD zu erkranken wird insbesondere durch Rauchen erhöht (vgl. Lingner et al. (2007)). Zudem ist das Auftreten von Atemwegserkrankungen starken saisonalen Schwankungen unterworfen, was einen Zusammenhang mit meteorologischen Parametern, wie z. B. Temperatur, Feuchtigkeit oder Luftdruck, nahelegt. Dieser Zusammenhang wurde in verschiedenen epidemiologischen Studien analysiert und belegt (vgl. z. B. Wen-Chao et al. (2007) und Michelozzi et al. (2009)). Teilweise wurden auch Interaktionseffekte zwischen Meteorologie- und Luftqualitätsparametern in die Analysen miteinbezogen (vgl. z. B. Ren und Tong (2006)).

Im Gegensatz zu vielen vorhergehenden, methodisch vergleichbaren Analysen, die sich mit der Mortalität in einer Subpopulation beschäftigen (vgl. z. B. Zanobetti et al. (2000) oder Welty et al. (2009)), befasst sich diese Arbeit ausschließlich mit der krankheitsspezifischen Morbidität, die in der Pilotregion Bayern (Deutschland) anhand von verschiedenen Ereignishäufigkeiten gemessen wird. Als potenzielle Einflussgrößen werden hier vor allem diverse Meteorologie- und Luftqualitätsparame-

ter betrachtet. Mithilfe adäquater statistischer Methoden sollen unmittelbar und zeitverzögert eintretende Effekte auf die Asthma- und COPD-Morbidität analysiert und charakterisiert werden. Eine besondere Herausforderung stellt dabei die zeitlich-räumliche Datenstruktur dar. Basierend auf den gewonnenen Erkenntnissen wird in der Folge der Versuch unternommen, das regionsspezifische Risiko des Auftretens akuter Symptome im Vorhinein abzuschätzen. Dazu werden Prognosemodelle konstruiert, die sich auf vergangene Beobachtungen und die Vorhersage von Wetter und Luftqualität stützen.

2 Vorstellung der Daten und Analyseziele

Diese Diplomarbeit ist Teil zweier Gemeinschaftsprojekte der Abteilung für Pneumologie der Medizinischen Klinik Innenstadt der LMU München, dem Deutschen Zentrum für Luft- und Raumfahrt (DLR) e. V. sowie dem Institut für Statistik der LMU München und baut auf der Bachelorarbeit von Teresa Exner (vgl. Exner (2009)) auf. Nach einer kurzen Vorstellung der Projekte (vgl. Abschnitt 2.1) werden in den Abschnitten 2.2 und 2.3 die von den Kooperationspartnern zur Verfügung gestellten Datensätze und ihre wesentlichen Eigenschaften dargestellt. Die für die weiteren Kapitel relevanten Daten werden in Abschnitt 2.4 im Hinblick auf die spätere Modellierung deskriptiv analysiert. Abschnitt 2.5 erläutert zunächst einige Grundlagen statistischer Regressionsmodelle und befasst sich schließlich mit den Herausforderungen, die sich bei der Modellierung aus der vorliegenden Datenlage ergeben, sowie der inhaltlichen und methodischen Zielsetzung.

2.1 Rahmenprojekte

Die Grundlage für die Kooperation der beteiligten Institutionen und die Enstehung dieser Arbeit bilden das GENESIS-Projekt und das Projekt „Gesundheitswetter Bayern“ (vgl. die Projektübersicht in Abbildung 2.1). Im Folgenden sollen die Hintergründe und Ziele dieser Projekte kurz erläutert werden (vgl. Exner (2009)).

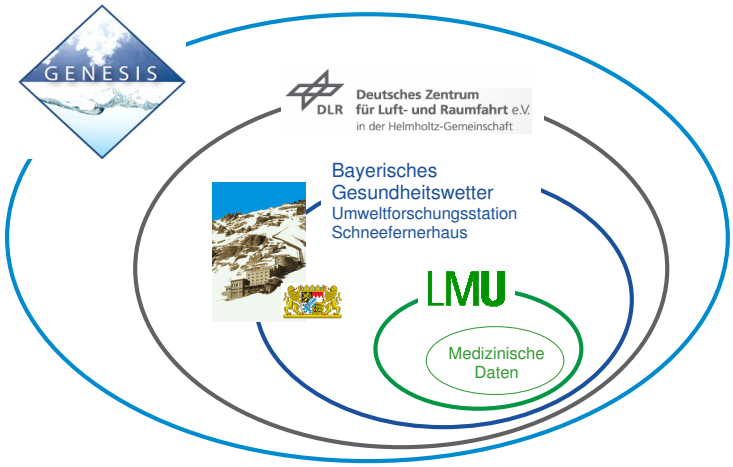


Abbildung 2.1: Überblick über die Gemeinschaftsprojekte der beteiligten Institutionen

GENESIS steht für „GENeric European Sustainable Information Space for Environment“ und ist ein im 7. EU-Rahmenprogramm gefördertes Projekt, an dem Wissen-

schaftler aus verschiedenen Nationen und unterschiedlichen Fachrichtungen beteiligt sind. Das Ziel von GENESIS ist die Entwicklung einer IT-Infrastruktur zur Zusammenführung und Darstellung unterschiedlicher Umweltdaten im Bereich Luft- und Wasserqualität und die Untersuchung deren Einfluss auf die Gesundheit. Für den Bereich Luftqualität dient neben Nizza (Frankreich) und London (Großbritannien) Bayern (Deutschland) als Pilotregion.

„Gesundheitswetter Bayern“ ist ein durch das Bayerische Staatsministerium für Wirtschaft, Infrastruktur, Verkehr und Technologie gefördertes Projekt und ist als Pilotstudie in GENESIS eingegliedert. Dabei ist der Schwerpunkt die Untersuchung des Einflusses geophysikalischer und klimatischer Faktoren auf verschiedene Gesundheitskenngrößen und die Entwicklung eines spezifischen Gesundheitsindex für pneumonale Erkrankungen. Dieser Index soll leicht interpretierbar sein und den Grad bestimmter Gesundheitsrisiken für bestimmte Patientengruppen möglichst gut abbilden. Eine regelmäßige Bestimmung und die Prognose des Gesundheitsindex könnten zur individuellen medizinischen Information spezieller Zielgruppen sowie der Bevölkerung im Ganzen genutzt werden.

Das „Gesundheitswetter“-Projekt konzentriert sich vor allem auf die Untersuchung der Morbidität von pulmonalen Erkrankungen, da der Schwerpunkt vieler früherer Studien auf der Mortalitätsrate lag. Dementsprechend sollen auch detaillierte Analysen zum individuellen Krankheitsverlauf (Veränderung der Lungenfunktion, bronchiale Obstruktion, subjektives Befinden etc.) erfolgen. Es ist bekannt, dass meteorologische Größen wie Temperatur und Luftdruck oder umweltspezifische Parameter wie bodennahe Ozonkonzentration, Feinstaubbelastung und Konzentrationen anderer Spurengase in der Atmosphäre Einfluss auf chronische Entzündungen der Atemwege haben. Die Ergebnisse bisheriger wissenschaftlicher Untersuchungen zu Charakter, Stärke und relativem Gewicht dieser Effekte sind jedoch in vielerlei Hinsicht unzureichend und teilweise widersprüchlich. Daher sollen medizinische Studien durchgeführt werden, um belastbare Aussagen zum Einfluss bestimmter äußerer Faktoren auf den Gesundheitszustand von Patienten treffen zu können.

Basierend auf den Modellergebnissen wird mitunter ein Online-Service angestrebt, der für verschiedene Regionen in Bayern, abhängig von der lokalen Vorhersage der ermittelten Einflussfaktoren, das Risiko für Lungenkranke prognostiziert und ggf. medizinische Verhaltensempfehlungen gibt. Dieser Service könnte durch die Verwendung von Satelliteninformationen über Meteorologie- und Luftqualitätsparameter praktisch europa- und weltweit ausgedehnt werden.

Zusätzlich soll im Rahmen einer kontrollierten Pilotstudie, die auf der Umweltforschungsstation Schneefernerhaus (UFS) durchgeführt wird, überprüft werden, ob sich die ermittelten Modellzusammenhänge an einem entsprechenden Patientenkollektiv validieren lassen. Dabei soll auf individueller Ebene untersucht werden, welche Konsequenzen die Veränderung bestimmter meteorologischer oder luftchemischer

Parameter auf die Lungenfunktion, die Herzfrequenz und das Befinden von Patienten mit COPD haben.

2.2 Medizinische Datensätze

Insgesamt stehen 4 verschiedene medizinische Datenquellen zur retrospektiven Erfassung der krankheitsspezifischen Morbidität auf Tagesbasis im Zeitraum 2006 bis 2008 zur Verfügung. Im Folgenden werden diese Datenquellen kurz vorgestellt und ihre Vor- und Nachteile im Bezug auf die Ziele der statistischen Modellierung diskutiert.

1. Call-Center-Daten für den Zeitraum 2006 bis 2008 der Kassenärztlichen Vereinigung Bayerns (KVB), welche den ärztlichen Bereitschaftsdienst organisiert: Die Vermittlungs- und Beratungszentralen der KVB sind für Patienten 24 Stunden am Tag und 365 Tage im Jahr erreichbar. Diese Daten beinhalten anrufbasierte Informationen (Datum, Uhrzeit, x-y-Koordinaten, PLZ, Landkreis, Alter, Meldebild, Versorgungs- bzw. Einsatzart und weitere Zusatzinformationen). Hier kam es in ganz Bayern im genannten Zeitraum zu insgesamt ca. 108000 Anrufen aufgrund der Diagnose „Atembeschwerden“. Es ist davon auszugehen, dass diese Größe die Asthma- und COPD-Morbidität in Bayern gut abbildet, eine Differenzierung zwischen den Krankheiten ist allerdings nicht möglich.

   Durch die ständige Erreichbarkeit der telefonischen Beratungsstellen ist eine optimale zeitliche Abdeckung garantiert. Wochenend-, Ferien- und Feiertagseffekte sind dennoch zu erwarten, da der Telefon-Service der KVB an Tagen, an denen die Arztpraxen geöffnet sind, deutlich seltener genutzt wird. Eine bestmögliche Erstdiagnose ist durch den Einsatz medizinischen Personals in der telefonischen Beratung sichergestellt. Dennoch bleibt eine gewisse Unsicherheit vorhanden, da die Zuordnung letztlich auf den Informationen der Anrufer beruht, die sich häufig in Stresssituationen befinden. Zudem sind die Altersangaben häufig ungenau oder beruhen sogar auf Schätzungen des Call-Center-Personals. Um eine stärkere Verzerrung zu vermeiden, wurde aus diesem Grund eine Diskretisierung in 5 Altersgruppen (0–20 Jahre, 21–40 Jahre, 41–60 Jahre, 61–80 Jahre, 81 Jahre und älter) vorgenommen. Zudem fehlen für den Zeitraum 19.10.2006 bis 15.12.2006 sämtliche Altersangaben im Datensatz. Um die beobachteten Fallzahlen nicht aus der Analyse ausschließen zu müssen, wurden die fehlenden Altersangaben nach folgendem Verfahren imputiert:

   a) Berechne die relative Häufigkeitsverteilung $\boldsymbol{p}_i = (p_{i1}, p_{i2}, p_{i3}, p_{i4}, p_{i5}, p_{\text{i,k.A.}})$ der Anrufe in den 5 Altersgruppen zwischen 1.1.2006 und 31.12.2008 separat für jeden Landkreis i ($i = 1, \ldots, 96$) mit einer eigenen Kategorie für fehlende Altersangaben (k.A.). Durch diese zusätzliche Kategorie wird

der „natürliche“ Anteil an fehlenden Werten bei der Imputation berücksichtigt.

b) Bestimme die tageweisen (Index t) Gesamt-Fallzahlen y_{it} für jeden Landkreis i im obigen Zeitraum und ziehe mit Hilfe der R-Funktion „sample“ die Altersangaben $\boldsymbol{a}_{it} = (a_{it,1}, \ldots, a_{it,y_{it}})$ mit $a \in \{1, 2, 3, 4, 5, \text{k.A.}\}$ basierend auf der Häufigkeitsverteilung $\boldsymbol{p}_i$.

c) Bestimme die gesuchten altersspezifischen Fallzahlen $(y_{i1t}, y_{i2t}, y_{i3t}, y_{i4t}, y_{i5t})$ für jeden Landkreis i $(i = 1, \ldots, 96)$ und jeden Tag t vom 19.10.2006 bis 15.12.2006 durch Abzählen der einzelnen Kategorien in $\boldsymbol{a}_{it}$.

Das Geschlecht der Patienten oder weitere personenbezogene Angaben wurden nicht erfasst. Ebenso problematisch sind die generell kleinen Fallzahlen. Dadurch enstehen nach Aufteilung der Anrufe in Landkreise und Altersgruppen viele nicht besetzte Zellen in der Häufigkeitstabelle, was eine zusätzliche Herausforderung bei der Modellierung bedeutet (vgl. Abschnitt 3.2). Die räumliche Auflösung ist bei den Call-Center-Daten im Vergleich zu den anderen Datenquellen sehr gut, es können sogar jedem Anruf genaue x-y-Koordinaten sowie die Postleitzahl zugeordnet werden. Für die Analysen in dieser Arbeit wurden dennoch nur die Landkreisangaben herangezogen, da die Meteorologie- und Luftqualitätsdaten nicht in kleinräumigerer Auflösung verfügbar waren (siehe Abschnitt 2.3). Ein wesentlicher Vorteil dieser Datenquelle besteht darin, dass bei den meisten Anrufen von einem akuten Zustand des Patienten auszugehen ist, so dass ein unmittelbarer Zusammenhang zur vorhergehenden und aktuellen Meteorologie- und Luftschadstoffsituation ableitbar ist. Des Weiteren stehen die Call-Center-Daten im Vergleich zu den anderen Datenquellen über den gesamten Zeitraum vom 1.1.2006 bis zum 31.12.2008 für ganz Bayern zur Verfügung. Aus diesen Gründen basiert ein erheblicher Teil der statistischen Analysen in dieser Arbeit auf dieser Datenquelle.

2. Abrechnungsdaten für den Zeitraum 2006 bis 2007 der KVB: Diese Daten beinhalten auf Landkreisebene die Tagessummen der Arztkonsultationen von Patienten wegen Asthma bronchiale und/oder COPD bei Allgemein-, Haus- und Fachärzten, getrennt nach Krankheitsbild, Geschlecht und Altersgruppen. Hier liegen in ganz Bayern für den genannten Zeitraum insgesamt ca. 371.5 Mio. Kontakte unabhängig von Arzt und Diagnose vor, davon ca. 23.6 Mio. Kontakte für die Diagnosen Asthma und/oder COPD.

   Der Abrechnungsdatensatz der KVB wurde bereits in der Bachelorarbeit von Teresa Exner (vgl. Exner (2009)) ausführlich analysiert. Ein großer Vorteil liegt hier in der relativ genauen Abgrenzung der Diagnosen Asthma bronchiale und COPD durch Verwendung der ICD-Codierung (International Classification of Diseases, 10. Version), wobei natürlich auch hier Fehldiagnosen nicht ausgeschlossen werden können. Zusätzlich zum Alter (in den 4 Kategorien 0–

20 Jahre, 21–40 Jahre, 41–60 Jahre und 61 Jahre und älter) steht hier das Geschlecht als weitere personenbezogene Kovariable zur Verfügung. Die Daten umfassen die abgerechneten Arztbesuche bei ca. 20500 Allgemein-, Haus- und Fachärzten in ganz Bayern. Somit ist eine gute Abdeckung der Pilotregion gewährleistet, wobei die Arztdichte zwischen städtischen und ländlichen Gebieten naturgemäß stark variiert. Im Vergleich zu den KVB-Call-Center-Daten kommen so trotz der Aufgliederung nach Alter und Geschlecht durchschnittlich größere Fallzahlen zustande. Durch die damit verbundene größere Variation der Fallzahlem lassen sich vorhandene Kovariableneffekte tendenziell besser erkennen. Die räumliche Auflösung (Landkreisebene) ist aufgrund der Kovariablensituation ausreichend. Administrative Kovariablen, wie Feiertage und Wochenenden, stellen bei dieser Datenquelle durch die eingeschränkten Öffnungszeiten der Arztpraxen einen sehr wichtigen Faktor bei der Modellierung dar. Leicht verzerrt werden die Fallzahlen möglicherweise durch die Tatsache, dass alle Arztkontakte innerhalb eines Quartals gezählt werden, sobald einmal die Diagnose Asthma bzw. COPD gestellt wurde, auch wenn die Arztbesuche andere Ursachen haben. Grundsätzlich sind die Abrechnungsdaten sehr gut geeignet, um die interessierenden Fragen mithilfe statistischer Modellierung zu beantworten, da die Arztbesuche in der Regel auf akute gesundheitliche Probleme zurückzuführen sind, die wiederum potenziell von der aktuellen Wetter- und Luftqualitätssituation ausgelöst werden können. Um also die aufgrund der geringen Anruferzahlen möglicherweise instabilen KVB-Call-Center-Modelle zu validieren, wurden die Analysen parallel auch mit den Abrechnungsdaten durchgeführt und die Ergebnisse zum Teil verglichen. Dabei gilt es jedoch auch den unterschiedlichen Untersuchungszeitraum zu beachten.

3. Arzneimitteldaten für den Zeitraum 2007 bis 2008 der Firma pharmafakt (GFD Gesellschaft für Datenverarbeitung mbH): Die GFD-Datenbasis umfasst zu 55% Ist-Daten des deutschen GKV-Verordnungsvolumens. Basis sind die Daten verschiedener Apothekenabrechnungszentren, wobei hier die Datenbasis „Bayern“ mit 90% abgedeckt ist. Es stehen sowohl die Daten des COPD- und Asthma-Markts (ca. 7.8 Mio. Rezepte) als auch des Antibiotika-Markts (ca. 3.4 Mio. Rezepte) zur Verfügung. Pro Tag werden für jeden Landkreis die Umsatzsumme, die Anzahl an eingelösten Rezepten und die Anzahl an Patienten, die an diesem Tag ein Rezept eingelöst haben, jeweils getrennt nach 5 Altersgruppen (analog zu den KVB-Call-Center-Daten) angegeben.

   Durch die Kenntnis der verschriebenen Medikamentengruppen ist hier eine gute Abgrenzung der Asthma- und COPD-Fälle von anderen Erkrankungen, die einer medikamentösen Therapie bedürfen, möglich. Bis auf die Altersgruppe stehen jedoch keine weiteren personenbezogenen Daten, wie z. B. das Geschlecht, zur genaueren Aufschlüsselung der täglichen Fallzahlen zur Verfügung. Des Weiteren sind die Daten dahingehend verzerrt, dass allen Rezepten, die

datumsmäßig nicht exakt zugeordnet werden können (z. B. aufgrund unleserlicher Handschrift), der 5. oder 25. des Monats als Abrechnungstag zugewiesen wird. Aus diesem Grund mussten die Fallzahlen an diesen Tagen des Monats durch ein Imputationsverfahren korrigiert werden. Fiel der 5. oder 25. des Monats beispielsweise auf einen Mittwoch, so wurde die korrigierte Fallzahl berechnet als Mittelwert aus den Fallzahlen des vorhergehenden und darauffolgenden Mittwochs. Ein weiterer Nachteil besteht in der mangelnden Abdeckung von Wochenend- und Feiertagen, an denen die Apotheken nicht oder nur eingeschränkt geöffnet sind und die Fallzahlen entsprechend gering ausfallen. Daneben gibt es Schulferien- und Quartalseffekte, die bei der Modellierung in jedem Fall berücksichtigt werden müssen. Die Hauptproblematik dieser Datenquelle besteht jedoch darin, dass die Akutheit des Zustandes eines Patienten am Tag der Rezepteinlösung, gerade bei chronisch kranken Personen, stark in Frage zu stellen ist. Da ein nicht zu vernachlässigender Teil der verschriebenen Asthma- und COPD-Medikamente die Dauermedikation darstellt, ist es sehr fragwürdig auf dieser Basis einen Zusammenhang zwischen dem akuten Auftreten von Krankheitssymptomen und vorangehenden Veränderungen äußerer Einflussfaktoren herzustellen. Aus diesem Grund wurde der Datensatz nicht für die statistischen Analysen herangezogen.

4. Einsatzdaten der Rettungsleitstellen für den Zeitraum 2006 bis 2008 des Bayerischen Roten Kreuz (BRK) und der integrierten Leitstelle München (betrieben durch die Berufsfeuerwehr München): In Bayern sind die Leitstellen rund um die Uhr unter der europaweiten Notrufnummer 112 zu erreichen, in einigen (vorwiegend ländlichen) Regionen unter 19222. Diese Daten beinhalten anrufbasierte Informationen jedes einzelnen Rettungsdiensteinsatzes (Datum, Uhrzeit, PLZ bzw. Straßenname, Meldebild und Versorgungs- bzw. Einsatzart). Im genannten Zeitraum wurden allein für die Stadt München ca. 31000 Anrufe registriert, die atmungsrelevante Meldebilder umfassen.

   Wie die Call-Center-Daten werden auch die Rettungsdienstdaten 24 Stunden am Tag und 365 Tage im Jahr erfasst. Die schnelle Verfügbarkeit von Einsatzpersonal ist durch ein dichtes Netz von Rettungsleitstellen gewährleistet. Auch hier werden die Anrufe von medizinischem Fachpersonal entgegengenommen, so dass im Allgemeinen eine korrekte Zuordnung sichergestellt ist. Diese hängt jedoch wiederum von der Qualität der Anruferinformationen ab, weswegen eine konkrete Abgrenzung der spezifischen Diagnose „Asthma/COPD“ häufig nicht möglich ist. Die geografischen Angaben sind bei den Rettungsdienstdaten sehr genau (Straßenname für die integrierte Rettungsleitstelle München, Postleitzahlen für die übrigen bayerischen Rettungsleitstellen), bei einer bayernweiten Modellierung muss man jedoch durch die gröbere Auflösung der Meteorologie- und Luftqualitätskovariablen auch hier auf Landkreisebene zurückgehen. Personenbezogene Kovariablen wie Alter und Geschlecht liegen hier aus Daten-

schutzgründen nicht vor. Bei Verwendung der Rettungsdienstdaten kann sicherlich größtenteils von einem akuten Zustand der Patienten ausgegangen werden, so dass ein direkter Zusammenhang zur vorherrschenden Kovariablensituation hergestellt werden kann. Jedoch muss auch berücksichtigt werden, dass man mit Rettungsdiensteinsätzen hauptsächlich schwere Fälle, die gegebenfalls eine ärztliche Intervention erfordern, abdeckt. Leichte Atembeschwerden oder Befindlichkeitsstörungen, die bei chronisch kranken Patienten häufig auftreten, werden dahingegen nicht erfasst. Insofern wäre hier zu prüfen, inwieweit die Anzahl der akuten Notfälle mit dem Auftreten solcher leichten Symptome korreliert ist. Wochenend-, Feiertags- und Ferieneffekte sind auch in diesem Zusammenhang zu erwarten und für die Modellierung relevant. Für die vorliegende Arbeit konnten die Rettungsdienstdaten nicht verwendet werden, da die Auswertungen für den Untersuchungszeitraum nicht flächendeckend für ganz Bayern vorlagen und somit keine räumliche Modellbildung möglich war. Die Daten der integrierten Rettungsleitstelle München werden in der Masterarbeit von Eva Wanka (vgl. Wanka (2010)) im Hinblick auf nichtlineare Kovariableneffekte (mittels Generalisierter Additiver Modelle) untersucht.

Die Verwendung des Bundeslandes Bayern als Pilotregion begründet sich neben der Verfügbarkeit der medizinischen Datenquellen auch darin, dass Bayern die deutsche und vermutlich auch europäische Bevölkerung gut repräsentiert (vgl. Wildner et al. (2005)). Diese Aussage stützt sich auf eine gute Mischung von Land- und Stadtbevölkerung sowie eine große Einkommens- und Bildungsspanne.

Um bei der bayernweiten Modellierung die Vergleichbarkeit der Landkreise zu gewährleisten, wurde die Einwohnerzahl der einzelnen Landkreise in der jeweiligen Altersgruppe (als Modell-Offset) in die Analysen miteinbezogen. Die dafür erforderlichen Daten stammen von der frei zugänglichen Online-Datenbank GENESIS des Bayerischen Landesamtes für Statistik und Datenverarbeitung (vgl. Bayer. Landesamt für Statistik und Datenverarbeitung). Ähnlich wurde bei der Analyse der KVB-Abrechnungsdaten mit der Anzahl der Ärzte pro Landkreis verfahren (Näheres dazu in Abschnitt 3.2.1).

Des Weiteren wurde in der vorliegenden Arbeit ein sogenannter Index Multipler Deprivation (IMD) auf Basis der Landkreise und kreisfreien Städte für Bayern eingesetzt (vgl. Maier et al. (2010)), welcher in Anlehnung an die im Vereinigten Königreich (UK) verwendeten Indizes Multipler Deprivation gebildet wurde (vgl. Noble et al. (2006)). Die Verwendung von Deprivationsindizes in epidemiologischen Studien ist dort seit Jahren Standard (vgl. Wild et al. (2008), Shack et al. (2007) und Connolly et al. (2000)). Deprivation beschreibt den Mangel an Ressourcen und die damit verbundene Benachteiligung, welche Individuen oder Personengruppen im Vergleich zur Gesamtpopulation erleiden (vgl. Townsend (1979)) und zeigt einen signifikanten, räumlich variierenden Zusammenhang mit dem Gesundheitszustand der Bevölkerung (vgl. Bayer. Landesamt für Gesundheit und Lebensmittelsicherheit

(2007) und Townsend et al. (1988)). Potenzielle räumliche Einflüsse auf die Gesundheit sind in den letzten Jahren wieder Gegenstand der aktuellen Forschung geworden (Macintyre et al. (2002)).

Der Status der materiellen und sozialen Deprivation in den bayerischen Kreisen und kreisfreien Städten wird durch den hier angewandten IMD (vgl. Maier et al. (2010)) abgebildet. Durch seine Miteinbeziehung als Kovariable sollen regionale Einflüsse auf die Asthma- und COPD-Inzidenz geprüft werden.

Der eingesetzte Index (vgl. Maier et al. (2010)) enthält Indikatoren, die aus Variablen der amtlichen Statistik für das Land Bayern auf Kreisbasis (Stand 2006) gebildet wurden und die zu thematischen Gruppen, den sogenannten Domänen, zusammengefasst wurden. Der Index enthält sieben Domänen: Einkommensdeprivation, Beschäftigungsdeprivation, Bildungsdeprivation, kommunale Einkommensdeprivation, Sozialkapitaldeprivation, Umweltdeprivation und Sicherheitsdeprivation. Aus den Scores der einzelnen Domänen wurde ein gewichteter Gesamtscore errechnet, der für die statistischen Analysen herangezogen wurde. Dieser ist normiert auf Werte von 1 bis 100, wobei große Werte für ein hohes Maß an Deprivation stehen. Die Score-Werte für die bayerischen Landkreise liegen zwischen 66.80 (Landkreis München) und 94.98 (Landkreis Wunsiedel im Fichtelgebirge).

2.3 Meteorologie- und Luftqualitätsdaten

Die verwendeten meteorologischen Daten stammen vom European Centre for Medium Range Weather Forecast (ECMWF, vgl. European Centre for Medium Range Weather Forecast) in Reading, UK. Es handelt sich dabei um Daten aus einem operationellem Analyselauf, die auf einem Gitter mit Auflösung $0.25° \times 0.25°$ erfasst werden. Um einen Wert pro Landkreis zu erhalten, wird ein flächenmäßig gewichtetes Mittel aller Werte im Landkreis gebildet. Die Messungen erfolgen um 0, 6, 12 und 18 Uhr UTC (Universal Time Coordinated, koordinierte Weltzeit). Um diese 4 Werte pro Parameter und Tag auf einen Wert zu reduzieren, wurde nach inhaltlichen Gesichtspunkten entweder der Mittelwert oder das Maximum der 4 Werte gebildet. Bei Feuchtigkeit, Luftdruck und Temperatur wurde zusätzlich eine Tagesrange (Maximum minus Minimum) berechnet, da man vermutet, dass auch starke Schwankungen dieser Größen einen Einfluss auf Lungenerkrankungen besitzen. Tabelle 2.1 gibt einen Überblick über die vorhandenen meteorologischen Parameter, die zugehörigen Variablennamen und die Einheit, in welcher der jeweilige Parameter gemessen wurde.

Die Niederschlagswerte (konvektiv und großskalig) sowie der Surface Stress wurden jeweils über den Zeitraum von 6 Stunden zwischen 2 Messungen akkumuliert. Die Windgeschwindigkeit und die Windrichtung (zunächst in Grad) wurden aus den Mittelwerten der gemessenen Ost-West- und Nord-Süd-Vektor-Komponenten

Parameter	Variablenname	Einheit
Bedeckungsgrad niedriger Bewölkung (Low Cloud Cover)	lcc	%
Bedeckungsgrad mittelhoher Bewölkung (Medium Cloud Cover)	mcc	%
Spezifische Luftfeuchtigkeit	q	$kg_{\text{Wasser}}/kg_{\text{Luft}}$
Logarithmierter Oberflächendruck	qnh	–
Bodendruck auf NN (nur für München)	qnh	hPa
Oberflächenwärmeleitung (Surface Sensible Heat Flux)	sshf	$W/(m^2 s)$
Zwei-Meter-Temperatur (Two Metre Temperature)	tmt	$°C$
Konvektiver Niederschlag (Convective Precipitation)	cp	m
Großskaliger Niederschlag (Long Scale Precipitation)	lsp	m
Oberflächenwind (Surface Stress)	stressspd	$N/(m^2 s)$
Windgeschwindigkeit in ca. 10 Metern Höhe (Wind Speed)	windspd	m/s
Windrichtung in 8 Kategorien	c.wdir	

Tabelle 2.1: Überblick über die Meteorologie-Parameter

errechnet. Aufgrund der Periodizität der Gradskala und der damit verbundenen Interpretationsprobleme bei der statistischen Modellierung, wurde die Windrichtung durch Zuordnung in 45°-Sektoren diskretisiert. Resultat ist eine nominale Variable mit den Ausprägungen Nord, Nordost, Ost, Südost, Süd, Südwest, West, Nordwest. Abbildung G.2 zeigt die Häufigkeiten der täglich vorherrschenden Windrichtungen in Bayern im Zeitraum 2006 bis 2008. Für die Darstellung wurden die tagesweisen Modi (Modus $\hat{=}$ häufigste Ausprägung) über alle 96 bayerischen Landkreise verwendet. Die häufigsten Windrichtungen sind Südwest und West, was das Vorliegen eines Westwind-Regimes bestätigt. Eine Besonderheit liegt bei der Variable Luftdruck (qnh) vor: Für alle bayernweiten Modelle wurde der logarithmierte Oberflächendruck verwendet. Bei den Modellen, die sich aufgrund computationaler Einschränkungen (vgl. dazu Abschnitt 3.1) nur auf die Pilotregion „Landeshauptstadt München“ beziehen, wurde der Bodendruck auf normal Null (NN) verwendet. Dieser lässt sich in hPa angeben, so dass die entsprechenden Modellparameter intuitiver interpretiert werden können, jedoch lag der Bodendruck auf NN nicht für die übrigen bayerischen Landkreise vor. Abbildung G.1 stellt den Zeitverlauf aller meteorologischen Größen im Untersuchungszeitraum 2006 bis 2008 dar. Dargestellt sind Tagesmittelwerte über alle bayerischen Landkreise (abgesehen von den letzten beiden Plots, die sich nur auf die Luftdruck-Werte in München beziehen). Anhand der rot eingezeichneten lowess-Glättung lassen sich der Jahreszeitenverlauf und die teilweise erheblichen saisonalen Schwankungen der jeweiligen Parameter gut erkennen, wobei Temperatur, Luftdruck und der Bewölkungsgrad niedriger Bewölkung die deutlichste Variation aufweisen. Der parallele Zeitverlauf der unterschiedlichen Luftdruckparameter für München und Bayern spricht dafür, dass beide Variablen hoch korreliert und bei der Modellierung auf München- bzw. Bayern-Ebene austauschbar sind.

Die Luftqualitätsdaten für Bayern wurden aus der European Air Quality Data Base der European Environmental Agency (EEA, vgl. European Environmental Agency) extrahiert. Dabei handelt es sich um Daten von Messstationen, die in die Kategorien „Background“, „Industrial“ und „Traffic“ eingeteilt sind. Tabelle 2.2 gibt einen Überblick über die verwendeten Luftschadstoffe.

Parameter	Variablenname	Einheit
Schwefeldioxid	SO2	$\mu g/m^3$
Feinstaub bis $10\mu m$ (Particulate Matter)	PM10	$\mu g/m^3$
Ozon	O3	$\mu g/m^3$
Stickstoffdioxid	NO2	$\mu g/m^3$
Kohlenstoffmonoxid	CO	$\mu g/m^3$

Tabelle 2.2: Überblick über die Luftqualitäts-Parameter

Abbildung 2.2 zeigt eine Karte, auf der alle bayerischen Messstationen, die zumindest einen der relevanten Luftqualitätsparameter messen, farblich getrennt nach Kategorie dargestellt sind. Zusätzlich sind die Zentroide aller Landkreise eingezeichnet.

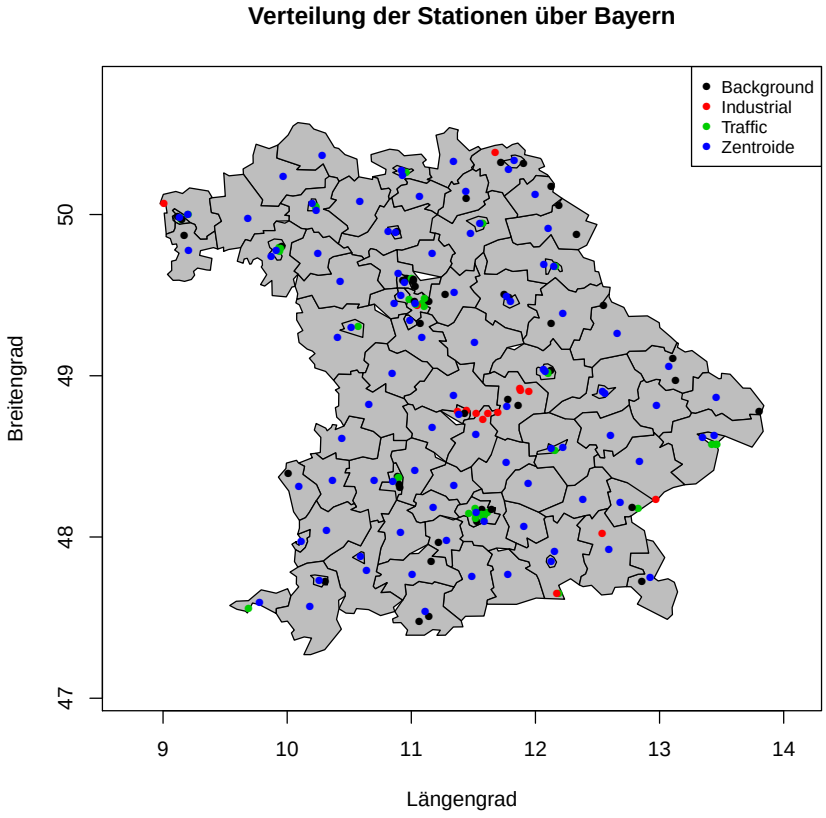


Abbildung 2.2: Verteilung der Messstationen über Bayern getrennt nach Kategorie und Landkreiszentroide

Für jede Station wurden zunächst die stundenweisen Mittelwerte durch Bildung des 95%-Quantils auf einen Tageswert reduziert. Auf diese Weise soll die tägliche Maximalbelastung unter Vernachlässigung einzelner Ausreißerwerte erfasst werden. Die Industrial-Stationen wurden aufgrund ihrer geringen Anzahl und ungleichmäßigen Verteilung über das Untersuchungsgebiet von vorneherein aus den Analysen aus-

geschlossen. Die Background-Stationen sind zwar zahlenmäßig stark vertreten und verhältnismäßig gut über Bayern verteilt, jedoch wird an den meisten Background-Stationen nur ein kleiner Teil der relevanten Schadstoffkonzentrationen gemessen. Aus diesem Grund beruhen die weiteren Analysen ausschließlich auf Werten der Traffic-Stationen, bei denen diese Problematik nicht auftritt. Dementsprechend konnte auch der ursprünglich geplante Vergleich zwischen auf Traffic- und Background-Daten basierenden Modellen nicht durchgeführt werden. Bei der gewählten Vorgehensweise ist allerdings zu beachten, dass die relevante durchschnittliche Exposition eines Individuums in der Zielpopulation wohl stark unter den an vielbefahrenen Straßen gemessenen Traffic-Werten liegt. Es wird jedoch angenommen, dass die Verwendung dieses Worst-Case-Szenarios bei der Modellierung eine untergeordnete Rolle spielt, da man von einer hohen Korrelation der tatsächlichen persönlichen Belastung mit den Traffic-Messwerten ausgehen kann.

Da nicht in allen bayerischen Landkreisen Messstationen zu finden sind und die räumliche Auflösung nicht noch weiter vergröbert werden soll, ist es entscheidend, einen möglichst realistischen landkreisspezifischen Messwert aus den vorhandenen Stationsdaten zu berechnen. Aufgrund des höchstdynamischen Verhaltens von Luftschadstoffen ist es nicht sinnvoll, weit entfernte Messstationen in die Bildung eines landkreisspezifischen Schadstoffwerts einzubeziehen. Nur die nächstgelegene Station zu verwenden wäre allerdings auch problematisch, da zum Teil fehlende Werte in den Messreihen auftreten und auch nicht jede Station alle relevanten Parameter erfasst. Um dieser Problematik beizukommen, wurde ein Gewichtungsverfahren verwendet, das auf der euklidischen Distanz der Stationen zum jeweiligen Landkreis-Zentroid beruht: Zunächst werden für jeden Landkreis und jeden der 5 Luftqualitätsparameter die 3 Stationen bestimmt, welche die geringste euklidische Distanz zum Landkreis-Zentroid besitzen und den betrachteten Parameter auch wirklich messen. Aus den 3 Distanzen d_1, d_2, d_3 werden dann mithilfe folgender Formel die normierten Gewichte w_1, w_2, w_3 für die einzelnen Stationen ermittelt (inverse Distanz-Gewichtung):

$$w_i = \frac{(1/d_i)^{2.5}}{\sum_{i=1}^{3}(1/d_i)^{2.5}}, \quad i = 1, 2, 3.$$

Der landkreisspezifische Schadstoffwert wird dann als gewichteter Mittelwert basierend auf den w_i berechnet. Auf diese Weise erhält die nächstgelegene Station das höchste Gewicht (vgl. nachfolgendes Rechenbeispiel und Abbildung 2.3). In der Literatur wird für die inverse Distanz-Gewichtung häufig der Exponent 2 verwendet (vgl. z. B. Hoek et al. (2002)), aufgrund des dynamischen Verhaltens der betrachteten Luftschadstoffe erschien es jedoch sinnvoll, weiter entfernten Stationen durch Erhöhung des Exponenten auf 2.5 weniger Einfluss beizumessen.

Rechenbeispiel zur inversen Distanz-Gewichtung:

Abbildung 2.3 zeigt exemplarisch die drei nächstgelegenen Stationen 1, 2 und 3 zum Zentroid des Landkreises Dingolfing. Die zugehörigen euklidischen Distanzen betragen $d_1 = 0.4582$, $d_2 = 0.5068$ und $d_3 = 0.8183$. Mit obiger Formel ergeben sich daraus die Gewichte $w_1 = 0.4970$, $w_2 = 0.3863$ und $w_3 = 0.1166$. Möchte man nun beispielsweise den landkreisspezifischen Tageswert für den Luftschadstoff Kohlenstoffmonoxid berechnen, dessen Konzentration an allen 3 Stationen gemessen wird, bildet man ein gewichtetes Mittel aus den Tageswerten der Stationen 1, 2 und 3 unter Verwendung von w_1, w_2 und w_3.

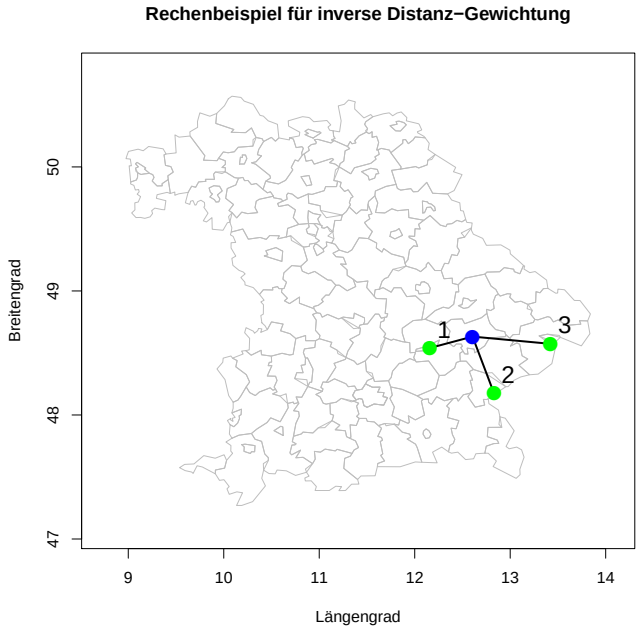


Abbildung 2.3: Rechenbeispiel zur inversen Distanz-Gewichtung (Landkreis Dingolfing)

Der Zeitverlauf der Luftqualitätsparameter kann Abbildung G.3 entnommen werden. Die Plots zeigen wiederum tagesweise Mittelwerte über alle bayerischen Landkreise zusammen mit einem lowess-Trend (rot eingezeichnet). Ähnlich wie bei den meteorologischen Parametern sind auch hier unterschiedlich ausgeprägte saisonale Schwankungen zu erkennen. Beispielsweise ist die Ozonkonzentration im Sommer deutlich stärker ausgeprägt als im Winter, während die Feinstaubkonzentration eher wenig über die Jahreszeiten variiert.

Tabelle 2.3 gibt schließlich einen zusammenfassenden Überblick über die im Zeitraum 2006 bis 2008 beobachteten Werte der verwendeten Meteorologie- und Luftqualitätsparameter und die Lagemaße, die verwendet wurden, um aus den 6-stündli-

chen (Meteorologie-Parameter) bzw. stündlichen (Luftqualitätsparameter) Messungen einen Tageswert zu bilden. Tabelliert sind 5-Punkte-Zusammenfassungen der über alle Landkreise gemittelten Tageswerte. Eine 5-Punkte-Zusammenfassung besteht aus Minimum, 1. Quartil (25%-Quantil), Median (50%-Quantil), 3. Quartil (75%-Quantil) und Maximum der beobachteten Stichprobenwerte und vermittelt somit einen guten Eindruck von Lage, Streuung und Symmetrie der empirischen Verteilung (vgl. Fahrmeir et al. (2004)).

Variable	Verw. Lagemaß pro Tag	Min.	1st Qu.	Median	3rd Qu.	Max.
lcc.ave	Arithm. Mittel	0.00000	0.01876	0.20280	0.48470	1.00000
mcc.ave	Arithm. Mittel	0.00000	0.06368	0.25500	0.48660	1.00000
q.ave	Arithm. Mittel	0.001001	0.004040	0.005654	0.008224	0.014730
q.mmm	Maximum – Minimum	0.000014	0.000699	0.001147	0.001766	0.008235
qnh.ave	Arithm. Mittel	11.37	11.46	11.47	11.48	11.53
qnh.mmm	Maximum – Minimum	0.000039	0.001836	0.003113	0.005262	0.021580
qnh.ave (Mü.)	Arithm. Mittel	986.1	1011.0	1016.0	1021.0	1040.0
qnh.mmm (Mü.)	Maximum – Minimum	0.2024	1.9000	3.1290	5.2440	20.1200
sshf.ave	Arithm. Mittel	-3442000	-1318000	-578900	43750	4056000
tmt.ave	Arithm. Mittel	-13.770	3.206	9.068	15.270	27.320
tmt.mmm	Maximum – Minimum	0.054	3.220	5.451	8.663	21.340
cp.max	Maximum	0.0000000	0.0000000	0.0002148	0.0014280	0.0300300
lsp.max	Maximum	0.0000000	0.0000000	0.0000733	0.0007105	0.0634900
stressspd.max	Maximum	135.4	5483.0	11280.0	22170.0	287500.0
windspd.max	Maximum	0.6465	2.8000	3.6980	5.0090	15.0200
SO2.q95	95%-Quantil	2.000	4.078	5.828	8.061	65.880
PM10.q95	95%-Quantil	2.228	27.420	37.880	52.100	420.300
O3.q95	95%-Quantil	3.00	36.35	57.33	80.03	183.70
NO2.q95	95%-Quantil	3.311	48.180	62.660	81.590	263.300
CO.q95	95%-Quantil	0.1038	0.5157	0.7296	1.0160	6.2130

Tabelle 2.3: 5-Punkte-Zusammenfassung der im Zeitraum 2006 bis 2008 beobachteten Meteorologie- und Luftqualitäts-Tageswerte (gemittelt über alle Landkreise)

2.4 Deskriptive Statistiken und Korrelationsanalyse

Dieser Abschnitt befasst sich mit den Eigenschaften der verwendeten KVB-Datensätze und stellt einen ersten Bezug zu den Meteorologie- und Luftqualitätsparametern her, jedoch zunächst ohne Anwendung statistischer Regressionsmodelle, nur durch Betrachtung von Korrelationen.

Tabelle 2.4 zeigt eine 5-Punkte-Zusammenfassung der pro Tag in Bayern beobachteten Gesamtfallzahlen, die sich durch Aufsummieren über alle Landkreise sowie über alle (Geschlechts- und) Altersgruppen ergeben. Bei den Abrechnungsdaten werden zudem die Diagnosen Asthma und COPD unterschieden. Die Bezeichnung „Arztbesuche gesamt“ meint hier und im Folgenden die Arztbesuche aufgrund von Asthma *oder* COPD. Summiert man die Anzahl der Arztbesuche wg. COPD und die Anzahl der Arztbesuche wg. Asthma auf, ergibt sich jedoch nicht zwangsläufig die „Anzahl gesamt“, da Personen gleichzeitig als Asthma- und COPD Patienten geführt werden können. Die Problematik der geringen Fallzahlen bei den KVB-Call-Center-Daten

(20 bis 360 Anrufe pro Tag in ganz Bayern) wird hier schnell deutlich, zumal in der Tabelle noch keine Aufteilung in Landkreise und Altersgruppen erfolgte.

Zielgröße	Min.	1st Qu.	Median	3rd Qu.	Max.
Call-Center-Anrufe	20	44	92	148	360
Arztbesuche gesamt	896	2588	38770	50610	111900
Arztbesuche COPD	492	1407	20610	26480	60590
Arztbesuche Asthma	461	1360	21010	28140	61270

Tabelle 2.4: 5-Punkte-Zusammenfassung der pro Tag beobachteten Gesamtfallzahlen in Bayern (KVB-Datenquellen)

In Appendix G sind weitere 5-Punkte-Zusammenfassungen der täglichen Fallzahlen getrennt nach Altersgruppen (Call-Center-Daten) bzw. Alters- und Geschlechtsgruppen (Abrechnungsdaten) tabellarisch dargestellt (vgl. Tabellen G.1 und G.2). Tendenziell fällt dort auf, dass die Anzahl der Anrufe bzw. Arztbesuche mit dem Alter zunimmt. Bei den Abrechnungsdaten lässt sich erkennen, dass in Altersgruppe 1 mehr Arztbesuche (gesamt) von männlichen Patienten registriert wurden als von weiblichen. Dies ist auf die bei Jungen größere Anzahl an Arztbesuchen wg. Asthma zurückzuführen. In allen anderen Altersgruppen ist das Geschlechtsverhältnis bezogen auf die Arztbesuche gesamt umgekehrt.

Abbildung 2.4 zeigt für beide Datenquellen den zeitlichen Verlauf der Fallzahlen über den jeweiligen Beobachtungszeitraum hinweg. Die Fallzahlen wurden dafür wiederum über alle (Geschlechts- und) Altersgruppen sowie die Landkreise summiert. Die glatten Kurven entstehen aus den Tagessummen, die durch einen lowess-Smoother geglättet werden. Lowess steht dabei für „locally weighted polynomial regression“ und beruht darauf, dass lokal ein Polynom niedrigen Grades an die Daten angepasst wird, wobei eine Beobachtung dann großes Gewicht erhält, wenn sie im Zentrum des aktuell betrachteten Datenabschnitts liegt (vgl. Cleveland (1979)). Bei den Abrechnungsdaten fällt der Rückgang der Arztbesuche Ende Mai/Anfang Juni sowie in den Monaten August, September und Dezember auf, was sicherlich zum großen Teil durch Schulferien, Urlaubszeit und die zahlreichen Feiertage zum Jahresende zu erklären ist. Bei den Call-Center-Daten treten die kleinsten Anruferzahlen generell zwischen Juli und Oktober auf, während zwischen Januar und Mai deutlich mehr Anrufe eintreffen. Die Schulferien lassen sich hier nicht so deutlich abgrenzen. Bei beiden Datenquellen lässt die deutliche Variation in den Fallzahlen über die Zeit hinweg eine Beeinflussung der Asthma- und COPD-Morbidität durch jahreszeitabhängige äußere Faktoren, wie z. B. Luftdruck oder Temperatur, vermuten.

In Appendix G ist der zeitliche Verlauf für beide Datenquellen nochmals getrennt nach (Geschlechts- und) Altersgruppen dargestellt (vgl. Abbildung G.4). Bei den Abrechnungsdaten ist hier nur die gesamte Anzahl der Arztbesuche (also wegen COPD *oder* Asthma) dargestellt. In beiden Darstellungen spiegelt sich die bereits in den

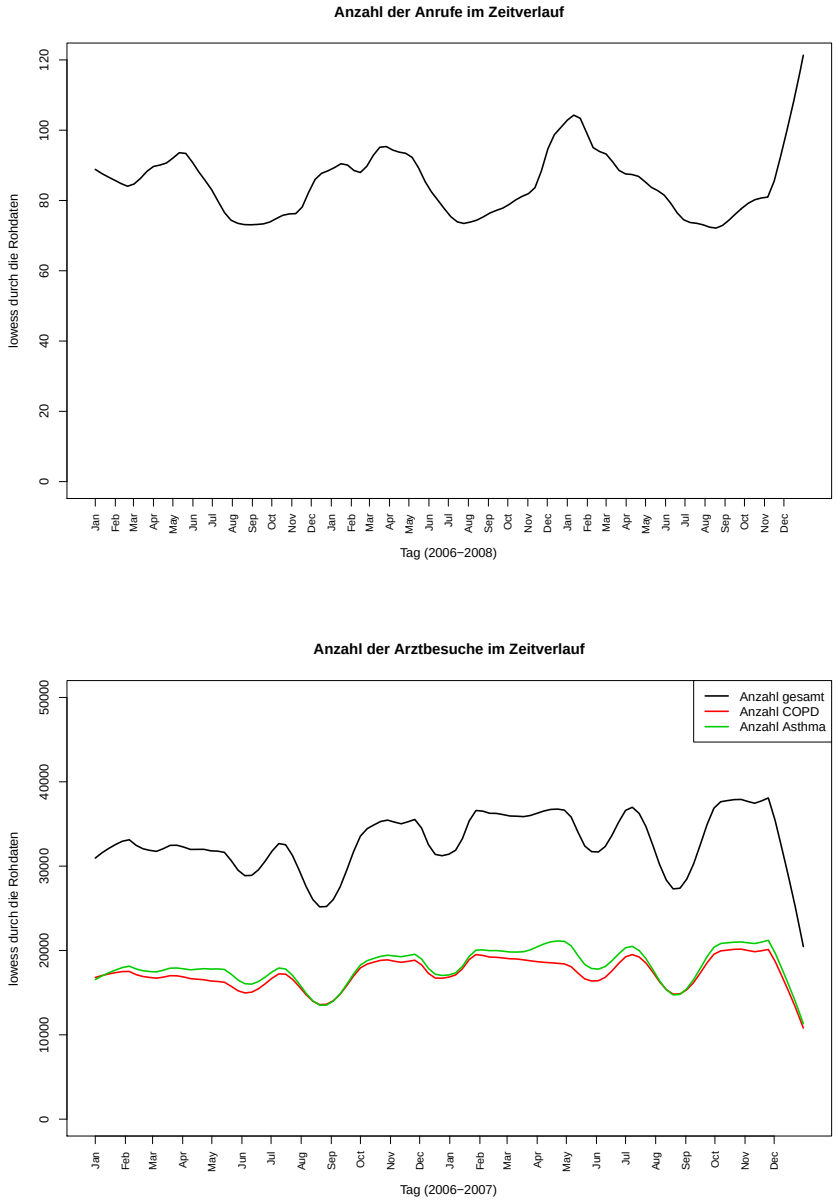


Abbildung 2.4: Anruferzahlen im zeitlichen Verlauf von 2006 bis 2008 (oben) und Arztbesuche im zeitlichen Verlauf getrennt nach Diagnose von 2006 bis 2007 (unten)

gruppenspezifischen 5-Punkte-Zusammenfassungen erkennbare Tendenz wider, dass die Fallzahl mit dem Alter zunimmt. Außerdem lassen sich in den höheren Altersgruppen deutlichere Schwankungen in den Fallzahlen erkennen, was dafür spricht, dass sich die Änderung von Umweltbedingungen bei älteren Menschen stärker aus-

wirkt. Grundsätzlich verlaufen die lowess-Kurven relativ parallel, besonders wenn man die beiden Geschlechtsgruppen bei den Abrechnungsdaten vergleicht.

Um auch die räumliche Verteilung der Fallzahlen zu betrachten sind in Abbildung 2.5 die landkreisspezifischen Summen der Anrufe bzw. Arztbesuche (gesamt) über alle (Geschlechts- und) Altersgruppen und über den gesamten Untersuchungszeitraum dargestellt. Um die Vergleichbarkeit der Landkreise zu gewährleisten, wurden diese Summen durch die jeweilige Einwohnerzahl geteilt. Bei den Call-Center-Daten lässt sich tendenziell ein Nordost-Südwest-Gefälle mit mehr Anrufen im Nordosten Bayerns ausmachen. Dieser Eindruck resultiert vor allem aus den niedrigen Anruferzahlen in den Regierungsbezirken Schwaben, Oberbayern und Unterfranken. Ein ähnliches Muster findet man bei der Betrachtung der Deprivationsscores in den bayerischen Landkreisen (vgl. Abbildung 2.6). Es soll unter anderem in den folgenden Abschnitten durch räumliche Regressionsmodelle geklärt werden, ob die stärkere Deprivation in den nordöstlichen Landkreisen Bayerns zur Erklärung der dort größeren Anruferzahlen beiträgt. Bei den Abrechnungsdaten lässt sich ein Nord-Süd-Gefälle erkennen mit mehr Arztbesuchen im Norden. Eine mögliche Erklärung dafür könnten die zwischen Nord- und Südbayern teils stark variierenden meteorologischen Bedingungen liefern. Auch diese Frage soll mithilfe räumlicher Regressionsmodelle beantwortet werden. Für die Abrechnungsdaten wurden die gleichen Darstellungen auch nochmal getrennt nach Diagnose erstellt (vgl. Abbildung G.5). Dabei lässt sich erkennen, dass das Nord-Süd-Gefälle bei der Diagnose COPD sehr deutlich ausgeprägt ist, während die Asthmafälle gleichmäßiger über ganz Bayern verteilt sind.

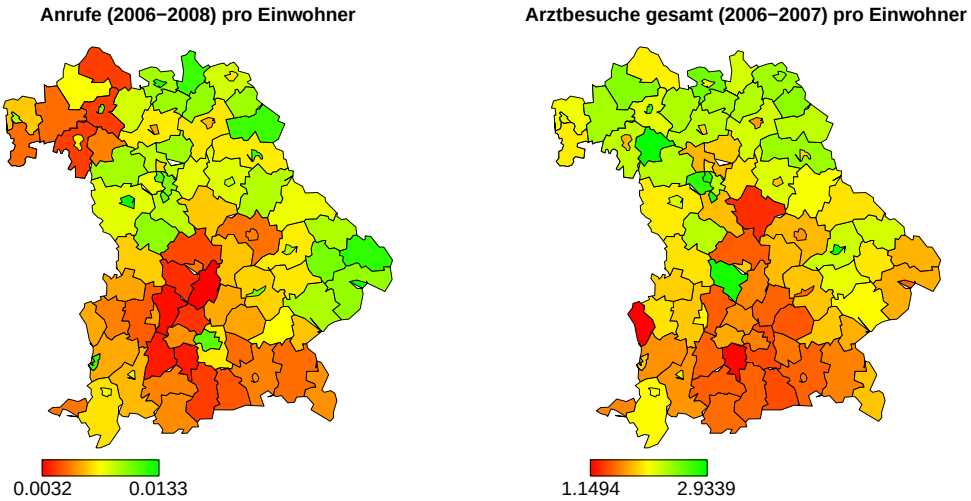


Abbildung 2.5: Räumliche Verteilung der Call-Center-Anrufe 2006 bis 2008 aggregiert über alle Altersgruppen (links) und der Arztbesuche (gesamt) von 2006 bis 2007 aggregiert über alle Geschlechts- und Altersgruppen (rechts) jeweils standardisiert durch die Einwohnerzahl

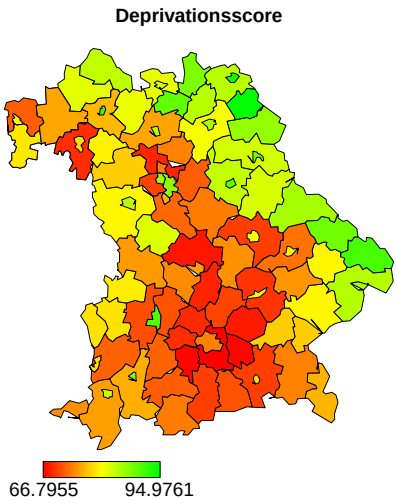


Abbildung 2.6: Deprivationsscore in den bayerischen Landkreisen

Wie bereits bei der Vorstellung der KVB-Datensätze erwähnt, ist in beiden Datenquellen mit erheblichen Wochenend- und Feiertagseffekten zu rechnen. Auch Schulferien, Brückentage und die ersten bzw. letzten Tage des Quartals müssen bei der Analyse der Fallzahlen berücksichtigt werden. Da der Effekt solcher administrativer Gegebenheiten erfahrungsgemäß weit stärker einzuschätzen ist als der Einfluss von Wetter und Luftqualität, ist die Berücksichtigung administrativer Kovariablen bei der Modellkonstruktion dringend erforderlich, um eine realistische Vorhersage der Fallzahlen zu erhalten, aber auch um Meteorologie- und Luftqualitätseffekte unverzerrt herausfiltern zu können.

Abbildung G.6 zeigt Balkendiagramme für alle verwendeten administrativen Kovariablen zur optischen Beurteilung der jeweiligen Effekte ohne Berücksichtigung des Einflusses weiterer Kovariablen. Dargestellt sind Mittelwerte der über alle (Geschlechts- und) Altersgruppen summierten Fallzahlen im jeweiligen Untersuchungszeitraum. Diese wurden getrennt nach den Faktorstufen der jeweiligen administrativen Kovariable berechnet.

Die Öffnungszeiten der Arztpraxen in Bayern spiegeln sich deutlich in den Fallzahlen an den Wochentagen wider. Während die Anrufe beim KVB-Call-Center an Samstagen und Sonntagen stark zunehmen, geht die Anzahl der Arztbesuche von im Schnitt etwa 40000 auf Werte unter 5000 (bei geöffneten Notfall-Praxen) zurück. Auch die in der Regel auf den Vormittag beschränkten Sprechzeiten an Mittwochen und Freitagen sind in beiden Plots zu erkennen. Während der Schulferien verringert sich die Anzahl der Arztbesuche durch die geschlossenen Praxen ebenfalls deutlich, die Anruferzahlen steigen wiederum. Ein ähnliches Bild zeigt sich an Feiertagen. Die Fallzahlen an Brückentagen dagegen unterscheiden sich vor allem bei der Anzahl der

Arztbesuche weniger deutlich von den gewöhnlichen Fallzahlen. In der ersten Quartalswoche sind überdurchschnittlich viele Arztbesuche verzeichnet. Dies lässt sich möglicherweise dadurch erklären, dass viele Patienten ihre Dauermedikation zu Beginn des Quartals abholen. In der letzten Quartalswoche dagegen werden weniger Arztbesuche verzeichnet als im Normalfall. Umgekehrt ist am Ende des Quartals ein deutlicher Anstieg der Anruferzahlen in den KVB-Call-Centern erkennbar. Auch in der ersten Woche eines neuen Quartals gehen mehr Anrufe als gewöhnlich ein, was möglicherweise auf die Angst vor vollen Wartezimmern zurückzuführen ist.

Um einen ersten Eindruck vom Zusammenhang zwischen den Fallzahlen und den vorhandenen Kovariablen zu erhalten, wurden die entsprechenden empirischen Korrelationen betrachtet (vgl. Tabelle 2.5). Die Tabellen enthalten den Korrelationskoeffizienten nach Bravais-Pearson, welcher den linearen Zusammenhang zwischen zwei Merkmalen misst und auf $[-1; 1]$ normiert ist (vgl. Fahrmeir et al. (2004)). Generell sind die Werte, vor allem bei den Call-Center-Daten, betragsmäßig relativ klein, was für einen schwach ausgeprägten Zusammenhang spricht. Hohe Anruferzahlen gehen beispielsweise tendenziell mit erhöhten Stickstoffdioxid-, Feinstaub- und Kohlenstoffmonoxid-Konzentrationen einher. Es gilt allerdings zu berücksichtigen, dass durch Betrachtung von Korrelationen der Einfluss der Kovariablen weder nachgewiesen noch quantifiziert werden kann, da sich Korrelationen auf die gemeinsame Verteilung der betrachteten Merkmale beziehen. Ein einseitiger Wirkungszusammenhang lässt sich erst durch Betrachtung der bedingten Verteilung der Fallzahlen gegeben die Kovariablen beurteilen (vgl. die in Abschnitt 2.5 beschriebene Struktur von Regressionsmodellen).

	Anrufe	Arztbesuche gesamt	Arztbesuche wg. COPD	Arztbesuche wg. Asthma
Deprivation	-0.0359	-0.0804	-0.0380	-0.1140
lcc.ave	-0.0092	-0.0218	-0.0122	-0.0283
mcc.ave	-0.0195	0.0150	0.0155	0.0145
q.ave	-0.0456	-0.0142	-0.0133	-0.0151
qnh.ave	0.0293	0.0442	0.0495	0.0407
sshf.ave	0.0215	0.0038	0.0084	0.0005
tmt.ave	-0.0333	-0.0088	-0.0127	-0.0059
cp.max	-0.0285	0.0137	0.0146	0.0126
lsp.max	-0.0075	0.0065	0.0061	0.0065
stressspd.max	-0.0020	-0.0100	-0.0061	-0.0130
windspd.max	0.0088	0.0147	0.0180	0.0127
SO2.q95	-0.0113	0.0763	0.0798	0.0717
PM10.q95	0.0399	0.1270	0.1179	0.1308
O3.q95	0.0056	-0.0838	-0.0851	-0.0828
NO2.q95	0.0430	0.2804	0.2596	0.2895
CO.q95	0.0213	0.1779	0.1729	0.1778

Tabelle 2.5: Pearson-Korrelationen der Anzahl der Anrufe bzw. der Arztbesuche mit den verfügbaren Meteorologie- und Luftqualitätsparametern

Des Weiteren ist es erforderlich die Wirkweise der potenziellen Einflussgrößen mithilfe von Regressionsmodellen simultan zu analysieren, da sich die verschiedenen Luftqualitäts- und Meteorologie-Parameter naturgemäß auch gegenseitig beeinflussen. Dies verdeutlichen die empirischen Korrelationsmatrizen der Meteorologie- und

Luftqualitätsparameter (vgl. Tabellen G.3 und G.4). Hohe Korrelationen treten z. B. zwischen Temperatur und Luftfeuchtigkeit (0.91) oder Schwefeldioxid und Kohlenstoffmonoxid (0.78) auf. Ebenso können Abhängigkeiten zwischen Meteorologie- und Luftqualitätsparametern beobachtet werden (vgl. Tabelle G.5). Beispielsweise beträgt der Korrelationskoeffizient zwischen Temperatur und Ozon 0.68. Die Tatsache, dass hohe Korrelationen zwischen Kovariablen existieren, kann bei der Modellierung Schwierigkeiten verursachen. Insbesondere treten Kollinearitäten zwischen Kovariablen auch bei der Betrachtung zeitlich verzögerter Effekte auf (mehr dazu in Abschnitt 4).

2.5 Herausforderungen bei der Modellierung und konkrete Zielsetzung

Dieser Abschnitt befasst sich zunächst mit der grundlegenden Struktur von statistischen Regressionsmodellen und definiert einige wichtige Begriffe in diesem Kontext. Anschließend werden die konkreten Fragestellungen formuliert, die in dieser Diplomarbeit anhand solcher Modelle beantwortet werden sollen. Dabei wird auch auf die Schwierigkeiten eingegangen, die sich aus der skizzierten Datenlage ergeben.

Grundsätzlich dienen Regressionsmodelle dazu, den Effekt einer oder mehrerer Kovariablen $x_1, \ldots, x_p$ auf eine Zielgröße y zu charakterisieren und zu quantifizieren (vgl. Fahrmeir et al. (2004)). Dazu nimmt man an, dass y in funktionaler Beziehung zu $x_1, \ldots, x_p$ steht, die durch einen zufälligen Fehler ε gestört wird:

$$y = f(x_1, \ldots, x_p) + \varepsilon. \tag{1}$$

Das einfachste Beispiel für solch eine funktionale Beziehung ist das lineare Modell

$$y = \beta_0 + \beta_1 x_1 + \ldots \beta_p x_p + \varepsilon.$$

Geht man nun davon aus, dass die $x_1, \ldots, x_p$ deterministisch bekannt sind und ε eine Zufallsverteilung besitzt, dann folgt aus Gleichung (1) die Zufälligkeit von $y|x_1, \ldots, x_p$ und die angenommene Zufallsverteilung überträgt sich von ε auf $y|x_1, \ldots, x_p$. Im linearen Modell nimmt man z. B. eine Normalverteilung für ε an, so dass auch $y|x_1, \ldots, x_p$ eine Normalverteilung besitzt.

Ziel ist es nun Kenngrößen dieser Verteilung anhand der vorliegenden Beobachtungen $(y_i, x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, n$, zu schätzen und somit das Verhalten der Zielgröße durch die realisierten Kovariablenwerte zu erklären. Bei der Erwartungswertregression beschränkt man sich darauf, den Erwartungswert $\mathbb{E}(y|x_1, \ldots, x_p)$ und die Varianz $\mathrm{Var}(y|x_1, \ldots, x_p)$ zu schätzen, es gibt jedoch auch allgemeinere Ansätze wie Quantil- oder Dichteregression (vgl. z. B. Koenker (2005) und Dunson et al. (2007)).

Parametrisiert wird die Verteilung von $y|x_1, \ldots, x_p$ durch die Regressionskoeffizienten $\boldsymbol{\beta}$. Für den Erwartungswert $\mathbb{E}(y|x_1, \ldots, x_p)$ nimmt man dann an, dass er durch eine Funktion $\xi(\boldsymbol{\beta})$ beschrieben werden kann. Im linearen Modell lautet diese Annahme beispielsweise

$$\mathbb{E}(y|x_1, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \ldots \beta_p x_p.$$

Die rechte Seite dieser Gleichung bezeichnet man in diesem Zusammenhang als linearen Prädiktor η.

Die Schätzung (Inferenz) der Regressionsparameter $\boldsymbol{\beta}$ hängt vom Typ des verwendeten Regressionsmodells ab und kann grundsätzlich entweder frequentistisch (Maximum-Likelihood-Theorie) oder bayesianisch erfolgen. Auf Basis der Schätzungen $\hat{\boldsymbol{\beta}}$ und $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})$ kann schließlich der Effekt der potenziellen Einflussgrößen untersucht und mithilfe von Parametertests auf Signifikanz geprüft werden. Zusätzlich kann für jede Beobachtung $(y_i, x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, n$, durch Einsetzen von $\hat{\boldsymbol{\beta}}$ in $\xi(\boldsymbol{\beta})$ eine Schätzung für $\hat{y}_i$ berechnet werden, die mit dem beobachteten Wert y_i verglichen werden kann. Durch Bildung der Differenz $y_i - \hat{y}_i$ erhält man eine beobachtungsspezifische Schätzung für den Fehlerterm, das sogenannte Residuum $\hat{\varepsilon}_i$. Basierend auf den Residuen aller Beobachtungen ist eine Beurteilung der Anpassungsgüte des Modells an die Daten möglich und es können verschiedene Maße zur Modelldiagnostik berechnet werden. Des Weiteren kann mithilfe von Regressionsmodellen der Wert der Zielvariable y auch für eine neue Beobachtung $(y_i^*, x_{i1}^*, \ldots, x_{ip}^*)$ ohne Kenntnis von y_i^* nur basierend auf den Kovariableninformationen $x_{i1}^*, \ldots, x_{ip}^*$ vorhergesagt werden. Dieses Vorgehen bezeichnet man als Prognose oder Prädiktion, wobei die Unsicherheit im Vergleich zur Schätzung zunimmt.

Longitudinale Regressionsmodelle unterscheiden sich von gewöhnlichen Regressionsmodellen dadurch, dass die Zielvariable y und die Kovariablen $x_1, \ldots, x_p$ wiederholt über die Zeit, zu den Zeitpunkten t $(t = 1, \ldots, T)$ gemessen werden. Im Wesentlichen gibt es zwei Ansätze zur Analyse longitudinaler Daten: marginale und bedingte Modelle. Der grundlegende Unterschied besteht darin, dass bedingte Modelle zur Erklärung der Zielvariable zum Zeitpunkt y_t auch vergangene Zielvariablenwerte $y_{t-1}, y_{t-2}, \ldots$ als Kovariablen in die Modellgleichung mitaufnehmen. Durch die Tatsache, dass y nicht mehr ausschließlich von $x_1, \ldots, x_p$ beeinflusst wird, ist die gewöhnliche intuitive und von y unabhängige Interpretation der Kovariablenparameter nicht mehr möglich. Aus diesem Grund wurden in dieser Arbeit ausschließlich marginale Modelle verwendet. Gleichwohl ist es in marginalen Modellen durchaus möglich, Kovariablenwerte $x_{1,t-1}, \ldots, x_{p,t-1}, x_{1,t-2}, \ldots, x_{p,t-2}, \ldots$ zu vergangenen Zeitpunkten (sogenannte Lags) in die Erklärung der Zielvariable zum Zeitpunkt t miteinzubeziehen. Man spricht in diesem Zusammenhang von verzögerten Kovariableneffekten (vgl. Abschnitt 4).

Es folgt eine konkrete, auf die vorliegende Problemstellung bezogene Definition der Zielvariablen y sowie der Kovariablen $x_1, \ldots, x_p$:

a) y_{ist} entspricht der Anzahl der Anrufe beim KVB-Call-Center
   - in Altersgruppe i ($i = 1 \mathrel{\hat{=}} \leq 20$ Jahre, $\ldots, I = 5 \mathrel{\hat{=}} \geq 81$ Jahre),
   - im Landkreis s ($s = 1 \mathrel{\hat{=}}$ Aichach-Friedberg, $\ldots, S = 96 \mathrel{\hat{=}}$ Würzburg),
   - am Tag t ($t = 1 \mathrel{\hat{=}}$ 1.1.2006, $\ldots, T = 1096 \mathrel{\hat{=}}$ 31.12.2008).

b) y_{ijst} entspricht der Anzahl der Arztbesuche gesamt (bzw. wegen COPD oder Asthma)
   - in Altersgruppe i ($i = 1 \mathrel{\hat{=}} \leq 20$ Jahre, $\ldots, I = 4 \mathrel{\hat{=}} \geq 61$ Jahre),
   - in der Geschlechtsgruppe j ($j = 1 \mathrel{\hat{=}}$ männlich, $J = 2 \mathrel{\hat{=}}$ weiblich),
   - im Landkreis s ($s = 1 \mathrel{\hat{=}}$ Aichach-Friedberg, $\ldots, S = 96 \mathrel{\hat{=}}$ Würzburg),
   - am Tag t ($t = 1 \mathrel{\hat{=}}$ 1.1.2006, $\ldots, T = 730 \mathrel{\hat{=}}$ 31.12.2007).

Der Zielvariablenvektor $\boldsymbol{y}$ enthält die Fallzahlen in allen Subgruppen und hat dementsprechend die Länge $n = I \cdot S \cdot T = 526080$ bzw. $n = I \cdot J \cdot S \cdot T = 560640$. Bei den KVB-Abrechnungsdaten wird zur Untersuchung der Kovariableneffekte in den Abschnitten 3 und 4 ausschließlich die Zielvariable Arztbesuche gesamt verwendet. Bei den Prognosemodellen in Abschnitt 5 werden Asthma und COPD zusätzlich separat betrachtet.

Die Kovariablen lassen sich unterteilen in:

a) Designbedingte Kovariablen:
   - Landkreis x_{district} (vgl. Abschnitt 3.5)
   - Alter x_{age} (1/.../4 bzw. 1/.../5)
   - Geschlecht x_{sex} (*1*/2)

b) Administrative Kovariablen:
   - Wochentag x_{dow} (Mo/Di/Mi/Do/Fr/Sa/So)
   - Schulferien x_{school} (*nein*/ja)
   - Feiertag x_{holiday} (*nein*/ja)
   - Brückentag x_{bridge} (*nein*/ja)
   - Quartalsanfang/-ende x_{quartal} (*Mitte*/Anfang/Ende)

c) Landkreisspezifische Deprivation $x_{\text{deprivation}}$

d) Landkreisspezifische meteorologische Kovariablen (vgl. Tabellen 2.1 und 2.3 für die Erklärung der Variablennamen):

$x_{\text{lcc.ave}}$, $x_{\text{mcc.ave}}$, $x_{\text{q.ave}}$, $x_{\text{q.mmm}}$, $x_{\text{qnh.ave}}$, $x_{\text{q.mmm}}$, $x_{\text{sshf.ave}}$, $x_{\text{tmt.ave}}$, $x_{\text{tmt.mmm}}$, $x_{\text{cp.max}}$, $x_{\text{lsp.max}}$, $x_{\text{stressspd.max}}$, $x_{\text{windspd.max}}$, $x_{\text{c.wdir}}$ (N/NO/NW/O/S/SO/SW/W)

e) Landkreisspezifische Luftqualitätskovariablen (vgl. Tabellen 2.2 und 2.3 für die Erklärung der Variablennamen):

$x_{\text{SO2.q95}}$, $x_{\text{PM10.q95}}$, $x_{\text{O3.q95}}$, $x_{\text{NO2.q95}}$, $x_{\text{CO.q95}}$

Für die kategorialen Variablen wurde nach inhaltlichen Gründen eine Referenz- oder Effektcodierung durchgeführt (vgl. Tutz (2000)). Referenzcodiert wurden Geschlecht, Schulferien, Feiertag, Brückentag und Quartalsanfang/-ende. Die jeweilige Referenzkategorie ist in obiger Übersicht kursiv markiert. Effektcodiert wurden Alter, Wochentag und Windrichtung. Da bei den KVB-Abrechnungsdaten das Verhältnis der Fallzahlen zwischen Männern und Frauen über die verschiedenen Altersstufen hinweg variiert (vgl. Tabelle G.2), wurde in alle Abrechnungsmodelle ein Interaktionsterm zwischen Alter und Geschlecht ($x_{\text{sex,age}}$) aufgenommen, um diese ungleichen Verhältnisse erfassen zu können. Dieser setzt sich aus dem Produkt der entsprechenden Dummy-Variablen zusammen. Die Kovariablenvektoren $\boldsymbol{x}_r$ der Länge n beinhalten die zu jeder Fallzahl y_{ist} bzw. y_{ijst} zugehörigen Werte der Kovariable r ($r = 1, \ldots, p$). Die $\boldsymbol{x}_r$ stellen die Spalten der Kovariablen- oder Designmatrix $\boldsymbol{X}$ mit Dimension $n \times p$ dar.

Der erste Schwerpunkt der vorliegenden Arbeit besteht darin, herauszufinden, welche der vorgestellten Meteorologie- und Luftqualitätsparameter einen nachweisbaren Einfluss auf die Asthma- und COPD-Morbidität besitzen. Darüber hinaus sollen die Kovariableneffekte mithilfe von geeigneten longitudinalen Regressionsmodellen charakterisiert und quantifiziert werden. Um vor allem im Hinblick auf die Prognose ein möglichst einfaches Modell zu konstruieren, welches das Verhalten der Zielvariable dennoch gut erklärt, kommen verschiedene Variablenselektions- bzw. Modellwahlverfahren zum Einsatz (vgl. Abschnitte 3.7 und 3.8). Auch der Einfluss von administrativen Kovariablen sowie Geschlechts- und Alterseffekte sollen simultan analysiert werden, um den jeweiligen Effekt klar von den Auswirkungen der übrigen Kovariablen abgrenzen zu können (vgl. Abschnitt 3). Schließlich soll unter Berücksichtigung der räumlichen Datenstruktur der Einfluss der landkreisspezifischen Deprivation auf die Fallzahl geprüft werden.

Eine wichtige Annahme klassischer Regressionsmodelle ist die Unkorreliertheit der Zielvariablenwerte $y_1, \ldots, y_n$. Diese Unkorreliertheit ist in dieser Situation in zweierlei Hinsicht fragwürdig. Zum einen ist eine zeitliche (serielle) Korrelation der Fallzahlen an aufeinanderfolgenden Tagen $y_t, y_{t-1}, y_{t-2}, \ldots$ zu erwarten. Abschnitt 3.3 zeigt verschiedene Möglichkeiten auf, autokorrelierte Zielvariablenwerte bei der Modellierung zu berücksichtigen. Zum anderen ist davon auszugehen, dass die Fallzahl y_s im Landkreis s nicht unabhängig von der Menge der Fallzahlen $\{y_{\tilde{s}} \mid \tilde{s} \in \delta_s\}$ in

den benachbarten Landkreisen ist (δ_s beschreibt die Menge aller zu Landkreis s benachbarten Landkreise). Solche räumlichen Abhängigkeiten können auf die geographische Lage zurückzuführen sein, aber auch durch nicht beobachtbare Einflüsse entstehen. Abschnitt 3.5 befasst sich mit der Modellierung räumlicher Effekte mithilfe bayesianischer gemischter Modelle.

Eine zusätzliche Herausforderung stellen Kovariableneffekte dar, die sich nicht konstant über den gesamten Wertebereich der Kovariable x verhalten. Nichtlineare Kovariableneffekte sind nicht Schwerpunkt dieser Arbeit, werden jedoch basierend auf den zuvor beschriebenen Datenquellen mithilfe von Generalisierten Additiven Modellen (vgl. Abschnitt 3.2.3) in Wanka (2010) ausführlich behandelt. Zudem wurde mithilfe sogenannter Bruchpunktmodelle der Frage nachgegangen, ob die Effekte von Temperatur, Luftdruck und Luftfeuchtigkeit ober- oder unterhalb eines bestimmten Schwellenwerts signifikant zu- oder abnehmen. Der Einfluss extremer Wetterbedingungen wurde in vorhergehenden Studien beispielsweise für die Temperatur nachgewiesen, allerdings hauptsächlich für die Mortalität als Zielgröße (vgl. z. B. Muggeo (2008a)). In Abschnitt 3.4 wird ein datengesteuertes Verfahren zur Festlegung der optimalen Bruchpunkte beschrieben. Die resultierenden Kovariablen sollen auf Signifikanz geprüft und in die Variablenselektion miteinbezogen werden.

Ein wichtiger methodischer Fokus dieser Diplomarbeit liegt auf der Analyse verzögerter Kovariableneffekte (Lag-Effekte). Dadurch sollen Fragen wie beispielsweise die folgende beantwortet werden können: „Hängt die heutige Anzahl der Anrufe beim KVB-Call-Center y_t von der Ozonkonzentration am Vortag $x_{\text{O3.q95,t-1}}$ ab?“ Ziel ist es, festzustellen, ob solche Effekte existieren, wie stark sie ausgeprägt sind und wie weit sie in die Vergangenheit zurückreichen. Dies führt direkt zur Schätzung der sogenannten „Distributed Lag Function“ (DLF), welche den Kovariableneffekt abhängig von der betrachteten Zeitverschiebung darstellt. Das wesentliche technische Problem bei der Schätzung dieser Funktion ist die hohe Korrelation benachbarter Lags, welche die Schätzung der Effekte instabil macht. Schwierigkeiten bereitet auch die hohe Anzahl an Modellparametern, die durch die Berücksichtigung von Lag-Kovariablen entsteht. Abschnitt 4 fasst verschiedene Methoden zur Schätzung der Distributed Lag Function zusammen, zeigt Vor- und Nachteile auf und beschreibt schließlich die Entwicklung eines neuen Regressionstools speziell für die Analyse verzögerter Kovariableneffekte. Darin ist auch eine neu entwickelte Methode zur Schätzung der DLF implementiert, die einige Vorteile gegenüber den bestehenden Ansätzen mit sich bringt.

Die aus den Abschnitten 3 und 4 gewonnenen Erkenntnisse sollen dann in die Entwicklung eines zeitlich-räumlichen Prognosemodells für die Fallzahlen, basierend auf allen verfügbaren Kovariableninformationen, einfließen (vgl. Abschnitt 5). Die bis zu 3 Tage im Voraus prognostizierten Fallzahlen y^*_{t+1}, y^*_{t+2} und y^*_{t+3} sollen letztendlich in einen mehrstufigen Prognoseindex umgesetzt werden, anhand dessen z. B. der Gefährdungsgrad von Risikopatienten vorhergesagt werden kann (vgl.

Abschnitt 5.4). Das Prognosemodell wird retrospektiv konstruiert, beginnend mit einem Trainingsmodell basierend auf den Daten des Jahres 2006, was den Vergleich zwischen wahren und prädiktierten Werten ermöglicht. Anhand von prädiktiven Maßen kann so auch die Prognosequalität verschiedener Kandidatenmodelle verglichen werden (vgl. Abschnitt 5.3). Durch einen fortlaufenden „Lernprozess“ soll eine stetige Verbesserung der Prognose erreicht werden. Angestrebtes Ziel ist schließlich der Einsatz des entwickelten Modells im tagesaktuellen Betrieb, basierend auf Prognosen der verwendeten Meteorologie- und Luftqualitätsparameter.

3 Modellbildung und Variablenselektion

Dieser Abschnitt beschäftigt sich mit der Wahl geeigneter Regressionsmodelle und der modellbasierten Analyse von Kovariableneffekten auf die Asthma- und COPD-Morbidität in Bayern. Dabei wird insbesondere auf die in Abschnitt 2.5 erwähnten Herausforderungen, die sich aus der vorliegenden Datenstruktur ergeben, und die Schwierigkeiten, welche die große Menge an verfügbaren Daten mit sich bringt, eingegangen.

3.1 Computationale Probleme und daraus resultierende Einschränkungen

Die komplette Datenanalyse sowie die frequentistische Modellierung erfolgte mit der frei zugänglichen Software R (vgl. R Development Core Team (2009)). Die bayesianischen Modelle wurden mit der ebenfalls frei verfügbaren Software BayesX (vgl. Belitz, Brezger, Kneib, und Lang (2009)) gefittet. Als problematisch erwies sich in beiden Programmen die Verarbeitung der vorhandenen großen Datenmengen.

Zum einen stand eine große Menge verfügbarer Kovariablen zur Verfügung. Die in Abschnitt 2.3 vorgestellten meteorologischen Kovariablen stellen bereits eine inhaltliche Auswahl aller verfügbaren Parameter dar. Jeder meteorologischen Kovariable wurde a priori, basierend auf medizinischen Erfahrungswerten, ein verzögerter Effekt bis zum Lag 3 unterstellt, jeder Luftqualitätskovariable sogar ein verzögerter Effekt bis zum Lag 14. Selbst unter Verwendung von parametersparenden Verfahren zur Analyse von Lag-Effekten (vgl. Abschnitt 4) ergibt sich somit eine große Anzahl p von Spalten in der Designmatrix $\boldsymbol{X}$. Zum anderen sind viele der Spalten, wie bereits in Abschnitt 2 erwähnt, hochkorreliert, so dass sich starke Konvergenzprobleme in Regressionsmodellen ergeben können. Diese haben ihre Ursache darin, dass zur Bestimmung der Schätzung die quadrierte Designmatrix $\boldsymbol{X}^\top \boldsymbol{X}$ der Dimension $p \times p$ (wiederholt) invertiert werden muss. Die Inversion ist für großes p sehr rechenintensiv und wird dadurch erschwert, dass durch (quasi-)kollineare Spalten in $\boldsymbol{X}$ eine Rangdefizienz der zu invertierenden Matrix entstehen kann.

Auch aus inhaltlicher Sicht ist eine Reduzierung der Kovariablenzahl p anzustreben, da sich Kovariablen, die keinen Einfluss auf die Zielvariable besitzen, störend auf deren Prädiktion auswirken können. Klassische Variablenselektionsverfahren, welche den Ausschluss von bedeutungslosen Kovariablen ermöglichen, beruhen jedoch darauf, viele Kovariablenkombinationen zu testen, um das optimale Modell zu finden. Gerade bei großem p ergibt sich so relativ schnell eine riesige Anzahl zu berechnender Modelle mit hochdimensionaler Designmatrix, so dass eine Modellwahl basierend auf solchen Verfahren nur realisierbar ist, wenn die einzelnen Modelle mit vertretbarem Zeitaufwand berechnet werden können.

Des Weiteren ergibt sich durch die Aufschlüsselung der täglichen Fallzahlen nach (Geschlecht,) Alter und Landkreis über 2 bzw. 3 Jahre hinweg eine große Beobachtungszahl n, so dass die Designmatrix nicht nur viele Spalten, sondern auch bei beiden Datenquellen über eine halbe Million Zeilen besitzt. Das führt dazu, dass statistische Modellierungsprozeduren (zu) viel Rechenzeit benötigen oder aus technischen Gründen (z. B. Arbeitsspeicherbegrenzung) überhaupt nicht verwendet werden können. Dies gilt wiederum insbesondere für schrittweise Variablenselektionsverfahren.

Aus diesem Grund ist eine Reduzierung der Beobachtungszahl n erforderlich, wobei nur so wenig wie möglich an Dateninformation verloren gehen soll. In dieser Arbeit wurden mehrere Reduktionsverfahren verwendet, um den unterschiedlich gelagerten Informationsverlust der einzelnen Verfahren zu kompensieren. Diese werden im Folgenden kurz vorgestellt:

a) Für einen Großteil der Analysen wurde der gesamte Datensatz auf die Landeshauptstadt München (ohne Landkreis München) als Pilotregion eingeschränkt. Diese Einschränkung begründet sich, neben der großen Einwohnerzahl Münchens (ca. 1.3 Millionen) im Vergleich zu den übrigen bayerischen Landkreisen, in der Verfügbarkeit zahlreicher Messstationen für die Luftqualitätsdaten, so dass diesbezüglich von einer hohen Zuverlässigkeit der Kovariableninformation auszugehen ist. Aufgrund der relativ geringen Fläche des Landkreises wurden die täglichen Luftqualitätswerte hier durch Mittelwertbildung über alle 9 Traffic-Messstationen im Stadtgebiet errechnet. Durch die hohe Arztdichte im Stadtgebiet von München ist im Vergleich zu ländlichen Regionen eine umfassende medizinische Versorgung der Patienten gewährleistet. Vernachlässigt wird durch diese Vorgehensweise jedoch die räumliche Struktur der Daten. Genauso ist es nicht möglich, den Einfluss der landkreisspezifischen Deprivation auf die Asthma- und COPD-Mortalität zu untersuchen. Die longitudinale Struktur der Daten in der exemplarisch betrachteten Pilotregion bleibt dagegen erhalten. Aus diesem Grund wurde insbesondere die Analyse verzögerter Kovariableneffekte, bei der die zeitliche Komponente im Vordergrund steht, basierend auf dieser Einschränkung durchgeführt. Ebenso wurden die auf die Landeshauptstadt München beschränkten Call-Center- und Abrechnungsdaten für die Variablenselektion in den Abschnitten 3.7 und 3.8 herangezogen. Die Zielvariablen reduzieren sich in der Notation von Abschnitt 2.5 auf y_{it} für die Call-Center-Daten bzw. y_{ijt} für die Abrechnungsdaten. Die Beobachtungszahl n geht damit auf $I \cdot T = 5480$ bzw. $I \cdot J \cdot T = 5840$ zurück. Tabelle 3.1 zeigt 5-Punkte-Zusammenfassungen der täglichen Fallzahlen für beide auf die Pilotregion München reduzierten KVB-Datensätze. Die kleinen Fallzahlen in den einzelnen Alterskategorien bei den KVB-Call-Center-Daten bedeuten eine zusätzliche Herausforderung bei der Modellierung (vgl. Abschnitt 3.2). Durch das Wegfallen aller übrigen Landkreise tritt zudem ein erheblicher Powerver-

lust bei Parametertests auf die Existenz von Kovariableneffekten (vgl. Abschnitt 3.2.1) ein.

Alter	Min.	1st Qu.	Mean	3rd Qu.	Max.
alle	2	10	13	17	41
<= 20 J.	0	0	0	0	5
21 - 40 J.	0	0	1	2	19
41 - 60 J.	0	1	2	3	13
61 - 80 J.	0	3	5	7	17
>= 81 J.	0	3	4	6	17

Geschlecht	Alter	Min.	1st Qu.	Mean	3rd Qu.	Max.
alle	alle	84	233	3331	4594	9095
männlich	<= 20 J.	0	15	149	214	389
männlich	21 - 40 J.	7	18	136	180	520
männlich	41 - 60 J.	9	29	315	429	890
männlich	>= 61 J.	9	42	745	999	2188
weiblich	<= 20 J.	0	10	100	144	245
weiblich	21 - 40 J.	14	28	229	313	655
weiblich	41 - 60 J.	11	32	506	711	1457
weiblich	>= 61 J.	14	65	1165	1570	3210

Tabelle 3.1: 5-Punkte-Zusammenfassungen der pro Tag in der Pilotregion München erfassten Anrufe beim KVB-Call-Center (oben) bzw. der Arztbesuche gesamt (unten)

b) Zur Kontrolle wurden sämtliche Analysen, die eingeschränkt auf die Pilotregion München durchgeführt wurden, basierend auf einem aggregierten Datensatz für Gesamtbayern validiert. Die gruppenspezifischen Fallzahlen wurden dazu über alle 96 bayerischen Landkreise addiert, die Werte der stetigen Kovariablen über alle Landkreise gemittelt. Die Windrichtung wurde tagesweise auf den Modus aller Landkreise reduziert. Durch diese Vorgehensweise fließt zwar die Ziel- und Kovariableninformation aller Landkreise in die Analysen mit ein, jedoch kann durch die Aggregation auch eine Abschwächung vorhandener Kovariableneffekte erfolgen. Aufgrund der unterschiedlichen geografischen Lage können die Meteorologie- und Luftqualitätsparameter deutlich über Bayern, insbesondere zwischen Nord- und Südbayern, variieren. Durch die Mittelung der Kovariablenwerte geht diese Variation komplett verloren, so dass signifikante Effekte verschwinden können. Außerdem ist es aufgrund der Mittelwertbildung sehr fraglich, ob die tatsächliche individuelle Exposition erfasst wurde und sich so ein Bezug zur Fallzahl herstellen lässt. Schließlich bewirkt auch das Wegfallen der Variation in den Fallzahlen einen Informationsverlust. Eine Untersuchung räumlicher Effekte und des Einflusses der Deprivation ist

auch bei diesem Reduktionsverfahren nicht möglich. Die notationelle Definition der Zielvariable sowie die Beobachtungszahl n unterscheiden sich nicht von Reduktionsverfahren a).

c) Um schließlich auch die räumliche Komponente der Daten und den Einfluss des Deprivationsscores zu untersuchen, wurden in einem weiteren Reduktionsverfahren die Fallzahlen über alle (Geschlechts- und) Altersgruppen summiert. Die longitudinale Struktur der Daten über alle Gruppen hinweg bleibt dabei zwar erhalten, die gruppenspezifische Zeitstruktur geht jedoch verloren. Da sich der Verlauf der Fallzahlen in den Subgruppen über den Beobachtungszeitraum hinweg durchaus unterscheidet (vgl. Abbildung G.4 und die in Abschnitt 3.3 beschriebenen altersspezifischen Zeitfunktionen), entsteht hier ein Informationsverlust im Bezug auf die zeitliche Struktur. Die so aggregierten Daten sind deswegen nur bedingt für Analysen zu gebrauchen, bei denen die zeitliche Komponente im Vordergrund steht, z. B. die Untersuchung von Lag-Effekten. In der Notation von Abschnitt 2.5 lassen sich die Zielvariablen für beide Datenquellen zu y_{st} vereinfachen. Die Beobachtungszahl n reduziert sich im Vergleich zu den vollen Datensätzen auf $S \cdot T = 105216$ für die Call-Center-Daten bzw. $S \cdot T = 70080$ für die Abrechnungsdaten. Der Rechenaufwand der räumlichen Modelle (vgl. Abschnitt 3.5), die basierend auf diesen Datensätzen gefittet werden, ist allerdings so groß, dass beispielsweise die Durchführung einer schrittweisen Variablenselektion, bei der mehrere Modelle mit unterschiedlichen Kovariablenkombinationen angepasst werden müssen, nicht mehr möglich ist. Für die Konstruktion der Trainingsmodelle in Abschnitt 5.1 wird der über alle (Geschlechts- und) Altersgruppen aggregierte Datensatz zusätzlich auf die Beobachtungen des Jahres 2006 reduziert. Basierend darauf soll eine Schätzung für die räumlichen Effekte erfolgen, die in die nachfolgenden Analysen der vollen Datensätze miteinfließt.

Alle reduzierten Datensätze (inkl. Kovariablen) und die vollen Datensätze (aus Speicherplatzgründen nur mit designbedingten und administrativen Kovariablen) sowie die kompletten Meteorologie- und Luftqualitätsdaten befinden sich auf der CD in Appendix I im Ordner „Datensätze“.

Grundsätzlich soll aufgrund der hohen Beobachtungs- und Variablenzahl der Versuch unternommen werden, die Struktur der verwendeten Modelle möglichst einfach zu halten, um die Konvergenz zu gewährleisten und die Rechenzeit zu begrenzen.

3.2 Generalisiertes Lineares Modell (GLM) und Diskussion alternativer Modelle

Dieser Abschnitt befasst sich zunächst mit dem Konzept des Generalisierten Linearen Modells (GLM) und zeigt verschiedene Erweiterungen auf, die aufgrund der

vorliegenden Datenstruktur erforderlich sind. Zudem werden erste Ergebnisse basierend auf den vorgestellten Modellen präsentiert.

3.2.1 Struktur des GLM und Quasi-Likelihood-Ansatz

Generalisierte Lineare Modelle basieren auf dem in Abschnitt 2.5 beschriebenen Grundkonzept statistischer Regressionsmodelle, also der Untersuchung der Verteilung von $y|\boldsymbol{x}_i$ ($\boldsymbol{x}_i = x_{i1}, \ldots, x_{ip}$). Sie wurden bereits 1972 von Nelder und Wedderburn eingeführt (vgl. Nelder und Wedderburn (1972)). Im Folgenden wird kurz der theoretische Hintergrund, der in dieser Arbeit verwendeten GLMs und der darauf aufbauenden Quasi-Likelihood-Modelle (vgl. McCullagh und Nelder (1989)) skizziert.

Die erste grundlegende Annahme (*Zufalls- oder Verteilungskomponente*) ist, dass $y|\boldsymbol{x}_i$ einer Verteilung aus der Exponentialfamilie folgt. Das bedeutet, dass die Dichte jeder Beobachtung $y_i|\boldsymbol{x}_i$ dargestellt werden kann als:

$$f(y_i|\theta(\mu_i), \varphi_i) = \exp\left(\frac{y_i\theta(\mu_i) - b(\theta(\mu_i))}{\varphi_i} + c(y_i, \varphi_i)\right) \tag{2}$$

mit $\mu_i = \mathbb{E}(y_i|\boldsymbol{x}_i)$. Dabei sind b und c von der konkreten Verteilung abhängige Funktionen, wobei c nicht von μ_i abhängt. $\theta_i = \theta(\mu_i)$ ist der natürliche/kanonische Parameter und φ_i der Dispersionsparameter der Verteilung aus der Exponentialfamilie.

Die zweite wesentliche Annahme (*strukturelle oder systematische Komponente*) ist, dass der Erwartungswert $\mu_i = \mathbb{E}(y_i|\boldsymbol{x}_i)$ durch eine zweimal stetig differenzierbare, streng monotone Responsefunktion h mit dem linearen Prädiktor $\eta_i = \eta(\boldsymbol{x}_i) = \boldsymbol{x}_i^\top\boldsymbol{\beta}$ verbunden ist, wobei $\boldsymbol{\beta}$ den Vektor der zu schätzenden Regressionskoeffizienten darstellt. Zusammengefasst ergibt sich:

$$\mu_i = \mathbb{E}(y_i|\boldsymbol{x}_i) = h(\eta(\boldsymbol{x}_i)) = h(\boldsymbol{x}_i^\top\boldsymbol{\beta}). \tag{3}$$

Diese Beziehung kann auch mithilfe der Linkfunktion $g = h^{-1}$ dargestellt werden:

$$g(\mu_i) = \eta(\boldsymbol{x}_i) = \boldsymbol{x}_i^\top\boldsymbol{\beta}. \tag{4}$$

Falls $g(\mu_i) = \theta(\mu_i)(= \eta_i)$ gilt, so bezeichnet man g als natürliche/kanonische Linkfunktion.

Grundsätzlich gilt innerhalb der Exponentialfamilie für die ersten beiden Momente der Verteilung von $y_i|\boldsymbol{x}_i$:

$$\mathbb{E}(y_i|\boldsymbol{x}_i) = \mu_i = \frac{\partial b(\theta_i)}{\partial\theta},$$
$$\mathrm{Var}(y_i|\boldsymbol{x}_i) = \varphi_i \underbrace{\frac{\partial^2 b(\theta_i)}{\partial\theta^2}}_{V(\mu_i)}.$$

Dabei ist $V(\mu_i)$ die verteilungsspezifische Varianzfunktion. Durch die konkrete Verteilungsannahme wird sowohl der Erwartungswert als auch die Varianz von $y_i|\boldsymbol{x}_i$ fixiert. Dies kann problematisch sein, wenn der Dispersionsparameter φ_i nicht durch einen eigenen Varianzparameter repräsentiert wird, sondern auf 1 fixiert ist, was für einige Mitglieder der Exponentialfamilie, z. B. die Poisson-Verteilung, gilt. In diesem Fall ist es möglich, dass die tatsächliche Varianz in den Daten von der theoretisch angenommenen Modellvarianz abweicht (Über- oder Unterdispersion). Dieses Phänomen kann dazu führen, dass die Standardfehler der Regressionskoeffizienten respektive unter- oder überschätzt werden, so dass eine fehlerhafte Beurteilung der Signifikanz von Kovariableneffekten erfolgt. Eine sinnvolle Alternative in beiden Situationen bietet der sogenannte Quasi-Likelihood-Ansatz, der eine Erweiterung des GLM-Prinzips darstellt.

Die Schätzung der Regressionsparameter $\boldsymbol{\beta}$ im GLM erfolgt basierend auf der logarithmierten Likelihood aller Beobachtungen. Fasst man die Dichtedarstellung (2) als Funktion im Parameter $\boldsymbol{\beta}$ auf, ergibt sich durch Logarithmieren der Log-Likelihood-Beitrag $l_i(\boldsymbol{\beta})$ der i-ten Beobachtung:

$$l_i(\boldsymbol{\beta}) = \frac{y_i\theta(\mu_i) - b(\theta(\mu_i))}{\varphi_i} + c(y_i, \varphi_i). \tag{5}$$

$\boldsymbol{\beta}$ tritt auf der rechten Seite der Gleichung in $\theta(\mu_i) = \theta(\mu(\boldsymbol{x}_i^\top\boldsymbol{\beta}))$ auf. Geht man von der Unkorreliertheit der Beobachtungen $y_i|\boldsymbol{x}_i$ aus, ergibt sich die Loglikelihood aller Beobachtungen $l(\boldsymbol{\beta})$ aus der Summe $\sum_{i=1}^n l_i(\boldsymbol{\beta})$. Die Schätzung der Regressionsparameter erfolgt nun durch Maximierung von $l(\boldsymbol{\beta})$ bezüglich $\boldsymbol{\beta}$. Im Allgemeinen ist dieses Maximierungsproblem analytisch nicht lösbar und muss näherungsweise durch iterative Algorithmen, wie das Newton-Raphson-Verfahren oder Fisher-Scoring, bestimmt werden. Beim Quasi-Likelihood-Verfahren erfolgt die Schätzung sehr ähnlich durch Nullsetzen der Quasi-Score-Funktion und iterative Lösung der daraus resultierenden Schätzgleichung (vgl. McCullagh und Nelder (1989) für weitere Details zur Schätzung im GLM und im Quasi-Likelihood-Modell).

Ein wichtiges Maß zur Beurteilung der Anpassungsqualität von GLMs stellt die sogenannte Devianz dar. Diese ist wie folgt definiert:

$$\begin{aligned} D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) &= -2\varphi\left[l(\boldsymbol{y}; \hat{\boldsymbol{\mu}}, \varphi) - l(\boldsymbol{y}; \boldsymbol{y}, \varphi)\right] \\ &= 2\sum_{i=1}^{n}\left[y_i(\theta(y_i) - \theta(\hat{\mu}_i)) - (b(\theta(y_i)) - b(\theta(\hat{\mu}_i)))\right]/a_i. \end{aligned} \tag{6}$$

Darin ist a_i die individuenspezifische Komponente des Dispersionsparameters, also $\varphi_i = \varphi a_i$. In der ersten Darstellung (obere Zeile von Gleichung (6)) wird deutlich, dass die Devianz auf der Log-Likelihood des geschätzten Modells $l(\boldsymbol{y}; \hat{\boldsymbol{\mu}}, \varphi)$ beruht. $l(\boldsymbol{y}; \boldsymbol{y}, \varphi)$ entspricht der Log-Likelihood des sogenannten saturierten Modells, die sich ergibt, wenn man die Beobachtungen $\boldsymbol{y}$ anstelle von $\hat{\boldsymbol{\mu}}$ einsetzt. Dadurch erhält man

ein perfekt an die Daten angepasstes Modell. Die skalierte Devianz $\hat{D}(\boldsymbol{y}, \hat{\boldsymbol{\mu}})$ ergibt sich aus $D(\boldsymbol{y}, \hat{\boldsymbol{\mu}})/\varphi$. Auf Basis der skalierten Devianz ist der Vergleich verschiedener Modelle möglich.

Es sei an dieser Stelle angemerkt, dass Generalisierte Lineare Modelle im Allgemeinen für longitudinale und räumliche Daten ungeeignet sind, da durch solche Strukturen die vorausgesetzte Unkorreliertheit der Beobachtungen beeinträchtigt und die Schätzung verzerrt werden kann. Um GLMs auch für die vorliegenden longitudinalen Daten nutzbar zu machen wird in Abschnitt 3.3 eine indirekte Methode zur Entkorrelierung der Beobachtungen vorgestellt. Der Einsatz von speziellen longitudinalen Regressionsmodellen wird ebenfalls diskutiert (vgl. Abschnitt 3.2.4).

Im Folgenden werden die drei in dieser Arbeit verwendeten GLM-Typen sowie das Quasi-Poisson-Modell kurz theoretisch vorgestellt:

a) Das einfachste GLM ist das bereits in Abschnitt 2.5 erwähnte *lineare Modell*. Für $y_i|\boldsymbol{x}_i$ wird dabei eine Normalverteilung der Gestalt $\mathrm{N}(\mu_i, \sigma^2)$ angenommen. Die Dichte $f(y_i|\boldsymbol{x}_i)$ lässt sich dann wie folgt in Exponentialfamiliengestalt bringen:

$$f(y_i|\mu_i, \sigma^2) = \exp\left(\frac{y_i\mu_i - \mu_i^2/2}{\sigma^2} + \left(-\log(\sqrt{2\pi\sigma^2}) - \frac{y_i^2}{2\sigma^2}\right)\right).$$

Vergleicht man diese Dichte mit der allgemeinen Dichte einer Verteilung aus der Exponentialfamilie (2), stellt man fest, dass $\theta(\mu_i) = \mu_i$ gilt. Die natürliche Linkfunktion θ entpricht damit der identischen Abbildung und es folgt $\theta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$. Der Dispersionsparameter φ_i ist gleich der Varianz der angenommenen Normalverteilung σ^2 und wird für alle Individuen i hinweg als konstant angenommen (Homoskedastizität). Die Funktionen b und c sind definiert als $b(\theta_i) = \theta_i^2/2$ und $c(y_i, \sigma^2) = -\log(\sqrt{2\pi\sigma^2}) - y_i^2/(2\sigma^2)$.

Die Devianz im linearen Modell reduziert sich auf die Quadratsumme der Residuen:

$$D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2.$$

Die Annahme einer Normalverteilung in Verbindung mit der Verwendung der natürlichen Linkfunktion impliziert eine symmetrische Verteilung der $y_i|\boldsymbol{x}_i$ und einen unbeschränkten Wertebereich der y_i. Bei den vorliegenden Zielvariablen handelt es sich jedoch um Fallzahlen, die naturgemäß ≥ 0 sind. Außerdem ist die Symmetrie der Verteilung der $y_i|\boldsymbol{x}_i$ vor allem bei den Call-Center-Daten in der Pilotregion München aufgrund der zahlreichen kleinen Fallzahlen nahe oder gleich 0 fraglich (vgl. Tabelle 3.1).

Eine sinnvolle Alternative für im Durchschnitt größere Fallzahlen stellt die Verwendung des sogenannten log-Links anstelle des natürlichen Links dar (*logli-*

neares Modell). Dementsprechend gilt für den Erwartungswert $\mu_i = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})$. Auf diese Weise ist sichergestellt, dass $\hat{y}_i = \exp(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}) \geq 0$ gilt. Die Verwendung des log-Links entspricht näherungsweise der Verwendung des identischen Links nach vorangehender Transformation der Zielvariablenwerte mit dem natürlichen Logarithmus (hier und im Folgenden mit log bezeichnet). Durch das Logarithmieren der Zielgröße kann in der Regel auch die fehlende Symmetrie der Verteilung der $y_i|\boldsymbol{x}_i$ korrigiert werden. Um Probleme mit auftretenden Nullen zu vermeiden, werden die Zielvariablenwerte in dieser Arbeit mit der Funktion $\log(y_i + 1)$ transformiert. Die Addition der konstanten 1 besitzt dabei keine Auswirkungen auf die Parameterschätzer und ihre Interpretation.

Das Problem der Überdispersion tritt beim (log)linearen Modell durch den eigenen, separat zu schätzenden Varianzparameter σ^2 und die damit verbundene Entkoppelung von Erwartungswert und Varianz nicht auf. Die Rechenzeit linearer Modelle ist sehr gering, da sich die Maximum-Likelihood-Schätzung der Regressionskoeffizienten $\hat{\boldsymbol{\beta}}$ ohne Verwendung iterativer Verfahren analytisch berechnen lässt.

b) Das *Poisson-Modell* ist das klassische und am weitesten verbreitete Modell für Zähldaten. Hier nimmt man eine Poisson-Verteilung für die Zielvariable gegeben die Kovariablen an, das heißt $y_i|\boldsymbol{x}_i \sim \mathrm{Po}(\mu_i)$. Die in Exponentialfamilienform transformierte Dichte $f(y_i|\boldsymbol{x}_i)$ lautet:

$$f(y_i|\mu_i) = \exp\left(\frac{y_i \log(\mu_i) - \mu_i}{1} + (-\log(y_i!))\right).$$

Zunächst lässt sich festhalten, dass die natürliche Link-Funktion $\theta(\mu_i) = \log(\mu_i)$ dem log-Link entspricht. Zudem gilt $b(\theta_i) = \exp(\theta_i)$ und $c(y_i, \varphi_i) = -\log(y_i!)$. Der Dispersionsparameter φ_i ist hier, wie bereits erwähnt, auf 1 fixiert. Des Weiteren gilt:

$$\mathbb{E}(y_i|\boldsymbol{x}_i) = \mathrm{Var}(y_i|\boldsymbol{x}_i) = \mu_i = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}).$$

Es existiert also kein eigener Parameter für die Varianzschätzung, so dass sich das Modell möglicherweise als zu unflexibel erweist, falls die empirische Varianz der Daten von der geschätzten Modellvarianz $\exp(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}})$ abweicht. In diesem Fall sind die nachfolgend vorgestellten Zähldaten-Modelle, wie z. B. das Negativ-Binomial- oder das Quasi-Poisson-Modell, besser geeignet.

Die Devianz des Poisson-Modells lautet:

$$D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = 2 \sum_{i=1}^{n} y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right).$$

Die Verwandtschaft zum loglinearen Modell zeigt sich darin, dass die für $y_i|\boldsymbol{x}_i$ angenommene Poisson-Verteilung für wachsendes μ_i gegen eine Normalverteilung der Form $\mathrm{N}(\mu_i, \mu_i)$ konvergiert. Das loglineare Modell unterscheidet

sich bei großen Fallzahlen also nur durch die flexiblere Varianzschätzung vom Poisson-Modell.

c) Das *Negativ-Binomial-Modell* ist ein weiteres Zähldaten-Modell im GLM-Kontext, das inbesondere für Datenstrukturen mit Überdispersion geeignet ist. Man nimmt hier eine Negativ-Binomial-Verteilung der Form $\mathrm{NB}(\nu, \mu_i)$ für $y_i|\boldsymbol{x}_i$ an ($\nu > 0$). Das Negativ-Binomial-Modell wird auch als Poisson-Gamma-Modell bezeichnet, da man analog eine Poissonverteilung $y_i|\boldsymbol{x}_i, b_i \sim \mathrm{Po}(b_i\mu_i)$ und eine Gammaverteilung der Form $\mathrm{Ga}(\nu, \nu)$ für b_i annehmen kann. Als Exponentialfamilie kann die Dichte der Negativbinomialverteilung wie folgt dargestellt werden:

$$f(y_i|\mu_i) = \exp\left(\frac{y_i \log\left(\frac{\mu_i}{\nu+\mu_i}\right) - (-\nu)\log\left(\frac{\nu}{\nu+\mu_i}\right)}{1} + \log\left(\frac{\Gamma(y_i+\nu)}{\Gamma(\nu)\Gamma(y_i+1)}\right)\right).$$

Die natürliche Linkfunktion $\theta(\mu_i)$ wird hier repräsentiert durch $\log(\mu_i/(\nu+\mu_i))$, in der Regel verwendet man jedoch auch hier den log-Link, also $\log(\mu_i) = \boldsymbol{x}_i^\top\boldsymbol{\beta}$. Daneben ergibt sich $b(\theta_i) = (-\nu)\log(1-\exp(\theta_i))$ sowie $c(y_i, \varphi_i) = \log(\Gamma(y_i + \nu)) - \log(\Gamma(\nu)) - \log(\Gamma(y_i+1))$. Der Dispersionsparameter ist wie bei der Poissonverteilung auf 1 fixiert, jedoch gilt für die Modellvarianz:

$$\mathrm{Var}(y_i|\boldsymbol{x}_i) = \mu_i + \frac{1}{\nu}\mu_i^2.$$

Daran lässt sich erkennen, dass die Varianzschätzung im Negativ-Binomial-Modell von dem zusätzlichen Parameter ν abhängt, was mehr Flexibilität bedeutet und das Problem der Überdispersion vermeidet. Genauso lässt sich jedoch erkennen, dass das Negativ-Binomial-Modell nur für Überdispersion geeignet ist, da die Modellvarianz aufgrund von $\nu > 0$ nach unten durch μ_i, also die Varianz des Poisson-Modells, beschränkt ist. Zudem lässt sich am Schätzwert für ν das Maß der Überdispersion erkennen. Große $\hat{\nu}$-Werte sprechen dafür, dass die Überdispersion vernachlässigbar ist und praktisch äquivalent ein Poisson-Modell eingesetzt werden kann.

In der R-Funktion „glm.nb“ im MASS-Paket ist die Schätzung von Negativ-Binomial-Modellen implementiert. Die Schätzung erfolgt durch ein iteratives Verfahren, bei dem abwechselnd $\hat{\boldsymbol{\beta}}$ und $\hat{\nu}$ aktualisiert werden. Die Schätzung von ν erwies sich dabei leider häufig als instabil und führte zur Divergenz des Verfahrens. Ähnliche Schwierigkeiten traten bei dem in BayesX implementierten Verfahren zur Berechnung von bayesianischen Negativ-Binomial-GLMs auf.

d) Der Quasi-Likelihood-Ansatz stellt insofern eine Erweiterung Generalisierter Linearer Modelle dar, dass anstelle der kompletten Verteilung von $y_i|\boldsymbol{x}_i$ nur

der Erwartungsswert $\mathbb{E}(y_i|\boldsymbol{x}_i)$ und die Varianz $\text{Var}(y_i|\boldsymbol{x}_i)$ getrennt voneinander spezifiziert werden. Auf diese Weise werden Varianz und Erwartungswert entkoppelt und man erreicht eine flexiblere Varianzschätzung. Somit können fehlerhafte Standardfehler aufgrund von Unter- oder Überdispersion vermieden werden. Im Vergleich zum später vorgestellten GEE-Ansatz für longitudinale Daten (vgl. Abschnitt 3.2.4), geht man hier von der korrekten Spezifikation beider Momente aus.

Für Zähldaten bietet sich in diesem Zusammenhang das sogenannte *Quasi-Poisson-Modell* an. Der Erwartungswert $\mathbb{E}(y_i|\boldsymbol{x}_i)$ wird dabei wie im regulären Poisson-GLM angenommen. Die zusätzliche Flexibilität bei der Varianzspezifikation

$$\text{Var}(y_i|\boldsymbol{x}_i) = \varphi_i V(\mu_i)$$

erreicht man im Vergleich zum Poisson-GLM dadurch, dass die Fixierung von φ_i auf 1 aufgehoben wird. Hier betrachtet man φ als unbekannten Parameter ($\varphi > 0$), der basierend auf allen Beobachtungen (in der Regel nimmt man $\varphi_i \equiv \varphi$ an) aus den Daten geschätzt wird. Zur Bestimmung von $\hat{\varphi}$ wird in der Regel ein gewöhnlicher Momentenschätzer verwendet. Alternativ wäre es auch denkbar, die Fixierung von φ auf 1 aufrecht zu erhalten, stattdessen aber die GLM-Annahme $V(\mu_i) = \mu_i$ aufzuheben und $V(\mu_i)$ durch eine geeignetere Varianzfunktion zu ersetzen.

Diese Vorgehensweise ermöglicht es auch, die Stärke der auftretenden Über- oder Unterdispersion zu beurteilen. Werte von $\hat{\varphi} \approx 1$ sprechen dafür, dass die Varianzannahme des Poisson-Modells korrekt ist, größere oder kleinere Werte erfordern den Einsatz alternativer Modelle.

Ein Nachteil des Quasi-Likelihood-Verfahrens ist jedoch, dass keine Likelihood im ursprünglichen Sinne existiert. Aus diesem Grund können Variablenselektionsverfahren, die auf likelihood-basierten Goodness-of-Fit-Kriterien beruhen, nicht verwendet werden. Zu diesen Kriterien gehört beispielsweise das später vorgestellte Akaike Informationskriterium (AIC, vgl. Abschnitt 3.4). Eine mögliche Lösung dieses Problems stellt die Verwendung der von Wedderburn (1974) entwickelten Quasi-Likelihood-Funktion dar. Darauf aufbauend können Likelihood-basierte Größen, wie z. B. die Quasi-Devianz (vgl. Chiou und Müller (1998)) oder das Quasi-AIC (vgl. Simonoff (2003)), berechnet werden. Das Quasi-Poisson-Modell kann ebenso wie die gewöhnlichen GLMs mithilfe der R-Funktion „glm“ gefittet werden.

Für einen Vergleich von Negativ-Binomial- und Quasi-Poisson-Modellen bei Daten mit Überdispersion sei auf den Artikel von Ver Hoef und Boveng (2007) verwiesen.

Zusammenfassend lässt sich festhalten, dass das loglineare Modell von allen Zähldaten-Modellen die einfachste Struktur besitzt und das Problem der Unter- oder Über-

dispersion keine Rolle spielt. Einzige Voraussetzung für einen sinnvollen Einsatz ist, dass die Fallzahlen groß genug sind, um die Normalverteilungsannahme für die logarithmierten Fallzahlen zu rechtfertigen. Diese Bedingung scheint bei allen Datenquellen, abgesehen von den nach Landkreis und Alter aufgeschlüsselten Call-Center-Daten, erfüllt zu sein. Die Call-Center-Daten für München wurden mithilfe von Quasi-Poisson- und Negativ-Binomial-Modellen auf das Auftreten und die Stärke von Über- oder Unterdispersion untersucht. Die Überdispersion erwies sich in diesem Fall als so gering ($\hat{\varphi} = 1.30$ bzw. $\hat{\nu} = 12.96$), dass für die Variablenselektion die näherungsweise Verwendung eines Poisson-Modells gerechtfertigt erschien. Der Einsatz von Poisson-GLMs bei der Variablenselektion begründet sich neben der Instabilität der Schätzung im Negativ-Binomial-Modell und der Nichtexistenz der klassischen Likelihood in Quasi-Poisson-Modellen auch in der verhältnismäßig einfachen Struktur von Poisson-Modellen. Für die bayernweiten Trainingsmodelle basierend auf den Call-Center-Daten im Jahr 2006 (vgl. Abschnitt 5.1) trat hingegen eine leichte Unterdispersion auf ($\hat{\varphi} \approx 0.55$ je nach verwendeter Kovariablenkombination). Aus diesem Grund wurden hier die Quasi-Poisson-Modelle bevorzugt.

Bei allen Modellen in dieser Arbeit wurden verschiedene Offset-Terme verwendet. Ein Offset ist eine Kovariable, deren Realisierungen unmittelbar in den linearen Prädiktor miteingehen. Dies geschieht etwa bei der Poisson-Regression, um unterschiedliche Ausgangsbedingungen für die einzelnen Fallzahlen y_i zu berücksichtigen. So ist beispielsweise in der Pilotregion München die tägliche Anzahl von KVB-Call-Center-Anrufen ≤ 20 Jahre nicht direkt mit der Anzahl in der Altersgruppe 21 bis 40 Jahre zu vergleichen, da sich die Einwohnerzahlen in den jeweiligen Altersgruppen stark unterscheiden. In diesem konkreten Fall gab es am 31.12.2007 231323 Einwohner in Altersgruppe 1 und 441385 in Altersgruppe 2. Der zur Offset-Kovariable zugehörige Regressionskoeffizient β wird nicht geschätzt, sondern a priori auf 1 festgelegt. Durch die Miteinbeziehung des Offsets werden die Parameterschätzer der übrigen Kovariablen bezüglich dieser Größe korrigiert. Bei der Bestimmung der Schätzung $\hat{y}_i$ bzw. der Prognose y_i^* wird der Offset allerdings nicht miteingerechnet. Neben den (geschlechts- und) altersspezifischen Einwohnerzahlen zum 31.12.2007 ($x_{\text{inhabitants}}$) wurde bei der bayernweiten Modellierung der Arztbesuche auch die Arztdichte im jeweiligen Landkreis miteinbezogen (x_{doctors}). Auch der im bayesianischen Trainingsmodell geschätzte, strukturelle räumliche Effekt (x_{district}) fließt als Offset in die laufende Prädiktion ein (vgl. dazu die Abschnitte 3.5 und 5.1). Bei Verwendung des log-Links werden die Offset-Werte in der Regel analog zu den Zielvariablenwerten logarithmiert, damit beide Größen auf der gleichen Skala bleiben.

Als nächster Schritt werden an dieser Stelle erste Modellierungsergebnisse basierend auf der Pilotregion München und dem über alle Landkreise aggregierten Datensatz für Gesamtbayern zusammengefasst. Die hier vorgestellten Modelle beinhalten noch keine verzögerten Kovariableneffekte, jedoch bereits altersspezifische Zeittrends zur Entkorrelierung der $y_i|\boldsymbol{x}_i$ (vgl. Abschnitt 3.3) und Cutpoint-Variablen für Luftfeuch-

tigkeit, Luftdruck und Temperatur (vgl. Abschnitt 3.4). Auf die geschätzten Koeffizienten für Zeittrends und Cutpoint-Variablen wird in den zugehörigen Abschnitten eingegangen. Der Übersichtlichkeit halber wurden diese in den hier gezeigten Schätzertabellen ausgelassen.

Zunächst wird das loglineare Modell für die über alle bayerischen Landkreise aggregierten Anrufe beim KVB-Call-Center (Reduktionsverfahren b)) betrachtet. Die dazugehörige Modellgleichung für den Erwartungswert $\mu_{it} = \mathbb{E}(y_{it}|\boldsymbol{x}_{it})$ kann Abbildung G.7 entnommen werden. Um die grundlegende Modellannahme der Normalität der $y_i|\boldsymbol{x}_i$ zu überprüfen, wurde ein Histogramm und ein sogenannter Normal-Quantil-Plot der Residuen $\hat{\varepsilon}_i$ erstellt (vgl. Abbildung 3.1). Das Histogramm stellt eine Dichteschätzung für die Verteilung der beobachteten Residuen dar. Dazu wurden die absoluten Häufigkeiten für jedes Intervall der Breite 0.1 so standardisiert, dass die Summe der Flächen aller Balken 1 ergibt. Die zum Vergleich eingezeichnete rote Linie stellt die Dichte der theoretisch angenommenen Normalverteilung $\mathrm{N}(0, \sigma^2)$ für die Residuen dar. Anstelle der unbekannten Residualvarianz σ^2 wurde die Modellschätzung $\hat{\sigma}^2$ eingesetzt. Der Plot zeigt eine gute Übereinstimmung zwischen der empirischen Verteilung der $y_i|\boldsymbol{x}_i$, repräsentiert durch die Residuen, und der theoretischen Verteilungsannahme des Modells. Bei einem Normal-Quantil-Plot werden die theoretischen Quantile einer Standardnormalverteilung gegen die Stichproben-Quantile der beobachteten Residuen aufgetragen. Weichen die Punkte stark von der rot eingezeichneten Normal-Quantil-Linie ab, so ist die Erfüllung der Normalverteilungsannahme für die $y_i|\boldsymbol{x}_i$ fraglich. Im vorliegenden Fall erscheint diese Annahme jedoch zumindest näherungsweise erfüllt zu sein.

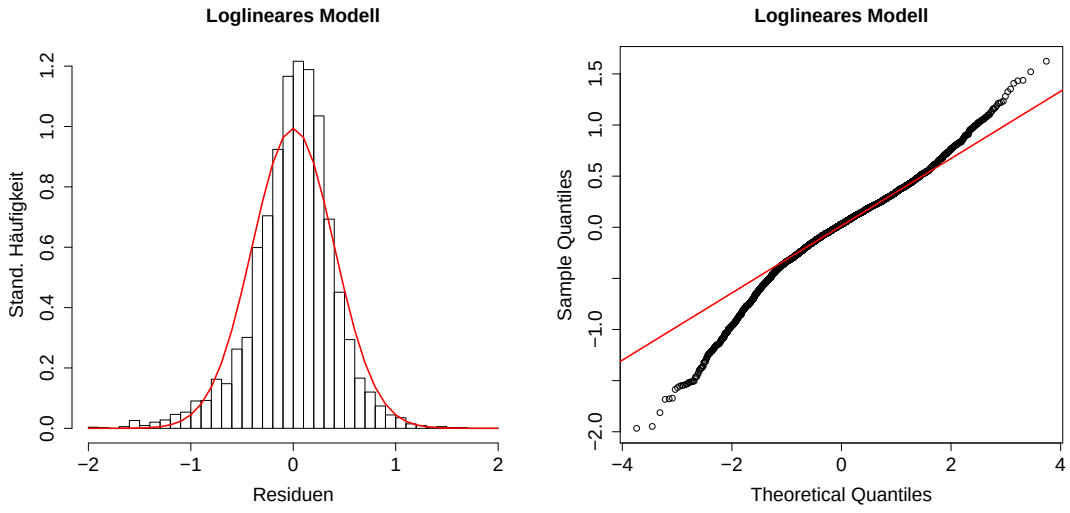


Abbildung 3.1: Histogramm und Normal-Quantil-Plot für die KVB-Call-Center-Daten aggregiert über ganz Bayern

Tabelle G.6 zeigt die geschätzten Regressionskoeffizienten $\hat{\boldsymbol{\beta}}$ des verwendeten loglinearen Modells (roh und exponentiell transformiert), zusammen mit den Standardfehlern $\text{se}(\hat{\beta}_r)$ der jeweiligen Komponenten $\hat{\beta}_r$. Diese ergeben sich durch

$$\text{se}(\hat{\beta}_r) = \sqrt{\widehat{\text{Var}}(\hat{\beta}_r)}$$

und bilden die Grundlage zur Beurteilung der Signifikanz des Kovariableneffekts. Letztere erfolgt basierend auf Signifikanztests mit den Hypothesen

$$H_0: \ \beta_r = 0 \quad \text{vs.} \quad H_1: \ \beta_r \neq 0.$$

Der Test besitzt die Teststatistik $T_r = \hat{\beta}/\,\text{se}(\hat{\beta}_r)$, die asymptotisch unter Gültigkeit von H_0 einer Standardnormalverteilung folgt. Anhand der T_r lässt sich die Stärke der verschiedenen Effekte vergleichen, da alle Kovariablen durch die vorgenommene Standardisierung auf die gleiche Skala gebracht werden. Aufgrund der ausreichend großen Fallzahl n der vorliegenden Datensätze kann man bei allen in dieser Arbeit verwendeten frequentistischen Modellen davon ausgehen, dass die asymptotische Verteilungsannahme für die T_r näherungsweise erfüllt ist. Basierend darauf lassen sich dann p-Werte und Konfidenzintervalle für die einzelnen β_r angeben, auf deren Basis dann (äquivalent) eine Signifikanzaussage getroffen werden kann. Als Signifikanzniveau wird durchgängig der Standardwert $\alpha = 0.05$ vorgegeben. Der p-Wert p_r für die r-te Kovariable errechnet sich aus

$$p_r = 2(1 - \Phi(|T_r|)),$$

wobei Φ die Verteilungsfunktion der Standardnormalverteilung darstellt. p-Werte kleiner als 0.05 sprechen für einen signifikanten Effekt der Kovariable auf dem vorgegebenen Signifikanzniveau α. Bei p-Werten größer 0.05 kann solch ein Effekt nicht nachgewiesen werden. Die 95%-Konfidenzintervalle für β_r ergeben sich zu:

$$\left[\hat{l}_\beta = \hat{\beta}_r - z_{0.975} \cdot \text{se}(\hat{\beta}_r); \hat{u}_\beta = \hat{\beta}_r + z_{0.975} \cdot \text{se}(\hat{\beta}_r)\right],$$

wobei $z_{0.975}$ das 97.5%-Quantil der Standardnormalverteilung bezeichnet. Ein signifikanter Effekt liegt dann vor, falls die 0 nicht im Konfidenzintervall liegt. Teststatistiken, p-Werte und 95%-Konfidenzintervalle sind ebenfalls in Tabelle G.6 dargestellt.

Die geschätzten Parameter selbst geben aufgrund ihres Vorzeichens Aufschluss über die Richtung des potenziell vorliegenden Effekts und können aufgrund der durchgängigen Verwendung des log-Links einheitlich interpretiert werden. Die Interpretation kann dabei direkt auf der log-Skala der Zielvariable oder nach exponentieller Transformation der $\hat{\beta}_r$ auf Ebene der unlogarithmierten Fallzahlen erfolgen. Die exponentiell transformierten Koeffizienten stellen eine Schätzung für das relative Risiko dar. Dieses kann bei Zähldaten als Fallzahlverhältnis zwischen 2 Subpopulationen betrachtet werden, die sich nur durch den Wert der betrachteten Kovariable unterscheiden (vgl. dazu die folgenden Interpretationsbeispiele). Es ist grundsätzlich bei

der Interpretation der Parameter zu beachten, dass Aussagen über eine Kovariable x_r nur unter der Voraussetzung gelten, dass alle anderen Kovariablen „festgehalten“ werden. In diesem und den folgenden Abschnitten wird häufig der Begriff „positiver/negativer Effekt“ verwendet. Diese Bezeichung bezieht sich ausschließlich auf das Vorzeichen des geschätzten Parameters und ist im Sinne von gleich-/gegensinnig zu verstehen, besitzt also keinerlei inhaltliche Wertung.

Interpretationsbeispiele für $\hat{\beta}$ (vgl. Tabelle G.6):

- x **stetig**:

  Nimmt die Windgeschwindigkeit $x_{\text{windspd.max}}$ um eine Einheit (in diesem Fall m/s) zu, so nimmt die logarithmierte Anruferzahl um den absoluten Wert $-\beta_{\text{windspd.max}} = 0.0071$ ab, die Anruferzahl selbst nimmt um den multiplikativen Faktor $\exp(\beta_{\text{windspd.max}}) = 0.9930$ ab. Das bedeutet, es sind in diesem Fall 0.70% weniger Fälle zu erwarten. Bei manchen stetigen Kovariablen macht es allerdings keinen Sinn die Zunahme um eine Einheit zu betrachten, da dieser Schritt entweder zu klein ist, um eine feststellbare Änderung der Zielvariable zu beobachten, z. B. bei $x_{\text{sshf.ave}}$, oder zu groß, wenn sich der gesamte Wertebereich der Kovariable nur über einen Bereich < 1 erstreckt, z. B. bei $x_{\text{lsp.max}}$ (vgl. Tabelle 2.3). In diesen Fällen muss die Schrittweite bei der Interpretation entsprechend angepasst werden. Die Formulierung lautet dann beispielsweise: Nimmt die Oberflächenwärmeleitung um 100000 Einheiten zu, so nimmt die Fallzahl um den multiplikativen Faktor $\exp(100000\beta_{\text{lcc.ave}}) = 1.0152$ zu. Die in dieser Arbeit gezeigten Schätzertabellen beinhalten nicht den ursprünglichen Schätzwert, sondern den mit der zur Interpretation verwendeten Schrittweite multiplizierten Schätzwert. Der jeweilige Multiplikator ist jeweils in Klammern hinter dem Namen der Kovariable angegeben. Konsistenterweise wurden die entsprechenden Standardfehler sowie die Grenzen der Konfidenzintervalle derselben Transformation unterzogen wie der Schätzwert.

- x **kategorial und referenzcodiert**:

  An Schulferientagen nimmt im Vergleich zu allen übrigen Tagen die logarithmierte Anzahl der Anrufe um den Wert $\beta_{\text{school=ja}} = 0.0100$ zu und die Anruferzahl selbst um den multiplikativen Faktor $\exp(\beta_{\text{school=ja}}) = 1.1051$. An Schulferientagen sind also um 10.51% mehr Fälle zu erwarten als an den übrigen Tagen des Jahres.

- x **kategorial und effektcodiert**:

  Bei vorherrschender Windrichtung Ost nimmt die logarithmierte Fallzahl im Vergleich zum Mittelwert über alle Windrichtungen um den absoluten Wert $-\beta_{\text{c.wdir=O}} = 0.0512$ ab, die Anruferzahl selbst nimmt um den multiplikativen Faktor $\exp(\beta_{\text{c.wdir=O}}) = 0.9501$ ab. Somit ist an Tagen mit vorherrschender Windrichtung Ost die erwartete Fallzahl im Vergleich zum Mittel über alle

Tage um 4.99% geringer.

Anmerkung:

Bei kategorialen Kovariablen mit C Kategorien gehen bei der Effektcodierung wie bei der Referenzcodierung aus Gründen der Identifizierbarkeit der Parameter nur $C-1$ Dummy-Variablen in die Modellgleichung ein. Demenstsprechend werden auch nur $C-1$ Regressionskoeffizienten $\beta_1, \ldots, \beta_{C-1}$ geschätzt. Der Koeffizient für die fehlende Kategorie β_C wird bei Referenzcodierung auf 0 gesetzt. Bei Effektcodierung gilt für den fehlenden Parameter aufgrund der Restriktion $\sum_{c=1}^{C} \beta_c = 0$:

$$\beta_C = -\sum_{c=1}^{C-1} \beta_c.$$

Aufgrund dieser Tatsache kann ein Standardfehler für den geschätzten Koeffizienten $\hat{\beta}_C$ wie folgt berechnet werden:

$$\text{se}(\hat{\beta}_C) = \sqrt{\sum_{c=1}^{C-1} \widehat{\text{Var}}(\hat{\beta}_c) + 2 \sum_{c=1}^{C-1} \sum_{c \neq \tilde{c}} \widehat{\text{Cov}}(\hat{\beta}_c, \hat{\beta}_{\tilde{c}})}.$$

Die so berechneten $\hat{\beta}_C$ und $\text{se}(\hat{\beta}_C)$ sind in allen gezeigten Schätzertabellen jeweils in Klammern angegeben.

Der geschätzte Intercept wird grundsätzlich in Schätzertabellen aufgeführt, besitzt aber, was die Interpretation der Effekte betrifft, keinerlei inhaltliche Bedeutung. Gleichwohl ist er ein wesentlicher Bestandteil des linearen Prädiktors und bei der Schätzung von $\hat{y}_i$ bzw. der Prognose y_i^* zu beachten.

Betrachtet man die Teststatistiken und p-Werte in Tabelle G.6, so fällt auf, dass die administrativen Kovariablen, wie zu erwarten, im Allgemeinen einen weit größeren Einfluss auf die Anzahl der Call-Center-Anrufe in Bayern besitzen als die Meteorologie- und Luftqualitätsparameter. Von den meteorologischen Parametern besitzen der Bewölkungsgrad niedriger Bewölkung, die Tagesrange des Luftdrucks sowie die Windrichtung einen signifikanten Einfluss auf die Fallzahl. Die Zunahme niedriger Bewölkung und eine höhere Tagesrange des Luftdrucks führen dem Modell zufolge zu einer Abnahme der Anruferzahl (-0.96% bei 10% mehr niedriger Bewölkung und -0.57% bei Erhöhung des logarithmierten Oberflächendrucks um 0.001). Bei vorherrschenden Windrichtungen Nordost und Ost ist mit weniger Anrufen als im Durchschnitt (-5.10% und -4.99%) zu rechnen, bei der Windrichtung Südost dagegen mit überdurchschnittlich vielen Anrufen ($+6.18\%$). Bei den Luftqualitätsparametern führt die Erhöhung der Feinstaubkonzentration um $1\mu g/m^3$ zu einer signifikanten Zunahme der Anrufe um 0.22%. Eine Erhöhung der NO_2-Konzentration dagegen führt zu einer signifikant kleineren Fallzahl (-0.31% pro Zunahme um $1\mu g/m^3$).

Solche aus medizinischer Sicht unplausibel erscheinenden Aussagen treten im Verlaufe der Analyse der Kovariableneffekte häufiger auf und sind vermutlich unter anderem auf die aufwendige Korrelationsstruktur innerhalb der Meteorologie- und Luftqualitätsparameter zurückzuführen (vgl. Tabellen G.3 und G.4). Um diese Korrelationsstrukturen bei der Modellierung zu berücksichtigen, wäre es erforderlich, Interaktionsterme zwischen den einzelnen Kovariablen in die Modellgleichung mitaufzunehmen. Rechnet man nur die stetigen Kovariablen (ohne Tagesranges, Bruchpunkte und Lag-Effekte) und berücksichtigt keine Interaktionen höherer Ordnung, ergäben sich jedoch bereits 105 Zweier-Interaktionsterme. Aufgrund der ohnehin bereits großen Menge an Kovariablen, war es nicht möglich, solche Interaktionseffekte in die Modellierung einzubeziehen. Des Weiteren bestehen möglicherweise auch Interaktionen zwischen Kovariablen und bisher nicht berücksichtigten Einflussgrößen auf die Asthma- und COPD-Morbidität, wie z. B. Pollenflug. Die getroffenen Aussagen sind daher generell vor diesem Hintergrund zu bewerten.

Die geschätzten Parameter für die administrativen Kovariablen spiegeln tendenziell die Erkenntnisse wider, die bereits die Mittelwertbetrachtung in Abbildung G.6 lieferte. Die Stärke der Effekte, beispielsweise dass laut Modell an Feiertagen ca. 2.6 mal so viele Fälle auftreten wie an allen anderen Tagen, ändert sich jedoch im Vergleich zur deskriptiven Analyse der Rohdaten, da im Modell auch die Effekte aller anderen Kovariablen berücksichtigt werden.

Für die KVB-Call-Center-Daten in der Pilotregion München wurden zunächst ein Quasi-Poisson- und ein Negativ-Binomial-Modell gefittet, um zu überprüfen, ob Unter- oder Überdispersion auftritt und um die Modellparameter und deren Standardfehler mit den entsprechenden Schätzwerten eines Poisson-GLMs vergleichen zu können. Tabelle G.7 stellt die geschätzten Regressionsparameter aller drei Modelle zusammen mit Standardfehlern und p-Werten gegenüber. Poisson- und Quasi-Poisson-Modell besitzen per Definition die gleichen Schätzwerte $\hat{\boldsymbol{\beta}}$, die Parameter des Negativ-Binomial-Modells weichen nur in geringem Maße (ab der 2. bis 3. Nachkommastelle) davon ab. Da nur eine schwache Überdispersion auftrat ($\hat{\varphi} = 1.30$ im Quasi-Poisson-Modell bzw. $\hat{\nu} = 12.96$ im Negativ-Binomial-Modell), unterscheiden sich die Standardfehler der Modelle nur unwesentlich und haben bei keiner der betrachteten Kovariablen einen Einfluss auf die Beurteilung der Signifikanz. Diese Beobachtung rechtfertigt den Einsatz von einfachen Poisson-Modellen bei der Variablenselektion (vgl. Abschnitte 3.7 und 3.8). Bei der Betrachtung der p-Werte ist festzustellen, dass hier aufgrund der Beschränkung auf die Pilotregion-München offensichtlich nicht ausreichend Power vorhanden ist, um signifikante Meteorologie- oder Luftqualitätseffekte nachzuweisen.

Vergleicht man die zwei loglinearen Modelle für die Anzahl der Arztbesuche in der Pilotregion München und aggregiert über ganz Bayern (vgl. Tabelle G.8), ist zu erkennen, dass die auf verschiedenen Reduktionsverfahren basierenden Modelle teilweise unterschiedliche und in einzelnen Punkten sogar widersprüchliche Aussagen zu

den Meteorologie- und Luftqualitätseffekten liefern. Im auf die Pilotregion München eingeschränkten Modell besitzt beispielsweise Feinstaub ($x_{\text{PM10.q95}}$) einen positiven Effekt auf die Anzahl der Arztbesuche, während Feinstaub im aggregierten Modell für Gesamtbayern einen negativen Effekt aufweist.

Generell gilt es zu beachten, dass sich die Interpretation der geschätzten Parameter stark auf den Datensatz bezieht, der dem Modell zugrunde liegt. Eine Verallgemeinerung der Aussagen zu den Kovariableneffekten ist daher auf Basis einzelner Modelle nicht möglich. Im zusammenfassenden Abschnitt 3.6 wird aus diesem Grund der Versuch unternommen, anhand der Gegenüberstellung der Ergebnisse aller Modelle basierend auf den reduzierten Datensätzen gemeinsame Tendenzen zu identifizieren, denen ein tatsächlicher Effekt zugrunde liegt.

Im Modell basierend auf der Landeshauptstadt München zeigen der Bewölkungsgrad niedriger Bewölkung (-0.72% pro 10% mehr niedriger Bewölkung) und die Feuchtigkeit ($+2.37\%$ pro Zunahme um $0.001 kg_{\text{Wasser}}/kg_{\text{Luft}}$) einen signifikant negativen Effekt auf die Anzahl der Arztbesuche. Bei vorherrschender Windrichtung Südwest gibt es 3.28% weniger Fälle als im Durchschnitt. Neben Feinstaub ($+0.09\%$ pro Zunahme um $1\mu g/m^3$) verursacht hier auch Stickstoffdioxid ($+0.32\%$ pro Zunahme um $1\mu g/m^3$) eine signifikante Zunahme der Fallzahl. Dagegen besitzen Ozon (-0.20% pro Zunahme um $1\mu g/m^3$) und Kohlenstoffmonoxid (-12.99% pro Zunahme um $1\mu g/m^3$) einen negativen Effekt auf die Fallzahl.

Das bayernweite Aggregationsmodell liefert hingegen folgende signifikante Effekte: Eine Zunahme der mittelhohen Bewölkung ($+0.69\%$ pro 10% mehr mittelhoher Bewölkung), des logarithmierten Oberflächendrucks ($+0.36\%$ pro Erhöhung des logarithmierten Oberflächendrucks um 0.001), der Zwei-Meter-Temperatur ($+0.96\%$ pro $1°C$) und der Windgeschwindigkeit ($+1.98\%$ pro Zunahme um $1m/s$) führt zu einer Steigerung der Fallzahl. Demgegenüber bewirkt eine höhere Tagesrange der spezifischen Luftfeuchtigkeit ($+1.86\%$ pro Zunahme um $0.001 kg_{\text{Wasser}}/kg_{\text{Luft}}$) eine Abnahme der Fallzahl. Bei den Windrichtungen Nordwest (-3.20%) und West (-4.58%) liegt die Fallzahl unter dem Durchschnitt, bei den Winrichtungen Süd ($+4.77\%$) und Ost ($+2.21\%$) wächst die Fallzahl dagegen überdurchschnittlich an. Schließlich haben Feinstaub (-0.09% pro Zunahme um $1\mu g/m^3$) und Ozon (-0.26% pro Zunahme um $1\mu g/m^3$) einen negativen Effekt auf die Fallzahl, während Stickstoffdioxid ($+0.39\%$ pro Zunahme um $1\mu g/m^3$) eine höhere Fallzahl begünstigt.

Die administrativen Effekte beider Modelle stehen im Gegensatz zu den Effekten von Wetter und Luftqualität in guter Übereinstimmung miteinander. Wie schon bei den KVB-Call-Center-Daten zu beobachten, ist der Einfluss dieser Kovariablen tendenziell deutlich größer als der Einfluss von Umweltfaktoren. Grundsätzlich bestätigen die Schätzwerte für die administrativen Effekte auch hier das Bild, das sich bereits beim deskriptiven Mittelwertvergleich (vgl. Abbildung G.6) ergab. Beispielsweise beträgt die Anzahl der Arztbesuche an Sonntagen nur 10.90% (München) bzw. 8.34%

(Aggregation über ganz Bayern) vom Durschnittswert über alle Tage, an Schulferientagen 75.69% bzw. 82.19%. Interessant ist die Tatsache, dass beide Modelle einen deutlichen Rückgang der Fallzahl an Brückentagen ergeben (−43.00% bzw. −37.16%), während das entsprechende Balkendiagramm in Abbildung G.6 sogar eine Zunahme der Anzahl der Arztbesuche zeigt. Hier wird wiederum die Bereinigung des Effekts einzelner Kovariablen vom Einfluss aller übrigen Größen deutlich.

Bei der Interpretation der Parameter für Alter und Geschlecht, gilt es die Interaktion zwischen beiden Kovariablen zu beachten. Bei den Männern (Referenzkategorie von x_{sex}) liegt beispielsweise die Anzahl der Arztbesuche in Altersgruppe 3 (41 bis 60 Jahre) 11.94% über dem Durchschnitt für alle Männer (München) bzw. um 18.48% darunter (Aggregation über ganz Bayern). Diese Differenz erklärt sich wohl auch durch die unterschiedliche Altersstruktur in München im Vergleich zu Gesamtbayern. Frauen weisen in beiden Modellen über alle Altersklassen hinweg mehr Arztbesuche auf (15.70% bzw. 2.05%) als Männer. In den einzelnen Altersgruppen können die Exponenten der Schätzer nicht direkt interpretiert werden: Die Anzahl der Arztbesuche von Frauen in Altersgruppe 3 ist zum Beispiel um den multiplikativen Faktor $\exp(\beta_{\text{sex=2}} + \beta_{\text{sex=2,age=3}}) = 1.4048$ bzw. 1.3278 höher als für Männer.

3.2.2 Zero-Inflated-Poisson-Modell (ZIP)

Das sogenannte Zero-Inflated-Poisson- (ZIP) oder Exzess-Nullen-Modell stellt eine zusätzliche Erweiterung des klassischen Poisson-GLMs dar und ist ebenfalls für die Modellierung von Zähldaten mit Überdispersion geeignet, insbesondere wenn die Zielvariable häufig den Wert 0 annimmt. Die folgenden und weitere Details zur Theorie und Schätzung von ZIP-Modellen können den Büchern von Winkelmann (2008) und Cameron und Trivedi (1998) entnommen werden.

Die grundlegende Annahme des ZIP-Modells besteht darin, dass sich die Gesamtzahl aller Beobachtungen $\boldsymbol{y}$ in zwei Gruppen, repräsentiert durch die binäre Zufallsvariable c_i, aufteilen lässt: Die eine Gruppe beinhaltetet die Fallzahlen, die einer gewöhnlichen Poisson-Verteilung mit Erwartungswert μ_i folgen ($c_i = 1$), die andere Gruppe umfasst die Beobachtungen, für die mit Wahrscheinlichkeit 1 $y_i = 0$ gilt ($c_i = 0$). Das heißt

$$\begin{aligned} &y_i|\boldsymbol{x}_i, c_i = 1 \sim \text{Po}(\mu_i), \\ &\mathbb{P}(y_i = y|\boldsymbol{x}_i, c_i = 0) = \begin{cases} 1 & y = 0 \\ 0 & y > 0 \end{cases}. \end{aligned}$$

Insgesamt nimmt man für die Verteilung der $y_i|\boldsymbol{x}_i$ die Mischverteilung

$$\mathbb{P}(y_i = y|\boldsymbol{x}_i) = \mathbb{P}(y_i = y|\boldsymbol{x}_i, c_i = 1) \cdot \pi_i + \mathbb{P}(y_i = y|\boldsymbol{x}_i, c_i = 0) \cdot (1 - \pi_i) \qquad (7)$$

mit unbekannter Mischungskomponente $\pi_i := \mathbb{P}(c_i = 1|\boldsymbol{z}_i)$ an. Dabei enthält $\boldsymbol{z}_i$ diejenigen Komponenten von $\boldsymbol{x}_i$, die einen potenziellen Einfluss auf die Gruppenzugehörigkeit der i-ten Beobachtung besitzen. Erwartungswert und Varianz von $y_i|\boldsymbol{x}_i$ ergeben sich mit dem Satz vom iterierten Erwartungswert und dem Satz von der iterierten Varianz (vgl. Grimmett und Stirzaker (2001)) zu

$$\begin{aligned} \mathbb{E}(y_i|\boldsymbol{x}_i) &= \pi_i\mu_i \quad \text{und} \\ \mathrm{Var}(y_i|\boldsymbol{x}_i) &= \pi_i\mu_i(1+\mu_i(1-\pi_i)). \end{aligned}$$

Das klassische Poisson-Modell kann als Spezialfall des ZIP-Modells mit $\pi_i = 1$ aufgefasst werden. Bei der Betrachtung der Modellvarianz lässt sich feststellen, dass Zero-Inflated-Poisson-Modelle nicht für Daten mit Unterdispersion geeignet sind, da aufgrund von $\pi_i < 1$ gilt: $\mathrm{Var}(y_i|\boldsymbol{x}_i) \geq \pi_i\mu_i = \mathbb{E}(y_i|\boldsymbol{x}_i)$.

Das Gesamtmodell setzt sich aus einem binären (z. B. logistischen) Regressionsmodell für π_i und einem loglinearen Poisson-Modell für μ_i zusammen, wobei π_i und μ_i separat modelliert werden:

$$\begin{aligned} \log\left(\frac{\pi_i}{1-\pi_i}\right) &= \boldsymbol{z}_i^\top\boldsymbol{\gamma}, \\ \log(\mu_i) &= \boldsymbol{x}_i^\top\boldsymbol{\beta}. \end{aligned}$$

Die Likelihood $L(\boldsymbol{\beta},\boldsymbol{\gamma})$ des Zero-Inflated-Poisson-Modells lässt sich aus der Wahrscheinlichkeitsfunktion der Mischverteilung (7) ableiten. Die Schätzung des Zero-Inflated-Poisson-Modells beruht wiederum auf der Maximierung der (Log-)Likelihood bzgl. $\boldsymbol{\beta}$ und $\boldsymbol{\gamma}$.

Generell sind ZIP-Modelle dann anwendbar, wenn $\boldsymbol{y}$ einen nicht zu vernachlässigenden Anteil an Nullen enthält. Bei den vorliegenden Datensätzen trifft dies insbesondere bei den auf die Landeshauptstadt München eingeschränkten KVB-Call-Center-Daten zu. Hier liegt der Anteil an Nullen von insgesamt $n = 5480$ Beobachtungen bei 25.97%. Allerdings ist zu beachten, dass ZIP-Modelle, genauso wie alle bisher vorgestellten Modelle, von der Unkorreliertheit der $y_i|\boldsymbol{x}_i$ ausgehen und daher grundsätzlich nicht zur Modellierung longitudinaler Daten mit serieller Korrelation geeignet sind. Durch die Miteinbeziehung altersspezifischer Zeitfunktionen (vgl. Abschnitt 3.3) konnte diese serielle Korrelation allerdings soweit reduziert werden, dass die Verwendung des ZIP-Modells für die Call-Center-Daten in der Pilotregion München durchaus gerechtfertigt erschien.

Im Folgenden werden die Ergebnisse des mit der R-Funktion „zeroinfl“ aus dem Paket „pscl“ gefitteten ZIP-Modells kurz zusammengefasst. Für die Modellierung der Wahrscheinlichkeit π_i wurde ausschließlich die designbedingte Kovariable x_{age} in den Kovariablenvektor $\boldsymbol{z}$ aufgenommen. Das logistische Regressionsmodell lautet

dann:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \gamma_{\text{Intercept}} + x_{\text{age=2},i}\gamma_{\text{age=2}} + \ldots + x_{\text{age=5},i}\gamma_{\text{age=5}}.$$

Die Wahrscheinlichkeiten π_i, dass die Fallzahlen in Altersgruppe i einer Poisson-Verteilung folgen, lassen sich dann wie folgt berechnen:

$$\pi_i = \frac{\exp(\gamma_{\text{Intercept}} + \gamma_{\text{age=i}})}{1 + \exp(\gamma_{\text{Intercept}} + \gamma_{\text{age=i}})}, \quad i = 1, \ldots, 5.$$

Die Schätzwerte für die altersspezifischen Wahrscheinlichkeiten π_i können folgender Tabelle entnommen werden:

i	p_i
1	(0.5572)
2	0.9282
3	0.9568
4	0.9974
5	0.9765

Der Wert für Altersgruppe 1 wurde aus der Restriktion $\sum_{i=1}^{5} \gamma_{\text{age=i}} = 0$ errechnet und ist aus diesem Grund in der Tabelle eingeklammert. Das logistische Modell schreibt der Kovariable Alter insgesamt einen signifikanten Effekt auf die Wahrscheinlichkeit π_i zu.

Tabelle G.9 zeigt die Schätzwerte $\boldsymbol{\beta}$ des ZIP-Modells für die KVB-Call-Center-Daten eingeschränkt auf die Pilotregion München zusammen mit Standardfehlern und p-Werten. Die Modellgleichung der Poisson-Modellkomponente entspricht der Modellgleichung des zuvor beschriebenen gewöhnlichen Poissonmodells (vgl. Abbildung G.7). Die Schätzwerte, Standardfehler und p-Werte unterscheiden sich nicht wesentlich von den Werten des Poisson-Modells (vgl. Tabelle G.7), an den Signifikanzaussagen ändert sich nichts. Dies liegt daran, dass die Schätzwerte für die Wahrscheinlichkeiten, dass die Fallzahlen einer Poisson-Verteilung folgen, zum großen Teil nahe bei 1 liegen und nur eine sehr schwache Überdispersion vorliegt (vgl. die Schätzungen der Überdispersionsparameter im Quasi-Poisson- und Negativ-Binomial-Modell). Die relativ gering eingeschätzte Wahrscheinlichkeit, dass die Fallzahlen in Altersgruppe 1 einer Poisson-Verteilung folgen (55.72%), spiegelt sich in der etwas stärkeren Abweichung des Schätzwerts $\hat{\beta}_{\text{age=1}} = -2.0873$ vom entsprechenden Schätzwert des Poisson-Modells ($\hat{\beta}_{\text{age=1}} = -2.5647$) wider.

Für das Trainingsmodell zur Prognose (vgl. Abschnitt 5.1) kommt das Zero-Inflated-Poisson-Modell aus zweierlei Gründen nicht infrage. Zum einen ist es mit der vorhandenen R-Funktion „zeroinfl“ nicht möglich, eine Penalisierung der Regressionsparameter durchzuführen. Diese Penalisierung ist jedoch erforderlich, um das in Abschnitt 4.4 vorgestellte Konzept der Penalized Distributed Lag Function zur Modellierung verzögerter Kovariableneffekte bei der Prognose umzusetzen. Zum anderen

tritt bei den auf die Beobachtungen des Jahres 2006 reduzierten Daten Unterdispersion auf, die wie beschrieben im Zero-Inflated-Poisson-Modell nicht berücksichtigt werden kann.

3.2.3 Regressionssplines und Generalisiertes Additives Modell (GAM)

Dieser Abschnitt befasst sich mit der Theorie von Regressionssplines und dem damit verbundenen Penalisierungskonzept. Die hier beschriebenen Techniken finden vor allem in den Abschnitten 3.3 und 4.4 Anwendung. Nähere Details zu diesem Thema finden sich in dem Buch von Fahrmeir und Tutz (2001).

Grundsätzlich geht man bei den bisher vorgestellten Modellen davon aus, dass der Effekt der stetigen Kovariablen auf den (mit der Link-Funktion transformierten Erwartungswert) linear ist. Das bedeutet der Effekt-Parameter β ist über den gesamten Wertebereich der Kovariable x hinweg konstant. Fasst man den Einfluss der Kovariable x als Funktion in x auf, ergibt sich die lineare Funktion $f(x) = x\beta$. Oft ist die funktionale Form von Kovariableneffekten jedoch weitaus komplexer. Eine einfache Möglichkeit, dies zu berücksichtigen, ist die Aufnahme polynomialer Terme in den linearen Prädiktor. Beispielsweise kann man zusätzlich zur Kovariable x noch einen quadratischen Term x^2 mit eigenem Regressionskoeffizient β_2 in den linearen Prädiktor aufnehmen. Dieser bleibt linear in den Parametern, jedoch wird der Kovariableneffekt nun durch ein quadratisches Polynom in x beschrieben ($f(x) = x\beta + x^2\beta_2$). Polynome besitzen jedoch die Problematik, dass die Schätzung der Funktion f an den Rändern des Wertebereichs verzerrt ist, da Polynome für $x \to \pm\infty$ gegen $+\infty$ oder $-\infty$ tendieren.

Einen weitaus flexibleren Ansatz zur Schätzung der funktionalen Form der Kovariableneffekte bieten Splinefunktionen. Dabei wird die Funktion $f(x)$ in Basisfunktionen entwickelt. Die Glattheit der Funktion $f(x)$ wird über die Anzahl der Basisfunktionen bzw. der zu den Basisfunktionen zugehörigen Knoten $\tau_1, \ldots, \tau_m$, die über den Wertebereich der Kovariable verteilt werden, gesteuert. Je mehr Basisfunktionen verwendet werden, desto genauer können lokale Änderungen in der Größe des Kovariableneffekts erfasst werden und desto rauher wird der Verlauf der resultierenden Funktion $f(x)$. Die Datentreue der Schätzung ist in diesem Fall sehr groß, jedoch wächst auch die Varianz der geschätzten Funktion. Verwendet man umgekehrt nur wenige Knoten, erhält man eine sehr glatte Funktion mit niedriger Varianz. Dabei werden allerdings möglicherweise wichtige Aspekte im Verlauf der unbekannten, wahren Funktion $f(x)$ übersehen und es entsteht ein Bias.

Häufig verwendete Basen sind die Truncated-Power-Series-Basis (TP-Basis), die B-Spline-Basis und die Radial-Basis. Die TP-Basis besteht aus einem polynomialen Anteil und einem Anteil von in den Knoten trunkierter Polynome. Die Basisfunk-

tionen lauten

$$\Phi_0 = 1, \Phi_1 = x, \ldots, \Phi_k = x^k, \Phi_{k+i} = (x-\tau_i)_+^k \quad \text{mit } (x-\tau_i)_+^k = \begin{cases} (x-\tau_i)^k & x \geq \tau_i \\ 0 & \text{sonst} \end{cases}$$

für $i = 1, \ldots, m$. $f(x)$ lässt sich dann in folgender Weise darstellen:

$$f(x) = \delta_0 + \delta_1 x + \ldots + \delta_k x^k + \sum_{i=1}^{m} \delta_{k+i}(x-\tau_i)_+^k.$$

Neben der Knotenwahl beeinflusst auch die Ordnung k der TP-Basis die Rauheit der resultierenden Funktion $f(x)$ nach dem Prinzip „je größer k, desto rauher die Funktion“. Die Dimension der Basis ist $k + m + 1$.

Eine für Regressionsmodelle numerisch stabilere Basis stellt die B-Spline-Basis dar. Eine B-Spline-Basis der Ordnung 1 (vom Grad 0) ist wie folgt definiert:

$$\Phi_{i,1}(x) = \begin{cases} 1 & x \in [\tau_i, \tau_{i+1}) \\ 0 & \text{sonst} \end{cases}.$$

B-Spline-Basen der Ordnung r (vom Grad $r-1$) werden rekursiv definiert:

$$\Phi_{i,r}(x) = \frac{x-\tau_i}{\tau_{i+r-1}-\tau_i} \cdot \Phi_{i,r-1}(x) + \frac{\tau_{i+r}-x}{\tau_{i+r}-\tau_{i+1}} \cdot \Phi_{i+1,r-1}(x) \quad (r \geq 2)$$

für $i = 1, \ldots, m$ (m ist die Dimension der Basis). Dementsprechend kann die Funktion $f(x)$ dargestellt werden als $\sum_{i=1}^{m} \Phi_{i,r}(x)\alpha_i$. Grundsätzlich gilt wiederum „je höher die Ordnung der Basis, desto rauher die Funktion“. In der Regel wählt man die Ordnung $r = 4$, um zwischen den Knoten kubische Polynome (Grad 3) zu erhalten. Für eine stabile Schätzung von $f(x)$ in den Randbereichen, wird häufig die Knotenmenge über $\min(x)$ und $\max(x)$ hinaus erweitert.

Die Basisfunktionen der Radial-Basis sind wie folgt definiert:

$$\Phi_i(x) = \exp\left(-\frac{(x-\tau_i)^2}{2h_i^2}\right)$$

für $i = 1, \ldots, m$. Die Basisfunktionen entsprechen also in den Knoten zentrierten Gauss-Kurven mit Varianz h_i^2. Die Glattheit der Funktion wird somit neben der Anzahl der Basisfunktionen durch h_i^2 gesteuert. $f(x)$ ergibt sich analog zur B-Spline-Basis.

Die Wahl der Basis ist im Vergleich zur Knotenwahl in der Regel nicht ausschlaggebend für eine gute Anpassung an die Daten. Die Funktion $f(x)$ bleibt bei allen verwendeten Basen linear in den Koeffizienten $\boldsymbol{\delta}$. $f(x)$ lässt sich somit ohne weiteres in den linearen Prädiktor der bisher vorgestellten Modelle integrieren, indem die Kovariable x zunächst mit den entsprechenden Basisfunktionen transformiert wird.

Die Interpretation der Parameterschätzer $\hat{\boldsymbol{\delta}}$ ist grafisch durch Darstellung der daraus resultierenden Funktion $\hat{f}(x)$ möglich. Durch die Berechnung von Konfidenzbändern wird auch eine optische Beurteilung möglich, in welchen Bereichen der Kovariableneffekt signifikant ist (mehr dazu in Abschnitt 3.3).

Generell ist das Ziel beim Einsatz von Splinefunktionen, einen möglichst geringen Bias bei der Schätzung von $f(x)$ zu erzielen, gleichzeitig aber eine möglichst glatte Funktion mit niedriger Varianz zu erhalten. Bei Verwendung der beschriebenen Basen kann dieser Ausgleich zwischen Glattheit und Datentreue manuell über die beschriebenen Steuerparameter erfolgen. In vielen Fällen ist es jedoch von Vorteil, diese Konfiguration automatisch bzw. datengesteuert vornehmen zu können. Diese Überlegung führt direkt zum Konzept der Penalisierung.

Zunächst wird an dieser Stelle die Idee der Penalisierung kurz anhand des linearen Modells $y_i = f(x_i) + \varepsilon_i$ erläutert. Die Schätzung der Funktion f beruht hier auf der Minimierung des Kleinste-Quadrate-Kriteriums (KQ-Kriterium) $\sum_{i=1}^{n}(y_i - f(x_i))^2$ bzgl. f. Bestraft man die Rauheit der geschätzten Funktion nicht, so wird das Kriterium minimal, wenn gilt $\hat{f}(x_i) = y_i$, also wenn die geschätzte Funktion durch jeden Datenpunkt läuft. Da dies eine extreme Überanpassung an die vorliegenden Daten bedeuten würde, kann man beispielsweise den folgenden Penalisierungsterm in das KQ-Kriterium einfügen:

$$\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \cdot \int (f''(x))^2 \, dx. \tag{8}$$

Der Penalisierungsterm besteht aus dem Penalisierungsparameter λ und der Fläche unter der quadrierten zweiten Ableitung der Funktion als Maß für die Rauheit von $f(x)$. Somit erreicht man den gewünschten Trade-Off zwischen Datentreue und Glattheit: Der erste Term wird umso kleiner, je geringer der Bias von $\hat{f}(x)$ ist, der zweite Term wird umso kleiner, je glatter die Funktion ist. λ ist für die Gewichtung der Terme zuständig: Für große Werte von λ wird entsprechend mehr Wert auf die Glattheit der Funktion gelegt, für kleine Werte von λ dagegen auf die Datentreue.

Die Minimierung des penalisierten KQ-Kriteriums (8) gelingt durch sogenannte natürliche kubische Splines, deren Knoten durch die Beobachtungen gegeben sind und die folgende Eigenschaften besitzen:

- $\hat{f}(x)$ ist ein kubisches Polynom zwischen den Knoten,
- $\hat{f}''(x)$ ist stetig in den Knoten,
- $\hat{f}''(x_{\min}) = \hat{f}''(x_{\max}) = 0$, das heißt am Rand des Wertebereichs der Kovariable x verschwindet die zweite Ableitung und die Kurve läuft linear weiter.

Letztere Eigenschaft ist vor allem für die Prognose von Werten außerhalb des ursprünglichen Datenbereichs attraktiv. Aus diesem Grund werden die in Abschnitt 3.3

näher beschriebenen, altersspezifischen Zeitfunktionen im Hinblick auf die Prädiktion an Folgetagen nach den Regeln von natürlichen kubischen Splines konstruiert. Die Knoten werden hier allerdings nicht durch die Beobachtungen definiert, sondern nach inhaltlichen Gesichtspunkten gesetzt. Die Verwendung einer B-Spline-Basis zur Konstruktion der natürlichen kubischen Splines ermöglicht die Integration in alle bisher vorgestellten Modelle.

Ein Glättungsverfahren, das in dieser Arbeit ebenfalls zum Einsatz kommt, sind sogenannte P-Splines (vgl. Eilers und Marx (1996)). Diese beruhen auf der Entwicklung der Funktion f in B-Spline-Basisfunktionen Φ (andere Basen sind ebenfalls möglich). Zusätzlich werden große Differenzen zwischen den Parametern benachbarter Knoten penalisiert, um so die Rauheit der Funktion f zu kontrollieren. Im linearen Modell (mit einer Kovariable x) lautet das penalisierte KQ-Kriterium wie folgt:

$$\sum_{i=1}^{n}\sum_{j=1}^{m}(y_i - \Phi_j(x_i)\delta_j)^2 + \lambda \cdot \sum_{d+1}^{m}(\Delta^d\delta_j)^2 \quad \rightarrow \quad \min_{\boldsymbol{\delta}}.$$

Δ^d bezeichnet darin Differenzen der Ordnung d:

$$\begin{aligned}\Delta^1\delta_j &= \delta_j - \delta_{j-1}\\ \Delta^2\delta_j &= \Delta^1(\Delta^1\delta_j) = \delta_j - 2\delta_{j-1} + \delta_{j-2}\\ &\vdots\end{aligned}$$

Für $d = 1$ werden die Differenzen in den Koeffizienten unmittelbar benachbarter Knoten bestraft, für $d = 2$ die Differenzen der Parameter von Knoten, die einen Zwischenknoten besitzen usw. Die Verwendung von Differenzen höherer Ordnung führt zu rauheren Funktionen. Genauer gesagt wird durch die verwendete Differenzenordnung die Glattheit der Funktionen nach unten beschränkt, da sich für $\lambda \rightarrow \infty$ ein Polynom vom Grad $d-1$ ergibt.

In Matrix-Form lautet das Minimierungsproblem

$$(\boldsymbol{y} - \boldsymbol{\Phi\delta})^\top(\boldsymbol{y} - \boldsymbol{\Phi\delta}) + \lambda \cdot \boldsymbol{\delta}^\top \underbrace{\boldsymbol{D}_d^\top \boldsymbol{D}_d}_{\boldsymbol{K}_d}\boldsymbol{\delta} \quad \rightarrow \quad \min_{\boldsymbol{\delta}}.$$

Darin ist

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \ \boldsymbol{\Phi} = \begin{pmatrix} \Phi_1(x_1) & \cdots & \Phi_m(x_1) \\ \vdots & \ddots & \vdots \\ \Phi_1(x_n) & \cdots & \Phi_m(x_n) \end{pmatrix} \text{ und } \boldsymbol{\delta} = \begin{pmatrix} \delta_1 \\ \vdots \\ \delta_m \end{pmatrix}.$$

Die Differenzenmatrix $\boldsymbol{K}_d$ setzt sich aus der Matrix $\boldsymbol{D}_d$ zusammen, für die wiederum gilt:

$$D_1 = \underbrace{\begin{pmatrix} -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{pmatrix}}_{(m-1)\times m}, \quad D_2 = \underbrace{\begin{pmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{pmatrix}}_{(m-2)\times m}, \ \ldots$$

Im Fall des linearen Modells ist die Lösung des Minimierungsproblems analytisch möglich und es ergibt sich der folgende penalisierte KQ-Schätzer:

$$\hat{\boldsymbol{\delta}}_\lambda = (\boldsymbol{\Phi}^\top \boldsymbol{\Phi} + \lambda \boldsymbol{K}_d)^{-1} \boldsymbol{\Phi}^\top \boldsymbol{y}.$$

Wendet man keine Penalisierung an ($\lambda = 0$), ergibt sich der klassische KQ-Schätzer des linearen Modells.

Der Penalisierungsparameter λ wird in der Regel datengesteuert durch Kreuzvalidierung (Cross Validation, CV) bestimmt. Dabei vergleicht man den Wert y_i mit der Schätzung $\hat{f}_\lambda^{-i}(x_i)$, die auf allen Beobachtungen außer der i-ten beruht. Insgesamt ergibt sich das Kreuzvalidierungskriterium zu

$$\mathrm{CV}(\lambda) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{f}_\lambda^{-i}(x_i))^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \hat{f}_\lambda(x_i)}{1 - s_{ii}} \right)^2 \quad \to \quad \min_\lambda .$$

Dabei ist s_{ii} das i-te Diagonalelement der sogenannten Smoothing-Matrix $\boldsymbol{S}$. Die zweite Formulierung ist vor allem deswegen interessant, da die Schätzung hier nur einmal bestimmt werden muss, während bei der ersten Formulierung $\hat{f}_\lambda^{-i}(x_i)$ für alle i bestimmt werden muss. Für die Smoothing-Matrix gilt hier:

$$\begin{pmatrix} \hat{f}(x_1) \\ \vdots \\ \hat{f}(x_n) \end{pmatrix} = \boldsymbol{S}\boldsymbol{y}.$$

$\boldsymbol{S}$ ergibt sich zu

$$\boldsymbol{S} = \boldsymbol{\Phi}(\boldsymbol{\Phi}^\top \boldsymbol{\Phi} + \lambda \boldsymbol{K}_d)^{-1} \boldsymbol{\Phi}^\top.$$

Anstelle des Kreuzvalidierungskriteriums wird in der Regel das sogenannte generalisierte Kreuzvalidierungskriterium GCV(λ) verwendet. Dies begründet sich unter anderem in der fehlenden Invarianz des CV-Kriteriums gegenüber orthogonalen Transformationen der Modellgleichung (vgl. Wood (2006)). Beim GCV-Kriterium wird das

i-te Diagonalelement der Smoothing-Matrix s_{ii} im Nenner des CV-Kriteriums durch den Durchschnittswert aller Diagonalelemente ersetzt:

$$\text{GCV}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{f}_\lambda(x_i)}{1 - \text{tr}(\boldsymbol{S})/n}\right)^2 \quad \rightarrow \quad \min_\lambda .$$

Anhand dieser Darstellung lässt sich der Trade-off zwischen Datentreue und Glattheit erneut gut erkennen: Der Zähler wird groß, wenn die geschätzte Funktion $\hat{f}_\lambda(x_i)$ nicht gut an die Daten y_i angepasst ist, der Nenner wird klein und damit der Kriteriumswert groß, wenn die Funktion sehr rauh ist, da die Smoothing-Matrix dann große Diagonaleinträge enthält.

Ein weiteres Kriterium zur Glättungsparameterwahl ist das sogenannte Mallow's C_p (vgl. Mallows (1973)), auch bekannt als UBRE (Unbiased Risk Estimator), das vor allem bei bekanntem Dispersionsparameter φ geeignet ist:

$$C_p(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}_\lambda(x_i))^2 - \varphi + 2\text{tr}(\boldsymbol{S})\frac{\varphi}{n} \quad \rightarrow \quad \min_\lambda .$$

Bei Generalisierten Linearen Modellen erfolgt die Schätzung der P-Splinekoeffizienten $\boldsymbol{\delta}$ nicht über die Minimierung des penalisierten KQ-Kriteriums, sondern über die Maximierung der entsprechend penalisierten Log-Likelihood $l_\lambda(\boldsymbol{\delta})$, bestehend aus den Komponenten $l_{\lambda,i}(\boldsymbol{\delta})$:

$$\sum_{i=1}^{n} l_{\lambda,i}(\boldsymbol{\delta}) - \frac{\lambda}{2}\cdot\boldsymbol{\delta}^\top\boldsymbol{K}_d\boldsymbol{\delta} \quad \rightarrow \quad \max_{\boldsymbol{\delta}} .$$

GCV und UBRE lassen sich ohne weiteres auf den generalisierten Fall übertragen, indem man den KQ-Term im Minimierungskriterium jeweils durch eine Approximation der Modelldevianz ersetzt (vgl. Wood (2006)).

Um ausreichend Flexibilität zu erhalten verwendet man in der Regel zwischen $m = 30$ und $m = 50$ äquidistante Knoten bzw. Basisfunktionen. In dieser Arbeit wird das P-Spline-Verfahren zur Schätzung der PDLF eingesetzt, einem neu entwickeltem Verfahren zur Berücksichtigung von verzögerten Kovariableneffekten (vgl. Abschnitt 4.4). Die Anordnung der Knoten wird hier zur Steuerung der Flexibilität der geschätzten Funktion in den unterschiedlichen Lag-Bereichen eingesetzt. Genauer gesagt werden die Knoten so gesetzt, dass der Abstand zwischen den Knoten quadratisch zunimmt, um eine große Flexibilität im niedrigen Lag-Bereich und eine glatte Funktion im höheren Lag-Bereich zu erhalten.

Sollen mehrere stetige Kovariablen mit nichtlinearem Einfluss im Modell berücksichtigt werden, trifft man in der Regel die Annahme, dass die einzelnen nichtlinearen

Kovariablenfunktionen eine additive Struktur besitzen. Daraus resultiert das Generalisierte Additive Modell (GAM), für das gilt:

$$\mathbb{E}(y_i|x_{i1},\ldots,x_{ip}) = h(\delta_0 + f_1(x_{i1}) + \ldots + f_p(x_{ip})).$$

Die Funktionen $f_1,\ldots,f_p$ müssen durch einen zyklischen Algorithmus iterativ angepasst werden. Dies geschieht durch den sogenannten Backfitting-Algorithmus (Gauss-Seidel-Algorithmus), der bei generalisierten Strukturen in das gewöhnliche Fisher-Scoring-Verfahren integriert werden kann. Die R-Funktion „gam“ im Paket „mgcv“ (vgl. Wood (2006)) ermöglicht das Fitten von Generalisierten Additiven Modellen. Als Standard-Glättungsverfahren werden hier sogenannte Thin-Plate-Regressionssplines verwendet. Diese bestehen aus der zuvor erwähnten Radial-Basis, wobei die Rauheit der Funktion durch ein geeignetes Penalisierungsverfahren kontrolliert wird. Für die in Abschnitt 4.4 beschriebene Methode zur Schätzung der Distributed Lag Function wurde manuell eine kubische B-Spline-Basis konstruiert. Differenzen in den Splinekoeffizienten wurden dann mithilfe der R-Funktion „gam“ unter Verwendung der beschriebenen Differenzenmatrix $\boldsymbol{K}$ penalisiert. Die resultierenden Modelle, die später auch für die Prognose eingesetzt werden, können somit als penalisierte GLMs betrachtet werden (vgl. Abschnitt 4.4 für weitere Details).

Generalisierte Additive Modelle sind also generell gut geeignet, um nichtlineare Kovariableneffekte zu identifizieren und bei der Modellierung zu berücksichtigen. Problematisch erweist sich allerdings häufig die große Anzahl an zu schätzenden Parametern, vor allem wenn nicht ausreichend Beobachtungen vorliegen. In der Masterarbeit von Wanka (2010) werden die hier vorgestellten Datensätze mithilfe von GAMs analysiert, um nichtlineare Meteorologie- und Luftqualitätseffekte herauszufiltern. In der vorliegenden Arbeit wurden nichtlineare Einflüsse nur am Rande betrachtet. Zum Beispiel wurde mithilfe von Bruchpunktmodellen versucht, nichtlineare Effekte der Kovariablen Temperatur, Luftfeuchtigkeit und Luftdruck zu identifizieren. Zur Bestimmung der Bruchpunkte (Cutpoints) wurden unter anderem GAMs nur mit den designbedingten und administrativen Kovariablen gefittet (vgl. Abschnitt 3.4). Ansonsten wurde den vorhandenen stetigen Kovariablen ein linearer Effekt unterstellt, um die Kovariablenstruktur für das angestrebte Prognosemodell möglichst einfach zu halten.

3.2.4 Generalized Estimating Equations (GEE) zur Analyse longitudinaler Daten

Das von Liang und Zeger 1986 vorgestellte GEE-Modell beruht auf dem in Abschnitt 2.5 beschriebenen marginalen Modellansatz zur Analyse longitudinaler Daten (vgl. Liang und Zeger (1986) für nähere Details zur GEE-Theorie). Der Name GEE leitet sich daraus ab, dass die Schätzung der Regressionsparameter durch die Lösung generalisierter Schätzgleichungen (Generalized Estimating Equations,

GEEs) erfolgt. Genau wie beim Quasi-Likelihood-Modell wird hier keine komplette Verteilung angenommen, sondern es werden lediglich der marginale Erwartungswert $\mathbb{E}(y_i|\boldsymbol{x}_i)$ und die marginale Varianz $\mathrm{Var}(y_i|\boldsymbol{x}_i)$ getrennt voneinander spezifiziert. Die Erweiterung im Vergleich zum Quasi-Likelihood-Modell besteht darin, dass man beim GEE die Kovarianzstruktur als Nuisance-Parameter betrachtet und zusätzlich eine sogenannte Arbeitskovarianzmatrix verwendet, die nicht notwendigerweise korrekt spezifiziert sein muss. Es lässt sich zeigen, dass die Regressionskoeffizienten auch bei falscher Spezifikation der Arbeitskovarianz konsistent geschätzt werden, allerdings verliert die Schätzung dadurch an Effizienz. Im Fall der Zähldaten-Regression wird die Arbeitskovarianz in der Regel über die Arbeitskorrelation (parametrisiert durch den Parametervektor $\boldsymbol{\alpha}$) spezifiziert. Auf diesem Wege ist es möglich, spezielle Korrelationsstrukturen der $y_i|\boldsymbol{x}_i$ zu berücksichtigen, z. B. eine serielle Korrelation der Fallzahlen an aufeinanderfolgenden Tagen. Im Vergleich zu allen bisher betrachteten Modellen wird also nicht zwangsläufig die Unkorreliertheit der Beoachtungen vorausgesetzt.

Häufig wird das GEE bei wiederholten Messungen einer Zielgröße an verschiedenen Individuen eingesetzt. Dazu nimmt man an, dass die Messungen $y_{i1}, \ldots, y_{iT_i}$ innerhalb des i-ten Individuums zwar korreliert sind, die Individuen untereinander aber unabhängig sind. Im hier vorliegenden Fall entsprechen die Individuen den (Geschlechts-,) Alters- und Landkreis-Gruppen. Man geht also beispielsweise beim auf die Pilotregion München eingeschränkten Abrechnungsdatensatz davon aus, dass die Anzahlen der Arztbesuche in den einzelnen Geschlechts- und Altersgruppen voneinander unabhängige Zeitreihen $y_{ij1}, \ldots, y_{ijT}$ darstellen. Innerhalb der Zeitreihen wird versucht, eine passende serielle Korrelationsstruktur zu spezifizieren.

Zum Beispiel kommt bei den vorliegenden Daten eine autoregressive Korrelationsstruktur der Ordnung p (AR-p-Struktur) infrage. Die entsprechende Arbeitskorrelationsmatrix $\boldsymbol{R}_{ij}(\boldsymbol{\alpha}) = (r_{ij,uv}(\boldsymbol{\alpha}))_{u,v=1,\ldots,T}$ hat dann die folgende Form:

$$r_{ij,uv}(\boldsymbol{\alpha}) = \begin{cases} 1 & u = v \\ \alpha_1 & |u-v| = 1 \\ \vdots & \vdots \\ \alpha_p & |u-v| = p \\ 0 & |u-v| > p \end{cases}, \quad u, v = 1, \ldots, T.$$

Die gesamte Arbeitskorrelationsmatrix $\boldsymbol{R}(\boldsymbol{\alpha})$ setzt sich aus den Blöcken $\boldsymbol{R}_{ij}(\boldsymbol{\alpha})$ zu einer Blockdiagonalmatrix zusammen. In der Regel wählt man für alle Individuen die gleiche Korrelationsstruktur. Alternativ könnte auch die restriktivere Annahme einer exponentiellen Korrelationsstruktur getroffen werden, bei der die Korrelation mit dem zeitlichen Abstand zwischen 2 Beobachtungen exponentiell abnimmt:

$$r_{ij,uv}(\alpha) = \alpha^{|u-v|}, \quad u, v = 1, \ldots, T.$$

Der Vorteil der exponentiellen Korrelationsstruktur gegenüber der AR-p-Struktur ist, dass nur einer anstelle von p Parametern geschätzt werden muss. Generell erfolgt die Schätzung im GEE iterativ durch abwechselndes Anpassen des Kovariablenvektors $\boldsymbol{\beta}$ und der Kovarianz- bzw. Korrelationsparameter $\boldsymbol{\alpha}$.

Die Anwendung von GEEs gemäß der beschriebenen Vorgehensweise ist jedoch in der vorliegenden Datensituation fragwürdig, da die Annahme der Unabhängigkeit der Individuen, also in diesem Fall der einzelnen Geschlechts- und Altersgruppen, mit großer Wahrscheinlichkeit verletzt ist. Die einzelnen Zeitreihen verlieren spätestens dann ihre Unabhängigkeit, wenn man zum Gesamtdatensatz übergeht, der eine räumliche Korrelation zwischen den Landkreisen beinhaltet. Zudem kam der Einsatz von GEEs in dieser Arbeit aus rein technischen Gründen nicht infrage, da die Anzahl der Messungen pro Individuum mit $T = 730$ bzw. 1096 deutlich zu groß ist. Dies ist vor allem deswegen problematisch, da die Arbeitskovarianzmatrix, die entsprechend große Blöcke enthält, bei der Schätzung wiederholt invertiert werden muss. Dazu kommt die große Anzahl an Kovariablen, welche die Schätzung per se instabil macht. Der Versuch GEEs mit sinnvoller Korrelationsstruktur mithilfe der R-Funktion „geeglm“ aus dem Paket „geepack“ (vgl. Halekoh et al. (2006)) zu fitten, führte aus den genannten Gründen zu keinerlei Ergebnissen.

Aus diesem Grund wurde der Versuch unternommen, die Autokorrelation in den Fallzahlen durch den Einsatz altersspezifischer Zeitfunktionen soweit zu reduzieren, dass die Annahme der Unabhängigkeit der $y_i|\boldsymbol{x}_i$, welche den zuvor vorgestellten Modellen zugrunde liegt, zumindest näherungsweise erfüllt ist (vgl. den folgenden Abschnitt 3.3).

3.3 Umgang mit autokorrelierten Zielvariablenwerten

Dieser Abschnitt widmet sich der indirekten Berücksichtigung von autokorrelierten Zielvariablenwerten durch die Miteinbeziehung altersspezifischer Zeitfunktionen in den linearen Prädiktor der bereits vorgestellten Modelle. Zunächst wird beschrieben, wie die serielle Korrelation durch die Hinzunahme des Zeittrends reduziert werden kann. Anschließend folgt ein kurzer Absatz zur Schätzung der gruppenspezifischen Funktionen $f(t)$ und der dazugehörigen Konfidenzbänder. Die Anwendung von Zeitsplines in longitudinalen Regressionsmodellen wird im Buch von Fahrmeir und Kneib (2010, Kap. 4) ausführlich beschrieben. Im Vergleich zu den in diesem Abschnitt verwendeten frequentistischen Modellen, erfolgt die Modellinferenz dort allerdings nach bayesianischen Konzepten.

Wie bereits in den Abschnitten 2.5 und 3.2.1 erwähnt, wird die Modellannahme der Unabhängigkeit der $y_i|\boldsymbol{x}_i$ durch die Autokorrelation von Fallzahlen an aufeinanderfolgenden Tagen verletzt, so dass die Schätzung der Regressionskoeffizienten möglicherweise nicht mehr korrekt ist. Im Kontext von Zeitreihen und longitudinalen

Daten ist die sogenannte partielle Autokorrelationsfunktion (PACF) ein wichtiges Mittel, um solch eine serielle Korrelation zu identifizieren und zu quantifizieren (vgl. Venables und Ripley (2002, Kap. 14)). Die PACF stellt die empirischen partiellen Korrelationen der Fallzahlen abhängig vom zeitlichen Abstand der Beobachtungen (dem sogenannten Lag) dar. Die partielle Korrelation entspricht der Korrelation zwischen den Fallzahlen mit einem festen zeitlichen Abstand unter Herausrechnung des Einflusses aller dazwischenliegenden Werte. Abbildung 3.2 stellt exemplarisch die PACF der Anruferzahlen beim KVB-Call-Center in Altersgruppe 4 (61 bis 80 Jahre) in der Pilotregion München dar.

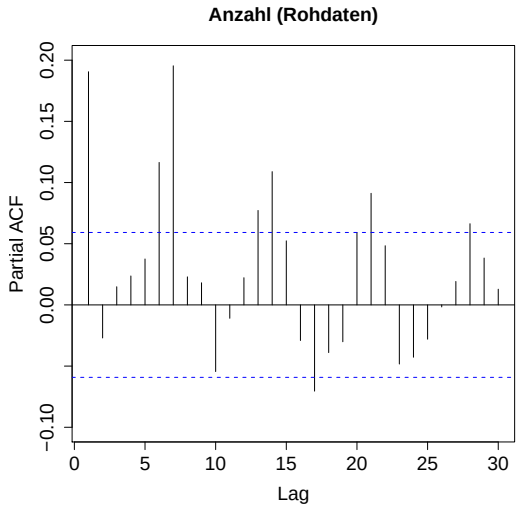


Abbildung 3.2: Partielle Autokorrelationsfunktion der Call-Center-Anrufe in der Pilotregion München in Altersgruppe 4

Die Höhe der einzelnen Linien ergibt sich durch das Fitten von autoregressiven Modellen (AR-Modelle). Möchte man beispielsweise die partielle Autokorrelation für das Lag 5 berechnen, wird ein AR-5-Modell gefittet, das bedeutet ein lineares Modell mit der Anzahl der Anrufe von heute y_t als Zielgröße und den Fallzahlen der jeweils letzten 5 Tage $y_{t-1}, \ldots, y_{t-5}$ als Kovariablen:

$$y_t = \alpha_0 + y_{t-1}\alpha_1 + \ldots + y_{t-5}\alpha_5 + \varepsilon_t.$$

Die empirische partielle Autokorrelation entspricht dann dem Schätzwert $\hat{\alpha}_5$. Durch die Verwendung des Regressionsmodells wird der Einfluss der Lags 1 bis 4 herausgerechnet. Die eingezeichneten gestrichelten Linien stellen 95%-Konfidenzintervalle für die partiellen Autokorrelationen dar. Diese leiten sich aus der approximativen Verteilung der partiellen Autokorrelationen ($\mathrm{N}(0, 1/n)$) ab und ergeben sich somit zu $[\mp z_{0.975}/\sqrt{n}]$. $z_{0.975}$ ist darin das 97.5%-Quantil der Standardnormalverteilung. In diesem Fall lautet das Konfidenzintervall $\left[\mp 1.9600/\sqrt{1096}\right] = [-0.0592, 0.0592]$.

Die PACF für die rohen Anruferzahlen in Altersgruppe 4 weist signifikant positive Autokorrelationen in den Lags 1, 6, 7, 13, 14, 20, 21 und 28 auf. Die Fallzahlen an aufeinanderfolgenden Tagen sind wie erwartet stark korreliert (partielle Korrelation von ca. 0.19). Die signifikanten partiellen Autokorrelationen in den höheren Lags enstehen wohl durch die erhöhten Fallzahlen an Samstagen und Sonntagen, worauf das periodische Muster hinweist, das sich im Abstand von 7 Tagen wiederholt.

Da sich die Unabhängigkeitsannahme nicht auf die marginale Verteilung der y_i, sondern auf die bedingte Verteilung gegeben die Kovariablen bezieht, ist es zur Überprüfung der Modellannahme sinnvoller, anstelle der Rohdaten die Residuen des gefitteten Modells zu betrachten. Verwendet man die Residuen des für die Call-Center-Daten in der Pilotregion München gefitteten Poisson-Modells (noch ohne altersspezifischen Zeittrend) zur Schätzung der PACF, ergibt sich der in Abbildung 3.3 dargestellte Plot. Darin lässt sich erkennen, dass das zyklische Muster durch Miteinbeziehung der Kovariablen, insbesondere wohl durch den Faktor Wochentag, verschwunden ist. In diesem Fall verschwindet sogar die signifikante partielle Korrelation in Lag 1.

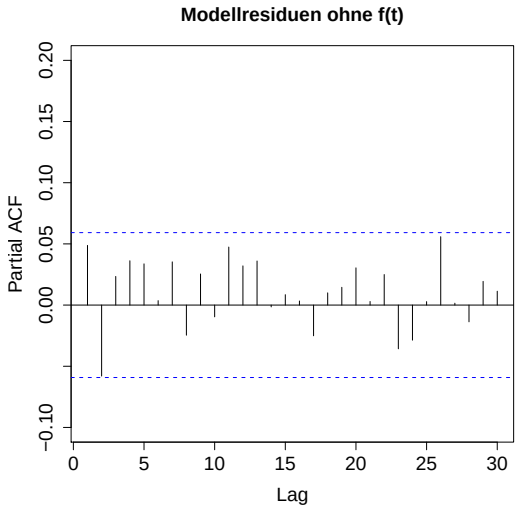


Abbildung 3.3: Partielle Autokorrelationsfunktion der Residuen des Poisson-Modells ohne Zeittrend für die Call-Center-Anrufe in der Pilotregion München (Altersgruppe 4)

Zur weiteren Reduktion der Autokorrelation wurden zusätzlich altersspezifische Zeitfunktionen in die Modelle integriert, deren Konstruktion nachfolgend beschrieben wird. Abbildung 3.4 zeigt die Modellresiduen in Altersgruppe 4 des so erweiterten Poisson-Modells für die Call-Center-Daten in der Pilotregion München. Generell konnten die partiellen Autokorrelationen dadurch weiter verringert werden. Es zeigt sich jedoch auch, dass bereits bestehende negative partielle Autokorrelationen durch

die Miteinbeziehung der Zeitfunktionen etwas stärker ausgeprägt sind, beispielsweise bei Lag 2.

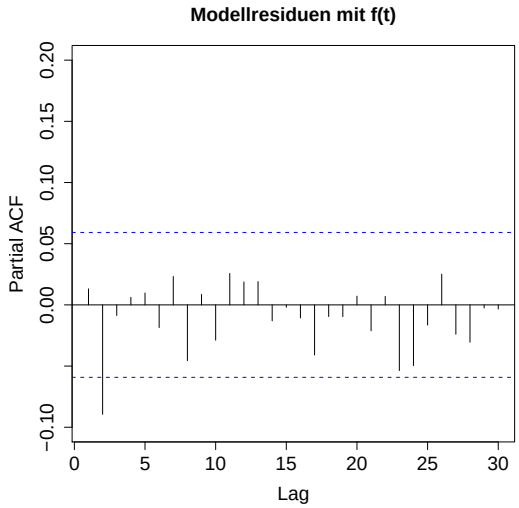


Abbildung 3.4: Partielle Autokorrelationsfunktion der Residuen des Poisson-Modells mit altersspezifischem Zeittrend für die Call-Center-Anrufe in der Pilotregion München (Altersgruppe 4)

Abbildung G.9 zeigt PACFs aller Altersgruppen für die KVB-Call-Center-Daten in München. Wie bereits für Altersgruppe 4 geschehen, werden darin die partiellen Autokorrelationen in den Rohdaten und in den Residuen der Modelle mit und ohne Zeittrend gegenübergestellt. Über alle Altersgruppen hinweg betrug die betragsmäßig größte partielle Autokorrelation in den Rohdaten ca. 0.23. Durch Einbeziehung aller Kovariablen ohne altersspezifischen Zeittrend konnte diese auf ca. 0.11 reduziert werden, mit altersspezifischem Zeittrend sogar auf ca. 0.08. Insgesamt treten nur noch wenige schwach signifikante Autokorrelationen meist im niedrigen Lag-Bereich auf (vgl. die Signifikanzgrenze bei ca. 0.06), so dass die GLM-Modellannahme der Unabhängigkeit der $y_i|\boldsymbol{x}_i$ hier näherungsweise erfüllt sein dürfte.

Genauso wurden für die Anruferzahlen aggregiert über ganz Bayern (b)) sowie die Anzahl der Arztbesuche (getrennt nach Geschlecht) in der Pilotregion München (a)) und aggregiert über die bayerischen Landkreise (b)) PACFs erstellt. Diese wurden wiederum mit den PACFs der Residuen der entsprechenden loglinearen Modelle ohne und mit altersspezifischem Zeittrend verglichen und finden sich auf der CD in Anhang I im Ordner „PACF“. In folgender Tabelle werden die betragsmäßig größten partiellen Autokorrelationen für diese Datensätze zusammengefasst:

Zeitreihe	CC-Daten (Bay. b))	AB-Daten (Mü. a))	AB-Daten (Bay. b))
Rohe Fallzahlen	0.57	0.67	0.64
Resid. (Mod. ohne $f(t)$)	0.22	0.48	0.37
Resid. (Mod. mit $f(t)$)	0.17	0.47	0.26
Signifikanzgrenze	0.0592	0.0725	0.0725

Vereinzelt verbleibende hohe partielle Autokorrelationen weisen meist das oben beschriebene zyklische Muster mit Periode 7 auf und können durch die Berücksichtigung der Kovariable Wochentag nicht gänzlich entfernt werden. Insgesamt konnte durch die Aufnahme von Kovariablen in die Regressionsmodelle, insbesondere durch das Anpassen der gruppenspezifischen Zeitfunktionen, die Autokorrelation in den Fallzahlen aber weitgehend reduziert werden, so dass die Anwendung von GLMs (mit den beschriebenen Erweiterungen) durchaus gerechtfertigt erscheint.

Die Zeitfunktion $f(t)$ wurde, wie in Abschnitt 3.2.3 beschrieben, in Basisfunktionen entwickelt. Die Glattheit der Funktion wurde hier nicht datengesteuert (etwa durch ein Penalisierungsverfahren) kontrolliert, sondern durch die Wahl einer moderaten Anzahl an Knoten. Diese wurden jeweils zu Beginn/Ende eines Quartals gesetzt. Für die Call-Center-Daten (2006-2008) wurden dementsprechend 11 innere Knoten plus zwei Randknoten ($m = 12$ Basisfunktionen) verwendet, für die Abrechnungsdaten (2006-2007) 7 innere Knoten plus 2 Randknoten ($m = 8$ Basisfunktionen). Auf diese Weise ergaben sich relativ gleichmäßig verlaufende Funktionen, die dennoch eine gute Anpassung an die Rohdaten zeigten (vgl. Abbildungen G.10 und G.11). Zur Konstruktion der Splinefunktionen wurde eine kubische B-Spline-Basis verwendet, die in einem natürlichen kubischen Spline gemäß den in Abschnitt 3.2.3 festgelegten Bedingungen resultiert. Neben der durchgehenden Stetigkeit und Differenzierbarkeit der geschätzten Splinefunktionen besitzen diese auch die Eigenschaft der linearen Fortsetzung außerhalb der Randknoten, was für die Prognose der Fallzahlen an zukünftigen Tagen von wesentlicher Bedeutung ist.

Zunächst wurde eine um die Mitte des Beobachtungszeitraums zentrierte Zeitvariable x_{time} konstruiert. Im Fall der KVB-Abrechnungsdaten besitzt diese zum Beispiel einen Wertebereich von -364.5 bis 364.5. Anschließend wurde x_{time} in die mithilfe der R-Funktion „ns“ im Paket „splines“ konstruierten Basisfunktionen eingesetzt. Aus jeder der Basisfunktionen $\Phi_{t'}$ ergibt sich so eine eigene Kovariable $\tilde{x}_{\text{time=t'}} = \Phi_{t'}(x_{\text{time}})$ mit zugehörigem Parameter $\beta_{\text{time=t'}}$ ($t' = 1, \ldots, m$). Durch diese Vorgehensweise konnten die Splinefunktionen ohne Schwierigkeiten in den linearen Prädiktor der verwendeten GLMs oder Quasi-Likelihood-Modelle eingebunden werden.

Da sich bei der Untersuchung des Zeitverlaufs der Rohdaten (vgl. Abbildung G.4) recht unterschiedliche Verläufe der Fallzahlen ergaben, wurden die Zeitfunktionen nicht global, sondern spezifisch für die einzelnen Altersgruppen gefittet. Dazu wurden Interaktionsterme zwischen den neu konstruierten Splinekovariablen $\tilde{x}_{\text{time=t'}}$ ($t' = 1, \ldots, m$) und den Dummy-Variablen für das Alter $x_{\text{age=i}}$ ($i = 2, \ldots, I$) in

die Modellgleichung aufgenommen. Bei den Abrechnungsdaten verliefen die lowess-Kurven durch die Rohdaten für Frauen und Männer weitgehend parallel (vgl. Abbildung G.4), so dass die Zeitfunktionen auch hier nur alters- und nicht geschlechtsspezifisch angepasst wurden.

Am Beispiel der KVB-Abrechnungsdaten für die Pilotregion München soll nun verdeutlicht werden, wie sich die globale Zeitfunktion $\hat{f}(t)$ und die Zeitfunktionen $\hat{f}_i(t)$ in Altersgruppe i zusammensetzen und grafisch dargestellt werden können.

Der geschätzte globale Zeittrend $\hat{f}(t)$ ergibt sich, indem man die Sequenz

$$\boldsymbol{x}_{\text{grid}} = (\min(x_{\text{time}}), \min(x_{\text{time}}) + 1, \ldots, \max(x_{\text{time}}))^\top$$

in die Basisfunktionen einsetzt und die entstehende (730×8)-Matrix $\boldsymbol{\Phi}(\boldsymbol{x}_{\text{grid}})$ mit den geschätzten Koeffizienten $\hat{\boldsymbol{\beta}}_{\text{time}} = (\hat{\beta}_{\text{time=1}}, \ldots, \hat{\beta}_{\text{time=8}})^\top$ multipliziert:

$$\hat{f}(t) = \boldsymbol{\Phi}(\boldsymbol{x}_{\text{grid}})\hat{\boldsymbol{\beta}}_{\text{time}}.$$

Die altersspezifischen Zeitfunktionen ergeben sich (zunächst ohne Berücksichtigung der zeitkonstanten Effekte von Alter und Geschlecht) durch Addition des Vektors der altersspezifischen Splineparameter $\hat{\boldsymbol{\beta}}_{\text{time,age=i}} = (\hat{\beta}_{\text{time=1,age=i}}, \ldots, \hat{\beta}_{\text{time=8,age=i}})^\top$:

$$\hat{f}_i(t) = \boldsymbol{\Phi}(\boldsymbol{x}_{\text{grid}})(\hat{\boldsymbol{\beta}}_{\text{time}} + \hat{\boldsymbol{\beta}}_{\text{time,age=i}}).$$

Die Koeffizienten $\hat{\beta}_{\text{time=t',age=1}}$ zur Konstruktion von $\hat{f}_1(t)$ ergeben sich aufgrund der Effektcodierung für die Altersvariable zu

$$\hat{\beta}_{\text{time=t',age=1}} = -\sum_{i=2}^{4} \hat{\beta}_{\text{time=t',age=i}}, \quad t' = 1, \ldots, 8.$$

Für den globalen und die altersspezifischen Zeittrends gilt aufgrund ihrer Konstruktionsweise $\hat{f}(0) = \hat{f}_i(0) = 0$. Der Einfluss der Zeitkomponente ist also zu Beginn des Beobachtungszeitraums auf 0 fixiert. Damit stellt die Fallzahl am 1.1.2006 den Referenzwert für die Interpretation aller nachfolgenden Funktionswerte dar.

Zusätzlich wurden für die grafische Darstellung die festen/zeitkonstanten Kovariableneffekte $\hat{\beta}_{\text{sex}}$, $\hat{\boldsymbol{\beta}}_{\text{age}}$ und $\hat{\boldsymbol{\beta}}_{\text{sex,age}}$ (vgl. Tabelle G.8) durch Addition der entsprechenden Parameter in die Zeitfunktion miteinbezogen. Beispielsweise gilt für den geschätzten Zeitverlauf bei Frauen in Altersgruppe 2 $\hat{f}_{\text{age=2,sex=2}}(t)$:

$$\hat{f}_{\text{age=2,sex=2}}(t) = \hat{\beta}_{\text{sex=2}} + \hat{\beta}_{\text{age=2}} + \hat{\beta}_{\text{sex=2,age=2}} + \hat{f}_2(t).$$

Die Funktionen $\hat{f}_{\text{sex=j}}(t)$ und $\hat{f}_{\text{age=i,sex=j}}(t)$ entstehen also durch eine gruppenspezifische Verschiebung (Shift) von $\hat{f}(t)$ und $\hat{f}_i(t)$ nach unten oder nach oben.

Um optisch die Signifikanz von Schwankungen im zeitlichen Verlauf beurteilen zu können, wurden zusätzlich punktweise 95%-Konfidenzbänder für die Altersfunktionen berechnet. Für das Konfidenzband um die globale Zeitfunktion $\hat{f}(t)$ wird zunächst eine Schätzung für die Kovarianz $\mathrm{Cov}(\hat{f}(t))$ benötigt. Diese ergibt sich aus

$$\widehat{\mathrm{Cov}}(\hat{f}(\boldsymbol{x}_{\mathrm{grid}})) = \widehat{\mathrm{Cov}}(\boldsymbol{\Phi}(\boldsymbol{x}_{\mathrm{grid}})\hat{\boldsymbol{\beta}}_{\mathrm{time}}) = \boldsymbol{\Phi}(\boldsymbol{x}_{\mathrm{grid}})\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}_{\mathrm{time}})\boldsymbol{\Phi}(\boldsymbol{x}_{\mathrm{grid}})^{\top}.$$

Die Grenzen des punktweisen Konfidenzbands berechnen sich dann wie folgt für ein beliebiges Element $\tau \in \{\min(\boldsymbol{x}_{\mathrm{grid}}), \ldots, \max(\boldsymbol{x}_{\mathrm{grid}})\}$:

$$\hat{f}(\tau) \mp z_{0.975} \cdot \sqrt{\widehat{\mathrm{Var}}(\hat{f}(\tau))}.$$

Darin ist $z_{0.975}$ wiederum das 97.5%-Quantil der Standardnormalverteilung und $\widehat{\mathrm{Var}}(\hat{f}(\tau))$ das zu τ zugehörige Diagonalelement der geschätzten Kovarianzmatrix von $f(\boldsymbol{x}_{\mathrm{grid}})$.

Bei den altersspezifischen Zeitfunktionen (zunächst wieder ohne zeitkonstante Effekte) muss die Korrelation zwischen globalen und altersspezifischen Splineparametern berücksichtigt werden. Die Schätzung für die Kovarianz $\mathrm{Cov}(\hat{f}_i(t))$ lautet dann für $i = 2, \ldots, 4$

$$\begin{aligned}
\widehat{\mathrm{Cov}}(\hat{f}_i(t)) &= \widehat{\mathrm{Cov}}\left[\boldsymbol{\Phi}(\boldsymbol{x}_{\mathrm{grid}})(\hat{\boldsymbol{\beta}}_{\mathrm{time}} + \hat{\boldsymbol{\beta}}_{\mathrm{time,age=i}})\right] \\
&= \boldsymbol{\Phi}(\boldsymbol{x}_{\mathrm{grid}})\left[\mathrm{Cov}(\hat{\boldsymbol{\beta}}_{\mathrm{time}}) + \mathrm{Cov}(\hat{\boldsymbol{\beta}}_{\mathrm{time,age=i}}) + 2\,\mathrm{Cov}(\hat{\boldsymbol{\beta}}_{\mathrm{time}}, \hat{\boldsymbol{\beta}}_{\mathrm{time,age=i}})\right]\boldsymbol{\Phi}(\boldsymbol{x}_{\mathrm{grid}})^{\top}
\end{aligned}$$

und für $i = 1$

$$\begin{aligned}
\widehat{\mathrm{Cov}}(\hat{f}_1(t)) &= \widehat{\mathrm{Cov}}\left[\boldsymbol{\Phi}(\boldsymbol{x}_{\mathrm{grid}})(\hat{\boldsymbol{\beta}}_{\mathrm{time}} + \hat{\boldsymbol{\beta}}_{\mathrm{time,age=1}})\right] \\
&= \boldsymbol{\Phi}(\boldsymbol{x}_{\mathrm{grid}})\widehat{\mathrm{Cov}}\left[\hat{\boldsymbol{\beta}}_{\mathrm{time}} - \left(\hat{\boldsymbol{\beta}}_{\mathrm{time,age=2}} + \hat{\boldsymbol{\beta}}_{\mathrm{time,age=3}} + \hat{\boldsymbol{\beta}}_{\mathrm{time,age=4}}\right)\right]\boldsymbol{\Phi}(\boldsymbol{x}_{\mathrm{grid}})^{\top} \\
&= \boldsymbol{\Phi}(\boldsymbol{x}_{\mathrm{grid}})\Big[\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}_{\mathrm{time}}) + \widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}_{\mathrm{time,age=2}}) + \widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}_{\mathrm{time,age=3}}) + \widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}_{\mathrm{time,age=4}}) \\
&\quad - 2\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}_{\mathrm{time}}, \hat{\boldsymbol{\beta}}_{\mathrm{time,age=2}}) - 2\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}_{\mathrm{time}}, \hat{\boldsymbol{\beta}}_{\mathrm{time,age=3}}) \\
&\quad - 2\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}_{\mathrm{time}}, \hat{\boldsymbol{\beta}}_{\mathrm{time,age=4}}) + 2\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}_{\mathrm{time,age=2}}, \hat{\boldsymbol{\beta}}_{\mathrm{time,age=3}}) \\
&\quad + 2\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}_{\mathrm{time,age=2}}, \hat{\boldsymbol{\beta}}_{\mathrm{time,age=4}}) + 2\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}_{\mathrm{time,age=3}}, \hat{\boldsymbol{\beta}}_{\mathrm{time,age=4}})\Big]\boldsymbol{\Phi}(\boldsymbol{x}_{\mathrm{grid}})^{\top}.
\end{aligned}$$

Zu Beginn des Beobachtungszeitraums besitzt das Konfidenzband die Breite 0, da $\hat{f}(t)$ und $\hat{f}_i(t)$ zu Beginn des Beobachtungszeitraums auf 0 fixiert sind. Nach einer Phase von ca. 1 bis 2 Monaten pendelt sich die Breite des Intervalls jedoch auf einen konstanten Wert ein.

Werden die zeitkonstanten Effekte mit hinzugenommen, kommt bei der Konstruktion der Konfidenzbänder zusätzlich die Unsicherheit bei deren Schätzung dazu. Soll

beispielsweise ein Konfidenzband für obiges Beispiel (Altersgruppe 2, weiblich) konstruiert werden, gilt es folgende Kovarianz zu berechnen:

$$\widehat{\mathrm{Cov}}(\hat{f}_{\mathrm{age=2,sex=2}}(t)) = \widehat{\mathrm{Cov}}(\hat{\beta}_{\mathrm{sex=2}} + \hat{\beta}_{\mathrm{age=2}} + \hat{\beta}_{\mathrm{sex=2,age=2}} + \hat{f}_2(t)).$$

Theoretisch müsste dazu die Kovarianz zwischen den designbedingten Variablen Alter und Geschlecht und den Splinevariablen berücksichtigt werden. Um nicht noch kompliziertere Terme zu bekommen, wurde allerdings die vereinfachende Annahme getroffen, dass die zeitkonstanten Effekte nicht mit den Splinekoeffizienten korreliert sind. Somit gilt für die gesuchte Kovarianz:

$$\begin{aligned}
\widehat{\mathrm{Cov}}(\hat{f}_{\mathrm{age=2,sex=2}}(t)) &= \widehat{\mathrm{Cov}}(\hat{\beta}_{\mathrm{sex=2}} + \hat{\beta}_{\mathrm{age=2}} + \hat{\beta}_{\mathrm{sex=2,age=2}}) + \widehat{\mathrm{Cov}}(\hat{f}_2(t)) \\
&= \widehat{\mathrm{Cov}}(\hat{\beta}_{\mathrm{sex=2}}) + \widehat{\mathrm{Cov}}(\hat{\beta}_{\mathrm{age=2}}) + \widehat{\mathrm{Cov}}(\hat{\beta}_{\mathrm{sex=2,age=2}}) \\
&\quad + 2\widehat{\mathrm{Cov}}(\hat{\beta}_{\mathrm{sex=2}}, \hat{\beta}_{\mathrm{age=2}}) + 2\widehat{\mathrm{Cov}}(\hat{\beta}_{\mathrm{sex=2}}, \hat{\beta}_{\mathrm{sex=2,age=2}}) \\
&\quad + 2\widehat{\mathrm{Cov}}(\hat{\beta}_{\mathrm{age=2}}, \hat{\beta}_{\mathrm{sex=2,age=2}}) + \widehat{\mathrm{Cov}}(\hat{f}_2(t)).
\end{aligned}$$

Abbildung 3.5 zeigt den globalen Zeittrend $\hat{f}(t)$ (links, schwarz) und exemplarisch den altersspezifischen Zeittrend $\hat{f}_2(t)$ (rechts, schwarz) mit den beschriebenen 95%-Konfidenzbändern (grau hinterlegte Fläche) für die KVB-Abrechnungsdaten in der Pilotregion München von 2006 bis 2007. Zusätzlich sind die, um die festen Alters- und Geschlechtseffekte verschobenen, gruppenspezifischen Funktionen $\hat{f}_{\mathrm{sex=1}}(t)$ (links, blau) und $\hat{f}_{\mathrm{sex=2}}(t)$ (links, rot) sowie $\hat{f}_{\mathrm{age=2,sex=1}}(t)$ (rechts, blau) und $\hat{f}_{\mathrm{age=2,sex=2}}(t)$ (rechts, rot) mit entsprechenden Konfidenzbändern (blau/rot gestrichelt) eingezeichnet. Die vertikalen gestrichelten Linien entsprechen den gesetzten Knoten. Im Fall der globalen Zeitfunktion entspricht die Kurve für die Männer genau dem Zeittrend der Gesamtpopulation, da $\beta_{\mathrm{sex=1}}$ als Referenzkategorie der dummycodierten Geschlechtsvariable auf 0 fixiert wurde. Zum Vergleich ist oberhalb der gefitteten Zeittrends jeweils eine lowess-Kurve durch die logarithmierten Rohdaten (vgl. Abschnitt 2.4) abgebildet. Generell wurde der zeitliche Verlauf in den Rohdaten durch den Zeitspline mit quartalsweisen Knoten gut erfasst. In Appendix H.1 ist der R-Code zur Erstellung von Abbildung 3.5 exemplarisch dargestellt.

Zum Vergleich wurden die gleichen Plots noch einmal mit einem Knoten nach jeweils 2 Monaten, insgesamt also 11 inneren und 2 Randknoten (12 Basisfunktionen anstelle von 8) erstellt (vgl. Abbildung 3.6). Die gefitteten Kurven sind nun noch besser an die Rohdaten angepasst, zeigen jedoch einen wesentlich unruhigeren Verlauf. Um vor allem im Hinblick auf die Prognose keine Überanpassung an die Trainingsdaten zu erzielen, wurden die Zeitfunktionen mit quartalsweisen Knoten bevorzugt.

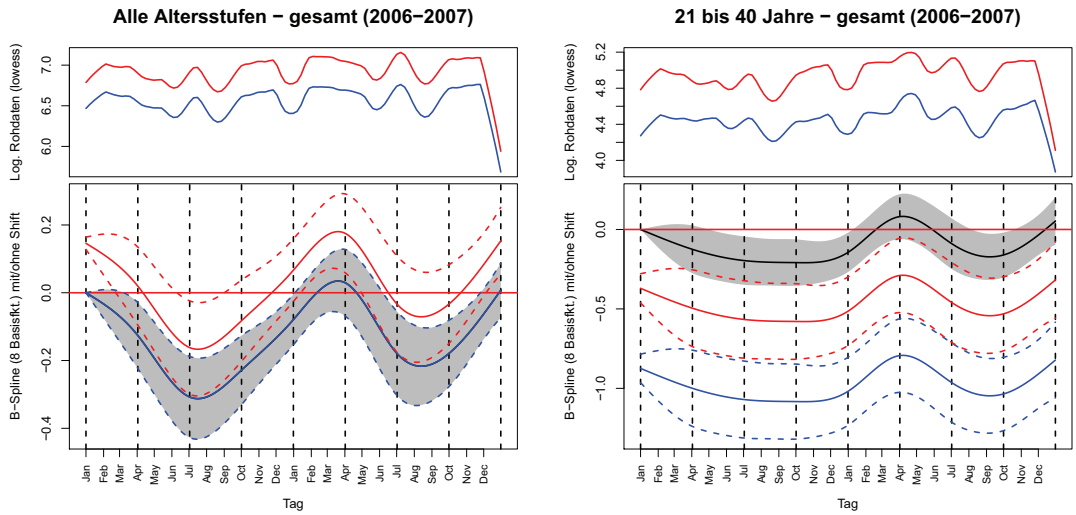


Abbildung 3.5: Darstellung der geschätzten (altersspezifischen) Zeittrends mit je 8 Basisfunktionen für die KVB-Abrechnungsdaten in der Pilotregion München zusammen mit zugehörigen punktweisen Konfidenzbändern: $\hat{f}(t)$ bzw. $\hat{f}_2(t)$ (schwarz), $\hat{f}_{\text{sex}=1}(t)$ bzw. $\hat{f}_{\text{age}=2,\text{sex}=1}(t)$ (blau), $\hat{f}_{\text{sex}=2}(t)$ bzw. $\hat{f}_{\text{age}=2,\text{sex}=2}(t)$ (rot)

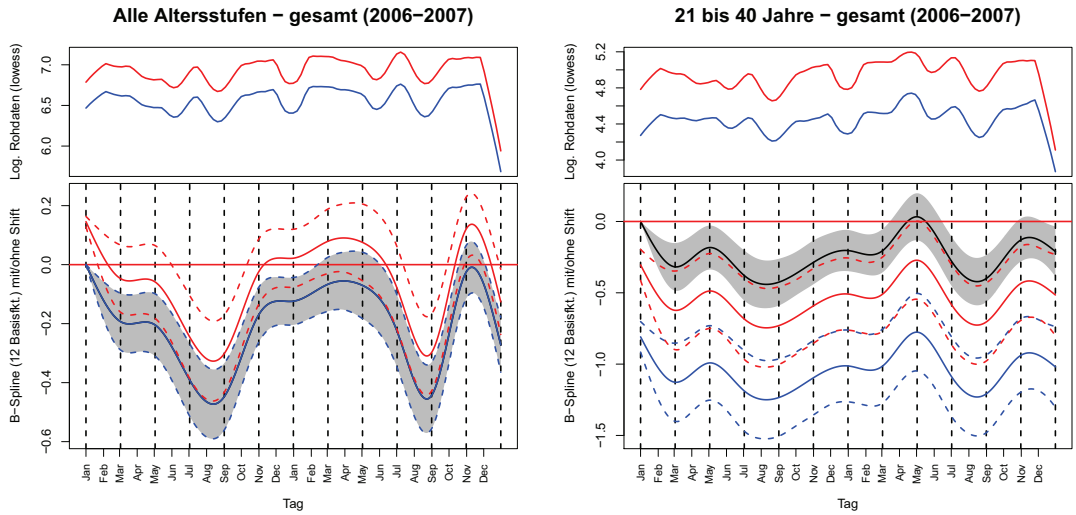


Abbildung 3.6: Darstellung der geschätzten (altersspezifischen) Zeittrends mit je 12 Basisfunktionen für die KVB-Abrechnungsdaten in der Pilotregion München zusammen mit zugehörigen punktweisen Konfidenzbändern: $\hat{f}(t)$ bzw. $\hat{f}_2(t)$ (schwarz), $\hat{f}_{\text{sex}=1}(t)$ bzw. $\hat{f}_{\text{age}=2,\text{sex}=1}(t)$ (blau), $\hat{f}_{\text{sex}=2}(t)$ bzw. $\hat{f}_{\text{age}=2,\text{sex}=2}(t)$ (rot)

Alle geschätzten Zeitfunktionen des loglinearen Modells für die Anzahl der Arztbesuche (gesamt) in der Pilotregion München sowie des Poisson-Modells für die Anzahl

der Anrufe in der Pilotregion München können den Abbildungen G.10 und G.11 entnommen werden. Die entsprechenden Abbildungen für die über die bayerischen Landkreise aggregierten Datensätze finden sich auf der CD in Appendix I im Ordner „Zeittrend“.

Die saisonalen Schwankungen, die bereits in den Rohdaten zu erkennen sind und in Abschnitt 2.4 (vgl. Abbildung G.4) beschrieben wurden, finden sich tendenziell auch in den geschätzten Splinekurven wieder. Die Tatsache, dass die eingezeichneten punktweisen Konfidenzbänder die Nulllinie nur zum Teil überdecken, belegt, dass die Fallzahlen unter Berücksichtigung der übrigen Kovariableneffekte eine signifikante Variation im zeitlichen Verlauf aufweisen. Signifikant erhöhte Fallzahlen sind über alle Datenquellen und Altersgruppen hinweg eher in den Monaten Dezember bis April zu beobachten, signifikant kleinere Fallzahlen dagegen in den Monaten Mai bis Oktober. Die betrachteten Meteorologie- und Luftqualitätsparameter können die saisonalen Schwankungen also nur zum Teil erklären. Die altersspezifischen Funktionen unterscheiden sich größtenteils recht deutlich von der globalen Funktion $f(t)$. Dies rechtfertigt die deutlich parameterintensivere Aufteilung der Splinefunktionen nach Altersgruppen, vor allem da die Zeittrends einen wesentlichen Bestandteil des Prognosemodells für die Fallzahlen darstellen.

3.4 Umgang mit nichtlinearen Kovariableneffekten (Bruchpunktmodelle)

In vorangehenden Studien wurden nichtlineare Effekte meteorologischer Parameter, in erster Linie der Temperatur, auf den Gesundheitszustand der Bevölkerung nachgewiesen (vgl. z. B. Muggeo (2008a), Zanobetti et al. (2000) und Ishigami et al. (2008)). In diesem Zusammenhang ist häufig die Rede von einer sogenannten „Comfort Range“, das bedeutet, eine überdurchschnittliche Zunahme der Fallzahl unter extremen Witterungsbedingungen. In dieser Arbeit wurde für die Kovariablen Temperatur, Luftdruck und Luftfeuchtigkeit untersucht, ob man unter- oder oberhalb gewisser Bruchpunkte (Cutpoints) eine signifikante Veränderung des Kovariableneffekts beobachten kann. Das R-Paket „segmented“(vgl. Muggeo (2008b)) bietet die Möglichkeit, solche Bruchpunkteffekte im Kontext Generalisierter Linearer Modelle zu analysieren. Ein weiteres R-Paket von Muggeo, „modTempEff“ (vgl. Muggeo (2008c)), bietet eine eigene Funktion, um Bruchpunkteffekte in Kombination mit verzögerten Kovariableneffekten zu untersuchen. Diese wurde allerdings in dieser Arbeit nicht verwendet, da bei mehreren Kovariablen mit zeitlich verzögertem bzw. nichtlinearem Effekt Probleme auftraten.

Zunächst wurde mithilfe einfacher Scatterplots und der oben beschriebenen GAMs versucht, die Cutpoints im Verlauf des Kovariableneffekts optisch zu bestimmen. Die Bestimmung erfolgte auf Basis der reduzierten Datensätze nach Reduktions-

verfahren a) und b). In den Scatterplots wurden die rohen Fallzahlen gegen die beobachteten Werte der jeweiligen Kovariable aufgetragen. Um bei den Scatterplots eine Verzerrung durch administrative Effekte zu vermeiden, wurden hierfür aus dem jeweiligen Datensatz zusätzlich alle Samstage und Sonntage, Feiertage, Schulferientage, Brückentage sowie die jeweils ersten und letzten Quartalswochen entfernt. Die GAMs beinhalten die beschriebenen Zeitfunktionen sowie alle designbedingten und administrativen Kovariablen, die erwartungsgemäß die Fallzahlen am stärksten beeinflussen. Die Variablen Feuchtigkeit, Luftdruck und Temperatur wurden jeweils einzeln ins Modell aufgenommen. Abbildung 3.7 zeigt exemplarisch den Scatterplot der Rohdaten sowie die Residuen des GAMs für die Temperatur basierend auf den KVB-Call-Center-Daten für München, jeweils zusammen mit einer rot eingezeichneten lowess-Kurve, um möglicherweise vorhandene Bruchpunkte besser erkennen zu können.

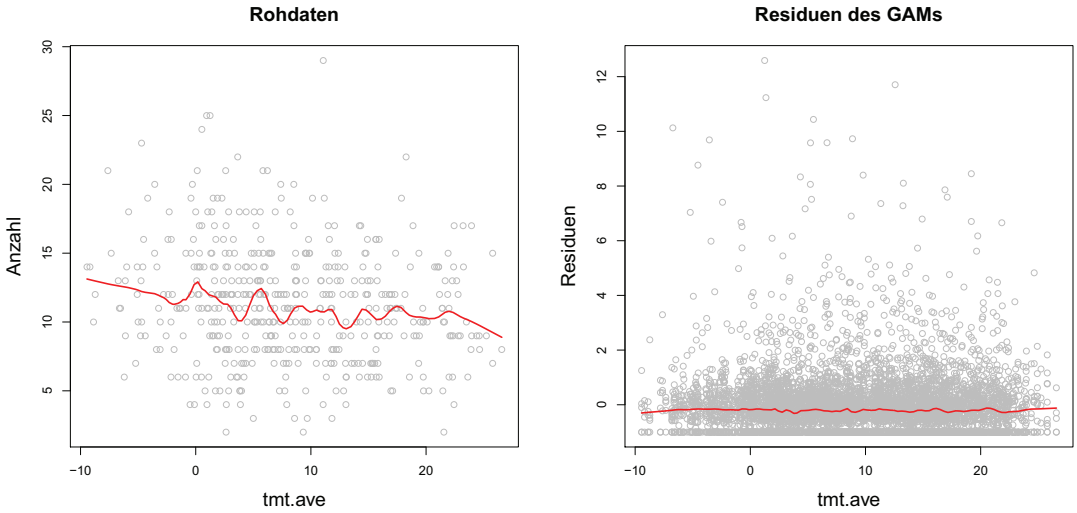


Abbildung 3.7: Scatterplot der Rohdaten und Residuen des GAMs für die Temperatur bei den KVB-Call-Center-Daten in den Pilotregion München

Anhand der Rohdaten lässt sich möglicherweise ein unterer Cutpoint bei etwa $-2°C$ Tagesdurchschnittstemperatur feststellen. Ein oberer Cutpoint ist dagegen schwierig festzulegen. Zudem ermöglichen Scatterplots keinerlei Signifikanzaussagen zur Veränderung des Kovariableneffekts. Basierend auf den Residuen des GAMs ist eine optische Festlegung der Bruchpunkte praktisch unmöglich. Aus diesem Grund wurde ein Verfahren angewandt, das mehrere Cutpoint-Modelle mit unterschiedlichen Kandidatenwerten für obere und untere Cutpoints vergleicht. Zur Modellselektion wurden einfache GLMs mit designbedingten und administrativen Kovariablen sowie altersspezifischem Zeittrend herangezogen, in deren linearen Prädiktor manuell konstruierte Cutpoint-Variablen integriert wurden.

Als Kandidatenwerte für den unteren Cutpoint wurden alle Perzentile zwischen dem 5%- und dem 20%-Quantil der jeweiligen Kovariable betrachtet. Analog wurden als Kandidatenwerte für den oberen Cutpoint alle Perzentile zwischen dem 80%- und dem 95%-Quantil verwendet. Zentralere Quantile wurden nicht betrachtet, da angenommen wurde, dass nur extreme Wetterbedingungen eine Änderung des Kovariableneffekts hervorrufen. Noch extremere Quantile wurden ausgeschlossen, um die Schätzung nicht von einzelnen Beobachtungen abhängig zu machen. Insgesamt müssen also für die oberen und unteren Bruchpunkte jeweils 16 verschiedene Kandidatenwerte miteinander verglichen werden.

Die Konstruktion der unteren (lcp) und oberen (ucp) Cutpoint-Variablen für die Kovariable x beruht auf folgendem Prinzip:

$$\begin{aligned} x_{i,\text{lcp}} &= \mathbb{1}(x_i < q) \cdot (x_i - q), \\ x_{i,\text{ucp}} &= \mathbb{1}(x_i > q) \cdot (x_i - q). \end{aligned}$$

q stellt dabei das verwendete Quantil der Kovariable x und $\mathbb{1}$ die Indikatorfunktion dar. Die so konstruierten Cutpoints können ohne weiteres in den linearen Prädiktor der verwendeten Modelle aufgenommen werden. Für die Modellselektion zur Identifikation der Cutpoints wurden die Cutpoint-Effekte zunächst einzeln und ohne den dazugehörigen linearen Haupteffekt ins Modell aufgenommen. oberhalb des unteren bzw. unterhalb des oberen Cutpoints wurde also überhaupt kein Effekt angenommen. In dieser Situation bedeutet ein negativer Koeffizient für den unteren Cutpoint eine Zunahme der Fallzahl bei Abnahme des Kovariablenwerts unterhalb des Cutpoints. Genauso bedeutet ein positiver Koeffizient eine Zunahme der Fallzahl bei wachsendem Kovariablenwert oberhalb des Cutpoints. Der Vergleich der Modelle mit den, je nach verwendetem Quantil, unterschiedlichen Cutpoint-Variablen erfolgte mithilfe des sogenannten Akaiken Informationskriteriums (AIC, vgl. Akaike (1974)). Dieses ist wie folgt definiert:

$$AIC = -2l(\hat{\boldsymbol{\beta}}) + 2p. \tag{9}$$

p ist darin die Anzahl der Modellparameter, $l(\hat{\boldsymbol{\beta}})$ die Log-Likelihood des geschätzten Modells im Maximum-Likelihood-Schätzer $\hat{\boldsymbol{\beta}}$. Es wird also sowohl die Anpassung an die Daten als auch die Komplexität des Modells berücksichtigt. Generell gilt die Regel „je kleiner das AIC, desto besser ist das Modell“. Nachdem p für alle Kandidatenmodelle gleich ist, reduziert sich das AIC im vorliegenden Fall auf einen Vergleich der Log-Likelihood der Modelle. Äquivalent könnte auch die Modelldevianz verwendet werden.

Durch diese modellbasierte Vorgehensweise kann neben der Bestimmung der optimalen Bruchpunkte gleichzeitig eine Beurteilung der Signifkanz des Bruchpunkteffekts erfolgen, das heißt, ob unter- oder oberhalb des jeweiligen Cutpoints signifikant mehr oder weniger Fälle auftreten. Dabei gilt es zu beachten, dass die geschätzten Modelle

noch keine Haupteffekte der Meteorologie-Parameter beinhalten. Nimmt man den entsprechenden Haupteffekt mit ins Modell auf (wie in den Modellen in Abschnitt 3.2 geschehen), ändert sich die Interpretation der Cutpoint-Koeffizienten. Diese geben nun an, ob und in welche Richtung sich der Kovariableneffekt verändert. Um den Gesamteffekt des betrachteten Meteorologie-Parameters zu erhalten, müssen die Cutpoint-Effekte außerhalb der Comfort Range zum Haupteffekt addiert werden, oder anders formuliert, der Haupteffekt allein gilt nur innerhalb der Comfort Range, wo die Cutpoint-Effekte gleich 0 sind.

Tabelle 3.2 beinhaltet die berechneten AIC-Werte der unteren und oberen Kandidatenmodelle für die Temperatur-Cutpoints bei den KVB-Call-Center-Daten in der Landeshauptsstadt München zusammen mit den p-Werten der Cutpoint-Variable im jeweiligen Selektionsmodell. Demnach liegt der AIC-optimale untere Cutpoint für die Tagesdurchschnittstemperatur bei $2.2540°C$, der AIC-optimale obere Cutpoint bei $16.8158°C$. In beiden Fällen ist die entsprechende Cutpoint-Variable signifikant.

Untere Quantile	$\hat{\beta}_{\text{lcp}}$	p-Wert	AIC	Obere Quantile	β_{ucp}	p-Wert	AIC
-3.2223 (5%)	-0.0142	0.3752	19168.66	16.8158 (80%)	0.0154	0.0260	19164.53
-2.5698 (6%)	-0.0138	0.3068	19168.40	17.0138 (81%)	0.0155	0.0303	19164.79
-2.1413 (7%)	-0.0134	0.2696	19168.23	17.2305 (82%)	0.0156	0.0358	19165.08
-1.4203 (8%)	-0.0130	0.2086	19167.87	17.3813 (83%)	0.0157	0.0399	19165.26
-0.7945 (9%)	-0.0116	0.1996	19167.81	17.7488 (84%)	0.0159	0.0515	19165.69
-0.4023 (10%)	-0.0114	0.1721	19167.59	17.9915 (85%)	0.0160	0.0605	19165.95
0.0035 (11%)	-0.0114	0.1426	19167.30	18.3640 (86%)	0.0161	0.0805	19166.42
0.1920 (12%)	-0.0113	0.1325	19167.19	18.6360 (87%)	0.0161	0.1004	19166.77
0.5008 (13%)	-0.0114	0.1099	19166.90	19.1595 (88%)	0.0154	0.1668	19167.55
0.9738 (14%)	-0.0118	0.0767	19166.33	19.3695 (89%)	0.0155	0.1848	19167.70
1.2493 (15%)	-0.0118	0.0641	19166.03	19.7423 (90%)	0.0162	0.2087	19167.88
1.5173 (16%)	-0.0120	0.0509	19165.65	19.9055 (91%)	0.0165	0.2215	19167.96
1.6998 (17%)	-0.0121	0.0429	19165.37	20.2270 (92%)	0.0175	0.2399	19168.08
1.8828 (18%)	-0.0123	0.0362	19165.08	20.7398 (93%)	0.0184	0.2930	19168.35
2.0140 (19%)	-0.0124	0.0316	19164.85	21.1718 (94%)	0.0198	0.3268	19168.49
2.2540 (20%)	-0.0126	0.0240	19164.38	21.5028 (95%)	0.0231	0.3071	19168.42

Tabelle 3.2: Kandidatenwerte für untere und obere Temperatur-Cutpoints bei den KVB-Call-Center-Daten in der Pilotregion München zusammen mit geschätztem Modellkoeffizient, p-Wert und Modell-AIC des jeweiligen Poisson-GLMs

Abbildung 3.8 verdeutlicht die vom Selektionsmodell unterstellte Form des Temperatureffekts. Zum Vergleich sind die logarithmierten Rohdaten eingezeichnet. Pro Grad Temperaturzunahme oberhalb des oberen Cutpoints wächst demnach die Fallzahl um 1.56%, pro Grad Temperaturabnahme unterhalb des unteren Cutpoints wächst die Fallzahl um 1.27%.

Weitere Visualisierungen der Bruchpunkteffekte analog zu Abbildung 3.8 finden sich auf der CD in Appendix I im Ordner „Cutpoints“. Eine signifikante Zu- oder Abnahme der Fallzahl außerhalb der Cutpoints ist in den Plots rot markiert. Derselbe Ordner beinhaltet auch die kompletten Ergebnisse der AIC-Selektion für alle betrachteten Variablen und Datensätze in tabellarischer Darstellung. Appendix H.2 beinhaltet exemplarisch für die Call-Center-Daten in der Landeshauptstadt München die Im-

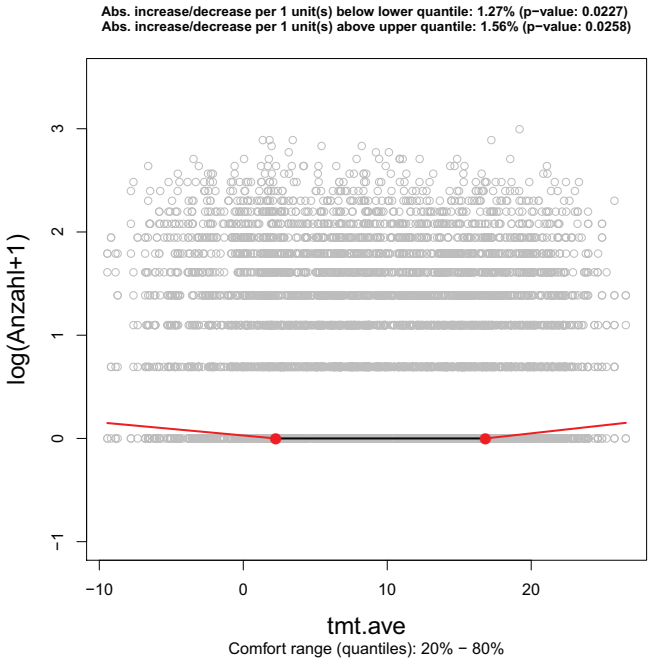


Abbildung 3.8: Temperatureffekt des durch AIC-Selektion bestimmten Cutpoint-Modells für die Call-Center-Daten mit logarithmierten Anruferzahlen (grau)

plementierung der AIC-Selektion in R sowie die Plot-Funktion zur Darstellung der Cutpoint-Effekte im Selektionsmodell.

Tabelle G.10 fasst die per AIC-Selektion ermittelten Luftfeuchtigkeits-, Luftdruck- und Temperatur-Cutpoints für die Call-Center- und Abrechnungsdaten in der Pilotregion München (Reduktionsverfahren a)) sowie aggregiert über die bayerischen Landkreise (Reduktionsverfahren b)) zusammen. Zusätzlich sind die Modellkoeffizienten (roh und exponentiell transformiert) und p-Werte des jeweiligen Selektionsmodells ohne Haupteffekt tabelliert. Gemäß des zuvor beschriebenen Interpretationsschemas wurde bei den unteren Cutpoints jeweils $\exp(-\hat{\beta}_{\text{lcp}})$, bei den oberen Cutpoints dagegen $\exp(+\hat{\beta}_{\text{ucp}})$ berechnet. Des Weiteren beinhaltet die Tabelle die geschätzten Cutpoint-Effekte mit zugehörigen p-Werten der in Abschnitt 3.2.1 beschriebenen GLMs, die alle Meteorologie- und Luftqualitätsparameter inklusive der linearen Haupteffekte der jeweiligen Cutpoint-Variable beinhalten. Zur Einordnung des Gesamteffekts ist neben dem Cutpoint-Effekt der zugehörige Haupteffekt angegeben. Anhand dieser Werte kann beurteilt werden, ob und in welche Richtung sich der Kovariableneffekt ober- oder unterhalb der Cutpoints signifikant verändert. Ein direkter Vergleich der Koeffizienten im Selektionsmodell und im vollen GLM ist nicht möglich, da die Parameter, wie erwähnt, verschieden interpretiert werden müssen. Hinter der jeweiligen Variable ist in Klammern wiederum der Faktor angegeben, mit dem die Koeffizienten zur sinnvolleren Interpretation multipliziert wurden.

Bei den KVB-Call-Center-Daten erwiesen sich im Selektionsmodell ohne Haupteffekte zahlreiche Cutpoint-Variablen als signifikant. Beispielsweise nahm in der Pilotregion München die Anzahl der Anrufe bei Abnahme des Luftdrucks um $1hPa$ unterhalb des Schwellenwerts von $1002.063hPa$ um 1.83% zu. Für den aggregierten Datensatz über ganz Bayern ergab sich eine Comfort Range der Temperatur zwischen $1.8018°C$ und $19.3971°C$. Darunter nimmt die Anzahl der Anrufe um 1.53% pro Abnahme der Temperatur um $1°C$ zu, darüber nimmt die Anzahl der Anrufe um 3.73% bei Anstieg der Temperatur um $1°C$ zu. Diese Effekte blieben auch im vollen GLM mit Haupteffekten erhalten, der Kovariableneffekt verändert sich hier jeweils signifikant unter- oder oberhalb der Cutpoints.

In den Selektionsmodellen für die Abrechnungsdaten in der Landeshauptstadt München ergaben sich beispielsweise signifikante Bruchpunkteffekte im oberen Luftdruckbereich (-0.49% Arztbesuche bei $1hPa$ Druckzunahme über $1023.284hPa$) und im unteren Temperaturbereich (-2.20% Arztbesuche bei $1°C$ Temperaturabnahme unter $2.2315°C$). Im jeweiligen vollen GLM konnte eine signifikante Änderung des Effekts bestätigt werden. Demnach steigt etwa die Anzahl der Arztbesuche im mittleren Luftdruckbereich um 0.24% pro Zunahme des Luftdrucks um $1hPa$, oberhalb des oberen Cutpoints dagegen sinkt die Fallzahl um 0.58% ($\exp(0.0024 - 0.0082)$) pro Zunahme des Luftdrucks um $1hPa$.

Im über die bayerischen Landkreise aggregierten Abrechnungsdatensatz konnte eine Comfort Range der Temperatur zwischen -4.3516 und $16.5372°C$ Tagesdurchschnittstemperatur nachgewiesen werden. Die Selektionsmodelle ergaben eine Zunahme der Arztbesuche um 20.59% pro $1°C$ Temperaturabnahme unterhalb des unteren Cutpoints und eine Zunahme von 4.24% pro $1°C$ Temperaturzunahme oberhalb des oberen Cutpoints. In beiden Bruchpunkten tritt eine signifikante Änderung des im vollen GLM betrachteten Gesamteffekts auf. Im Selektionsmodell ergaben sich weitere signifikante Bruchpunkte für den unteren ($0.0025kg_{\text{Wasser}}/kg_{\text{Luft}}$) und oberen ($0.0099kg_{\text{Wasser}}/kg_{\text{Luft}}$) Luftfeuchtigkeitsbereich sowie für den oberen Luftdruckbereich (11.4778 logarithmierter Oberflächendruck). Der starke Anstieg der Arztbesuche bei abnehmender Luftfeuchtigkeit unterhalb des unteren Cutpoints (+78.12% pro Einheit) steht allerdings im Widerspruch zum im vollen GLM berechneten positiven Gesamteffekt ($-0.0101+0.1726 = 0.1625$), der eine deutliche Abnahme der Fallzahl bei sinkender Luftfeuchtigkeit ($\exp(-0.1625) = 0.8500 \hat{=} -15.00\%$ pro Einheit) unterhalb des unteren Cutpoints bedeutet. Der obere Luftfeuchtigkeits-Cutpoint wurde im GLM dagegen tendenziell bestätigt, auch wenn der Gesamteffekt im vollen GLM (+4.20% Arztbesuche pro 1 Einheit Zunahme über dem Cutpoint) deutlich schwächer ausfällt als der reine Cutpoint-Effekt im Selektionsmodell (+20.51% Arztbesuche pro 1 Einheit Zunahme über dem Cutpoint). Beim oberen Luftdruck-Cutpoint konnte im vollen GLM keine signifikante Veränderung des Kovariableneffekts nachgewiesen werden. Umgekehrt erwies sich die Änderung des Kovariableneffekts im unteren Luftdruck-Cutpoint als signifikant, obwohl an die-

ser Stelle im Selektionsmodell kein signifikanter Cutpoint-Effekt festgestellt werden konnte.

Vergleicht man die Luftfeuchtigkeits- und Temperatur-Cutpoints des auf die Pilotregion München eingeschränkten Datensatzes mit denen des aggregierten Datensatzes für ganz Bayern, ergeben sich bei den KVB-Call-Center-Daten relativ übereinstimmende Ergebnisse, abgesehen vom oberen Cutpoint für die Temperatur. Auch die Effekte sind tendenziell vergleichbar. Beim Abrechnungsdatensatz dagegen unterscheiden sich die Lage der entsprechenden Bruchpunkte und die resultierenden Effekte zum Teil stark, wobei die Luftdruck-Cutpoints aufgrund der unterschiedlichen zugrunde liegenden Variablen nicht direkt verglichen werden können.

Um die Stabilität der gefundenen Effekte zu überprüfen, werden die gefundenen Cutpoint-Variablen, genau wie die übrigen Meteorologie- und Luftqualitätsvariablen, in die später vorgestellten Variablenselektionsverfahren (vgl. Abschnitte 3.7 und 3.8) einbezogen.

3.5 Berücksichtigung räumlicher Effekte mittels bayesianischer gemischter Modelle

Dieser Abschnitt befasst sich mit der Berücksichtigung der räumlichen Abhängigkeiten in den vorliegenden Datensätzen. Da die Analyse der räumlichen Effekte basierend auf einem bayesianischen Modellansatz erfolgt, wird zunächst ein kurzer Einblick in die Theorie bayesianischer Regressionsmodelle im Allgemeinen gegeben. Anschließend wird die Aufteilung des räumlichen Effekts in eine strukturelle und eine zufällige Komponente beschrieben. Schließlich werden ausgewählte Ergebnisse der räumlichen Modelle, basierend auf dem über alle (Geschlechts- und) Altersgruppen aggregierten Datensatz für Gesamtbayern (Reduktionsverfahren c)), präsentiert.

Der grundlegende Unterschied zwischen frequentistischen und bayesianischen Modellen besteht darin, dass der Vektor der Regressionsparameter $\boldsymbol{\beta}$ sowie alle weiteren unbekannten Modellparameter $\boldsymbol{\theta}$ im bayesianischen Fall nicht als deterministisch angenommen werden, sondern Zufallsgrößen darstellen, für die eine Priori-Verteilungsannahme $p(\boldsymbol{\beta}, \boldsymbol{\theta})$ getroffen wird. Da im Regelfall kein Vorwissen über die unbekannten Parameter vorliegt, verwendet man nichtinformative, meist impropere Priori-Verteilungen. Nach Spezifikation der Likelihood $p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\theta})$ kann basierend auf dem Satz von Bayes die gemeinsame Posteriori-Verteilung $p(\boldsymbol{\beta}, \boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{X})$ der Modellparameter hergeleitet werden. Daraus können die vollbedingten Posteriori-Verteilungen der Modellparameter $p(\boldsymbol{\beta}|\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{X})$ und $p(\boldsymbol{\theta}|\boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{X})$ bestimmt werden. Die Modellparameter werden dann iterativ durch abwechelndes Ziehen von Zufallszahlen aus den vollbedingten Verteilungen angepasst. Die Prädiktion für eine neue Beobachtung y_i^* kann basierend auf der (Posteriori-)prädiktiven Verteilung $p(y_i^*|\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\theta})$ erfolgen.

Da die vollbedingten Posteriori-Verteilungen in vielen Fällen analytisch nicht zugänglich sind bzw. das Ziehen von Zufallszahlen daraus nicht möglich ist, verwendet man sogenannte Markov-Chain-Monte-Carlo-Verfahren (MCMC). Dabei wird eine Markov-Kette von abhängigen simulierbaren Zufallsvariablen konstruiert, deren Verteilung unter gewissen Voraussetzungen gegen die gesuchte Posteriori-Verteilung konvergiert. Nach einer ausreichend langen Burn-in-Phase geht man davon aus, dass die stationäre Verteilung der Markov-Kette, also die gesuchte Posteriori-Verteilung, erreicht ist. Die empirische Posteriori-Verteilung erhält man, indem man eine große Zahl von Zufallszahlen aus der Markov-Kette zieht. Um die Burn-in-Phase und die Abhängigkeit der Zufallszahlen zu berücksichtigen, wird in der Regel der erste Teil der Realisierungen der Markov-Kette verworfen und danach z. B. nur jeder 10. Wert zur Schätzung der Posteriori-Verteilung verwendet (Ausdünnung). Die Modellparameter werden dann beispielsweise aus dem Mittelwert der verbliebenen Realisierungen als empirisches Pendant zum Posteriori-Erwartungswert geschätzt. Um Konvergenz und Unabhängigkeit optisch zu überprüfen, betrachtet man den Verlauf der gezogenen Zufallszahlen (Trace-Plots) und der Autokorrelationsfunktionen der ausgedünnten Werte.

Bei der empirischen Bayes-Inferenz wird im Vergleich zur vollen Bayes-Inferenz nicht für alle unbekannten (Hyper-)Parameter eine Priori-Verteilung angenommen. Stattdessen schätzt man diese Parameter mithilfe von Verfahren aus anderen Inferenzdisziplinen, z. B. frequentistischen Verfahren. Bei den in diesem Abschnitt beschriebenen Regressionsmodellen handelt es sich um volle Bayes-Verfahren. Weitere Details zur Theorie der Bayes-Inferenz im Allgemeinen können dem Buch von Held (2008) entnommen werden.

Im Folgenden wird am Beispiel des linearen Modells kurz skizziert, wie die in Abschnitt 3.2.1 beschriebenen Modelle aus bayesianischer Sicht dargestellt und geschätzt werden können (vgl. dazu Fahrmeir und Heumann (2009, Kap. 4)).

Im linearen Modell muss neben dem Parametervektor $\boldsymbol{\beta}$ der Varianzparameter $\theta = \sigma^2$ geschätzt werden. Wie im frequentistischen Kontext wird eine normalverteilte Likelihood angenommen:

$$\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{X} \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}).$$

Unter Verwendung der nichtinformativen Priori $p(\boldsymbol{\beta}, \sigma^2) \propto (\sigma^2)^{-1}$ erhält man die gemeinsame Posteriori

$$p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}, \boldsymbol{X}) \propto (\sigma^2)^{-n/2+1} \exp\left(-\frac{1}{2\sigma^2}\left(\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}} + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top\boldsymbol{X}^\top\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right)\right).$$

Darin ist $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ und $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$. Für die vollbedingten Verteilungen der Regressionsparameter gilt:

$$\begin{aligned} p(\boldsymbol{\beta}|\sigma^2, \boldsymbol{y}, \boldsymbol{X}) &\sim \mathrm{N}(\hat{\boldsymbol{\beta}}, \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}), \\ p(\sigma^2|\boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{X}) &\propto p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}, \boldsymbol{X}). \end{aligned}$$

Die Schätzung erfolgt dann mithilfe eines speziellen MCMC-Verfahrens (Gibbs-Sampling) durch abwechselndes Update von $\boldsymbol{\beta}$ und σ^2. In diesem Fall wäre auch ein direktes Simulationsverfahren ohne MCMC-Techniken möglich, da sich die marginale Posteriori $p(\sigma^2|\boldsymbol{y}, \boldsymbol{X})$ von σ^2 durch Herausintegrieren von $\boldsymbol{\beta}$ zu einer inversen Gamma-Verteilung ergibt. Die Schätzung ist somit auch durch abwechselndes Ziehen aus den analytisch zugänglichen Verteilungen von $p(\sigma^2|\boldsymbol{y}, \boldsymbol{X})$ und $p(\boldsymbol{\beta}|\sigma^2, \boldsymbol{y}, \boldsymbol{X})$ möglich.

Zur bayesianischen Schätzung im GLM wird der in Abschnitt 3.2.1 erwähnte Fisher-Scoring-Algorithmus zunächst in einen „Iteratively Weighted Least Squares“-Algorithmus (IWLS) umformuliert und dann mit einem speziellen MCMC-Verfahren (Metropolis-Hastings-Algorithmus) kombiniert. Neben den GLMs lassen sich beispielsweise auch Zero-Inflated-Poisson-Modelle in einen bayesianischen Kontext übertragen (vgl. z. B. Ghosh et al. (2006)). Die Dissertation von Echavarría (2004) bietet einen ausführlichen Überblick über die bayesianische Modellinferenz für alle in Abschnitt 3.2.1 beschriebenen und für weitere Zähldaten-Modelle mit Bezug zur hier verwendeten Software BayesX (vgl. Belitz et al. (2009)).

Die Berücksichtigung räumlicher Korrelationen erfolgt in der statistischen Modellierung grundsätzlich durch Einbeziehung zufälliger Effekte, für die eine eigene Verteilung, in der Regel eine Normalverteilung, angenommen wird. In diesem Zusammenhang spricht man von (Generalisierten) Gemischten Linearen Modellen ((Generalized) Linear Mixed Models, (G)LMMs). Im vorliegenden Fall integriert man einen zufälligen landkreisspezifischen Parameter (Random Intercept) für alle 96 Landkreise in den linearen Prädiktor. Dieser hat dann die folgende Form:

$$\eta_{st} = \boldsymbol{x}_{st}^\top \boldsymbol{\beta} + f(t) + \alpha_s, \quad \alpha_s \sim \mathrm{N}(0, \tau^2), \quad s = 1, \ldots, 96.$$

In dieser Arbeit wurde ein bayesianisches gemischtes Modell zur Berücksichtigung der räumlichen Datenstruktur verwendet, dessen grundlegende Idee es ist, den räumlichen Effekt in eine strukturelle und eine zufällige Komponente aufzuteilen (vgl. Besag et al. (1991)). Die strukturelle Komponente kann als „echter“, durch die geografische Lage bedingter, räumlicher Effekt aufgefasst werden. Die zufällige Komponente erfasst dagegen alle räumlichen Unterschiede in den Fallzahlen, die durch nicht beoachtbare Einflussgrößen entstehen. Im Folgenden werden die Annahmen, die für beide Komponenten getroffen werden, erläutert (vgl. das Methodology Manual zur BayesX-Software, Belitz et al. (2009, Kap. 4)).

Der Begriff „zufälliger Effekt“ erscheint im bayesianischen Kontext irreführend, da ohnehin alle Parameter als zufällig angenommen werden. Das allgemeine Konzept zufälliger Effekte kann dadurch aber sehr einfach in bayesianische Modelle integriert werden, indem man die für α_s angenommene Verteilung einfach als Priori-Verteilung spezifiziert:

$$\boldsymbol{\alpha} \sim \mathrm{N}(0, \tau^2 \boldsymbol{I}).$$

Für den Varianzparameter τ^2 wird eine Hyperpriori angenommen, in diesem Fall eine nicht- bzw. schwachinformative inverse Gamma-Verteilung $\text{IG}(a, b)$ mit $a = b = 0.001$. Für deren Dichte gilt:

$$p(\tau^2) \propto (\tau^2)^{-a-1} \exp\left(\frac{-b}{\tau^2}\right).$$

In BayesX sind mehrere Verfahren zur Schätzung der strukturellen räumlichen Komponente $f(x_{\text{district}})$ implementiert. Da die räumlichen Informationen hier nicht in Form von (quasistetigen) Gitterpunkten vorliegen, sondern in Form einer diskreten Landkreisstruktur, werden anstelle von zweidimensionalen P-Splines sogenannte Markov'sche Zufallsfelder (Markov Random Fields, MRF) verwendet. Diese entstammen der Theorie der stochastischen Prozesse und stellen eine diskrete zweidimensionale Irrfahrt dar, bei der Übergänge nur zu direkt benachbarten Landkreisen möglich sind. Ähnlich wie bei der zufälligen Komponente wird auch hier ein landkreisspezifischer Parameter in den linearen Prädiktor aufgenommen, das heißt $f(x_{\text{district}}) = \gamma_s$. Bei der Schätzung von γ_s soll berücksichtigt werden, dass benachbarte Landkreise ähnlicher sind als weiter voneinander entfernt liegende. Eine Möglichkeit, dies umzusetzen ist die Penalisierung von Differenzen in den Parametern benachbarter Landkreise. Hier werden Nachbarschaftsbeziehungen allerdings nicht durch Penalisierung berücksichtigt, sondern gemäß des MRF-Ansatzes durch die Annahme einer bedingten Priori für γ_s gegeben die Parameter in den benachbarten Landkreisen. Diese räumliche „Glättungspriori“ hat folgende Form:

$$\gamma_s | \gamma_{\tilde{s}}, \nu^2 \sim \text{N}\left(\frac{1}{N_s} \sum_{\tilde{s} \in \delta_s} \gamma_{\tilde{s}}, \frac{\nu^2}{N_s}\right) \quad (\tilde{s} \neq s).$$

Dabei ist N_s die Anzahl benachbarter Landkreise zu Landkreis s, δ_s eine Indexmenge, welche die Indizes aller benachbarten Landkreise beinhaltet, und ν^2 ein weiterer Hyperparameter, für den ebenso wie für den Varianzparameter des zufälligen Landkreiseffekts eine inverse Gamma-Verteilung mit Hyperparametern $a = b = 0.001$ angenommen wird.

Insgesamt kann der lineare Prädiktor für das räumliche bayesianische Regressionsmodell im vorliegenden Fall wie folgt formuliert werden:

$$\eta_{st} = \boldsymbol{x}_{st}^\top \boldsymbol{\beta} + f(t) + \alpha_s + \gamma_s.$$

Die Schätzung der so spezifizierten Modelle erfolgte mithilfe der BayesX-Funktion „bayesreg“ und beruht wiederum auf den vollbedingten Posteriori-Verteilungen aller unbekannten Parameter. Details zu den MCMC-Schätztechniken können dem Methodology Manual zur BayesX-Software (vgl. Belitz et al. (2009, Kap. 6)) sowie dem Buch von Fahrmeir und Kneib (2010) entnommen werden.

Der beschriebene bayesianische Ansatz wurde verwendet, da er eine intuitive Möglichkeit anbietet, die diskrete räumliche Datenstruktur in die Modellierung einzubeziehen. Zudem ist die Schätzung alternativer gemischter Modellansätze aufgrund der großen Datenmenge, insbesondere der großen Beobachtungszahl pro Landkreis, relativ instabil sowie extrem zeit- und rechenintensiv. Beispielsweise wurde versucht, mithilfe der R-Funktion „gamm“ aus dem Paket „mgcv“ (empirischer Bayes-Ansatz, vgl. Wood (2006)) ein äquivalentes gemischtes Modell zu schätzen, jedoch ohne räumlichen Term, nur mit entsprechendem Random Intercept. Der Rechenaufwand war dabei allerdings unverhältnismäßig höher als bei der vollbayesianischen BayesX-Prozedur.

Um die Kovariableneffekte unter Berücksichtigung der räumlichen Datenstruktur zu untersuchen, wurde der nach Verfahren c) reduzierte Datensatz herangezogen. Da sich bei den KVB-Call-Center-Daten auch nach Aggregation über alle Altersgruppen immer noch zu geringe Fallzahlen ergaben, um eine Normalverteilung für die $y_i|\boldsymbol{x}_i$ anzunehmen, wurde hier ein Poisson-GLMM verwendet. Zum Vergleich wurde ein Negativ-Binomial-GLMM gefittet, um eine mögliche Überdispersion feststellen zu können. Der Varianzparameter ν wurde auf $\hat{\nu} = 4.7686$ geschätzt, was für ein geringes Maß an Überdispersion spricht (vgl. Abschnitt 3.2.1). Dementsprechend unterscheiden sich die Parameterschätzer und Kredibilitätsintervalle des Poisson- und des Negativ-Binomial-Modells nicht besonders stark (vgl. Tabelle G.11) und die Verwendung des Poisson-GLMMs scheint durchaus gerechtfertigt. Zudem wurde versucht, ein ebenfalls in BayesX implementiertes Zero-Inflated-Poisson-Modell zu fitten, wobei allerdings keine Konvergenz erzielt werden konnte. Für die Abrechnungsdaten (Arztbesuche gesamt) wurde ein loglineares gemischtes Modell verwendet. Die Modellgleichungen von Call-Center-Modellen und Abrechnungsmodell unterscheiden sich, durch das Wegfallen der Alters- und Geschlechtsterme, nur durch den zusätzlichen Offset-Term für die landkreisspezifische Arztdichte im Abrechnungsmodell. Der Zeittrend zur Entkorrelierung der Fallzahlen an aufeinanderfolgenden Tagen konnte dementsprechend nur global geschätzt werden. Die dafür benötigten Basisfunktionen wurden wie in den vorherigen Modellen manuell konstruiert. Im Vergleich zu den bisher betrachteten Modellen wurde zusätzlich die landkreisspezifische Deprivation als Kovariable in den Prädiktor aufgenommen, die Meteorologie- und Luftqualitätskovariablen bleiben unverändert. Die vollständige Modellgleichung kann Abbildung G.12 entnommen werden.

Zur Simulation der jeweiligen Posteriori-Verteilung wurde in einem MCMC-Verfahren eine geeignete Markov-Kette konstruiert, aus der insgesamt 12000 Zufallszahlen gezogen wurden. Die ersten 2000 davon wurden als Burn-in verworfen. Von den 10000 verbliebenen Werten gelangte nur jeder 10. Wert in die Schätzung der Posteriori-Verteilung, so dass insgesamt 1000 Zufallszahlen zur Schätzung der Parameter verwendet wurden. Die Tabellen G.11 und G.12 zeigen die geschätzten Effekte (hier die Mittelwerte der Realisierungen aus der Posteriori-Verteilung, roh und exponentiell

transformiert) der Call-Center-Modelle und des Abrechnungsmodells zusammen mit 95%-Kredibilitätsintervallen. Letztere bestehen aus dem 2.5%- und 97.5%-Quantil der geschätzten Posteriori-Verteilung. Bei den effektcodierten Variablen wurden die Werte für die fehlende Kategorie ergänzt, indem für jedes der 1000 Samples gemäß der bekannten Nullrestriktion ein Schätzwert berechnet wurde. Aus all diesen Schätzwerten wurden dann der Mittelwert sowie die 2.5%- und 97.5%-Quantile gebildet. Bei den Cutpoint-Variablen sind nur die rohen Schätzwerte angegeben, da der Exponens nur in Summe mit dem Haupteffekt sinnvoll interpretierbar ist. Die geschätzten Meteorologie- und Luftqualitätseffekte der GLMMs sind von allen bisherigen Modellen als am aussagekräftigsten zu bewerten, da die Berücksichtigung der räumlichen Struktur für deren Beurteilung eine wichtige Rolle spielt.

Bei den KVB-Call-Center-Daten (vgl. Tabelle G.11) erweist sich zunächst die Deprivation als signifikanter Einflussfaktor auf die Anzahl der Anrufe. Laut Poisson-GLMM steigt diese um 2.33% bei Zunahme des Scores um 1. Die Signifikanzaussagen der administrativen Effekte decken sich weitgehend mit den Ergebnissen der GLMs für die Pilotregion München und für den über die Landkreise aggregierten Datensatz. Hier konnte zusätzlich eine signifikante Abnahme der Anruferzahl in der ersten Quartalswoche festgestellt werden. Bei den meteorologischen Kovariablen konnten signifikante Effekte für die Tagesranges von Temperatur (-0.36% pro Zunahme um $1°C$) und Luftdruck (-0.57% pro Zunahme um eine Einheit) sowie für den Bedeckungsgrad mittlerer Bewölkung (-0.53% pro Zunahme um 10%) nachgewiesen werden. Eine signifikante Änderung des Effekts trat bei den unteren Cutpoints von Temperatur ($1.8018°C$) und Luftfeuchtigkeit ($0.0038 kg_{\text{Wasser}}/kg_{\text{Luft}}$) sowie beim oberen Luftdruck-Cutpoint (11.4776 logarithmierter Oberflächendruck) auf. Die Windrichtungen Nordost (-2.47%) und Ost (-4.43%) werden mit einer signifikant niedrigeren Fallzahl assoziiert, die Windrichtungen Nordwest($+2.58\%$), West($+2.02\%$) und Südost($+4.11\%$) dagegen mit einer Zunahme der Anruferzahlen in den Landkreisen. Schwefeldioxid zeigte einen signifikant negativen Effekt (-0.67% pro Zunahme um $1\mu g/m^3$), Feinstaub ($+0.18\%$ pro Zunahme um $1\mu g/m^3$) und Ozon ($+0.16\%$ pro Zunahme um $1\mu g/m^3$) bewirken dagegen eine signifikante Zunahme der Fallzahl.

Bei den Abrechnungsdaten (vgl. Tabelle G.12) konnte kein signifikanter Deprivationseffekt nachgewiesen werden. Bei den administrativen Effekten ergeben sich wiederum keine Abweichungen in den Aussagen zu Signifikanz und Richtung der Effekte im Vergleich zu den GLMs für die Datensätze nach Reduktionsverfahren a) und b). Signifikant positive Effekte ergaben sich hier für den Bewölkungsgrad mittlerer Bewölkung ($+0.39\%$ pro Zunahme um 10%), die Temperatur ($+0.66\%$ pro Zunahme um $1°C$), den Surface Stress ($+0.36\%$ pro Zunahme um $1N/(m^2 s)$) und die Windgeschwindigkeit ($+0.39\%$ pro Zunahme um $1m/s$). Signifikant negative Effekte konnten beobachtet werden bei der Tagesrange der Luftfeuchtigkeit (-1.11% pro Zunahme um $1 kg_{\text{Wasser}}/kg_{\text{Luft}}$) und dem Luftdruck ($-0.0012\%$ pro Zunahme um eine Einheit).

Eine signifikante Änderung der linearen Effekte trat im oberen Luftfeuchtigkeits- ($0.0099 kg_{\text{Wasser}}/kg_{\text{Luft}}$) und im unteren Luftdruck-Cutpoint (11.4606 logarithmierter Oberflächendruck) ein. Den vorherrschenden Windrichtungen Nord (-1.66%), Südost (-1.86%) und West (-1.93%) wird ein protektiver Effekt im Bezug auf die Fallzahl unterstellt, bei den Windrichtungen Süd ($+1.01\%$) und Südwest ($+3.59\%$) dagegen treffen signifikant mehr Anrufe ein. Eine Zunahme von Feinstaub ($+0.08\%$ pro Zunahme um $1\mu g/m^3$) und Stickstoffdioxid ($+0.15\%$ pro Zunahme um $1\mu g/m^3$) bewirkt eine signifikante Zunahme der Anruferzahlen, Ozon (-0.09% pro Zunahme um $1\mu g/m^3$) und Kohlenstoffmonoxid (-3.19% pro Zunahme um $1\mu g/m^3$) besitzen dagegen einen nachweisbaren negativen Effekt.

Abbildung G.13 zeigt schließlich die bayesianisch geschätzten räumlichen Effekte des Poisson-GLMMs für die Call-Center-Daten (links) und des loglinearen GLMMs für die Abrechnungsdaten (rechts). Die obere Reihe zeigt die jeweilige zufällige Komponente $\hat{\alpha}_s$, die mittlere Reihe die jeweilige strukturelle Komponente $\hat{\gamma}_s$. Die Plots in der untersten Reihe zeigen Signifikanz und Richtung der strukturellen räumlichen Effekte an. -1 steht dabei für einen signifikant negativen Effekt, $+1$ für einen signifikant positiven Effekt. Die Beurteilung der Signifikanz erfolgte auf Basis von 95%-Posteriori-Kredibilitätsintervallen für $\hat{\gamma}_s$.

Zunächst fällt auf, dass sich bei der zufälligen Komponente im Vergleich zur strukturellen Komponente ein wesentlich ungleichmäßigeres Bild ergibt, was auf die Verwendung der Glättungspriori zurückzuführen ist. Die geschätzte Varianz der zufälligen Effekte, die als Maß für die Heterogenität der Fallzahlen in den Landkreisen betrachtet werden kann, beträgt 0.0263 bei den Call-Center-Daten und 0.0125 bei den Abrechnungsdaten. Die zufälligen Effekte rangieren bei den Call-Center-Daten von einer Abnahme der Fallzahl um 24.91% bis zu einer Zunahme um 38.90%, bei den Abrechnungsdaten von -14.19% bis $+5.74\%$. Der strukturelle räumliche Effekt ist in beiden Fällen, vor allem bei den Abrechnungsdaten stärker ausgeprägt als der zufällige Effekt. Bei den Call-Center-Daten variieren die räumlichen Effekte von -32.42% bis 45.37%, bei den Abrechnungsdaten zwischen -96.30% bis $+218.90\%$.

Das in Abschnitt 2.4 beschriebene Nordost-Südwest-Gefälle in den Call-Center-Rohdaten findet sich auch beim entsprechenden strukturellen räumlichen Effekt wieder. Die Deprivation, die ein ähnliches räumliches Muster aufweist, trägt zwar signifikant zur Erklärung der Anruferzahlen bei, jedoch verbleibt eine auf die geografische Lage zurückzuführende räumliche Struktur. Signifikant positive räumliche Effekte treten vor allem im Raum Nürnberg-Erlangen-Fürth auf, signifikant weniger Anrufe dagegen an der Grenze zwischen Schwaben und Oberbayern sowie in Teilen Unterfrankens. Das Nord-Süd-Gefälle bei den Arztbesuchen ist in der entsprechenden Grafik für den strukturellen räumlichen Effekt tendenziell erkennbar und nicht gänzlich durch die verwendeten Kovariablen zu erklären. Deutlicher wird die Diskrepanz bei der Betrachtung der Signifikanzen. In Südbayern treten vermehrt signifikant negative räumliche Effekte auf, in Nordbayern dagegen kommt es zu signifikant mehr

Anrufen.

Das beschriebene räumliche bayesianische Regressionsmodell wurde auch für die Trainingsmodelle zur Prognose herangezogen. Für das fortlaufend lernende Prognosemodell konnte es allerdings aus technischen Gründen und aufgrund der immer noch zu großen Rechenzeit nicht verwendet werden. Um die räumliche Datenstruktur jedoch auch in den Prognosemodellen zu berücksichtigen, wurde der im Trainingsmodell geschätzte strukturelle räumliche Effekt als Offset in die Modellgleichung der Prädiktionsmodelle einbezogen. Ein weiterer Grund für die Nichtverwendung des bayesianischen Ansatzes war, dass das in Abschnitt 4.4 vorgestellte Penalisierungsverfahren für die Distributed Lag Function in die Prädiktion miteinbezogen werden sollte. Dazu ist eine Penalisierung einzelner Teile der Designmatrix erforderlich, die mit der BayesX-Funktion „bayesreg“ so nicht möglich war.

3.6 Zusammenfassung der Modellergebnisse basierend auf den reduzierten Datensätzen

In diesem Abschnitt werden die bisherigen Ergebnisse der modellbasierten Analyse von Kovariableneffekten nochmals kurz zusammengefasst. Da die einzelnen Reduktionsmodelle, wie erwähnt, unterschiedliche Schwerpunkte bei der Berücksichtigung der Korrelationsstruktur in den Daten besitzen, ist es wichtig, die Gesamtheit der Schätzwerte für einen Parameter zu betrachten, um sich ein umfassendes Bild von seiner Wirkweise machen zu können. Allein durch die Tatsache, dass unterschiedlich reduzierte Datensätze zum Einsatz kommen, sind nicht übereinstimmende Aussagen relativ wahrscheinlich. Des Weiteren sind die Ergebnisse von Call-Center- und Abrechnungsmodellen nur bedingt vergleichbar, da beide Datensätze, wie in Abschnitt 2 beschrieben, unterschiedliche Eigenschaften aufweisen. Für Meteorologie- und Luftqualitätsparameter sind zwar grundsätzlich ähnliche Tendenzen zu erwarten, da beide Zielvariablen ein Surrogat für die Lungen-Morbidität darstellen, jedoch ist fraglich, ob die in den unterschiedlichen Datensätzen erfassten Fälle beispielsweise hinsichtlich der Schwere der Erkrankung vergleichbar sind. Des Weiteren gilt es zu beachten, dass bis jetzt keine verzögerten Kovariableneffekte in die Modellierung einbezogen wurden, die möglicherweise auch eine wichtige Rolle bei der Erklärung der Fallzahlen spielen. Dennoch soll im Folgenden versucht werden, gemeinsame Tendenzen, die über mehrere Modelle hinweg erkennbar sind, zu identifizieren.

Tabelle G.13 zeigt einen Überblick über Richtung $(-/+)$ und Signifikanz $(0/1)$ der Effekte aller gefitteten Modelle. Ergänzend bildet Tabelle G.14 die Verteilung der geschätzten Parameter ab, genauer gesagt, kleinste, größte und mittlere Effekte getrennt nach Call-Center- und Abrechnungsmodellen. Bei der Berechnung der mittleren Effekte erhielt jeder reduzierte Datensatz (a), b) und c)) das Gewicht 1. Dementsprechend wurden bei den KVB-Call-Center-Daten die Poisson-Modelle und

die parallel gefitteten Alternativmodelle entsprechend heruntergewichtet. Wie erwartet, fallen bei der Gegenüberstellung der Modellergebnisse für die unterschiedlichen Datensätze zahlreiche abweichende und zum Teil widersprüchliche Ergebnisse auf.

Die Deprivation erwies sich als signifikant positiver Einflussfaktor für die Anzahl der Anrufe beim KVB-Call-Center. Die administrativen Effekte sind stark an die Öffnung der Arztpraxen und die Quartalseinteilung gebunden und unterscheiden sich dementsprechend zwischen Call-Center- und Abrechnungsdaten. Insgesamt stehen die administrativen Effekte in Einklang mit den bereits in den Rohdaten beobachteten Tendenzen (vgl. Abbildung G.6) und erwiesen sich weitgehend als signifikant. Genauso ergaben sich nachweisbare, der deskriptiven Betrachtung (vgl. Tabellen G.1 und G.2) entsprechende Geschlechts- und Alterseffekte, wobei die Fallzahlen ab einem Alter von 60 Jahren überdurchschnittlich anwachsen. Bei den Abrechnungsdaten variiert der Geschlechtseffekt signifikant auf den Faktorstufen des Alters.

Keine signifikanten Effekte konnten für die Niederschlagsvariablen $x_{\text{cp.max}}$ und $x_{\text{lsp.max}}$ nachgewiesen werden. Vereinzelt ergaben sich signifikante Effekte für den Bedeckungsgrad niedriger und mittelhoher Bewölkung, z. B. eine signifikante Zunahme der Fallzahl mit wachsender mittelhoher Bewölkung bei den bayernweiten Abrechnungsmodellen. Der Luftfeuchtigkeit wird von den Abrechnungsmodellen teilweise ein signifikant negativer Effekt unterstellt. Der mittlere Effekt liegt bei einer Abnahme der Fallzahl von 1.26% pro Zunahme der Feuchtigkeit um $0.001 kg_{\text{Wasser}}/kg_{\text{Luft}}$. Bei den bayernweiten Call-Center-Modellen gibt es Hinweise auf eine Veränderung des Effekts der Luftfeuchtigkeit im unteren Cutpoint ($0.0038 kg_{\text{Wasser}}/kg_{\text{Luft}}$), bei den bayernweiten Abrechnungsmodellen gilt das gleiche für den oberen Cutpoint ($0.0091 kg_{\text{Wasser}}/kg_{\text{Luft}}$). Beim Luftdruck zeigt sich generell die Tendenz einer Veränderung des Effekts im unteren Wertebereich. Keine bis wenig Hinweise auf einen signifikanten Effekt ergaben sich beim Oberflächenwärmefluss und dem Surface Stress. Die Windgeschwindigkeit ist in den bayernweiten Abrechnungsmodellen positiv signifikant (im Mittel +0.79% Arztbesuche pro Zunahme um $1 m/s$). Ebenso zeigt die Temperatur einen signifikant positiven Effekt in den bayernweiten Abrechnungsmodellen (im Mittel +0.76% Arztbesuche pro Zunahme um $1°C$). Über mehrere Modelle hinweg besteht ein Hinweis auf eine Zunahme der Fallzahlen (bezogen auf den hier nicht tabellierten Gesamteffekt) unterhalb des unteren Temperatur-Cutpoints, der bei ca. $2°C$ liegt. Die Windrichtung spielt hauptsächlich in den bayernweiten Modellen eine Rolle. Die Richtung der Effekte variiert jedoch stark zwischen Call-Center- und Abrechnungsdatensatz.

Bei den Luftqualitätsparametern zeigte Feinstaub insgesamt einen positiven Effekt (im Mittel +0.14% Anrufe bzw. +0.03% Arztbesuche pro Zunahme um $1 \mu g/m^3$). Bei den Call-Center-Modellen erwies sich Ozon zum Teil als Risikofaktor für eine Zunahme der Anrufe (im Mittel +0.12% pro Zunahme um $1 \mu g/m^3$). Bezogen auf die Abrechnungsdaten ergaben sich negative Effekte für Ozon (im Mittel −0.18% pro Zunahme um $1 \mu g/m^3$) und Kohlenstoffmonoxid (im Mittel −7.15% pro Zunahme

um $1\mu g/m^3$), die möglicherweise auf die komplexe Korrelationsstruktur innerhalb der Meteorologie- und Luftqualitätseffekte zurückzuführen sind. Dagegen war eine signifikante Zunahme der Arztbesuche über alle Abrechnungsmodelle hinweg bei der Zunahme von Stickstoffdioxid (im Mittel +0.28% pro Zunahme um $1\mu g/m^3$) festzustellen.

3.7 Schrittweise Variablenselektionsverfahren

Um die Stabilität der bisher gefundenen Kovariableneffekte zu überprüfen und um im Hinblick auf die Prognose einen möglichst einfach strukturierten linearen Prädiktor zu erhalten, der nur die für die Fallzahlen relevanten Kovariableneffekte beinhaltet, wurden 2 verschiedene Variablenselektionsverfahren, basierend auf den nach Verfahren a) und b) reduzierten Datensätzen, durchgeführt. In diesem Abschnitt werden die Ergebnisse einer Standardmethode zur Variablenselektion, der schrittweisen AIC-Selektion, präsentiert (vgl. Burnham und Anderson (1998)). Das verwendete Verfahren basiert auf dem bereits in Abschnitt 3.4 zur Bruchpunktselektion verwendeten Akaike Informationskriterium zum Modellvergleich (vgl. (9) und Akaike (1974)). Für jeden der Datensätze wurde jeweils eine Vorwärts- und eine Rückwärtsselektion durchgeführt. Generell ist bei der AIC-Selektion zu beachten, dass ausgewählte Kovariablen nicht zwangsläufig einen signifkanten Effekt aufweisen müssen.

Bei der Vorwärtsselektion nimmt man im ersten Schritt, ausgehend von einem Basis-Modell, das alle festen Bestandteile des linearen Prädiktors enthält, jede der zur Verfügung stehenden Kovariablen einzeln in das Modell auf und berechnet das AIC aller sich ergebenden Modelle. Diejenige Kovariable, die zum AIC-optimalen Modell führt, wird dann in das Basismodell für den nächsten Schritt integriert, falls das AIC des resultierenden Modells geringer ist als das AIC des Ausgangsmodells. Dieses Verfahren wendet man solange an, bis das Ausgangsmodell durch Hinzunahme weiterer Kovariablen nicht mehr hinsichtlich des AIC verbessert werden kann.

Umgekehrt geht man bei der Rückwärtsselektion vom vollen Modell mit allen möglichen Kovariablen aus. Im ersten Schritt wird jede der zu selektierenden Kovariablen einzeln aus dem linearen Prädiktor entfernt. Anschließend wird diejenige Kovariable ausgeschlossen, die zum Modell mit dem höchsten AIC führt, falls das AIC des resultierenden Modells höher ist als das AIC des Ausgangsmodells. Das Verfahren bricht ab, wenn alle berechneten Kandidatenmodelle ein geringeres AIC besitzen als das Ausgangsmodell.

Die designbedingten und administrativen Kovariablen sowie die altersspezifischen Zeittrends wurden als Bestandteile des Basis-Modells betrachtet, dessen Komponenten nicht der Variablenselektion unterzogen wurden. Für das volle Modell wurde der lineare Prädiktor der in Abschnitt 3.2.1 vorgestellten Modelle um zusätzliche Mo-

dellterme zur Berücksichtigung verzögerter Effekte von Meteorologie- und Luftqualitätsparametern erweitert. Da man den Meteorologie-Parametern einen vergleichsweise kurzfristigen Effekt auf den menschlichen Organismus (vgl. z. B. Kassomenos et al. (2007)) von bis zu 3 Tagen unterstellt, wurden jeweils nur die ersten drei Lags x_{t-1}, x_{t-2} und x_{t-3} verwendet. Bei den Luftqualitätsparametern dagegen rechnet man eher mit einem länger anhaltenden Effekt von bis zu 2 Wochen (vgl. z. B. Welty et al. (2009)). Um die Anzahl der Modellparameter nicht zu stark anwachsen zu lassen, wurden hier anstelle der einzelnen Lags $x_{t-1}, \ldots, x_{t-14}$ durchschnittliche Lag-Kovariablen x_{st}, x_{mt} und x_{lt} verwendet, um die Existenz kurz-, mittel- und langfristiger Effekte zu ergründen:

$$\begin{aligned} x_{\text{st}} &= (x_{t-1} + \ldots + x_{t-3})/3, \\ x_{\text{mt}} &= (x_{t-4} + \ldots + x_{t-6})/3, \\ x_{\text{lt}} &= (x_{t-7} + \ldots + x_{t-14})/8. \end{aligned}$$

Wie in Abschnitt 4.1 beschrieben, handelt es sich bei diesem Verfahren um einen eher naiven Ansatz zur Berücksichtigung verzögerter Kovariableneffekte. Dieser wird dennoch angewandt, da an dieser Stelle nur ein erster Eindruck davon gewonnen werden soll, ob überhaupt Lag-Effekte existieren und wie weit diese in die Vergangenheit zurückreichen. Detaillierte Analysen basierend auf geeigneteren Modellierungsansätzen für zeitlich verzögerte Effekte werden im weiteren Verlauf der Arbeit (vgl. Abschnitt 4) präsentiert.

Nach Durchführung der schrittweisen AIC-Selektion wurde eine logische Korrektur bezüglich der Lag- und Cutpoint-Variablen durchgeführt. Falls beispielsweise für eine Kovariable nur das 2. Lag ausgewählt wurde, wurden das 1. Lag und der zugehörige Haupteffekt wieder ins Modell aufgenommen. Ebenso wurde bei den Cutpoints verfahren. Die CD in Appendix I (Ordner „Variablenselektion“) enthält für alle Datensätze eine vollständige Zusammenfassung von Schätz- und p-Werten der vollen Modelle, der Selektionsmodelle und der korrigierten Selektionsmodelle. Die Schätzwerte der Meteorologie-Parameter wurden dabei nicht, wie in den bisher gezeigten Schätzertabellen, mit 10er-Potenzen multipliziert. Durch die Korrektur wieder aufgenommene Terme sind grün markiert, Komponenten des Basis-Modells blau.

Tabelle G.15 bietet einen zusammenfassenden Überblick über die Signifikanz und Richtung der Effekte in den korrigierten Modellen. Dabei steht $-1/+1$ für einen signifikant negativen/positiven Effekt und $-0/+0$ für nichtsignifikante (negative/positive) Effekte. Außerdem wird in Klammern das maximal ins Modell aufgenommene Lag sowie das maximal signifikante Lag angegeben, um Existenz und Dauer zeitverzögerter Kovariableneffekte festzustellen. Für die KVB-Call-Center-Daten wurden aufgrund der geringen Überdispersion (vgl. Abschnitt 3.2.1) Poisson-Modelle zur Variablenselektion verwendet, für alle übrigen Datensätze loglineare Modelle. Die Kategorien der Variable Windrichtung sind in der Tabelle zwar einzeln aufgeführt,

in der Variablenselektion wurde der Faktor jedoch als Ganzes betrachtet und entweder komplett aufgenommen oder entfernt. Ebenfalls in der Tabelle angegeben ist die Parameterzahl sowie das AIC der Modelle. Letzteres kann zum Vergleich zwischen dem jeweiligen Vorwärts- und Rückwärtsselektionsmodell verwendet werden, allerdings nicht zum Vergleich der auf unterschiedlichen Datensätzen beruhenden Modelle.

Die korrigierten Call-Center-Modelle enthalten durchgehend die Temperatur und den Luftdruck zusammen mit den entsprechenden unteren Cutpoints sowie die Luftfeuchtigkeit und den Oberflächenwärmefluss. Häufig gelangten auch der großskalige Niederschlag, die Bewölkungsvariablen, der Surface Stress, die Windgeschwindigkeit und die Tagesranges von Temperatur und Luftdruck in die resultierenden Modelle. Viele dieser Parameter wurden allerdings nur aufgrund von verzögerten Kovariableneffekten ergänzt. Signifikante Lag-Effekte ergaben sich für die Variablen Luftfeuchtigkeit (bis Lag 1), Luftdruck (bis Lag 3), Oberflächenwärmefluss (bis Lag 3), Surface Stress (bis Lag 3), Temperatur (bis Lag 2) und Windgeschwindigkeit (bis Lag 3). Von den Luftqualitätsparametern wurde nur Stickstoffdioxid (in der Pilotregion München) und Feinstaub (im bayernweiten Modell) aufgrund der Signifikanz des Haupteffekts in die finale Variablenauswahl aufgenommen. Die Aufnahme der übrigen Luftqualitätsparameter resultiert ebenfalls aus der Existenz von in der Regel langfristigen und größtenteils signifikanten Lag-Effekten.

Bei den Abrechnungsdaten wurden die Niederschlags- und Bewölkungsvariablen sowie Luftdruck, Temperatur (jeweils mit unterem Cutpoint) und Luftfeuchtigkeit ohne Ausnahme in die Modelle aufgenommen. In den meisten Selektionsmodellen sind auch Surface Stress und Windgeschwindigkeit sowie die verbleibenden Cutpoint-Variablen vertreten. Die Berücksichtigung der Haupteffekte beruht wiederum zum großen Teil auf der logischen Korrektur bezüglich verzögerter Kovariableneffekte. Signifikante Lag-Effekte traten bei den Niederschlags- und Bewölkungsvariablen (jeweils bis Lag 3), der Luftfeuchtigkeit (bis Lag 3), dem Luftdruck (bis Lag 2), dem Oberflächenwärmefluss (bis Lag 3), dem Surface Stress (bis Lag 3), der Temperatur (bis Lag 2) und der Windgeschwindigkeit (bis Lag 2) auf. Die Luftqualitätsparameter wurden komplett in die korrigierten Selektionsmodelle aufgenommen und weisen, neben meist signifikanten Haupteffekten, nachweisbare mittel- bis langfristige Lag-Effekte auf.

Bei beiden Datenquellen wurde die Windrichtung in die bayernweiten Selektionsmodelle aufgenommen, in die Selektionsmodelle der Pilotregion München dagegen nicht. Die Ergebnisse von korrigierten Vorwärts- und Rückwärtsselektionsmodellen sind grundsätzlich relativ ähnlich, auch im Bezug auf Parameterzahl und AIC. Mit Ausnahme des korrigierten Selektionsmodells für die Abrechnungsdaten in der Pilotregion München sind die Vorwärtsselektionsmodelle etwas sparsamer. Bei den KVB-Call-Center-Daten weisen die korrigierten Vorwärtsselektionsmodelle das geringere AIC auf, bei den Abrechnungsdaten die korrigierten Rückwärtsselektionsmodelle.

3.8 Implizite Variablenselektion mithilfe von Shrinkage Priors (Bayesianische Lasso- und Ridge-Regression)

Eine weitere Möglichkeit, (implizit) wichtige Kovariablen zu selektieren, bieten sogenannte Shrinkage-Modelle. Deren Idee ist es, den Wert der Regressionsparameter $\boldsymbol{\beta}$ bzw. einer Teilmenge $\tilde{\boldsymbol{\beta}}$ in der Summe zu beschränken, damit sich essentielle Kovariableneffekte deutlicher gegenüber nicht relevanten abzeichnen. Man bezeichnet diese Vorgehensweise auch als Regularisierung. Bei der Lasso-Regression („least absolute shrinkage and selection operator“, vgl. Tibshirani (1996)) beschränkt man die Summe der absoluten Regressionskoeffizienten, bei der Ridge-Regression (vgl. Hoerl und Kennard (2000)) die Summe der quadratischen Regressionskoeffizienten. Die zugehörigen Restriktionen lauten also

$$\sum_{j=1}^{p} |\beta_j| \leq c \quad \text{bzw.} \quad \sum_{j=1}^{p} \beta_j^2 \leq c.$$

In der frequentistischen Inferenz beruht die Schätzung von generalisierten Shrinkage-Modellen auf der Maximierung der Log-Likelihood unter diesen Nebenbedingungen. Äquivalent kann das Maximierungsproblem, beispielsweise bei der Lasso-Regression, auch wie folgt formuliert werden:

$$l(\boldsymbol{\beta}) - \lambda \sum_{j=1}^{p} |\beta_j| \quad \rightarrow \quad \max_{\boldsymbol{\beta}}.$$

In dieser Darstellung ist zu erkennen, dass die Beschränkung der Koeffizienten einer Penalisierung großer, in diesem Fall absoluter, Koeffizienten entspricht. Im Lasso-Modell wird bei der frequentistischen Schätzung ein Teil der Koeffizienten direkt auf 0 geshrinkt, also im Kontext der Variablenselektion implizit ausgeschlossen. Bei der Ridge-Regression kann der Einfluss der Variablen beispielsweise mithilfe der aus den geshrinkten Koeffizienten resultierenden p-Werte beurteilt werden. Um jeder Kovariable das gleiche Gewicht zu geben, ist eine vorangehende Standardisierung notwendig, so dass alle Kovariablen um die 0 zentriert sind und Varianz 1 besitzen. Ein wesentlicher Vorteil solcher Shrinkage-Modelle besteht darin, dass sie auch dann eingesetzt werden können, wenn die Zahl der Kovariablen p die Anzahl der Beobachtungen n übersteigt.

In bayesianischen Regressionsmodellen wird das Shrinkage-Konzept nicht durch Penalisierung der Parameter umgesetzt, sondern durch Verwendung informativer Priori-Verteilungen (vgl. Griffin und Brown (2005), Fahrmeir und Kneib (2010)). Für die Regressionskoeffizienten $\boldsymbol{\beta}$ bzw. $\tilde{\boldsymbol{\beta}}$ nimmt man zunächst wie in nicht regularisierten Modellen eine Normalverteilung an:

$$\tilde{\beta}_j | \tau_j^2 \sim \mathrm{N}(0, \tau_j^2).$$

Das Parameter-Shrinkage wird durch die Priori-Annahme für den Varianzparameter τ_j^2 gesteuert. Anstelle von $\tau_j^2 = \infty$, was zu einer nichtinformativen „flachen“ Priori für $\tilde{\beta}_j$ führt, wird für τ_j^2 beim bayesianischen Lasso-Ansatz folgende Priori-Verteilungsannahme in Abhängigkeit vom Shrinkage-Parameter λ getroffen:

$$\tau_j^2|\lambda^2 \sim \text{Exp}(\lambda^2/2).$$

Für λ^2 wiederum wird eine Gamma-Priori der Form $\text{Ga}(0.001, 0.001)$ spezifiziert. Bei großem λ erhalten auf diese Weise kleine τ_j^2 (und damit betragsmäßig kleine $\tilde{\beta}_j$) eine größere Priori-Wahrscheinlichkeit.

In gleicher Weise spezifiziert man im bayesianischen Ridge-Modell

$$p(\tau_j^2|\lambda) = \begin{cases} 1 & \text{für } \tau_j^2 = 1/(2\lambda) \\ 0 & \text{für } \tau_j^2 \neq 1/(2\lambda) \end{cases},$$

also Punktmassen in $1/(2\lambda)$. Für λ wird wiederum eine nichtinformative Gamma-Verteilung $\text{Ga}(0.001, 0.001)$ angenommen. Große λ-Werte führen so zu einer kleineren Priori-Varianz der $\tilde{\beta}_j$.

Eine weitere Möglichkeit der Regularisierung besteht in der Spezifikation einer Mischung inverser Gamma-Verteilungen als Priori für τ_j^2. Weitere Details hierzu sowie zu Theorie und Schätzung bayesianischer Lasso- und Ridge-Modelle, können im Methodology Manual zur BayesX-Software (vgl. Belitz et al. (2009)) nachgelesen werden.

Zur Kontrolle der schrittweisen AIC-Selektionsmodelle wurden für die nach Schema a) und b) reduzierten Datensätze jeweils Lasso- und Ridge-Shrinkage-Modelle mit der gleichen Kovariablenkonfiguration gefittet. Wie im vorhergehenden Abschnitt wurden auch hier die administrativen und designbedingten Effekte von der Variablenselektion ausgenommen, also nicht in den Vektor der zu shrinkenden Parameter $\tilde{\boldsymbol{\beta}}$ einbezogen. Die Schätzung der Parameter erfolgte durch Gibbs-Sampling. Es wurden wiederum 12000 Zufallszahlen gezogen, die ersten 2000 verworfen und nur jede zehnte der verbleibenden 10000 in die Posteriori-Verteilung einbezogen. Da hier anstelle des Posteriori-Modus das arithmetische Mittel zur Parameterschätzung verwendet wurde, wurden beim Lasso-Modell die Effekte unwichtiger Kovariablen nicht exakt auf 0 geshrinkt (vgl. Fahrmeir und Kneib (2010)), was im frequentistischen Lasso-Modell der Fall ist. Für die effektcodierten Variablen wurden die Schätzwerte der jeweils fehlenden Kategorie wie bei den bayesianischen GLMMs (vgl. Abschnitt 3.5) berechnet.

Um vor dem Hintergrund der Variablenselektion ein Vergleichsmaß für das Gewicht der Kovariableneffekte zu erhalten, wurde für jeden Parameter die sogenannte „posterior probability“, eine bayesianische Variante des p-Werts berechnet. Diese ergibt sich für β_j aus

$$2 \min \left\{ \hat{\mathbb{P}}(\beta_j < 0), \hat{\mathbb{P}}(\beta_j > 0) \right\}.$$

Die enthaltenen Wahrscheinlichkeiten werden aus dem Anteil positiver bzw. negativer Posteriori-Realisierungen von β_j geschätzt. Kleine posterior probabilities stehen dementsprechend für einen stark ausgeprägten Effekt. Um wie bei den schrittweisen Selektionsmodellen ein Ergebnismodell zu erhalten, wurden alle Variablen mit einer posterior probability größer als 0.05 aus dem jeweiligen Modell ausgeschlossen, wobei wie in Abschnitt 3.7 eine logische Korrektur bezüglich der verzögerten Kovariableneffekte und der Cutpoints erfolgte. Mit den verbleibenden Kovariablen wurden dann bayesianische GLMs (ohne Shrinkage) gefittet.

Die vollständigen Ergebnisse der Shrinkage-Modelle finden sich auf der CD in Appendix I im Ordner „Variablenselektion“. Diese wurden analog zu den Ergebnissen der schrittweisen Selektion aufbereitet. Allerdings gilt es zu beachten, dass die Parameterschätzwerte nicht vergleichbar sind, da für die Shrinkage-Modelle eine Standardisierung der stetigen Kovariablen durchgeführt werden musste. Die Parameter in den Lasso-Modellen werden tendenziell etwas stärker geshrinkt als in den Ridge-Modellen. Insgesamt sind die Ergebnisse von Lasso- und Ridge-Regression, vor allem im Bezug auf Signifikanzaussagen und den Ausschluss von Variablen, durchaus vergleichbar. Tabelle G.16 bietet einen zusammenfassenden Überblick über die ausgewählten Meteorologie- und Luftqualitätsparameter der Shrinkage-basierten Variablenselektion und die Signifikanz der zugehörigen Effekte. $-1/+1$ steht wiederum für einen signifikant negativen/positiven Effekt und $-0/+0$ für nichtsignifikante (negative/positive) Effekte. Die eingeklammerten Werte geben Aufschluss darüber, bis zu welchem Lag die verzögerten Kovariableneffekte in die korrigierten Modelle aufgenommen wurden und bis zu welchem Lag ein signifikanter Einfluss vorliegt. Vergleicht man die Ergebnisse von Shrinkage-basierter Variablenselektion und schrittweiser AIC-Selektion ergeben sich sowohl übereinstimmende Ergebnisse, wie beispielsweise die nur in den bayernweiten Modellen ausgewählte Windrichtung, als auch Unterschiede. Für die Konstruktion der Prognosemodelle in Abschnitt 5 wurden daher die Resultate beider Verfahren berücksichtigt.

Bei den Call-Center-Daten wurde nur der Luftdruck aufgrund der Signifikanz des dritten Lags durchgehend in die Modelle aufgenommen. In der Pilotregion München gelangten der Surface Stress und die Windgeschwindigkeit ebenfalls aufgrund der Signifikanz des dritten Lags in die korrigierten Modelle. Von den Luftschadstoffen wurden Schwefel- und Stickstoffdioxid häufig ausgewählt. Feinstaub und Kohlenstoffmonoxid weisen einen signifikanten Haupteffekt im resultierenden Modell für den über die bayerischen Landkreise aggregierten Datensatz auf. Signifikante Lag-Effekte treten für alle erwähnten Kovariablen mit Ausnahme von Feinstaub auf.

Die korrigierten Abrechnungsmodelle enthalten grundsätzlich deutlich mehr Variablen als die Call-Center-Modelle. Ohne Ausnahme wurden Luftdruck, Luftfeuchtigkeit und Temperatur sowie die Bewölkungsvariablen, der großskalige Niederschlag und der Oberflächenwärmefluss ausgewählt. Bis auf Schwefeldioxid in der Pilotregion München enthalten die resultierenden Modelle alle Luftqualitätsparameter. Signifi-

kante Lag-Effekte besitzen in der Pilotregion München beispielsweise der großskalige Niederschlag (bis Lag 3), die mittelhohe Bewölkung (bis Lag 3), der Luftdruck (bis Lag 1) und der Oberflächenwärmefluss (bis Lag 3). Ebenso besitzt Feinstaub einen mittelfristigen Lag-Effekt. Für den über ganz Bayern aggregierten Datensatz ergaben sich zusätzliche signifikante Lag-Effekte für die Luftfeuchtigkeit (bis Lag 2), die niedrige Bewölkung (bis Lag 3), den Surface Stress (bis Lag 2) und die Temperatur (bis Lag 2). Alle Luftqualitätsparameter besitzen hier einen signifikanten, mittelfristig verzögerten Kovariableneffekt.

Wie bereits bei der Zusammenfassung der Kovariableneffekte in Abschnitt 3.6 beobachtet, weichen auch bei der Variablenselektion die Modellergebnisse, basierend auf den unterschiedlichen Datensätzen, zum großen Teil stark voneinander ab. Die Kovariablenkonfiguration der Prognosemodelle (vgl. Tabelle G.18) beruht auf den Shrinkage- und AIC-Selektionsergebnissen in der Pilotregion München. Dies begründet sich zum einen in der besseren Kovariablenqualität in der Pilotregion München im Vergleich zu den über die bayerischen Landkreise aggregierten Datensätzen. Zum anderen ist durch die kleinräumige Betrachtung, gerade im Hinblick auf das hochdynamische Verhalten der Luftschadstoffe, eher gewährleistet, dass die tatsächliche Exposition der Bevölkerung erfasst wurde.

4 Analyse von verzögerten Kovariableneffekten

Dieser Abschnitt befasst sich mit der Modellierung verzögerter Kovariableneffekte und deren Einbettung in die in Abschnitt 3 beschriebenen Modelle. Nach einer kurzen Einführung werden zunächst einige naheliegende Modellierungsansätze erläutert, die allerdings nicht zu einer befriedigenden Lösung führen (vgl. Abschnitt 4.1). Im darauffolgenden Abschnitt wird ein bayesianischer Ansatz nach Welty et al. (2009) diskutiert. Anschließend werden das sogenannte Almon-Lag-Modell (vgl. Almon (1965)) und eine selbst entwickelte P-Spline-basierte Methode zur Berücksichtigung von Lag-Effekten vorgestellt. Beide Verfahren wurden in der R-Funktion „lag_regress“ implementiert, die in Abschnitt 4.5 präsentiert wird. Die beschriebenen Methoden werden hinsichtlich ihrer Vor- und Nachteile untersucht und anhand von Ergebnissen, basierend auf den nach Schema a) und b) reduzierten Datensätzen, veranschaulicht.

Im Kontext longitudinaler Daten bezeichnet man den Einfluss β_{t-l} der Kovariable x_{t-l} auf die Zielvariable y_t als Lag-Effekt, die Zeitverschiebung wird als Lag l bezeichnet. In der Regel möchte man herausfinden, mit welcher Zeitverzögerung solche Effekte auftreten, bei welchem Lag der stärkste verzögerte Effekt auftritt und bis zu welchem Lag der Einfluss der Kovariable nachgewiesen werden kann. Die Distributed Lag Function (DLF) beantwortet all diese Fragen zur Form der verzögerten Wirkung der Kovariable, indem die Lags gegen die jeweils geschätzten Lag-Effekte geplottet werden. Es gilt zu beachten, dass die DLF streng genommen gar keine Funktion darstellt, sondern nur eine diskrete Folge von geschätzten Koeffizienten. Um den zeitlichen Verlaufscharakter optisch zu unterstreichen, werden die Punkte jedoch in der Regel miteinander verbunden. Abbildung 4.1 zeigt typische Verläufe von DLFs. Bei Funktion a) tritt der größte Einfluss noch am Tag der Beobachtung des Kovariablenwerts auf. Danach tendiert der Effekt kontinuierlich gegen 0. Bei Funktion b) dagegen wird der maximale Effekt erst bei Lag 2 erreicht. Genauso sind negative verzögerte Kovariableneffekte denkbar (Funktionen c) und d)).

Ein Problem, das in diesem Kontext häufig auftritt, sind sogenannte Harvesting- oder Rebound-Effekte, insbesondere wenn die Mortalität als Zielgröße untersucht wird (vgl. Zanobetti et al. (2000)). Dabei geht man davon aus, dass es einen bestimmten Pool an besonders gefährdeten Individuen gibt. Wächst der Wert eines externen Risikofaktors über mehrere Zeiteinheiten hinweg an (wie z. B. bei einer Hitzewelle), stirbt ein Großteil dieser Individuen bereits nach kurzer Zeit. Da der Risiko-Pool nach wenigen Lags dementsprechend leer ist, ist das Risiko in der untersuchten Population insgesamt kleiner, so dass die mittelfristigen Effekte tendenziell unterschätzt werden. Im Extremfall wird Schadstoffen sogar fälschlicherweise ein protektiver Effekt im mittleren Lag-Bereich zugeschrieben. Gegen Ende des Lag-Bereichs, nachdem sich der Risiko-Pool wieder langsam gefüllt hat, kann man wieder von einer unverzerrten Schätzung der Lag-Koeffizienten ausgehen. Abbildung 4.2

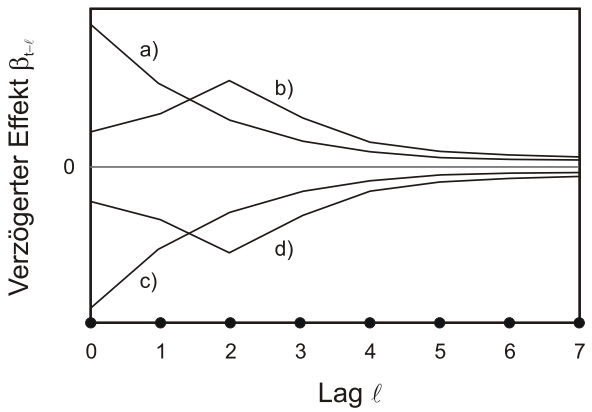


Abbildung 4.1: Typische Verläufe von DLFs

stellt den geschätzten Verlauf einer DLF beim Auftreten von Harvesting-Effekten (Funktion a)) im Kontrast zur wahren DLF (Funktion b)) dar. Bei der Untersuchung der Morbidität ist es fraglich, ob solche Effekte eine Rolle spielen, da ein Patient, der aufgrund einer akuten Erkrankung einen Arzt aufsucht oder beim KVB-Call-Center anruft, dies an den Folgetagen erneut tun kann und somit nicht aus dem Risiko-Pool fällt. Um diese Frage endgültig zu beantworten, müsste man die Wahrscheinlichkeit eines zweiten Ereignisses beim gleichen Patient an den Tagen nach dem ersten Ereignis untersuchen.

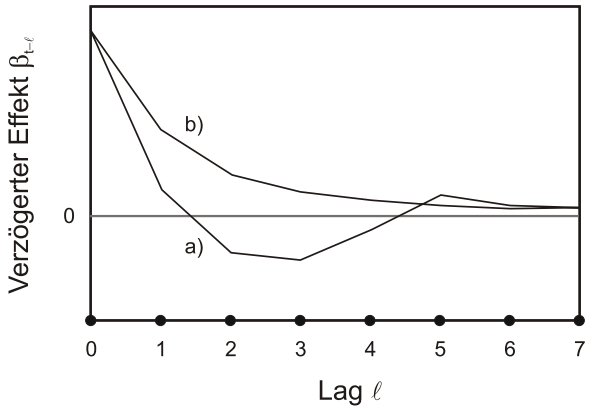


Abbildung 4.2: Verlauf der DLF beim Auftreten von Harvesting-Effekten (a)) im Vergleich zur wahren DLF (b))

Bei der Modellierung der DLF unterscheidet man zwischen finiten und infiniten Lag-Modellen. Finite Lag-Modelle unterstellen den Kovariablen einen zeitlich begrenzten Einfluss, das bedeutet die DLF wird ab einem zuvor definierten maxi-

malen Lag L auf 0 gesetzt. Bei infiniten Lag-Modellen geht man davon aus, dass der Lag-Effekt mit wachsendem Lag gegen 0 tendiert, jedoch wird der Grenzwert im Endlichen nicht angenommen. Dementsprechend ergeben sich Unterschiede in der Parametrisierung von finiten und infiniten Distributed Lag Modellen. In dieser Arbeit werden hauptsächlich finite Lag-Modelle betrachtet. Analog zu vorhergehenden Studien wurden, wie in Abschnitt 3.7 erläutert, Lag 3 als maximales Lag für die Meteorologie-Parameter und Lag 14 als maximales Lag für die Luftschadstoffe festgelegt.

Oftmals ist bei der Modellierung von Lag-Effekten auch der kumulative Effekt $\beta^* = \beta_t + \beta_{t-1} + \ldots + \beta_{t-L}$ einer Kovariable über alle betrachteten Lags hinweg von Interesse. Bei den betrachteten Zähldaten-Modellen mit log-Link lässt sich β^* wie folgt interpretieren: Eine Zunahme der Kovariable x zum Zeitpunkt t um eine Einheit bewirkt eine Zu-/Abnahme der Gesamtfallzahl im Zeitraum $t, \ldots, t+L$ um den multiplikativen Faktor $\exp(\beta^*)$. Tabelle 4.1 zeigt die kumulativen Effekte (roh und exponentiell transformiert) aller Meteorologie- und Luftqualitätsparameter für Call-Center- und Abrechnungsdaten in der Pilotregion München und aggregiert über die bayerischen Landkreise, geschätzt mithilfe des in Abschnitt 4.1 beschriebenen unrestringierten Modells. Die verwendete Kovariablenkonfiguration ist identisch zu den in Abschnitt 3.2.1 beschriebenen GLMs. Die Berechnung der angegebenen p-Werte beruht auf der geschätzten Varianz

$$\widehat{\mathrm{Var}}(\hat{\beta}^*) = \sum_{l=0}^{L} \widehat{\mathrm{Var}}(\beta_l) + 2 \sum_{l=0}^{L} \sum_{l \neq \tilde{l}} \widehat{\mathrm{Cov}}(\beta_l, \beta_{\tilde{l}}).$$

Kovariable	CC a)	CC b)	AB a)	AB b)
cp.max (0.001)	0.9960/0.4605	1.0249/0.0001	1.0079/0.4160	0.9965/0.7326
lsp.max (0.001)	1.0041/0.1798	0.9982/0.7013	0.9959/0.5124	1.0014/0.8591
lcc.ave (0.1)	0.9975/0.6001	1.0118/0.0077	1.0016/0.8399	0.9885/0.1066
mcc.ave (0.1)	1.0081/0.1177	1.0244/0.0000	0.9879/0.1614	0.9940/0.4108
q.ave (0.001)	0.9515/0.0003	0.9482/0.0000	0.9837/0.4940	1.0116/0.5415
qnh.ave (Mü.)	1.0023/0.1900	–	1.0033/0.1221	–
qnh.ave (0.001)	–	1.0027/0.0208	–	1.0051/0.0022
sshf.ave (100000)	0.9952/0.0021	0.9937/0.0009	1.0024/0.3808	1.0064/0.0348
stressspd.max (10000)	1.0014/0.8511	0.9854/0.2998	0.9966/0.8216	1.0089/0.6880
tmt.ave	1.0196/0.0003	1.0302/0.0000	1.0108/0.2825	0.9903/0.2221
windspd.max	1.0079/0.2600	1.0394/0.0078	0.9796/0.1695	0.9697/0.1598
SO2.q95	0.9781/0.0000	1.0023/0.6538	1.0004/0.9633	0.9800/0.0281
PM10.q95	1.0039/0.0000	1.0030/0.0056	0.9982/0.2518	0.9979/0.1885
O3.q95	0.9983/0.1478	0.9978/0.0198	0.9991/0.6456	1.0018/0.2526
NO2.q95	1.0020/0.0723	1.0002/0.8959	1.0016/0.3698	1.0018/0.4885
CO.q95	0.8801/0.1738	0.8895/0.2983	1.0194/0.9038	1.3740/0.0603

Tabelle 4.1: Kumulative Effekte der Meteorologie- und Luftqualitätsparameter für Call-Center- (CC) und Abrechnungsdaten (AB) in der Pilotregion München (a)) und aggregiert über alle Landkreise (b)): exponentiell transformierte Schätzwerte/p-Wert

Bei den KVB-Call-Center-Daten traten zahlreiche signifikante kumulative Effekte auf. Ein Anstieg der Luftfeuchtigkeit um $0.001 kg_{\text{Wasser}}/kg_{\text{Luft}}$ verursacht z. B.

eine Abnahme der Gesamtfallzahl an den nächsten 3 Tagen um 4.85% (Pilotregion München) bzw. 5.18% (Aggregation über Gesamtbayern). Eine Zunahme der Feinstaub-Konzentration um $1\mu g/m^3$ bewirkt eine Zunahme der Gesamtfallzahl in den nächsten beiden Wochen um 0.39% (München) bzw. 0.30% (Bayern). Weitere signifikante kumulative Effekte konnten etwa für den Oberflächenwärmefluss oder die Temperatur nachgewiesen werden. Bei den Abrechnungsdaten ergaben sich insgesamt deutlich weniger signifikante kumulative Effekte. Zum Beispiel besitzt der Luftdruck bei der bayernweiten Betrachtung einen signifikant positiven Effekt auf die Gesamtfallzahl der nächsten 3 Tage (+0.51% Arztbesuche pro Zunahme des Luftdrucks um eine Einheit).

4.1 Einfache Ansätze zur Modellierung der Distributed Lag Function

Die einfachste Möglichkeit, verzögerte Kovariableneffekte in die beschriebenen GLMs einzubeziehen, ist es, die gelagten Kovariablen $x_t, \ldots, x_{t-L}$ direkt in den linearen Prädiktor aufzunehmen und die zugehörigen Koeffizienten $\beta_t + \ldots + \beta_{t-L}$ explizit zu schätzen. Die Modellgleichung lautet dann schematisch:

$$\log(\mu_t) = \ldots + x_t\beta_t + x_{t-1}\beta_{t-1} + \ldots + x_{t-L}\beta_{t-L} + \ldots \tag{10}$$

Für die Parameter werden hier keinerlei Restriktionen getroffen. Abbildung 4.3 zeigt die resultierenden DLFs für Temperatur und Kohlenstoffmonoxid am Beispiel der KVB-Call-Center-Daten in der Pilotregion München zusammen mit punktweisen Konfidenzbändern basierend auf $\widehat{\mathrm{Var}}(\hat{\beta}_l)$. Der maximale Kovariableneffekt wird jeweils einen Tag nach Beobachtung des entsprechenden Kovariablenwerts beobachtet. Ein signifikanter Lag-Effekt ergibt sich im gefitteten Poisson-GLM nur für das 1. Lag von Kohlenstoffmonoxid ($\beta_1 = 0.1589$). Das bedeutet, die Anruferzahl steigt bei Zunahme der Kohlenstoffmonoxid-Konzentration am Vortag um $1\mu g/m^3$ um den multiplikativen Faktor $\exp(\beta_1) = 1.1722$.

Diese Methode ermöglicht zwar eine sehr flexible Form der DLF, jedoch ergibt sich auch eine Reihe von Problemen. Für die Temperatur mit maximalem Lag $L = 3$ erscheint das verwendete Verfahren noch halbwegs geeignet. Bei Betrachtung der geschätzten DLF für Kohlenstoffmonoxid in Abbildung 4.3 fällt jedoch auf, dass die Funktion sehr unruhig verläuft und mehrfach zwischen positivem und negativem Effektbereich alterniert. Dieses Muster ist auf eine hohe serielle Korrelation in der Kovariablenzeitreihe zurückzuführen. Speziell beim Beispiel Kohlenstoffmonoxid in der Pilotregion München liegen alle empirischen (Auto-)Korrelationen ρ_l zwischen x_t und x_{t-l} oberhalb der aus der asymptotischen Verteilung $\rho_l \overset{a}{\sim} \mathrm{N}(0, 1/n)$ berechneten Signifikanzgrenze von 0.0592 (vgl. Tabelle 4.2). Aus diesem Grund ist die Schätzung aufeinanderfolgender Lag-Koeffizienten relativ instabil, so dass der wahre Verlauf der verzögerten Effekte in der DLF nicht zu erkennen ist.

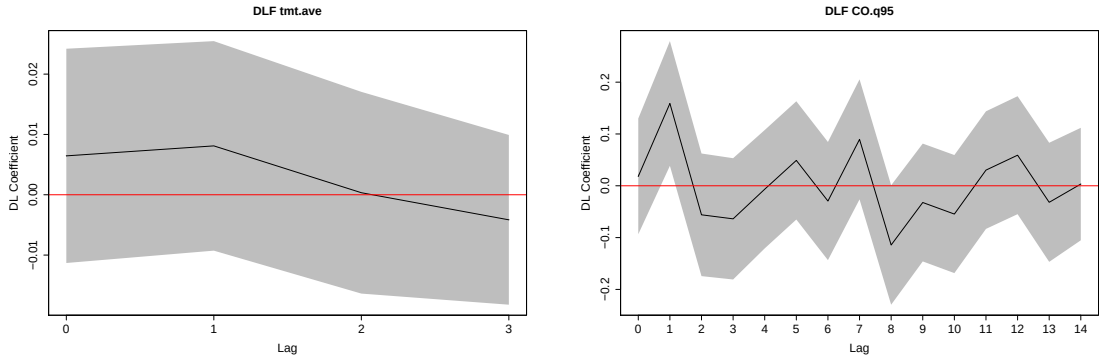


Abbildung 4.3: DLFs von Temperatur (links) und Kohlenstoffmonoxid (rechts) im unrestringierten Modell für die Call-Center-Daten in der Pilotregion München mit punktweisen 95%-Konfidenzbändern (grau)

l	1	2	3	4	5	6	7	8	9	10	11	12	13	14
ρ_l	0.60	0.32	0.23	0.18	0.19	0.28	0.31	0.26	0.21	0.18	0.15	0.20	0.25	0.30

Tabelle 4.2: Empirische Autokorrelationen $\rho_l = \mathrm{Cor}(x_t, x_{t-l})$ in der Kohlenstoffmonoxid-Zeitreihe (2006-2008) in der Pilotregion München

Besonders im hinteren Lag-Bereich, nach Erreichen des maximalen Effekts, erwartet man theoretisch, dass die DLF einigermaßen gleichmäßig gegen 0 tendiert. Schätzt man auch die Koeffizienten höherer Lags ohne Restriktionen, ist ein dahingehend realistischer Verlauf nicht gewährleistet, wie die DLFs für Temperatur und Kohlenstoffmonoxid in der Pilotregion München belegen. Ein geeignetes Schätzverfahren für die DLF sollte zudem berücksichtigen, dass die Varianz der DLF mit zunehmenden Lag abnimmt, zumal man im finiten Lag-Modell davon ausgeht, dass die DLF nach dem maximalen Lag auf 0 fixiert ist und überhaupt keine Variablität mehr aufweist.

Des Weiteren ist die unrestringierte Schätzung der DLF sehr parameterintensiv, was insbesondere ins Gewicht fällt, wenn die verzögerten Effekte aller Meteorologie- und Luftqualitätsparameter simultan analysiert werden sollen. Bei den verwendeten maximalen Lags von 3 für die Meteorologie-Parameter und 14 für die Luftschadstoffe ergeben sich insgesamt 115 Modellparameter zur Schätzung aller DLFs.

Um der Instabilität der Schätzung der Lag-Effekte, verursacht durch die hohen Korrelationen aufeinanderfolgender Lag-Kovariablen x_{t-l} und x_{t-l-1}, zu begegnen, können auch $L+1$ separate Modelle gefittet werden, in die jeweils die Kovariablen x_t bis x_{t-L} einzeln aufgenommen werden. Abbildung 4.4 zeigt die auf diese Weise geschätzte DLF für Kohlenstoffmonoxid in der Pilotregion München. Die erhoffte Glättung der DLF konnte auch durch die separate Schätzung der Koeffizienten nicht erreicht werden.

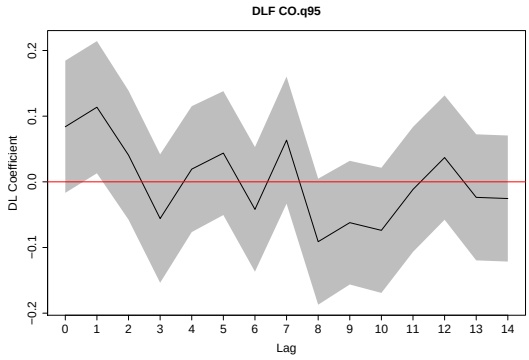


Abbildung 4.4: DLF von Kohlenstoffmonoxid im Modell mit durchschnittlichen Lags für die Call-Center-Daten in der Pilotregion München mit punktweisen 95%-Konfidenzbändern (grau)

Eine Möglichkeit, die Parameterzahl von Distributed Lag Modellen mit großem maximalen Lag zu reduzieren, besteht in der Verwendung durchschnittlicher Lags. Diese Methode wurde bereits in den Abschnitten 3.7 und 3.8 zur Variablenselektion eingesetzt, um einen ersten Eindruck von den Lag-Effekten der Luftschadstoffe zu erhalten. Die Modellgleichung hat hier folgende Gestalt:

$$\begin{aligned}\log(\mu_t) &= \ldots + x_t\beta_t + \underbrace{(x_{t-1} + \ldots + x_{t-3})/3}_{\bar{x}_{\text{st}}}\beta_{\text{st}} + \underbrace{(x_{t-4} + \ldots + x_{t-6})/3}_{\bar{x}_{\text{mt}}}\beta_{\text{mt}} \\ &\quad + \underbrace{(x_{t-7} + \ldots + x_{t-4})/8}_{\bar{x}_{\text{lt}}}\beta_{\text{lt}} + \ldots\end{aligned}$$

Die Einteilung in kurz-, mittel- und langfristigen Effekt wurde nach inhaltlichen Gesichtspunkten vorgenommen, grundsätzlich ist aber eine beliebige Einteilung möglich. Sinnvoll ist die Mittelung über mehrere Lags vor allem bei großem maximalen Lag L, etwa bei den Luftschadstoffen. Die resultierenden gemittelten Lag-Kovariablen $\bar{x}_1, \ldots, \bar{x}_M$ $(M < L)$ sind untereinander weniger stark korreliert als die ursprünglichen Lag-Kovariablen, so dass eine stabilere Schätzung möglich ist. Gemäß der vorliegenden Einteilung in Lag-Bereiche ergeben sich die Funktionswerte der DLF aus den folgenden Restriktionen:

$$\begin{aligned}\beta_0 &= \beta, \\ \beta_1 = \ldots = \beta_3 &= \beta_{\text{st}}, \\ \beta_4 = \ldots = \beta_6 &= \beta_{\text{mt}}, \\ \beta_7 = \ldots = \beta_{14} &= \beta_{\text{lt}}.\end{aligned}$$

Die resultierende DLF verläuft also konstant innerhalb der Lag-Bereiche. Abbildung 4.5 zeigt die in kurz-, mittel- und langfristigen Effekt eingeteilte DLF für Kohlenstoffmonoxid in der Pilotregion München.

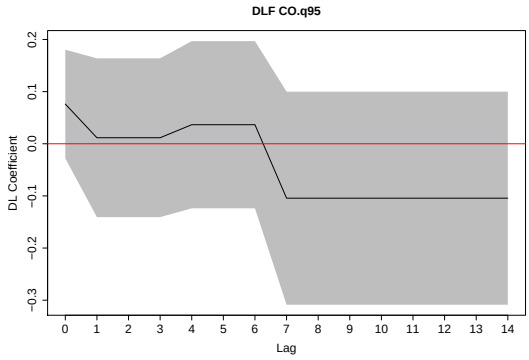


Abbildung 4.5: DLF von Kohlenstoffmonoxid im Modell mit durchschnittlichen Lags für die Call-Center-Daten in der Pilotregion München mit punktweisen 95%-Konfidenzbändern (grau)

Besonders problematisch ist hier, dass durch die Mittelwertbildung bestehende Lag-Effekte verschwinden können. Zudem erscheint die vorgenommene Einteilung in verschiedene Lag-Bereiche relativ willkürlich und führt zu einer wenig flexiblen Schätzung der DLF. Eine datengesteuerte Einteilung wäre in jedem Fall besser geeignet, ist jedoch schwer zu realisieren. Ebenso wie bei der unrestringierten Modellierung findet kein Shrinkage höherer Lag-Koeffizienten gegen 0 statt. Mehr als ein grober Eindruck von Existenz und Dauer zeitverzögerter Effekte ist also auch von diesem Verfahren nicht zu erwarten.

Einen wesentlich weiter entwickelten Ansatz, der besonders bei langfristigen Lag-Effekten geeignet ist, bietet das (infinite) Lag-Modell nach Solow (vgl. Solow (1960)). Idee dieses Verfahrens ist es, dass sich der Lag-Koeffizient β_{t-l} $(l = 1, 2, \ldots)$ aus einer festen Komponente β und einem vom Lag l abhängigen Gewicht w_l zusammensetzt. Die theoretische Regressionsgleichung lautet dann

$$\log(\mu_t) = \ldots + \sum_{l=0}^{\infty} x_{t-l} \underbrace{w_l \beta}_{\beta_{t-l}} + \ldots \tag{11}$$

Solow konstruiert die Gewichte w_l aus der Wahrscheinlichkeitsfunktion einer negativen Binomialverteilung mit Parametern k und λ, das heißt es gilt folgende Restriktion:

$$w_l = \begin{pmatrix} k+l-1 \\ l \end{pmatrix} (1-\lambda)^k \lambda^l, \quad k \in \mathbb{N}^+, \quad 0 < \lambda < 1.$$

Somit summieren sich die Gewichte insgesamt zu 1 auf. Betrachtet man die Form der Wahrscheinlichkeitsfunktion der negativen Binomialverteilung, beispielsweise für $\lambda = 0.7$ und $k = 2$ bzw. $\lambda = 0.4$ und $k = 3$ (vgl. Abbildung 4.6 links bzw. rechts), wird deutlich, wie durch die Gewichte die Form der DLF gesteuert wird. Bei Verwendung der ersten Variante wird Lag 0 am stärksten gewichtet. Danach werden

die Gewichte immer kleiner, man erhält also eine streng monotone DLF. Bei der zweiten Variante dagegen wird der betragsmäßig größte Effekt bei Lag 2-3 erreicht und erst danach tendieren die w_l und damit auch die β_{t-l} gegen 0. Die Wahrscheinlichkeitsfunktion der negativen Binomialverteilung bietet also genügend Flexibiliät, um alle typischen Verläufe einer DLF (vgl. Abbildung 4.1) zu erfassen.

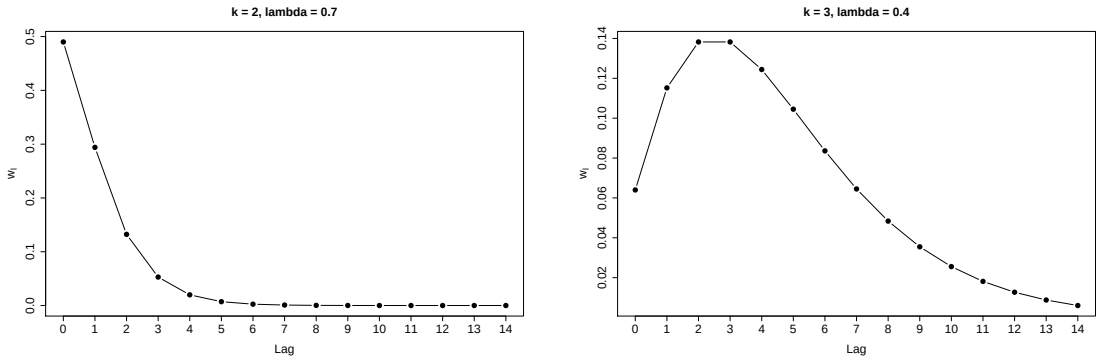


Abbildung 4.6: Wahrscheinlichkeitsfunktion der Negativ-Binomial-Verteilung für $\lambda = 0.7$ und $k = 2$ (links) bzw. $\lambda = 0.4$ und $k = 3$ (rechts)

Als Spezialfall des Solow-Modells ergibt sich für $k = 1$ das Koyck-Modell (vgl. Koyck (1954)). Die Gewichte berechnen sich dann aus der Verteilungsfunktion einer geometrischen Verteilung, also $w_l = (1 - \lambda)\lambda^l$. Damit gilt allerdings die Restriktion $\beta_{t-l-1} = \lambda\beta_{t-l}$, so dass nur streng monotone Verläufe der DLF möglich sind.

Die Schätzung im Solow-Modell erfolgt iterativ, da die Parameter β, k und λ gleichzeitig angepasst werden müssen. Ravines et al. (2006) erläutert wie das Solow-Modell als hierarchisches Modell formuliert und bayesianisch geschätzt werden kann. Betrachtet man die Parameter k und λ als gegeben, kann β durch ein gewöhnliches GLM geschätzt werden. Dazu zieht man in der Regressionsgleichung (11) β vor die Summe und berechnet aus den analytisch bestimmbaren Gewichten w_l und den gelagten Kovariablen x_{t-l} die neue Kovariable $\tilde{x}$:

$$\log(\mu_t) = \ldots + \beta \underbrace{\sum_{l=0}^{\infty} x_{t-l} w_l}_{\tilde{x}} + \ldots$$

Zur Approximation der unendlichen Summe genügt es, eine ausreichend große Anzahl L von Summanden zu verwenden, da diese aufgrund der Gewichtung für großes l an Bedeutung verlieren. Benutzt man wie oben $\lambda = 0.7$ und $k = 2$ bzw. $\lambda = 0.4$ und $k = 3$, ergeben sich die in Abbildung 4.7 dargestellten DLFs für Kohlenstoffmonoxid in der Pilotregion München. β wurde bei der ersten Variante auf 0.1377, bei der zweiten Variante auf 0.0900 geschätzt. $\tilde{x}$ wurde aus $L = 14$ Summanden konstruiert.

Die Berechnung der grau eingezeichneten punktweisen Konfidenzbänder beruht auf

$$\widehat{\mathrm{Var}}(\hat{\beta}_{t-l}) = \widehat{\mathrm{Var}}(w_l\hat{\beta}) = w_l^2\widehat{\mathrm{Var}}(\hat{\beta}), \quad l = 1, \ldots, L.$$

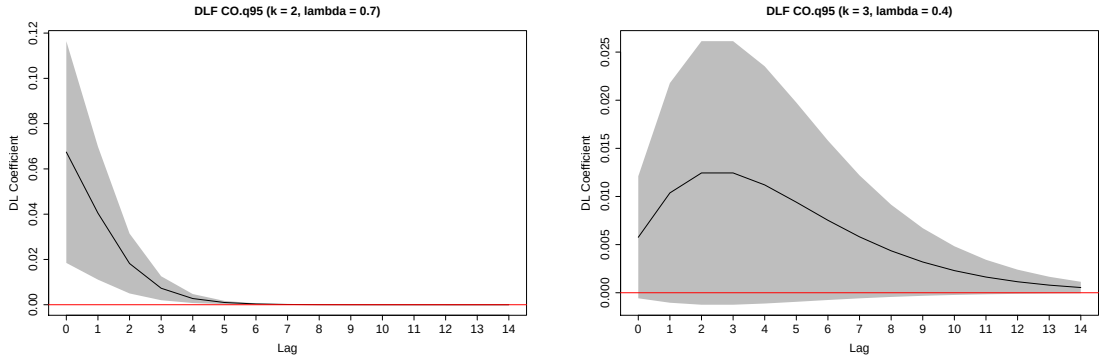


Abbildung 4.7: DLFs von Kohlenstoffmonoxid in unterschiedlich parametrisierten Solow-Modellen für die Call-Center-Daten in der Pilotregion München mit punktweisen 95%-Konfidenzbändern (grau): $\lambda = 0.7$ und $k = 2$ (links) bzw. $\lambda = 0.4$ und $k = 3$ (rechts)

Naturgemäß nehmen die DLFs die von den entsprechenden Gewichtsfunktionen vorgegebene Form an. Schätzt man auch die dafür zuständigen Parameter k und λ aus den Daten, erhält man eine relativ flexible Schätzung der DLF. Nicht möglich sind allerdings mehrmodale Formen oder ein Vorzeichenwechsel der DLF. Im Vergleich zu den bisher betrachteten Ansätzen ergibt sich in jedem Fall ein wesentlich gleichmäßigerer Verlauf, der auf die Restriktion der Koeffizienten zurückzuführen ist. Generell wird die Datentreue der Schätzung durch das Aufstellen von Restriktionen für die β_{t-l} reduziert. Dies nimmt man jedoch in Kauf, da sich die Varianz der Schätzung dadurch gleichzeitig verringert. Ein weiterer, entscheidender Vorteil des Solow-Modells ist, dass die geschätzte DLF, genauso wie die geschätzte Varianz der Lag-Koeffizienten, mit wachsendem l gegen 0 konvergieren. Außerdem müssen insgesamt nur drei Parameter pro Kovariable geschätzt werden. Allerdings ist die Schätzung des Solow-Modells für unbekanntes k und λ in keinem der gängigen Statistik-Programmpakete implementiert.

4.2 Bayesian Distributed Lag Models

Einen neuen, weitgehend datengesteuerten, Ansatz zur Schätzung der DLF bei großem maximalen Lag liefert das Bayesian Distributed Lag Model (BDLM) von Welty et al. (2009). Im Folgenden werden die zugrunde liegenden Annahmen des Modells in groben Zügen erläutert.

Die Steuerung der Form der DLF erfolgt grundsätzlich dadurch, dass man die Einträge der Kovarianzmatrix $\boldsymbol{\Sigma}$ der Lag-Koeffizienten $\boldsymbol{\beta} = (\beta_t, \ldots, \beta_{t-L})$ kontrolliert. Genauer gesagt lässt man die Varianzen der β_{t-l}, also die Diagonaleinträge von $\boldsymbol{\Sigma}$, für wachsendes l gegen 0 und die Kovarianzen von β_{t-l-1} und β_{t-l} für wachsendes l gegen 1 gehen. Dadurch erhält man eine flexible Schätzung der DLF im vorderen Lag-Bereich, dagegen eine glattere, sich der 0 annäherende Funktion im hinteren Lag-Bereich. Umgesetzt wird dieses Prinzip durch die Wahl geeigneter Priori-Verteilungen für die β_l, die mit wachsendem Lag an Informativität gewinnen.

$\boldsymbol{\Sigma} = (\sigma_{uv})_{u,v=1,\ldots,L+1}$ wird durch die Hyperparameter $\sigma^2 = \mathrm{Var}(\beta_0) \geq 0$, $\eta_1 \leq 0$ und $\eta_2 \leq 0$ in folgender Weise parametrisiert:

$$\sigma_{uv} = \begin{cases} \sigma^2 \exp(\eta_1 \cdot (u-1)) & \text{für } u = v \\ [1 - \exp(\eta_2 \cdot (u-1))]\,[1 - \exp(\eta_2 \cdot (v-1))] & \text{für } u \neq v \end{cases} .$$

Man sieht leicht, dass die Diagonalelemente $\sigma^2 \exp(\eta_1 \cdot l)$ von $\boldsymbol{\Sigma}$ aufgrund von $\eta_1 \leq 0$ mit wachsendem l gegen 0 tendieren. Je kleiner η_1 ist, desto schneller konvergieren die Varianzen der β_{t-l} und damit die DLF gegen 0. η_2 steuert dagegen die Zunahme der Kovarianzen der Lag-Koeffizienten bei wachsendem Lag und damit die Zunahme der Glattheit der DLF. Je kleiner η_2 gewählt wird, desto schneller nähern sich die Kovarianzkomponenten $[1 - \exp(\eta_2 \cdot l)]$ bei wachsendem l der 1.

Der Varianzparameter σ^2 wurde in diesem Fall nicht mit einer Priori-Verteilung belegt, sondern aus dem zuvor beschriebenen unrestringierten GLM mit geschätztem Parametervektor $\boldsymbol{\beta}_{\text{unres}}$ geschätzt (Empirisches Bayes-Verfahren). Gemäß der Vorgehensweise von Welty et al. (2009) wurde hier $\hat{\sigma}^2 = 10 \cdot \widehat{\mathrm{Var}}(\hat{\beta}_{\text{unres},t})$ verwendet. Größere Werte von σ^2 führen zu einer langsameren Konvergenz der DLF gegen 0, insgesamt besitzt die Wahl von σ^2 aber keinen wesentlichen Einfluss auf die Form der resultierenden DLF. Für die Parameter η_1 und η_2 wird jeweils eine diskrete Gleichverteilungspriori angenommen, d.h. man gibt ein Gitter möglicher Werte vor. An dieser Stelle ist also ein geringes Maß an manuellem Tuning notwendig. Für die Call-Center-Daten wurde der Wertebereich von η_1 auf $[-1.00, -0.05]$ und für η_2 auf $[-0.80, -0.00]$ festgelegt, bei den Abrechnungsdaten auf $[-1.00, -0.50]$ für η_1 und auf $[-0.80, -0.70]$ für η_2.

Die Schätzung von $\boldsymbol{\beta}$ und $\boldsymbol{\eta}$ erfolgt im generalisierten Fall durch ein Gibbs-Sampling-Verfahren. Dieses wurde mithilfe der R-Funktion „bayesDLM“, unter Annahme einer Poisson-Verteilung für die $y_i|\boldsymbol{x}_i$, durchgeführt. Die R-Funktion ist auf der Homepage des Internet-based Health and Air Pollution Surveillance System (IHAPSS) unter http://www.ihapss.jhsph.edu/software/BayesDLM erhältlich. Insgesamt wurden 20000 Realisierungen aus den vollbedingten Posteriori-Verteilungen der Modellparameter gezogen, davon die ersten 5000 verworfen und von den verbleibenden 15000 Werten nur jeder 10. verwendet. Die Trace-Plots der einzelnen Parameter zeigen eine gute Konvergenz der Markov-Kette nach Ablauf der Burn-in-Phase. Als Schätzwerte für die Parameter wurden die Posteriori-Mittelwerte herangezogen. Es gilt

zu beachten, dass bei der Implementierung der BDLMs nach Welty et al. (2009) nur ein Luftschadstoff pro Modell betrachtet werden kann, so dass mehrere separate Modelle gefittet werden mussten. Bei der Interpretation der Ergebnisse muss dementsprechend beachtet werden, dass nicht auf die jeweils anderen Luftschadstoffe bedingt wurde, was sich aufgrund der hohen Korrelation der Schadstoffwerte als problematisch erweisen kann.

Ansonsten wurden die üblichen designbedingten und administrativen Kovariablen, sowie der in Abschnitt 3.3 beschriebene Zeittrend in den linearen Prädiktor der BDLMs aufgenommen. Um gleichzeitig mögliche Lag-Effekte der Meteorologie-Parameter zu berücksichtigen, wurden die gelagten Wetter-Kovariablen gemäß des in Abschnitt 4.3 beschriebenen Almon-Verfahrens polynomial transformiert und in den linearen Prädiktor integriert.

Tabelle 4.3 fasst die Schätzwerte für η_1 und η_2 der gefitteten BDLMs für die Luftschadstoffe basierend auf den Call-Center-Daten in der Pilotregion München zusammen. Abbildung G.14 zeigt die dazugehörigen geschätzten DLFs mit 95%-Kredibilitätsbändern, die aus den 2.5%- und 97.5%-Quantilen der geschätzten Posteriori-Verteilungen ermittelt wurden.

Kovariable	η_1	η_2
SO2.q95	-0.9332	-0.7000
PM10.q95	-0.9041	-0.6974
O3.q95	-0.8722	-0.6857
NO2.q95	-0.9346	-0.7041
CO.q95	-0.6053	-0.7251

Tabelle 4.3: Schätzwerte für η_1 und η_2 in den BDLMs für die Luftschadstoffe bei den Call-Center-Daten in der Pilotregion München

Die geschätzten Werte von η_1 und η_2 bewegen sich tendenziell am unteren Rand des vorgegebenen Wertebereichs. Es erfolgt also ein relativ starkes Shrinkage der DLF gegen 0 sowie eine relativ starke Glättung. Die geschätzten DLFs zeigen keine signifikanten verzögerten Kovariableneffekte. Der Verlauf der Kurven in Abbildung G.14 erscheint jedoch wesentlich realistischer als beispielsweise im unrestringierten Modell (vgl. die DLF für Kohlenstoffmonoxid in Abbildung 4.3).

Weitere Luftschadstoff-BDLMs wurden gefittet für die über ganz Bayern aggregierten Call-Center-Daten und die nach Verfahren a) und b) reduzierten Abrechnungsdatensätze. Alle geschätzten DLFs sowie eine Tabelle der jeweiligen Schätzwerte für η_1 und η_2 finden sich auf der CD in Appendix I im Ordner „BDLM“. Zusätzlich zur geschätzten DLF enthalten die Dateien Kerndichteschätzungen für die Posteriori-Dichten der einzelnen Parameter sowie Trace-Plots der Realisierungen der Markov-Kette.

Ein wesentlicher Nachteil der BDLMs ist ihre lange Rechenzeit (ca. 2 Tage pro Modell). Außerdem besteht durch die Annahme einer Poisson-Verteilung die Gefahr ei-

ner verfälschten Varianzschätzung durch mögliche Unter- oder Überdispersion. Eine Lösung für diese Probleme bietet die ebenfalls unter obiger Adresse erhältliche Funktion „bayesDLMapprox“, die auf einer Normalapproximation der Poisson-Verteilung beruht. Aufgrund der Tatsache, dass die vollbedingten Posteriori-Verteilungen für $\boldsymbol{\beta}$ und $\boldsymbol{\eta}$ in diesem Fall analytisch zugänglich sind, verkürzt sich die Rechenzeit auf wenige Sekunden. Für die KVB-Call-Center-Daten in der Pilotregion München ist die Normalapproximation der Poisson-Verteilung aufgrund der geringen Fallzahlen allerdings fragwürdig.

Abschließend lässt sich festhalten, dass die BDLMs ein gutes Werkzeug sind, um die zeitverzögerten Effekte einzelner Kovariablen, insbesondere bei großem maximalen Lag, datengesteuert zu analysieren. Für die Realisierung des angestrebten fortlaufenden Prognosemodells kommt diese Methode jedoch nicht in Frage, da mit der zur Verfügung stehenden Software die Rechenzeit der einzelnen Modelle deutlich zu lang ist und keine parallele Betrachtung mehrerer Kovariablen mit zeitverzögerten Effekten möglich ist.

4.3 Polynomiale Struktur der Distributed Lag Function (Almon-Modell)

Eine weitere Möglichkeit, die Form der DLF durch Restriktion der Lag-Koeffizienten zu kontrollieren, bietet das Almon-Lag-Modell (vgl. Almon (1965)). Eine attraktive Eigenschaft der Almon-Methode ist, dass mehrere Kovariablen mit zeitverzögertem Effekt gleichzeitig betrachtet werden können und die Anzahl der zu schätzenden Parameter dabei überschaubar bleibt. Zudem lässt sich das Verfahren sehr intuitiv in die in Abschnitt 3 vorgestellten Regressionsmodelle einbetten.

Die wesentliche Annahme des Almon-Modells besteht darin, dass die β_{t-l} auf einem Polynom niedrigen Grades liegen und sich demzufolge darstellen lassen als

$$\beta_{t-l} = \gamma_0 + l \cdot \gamma_1 + l^2 \cdot \gamma_2 + \ldots + l^d \cdot \gamma_d. \tag{12}$$

Die Steuerung der Glattheit der DLF erfolgt über den Polynomgrad d. Je kleiner man diesen wählt, desto gleichmäßiger verläuft die resultierende Funktion. Setzt man die Restriktion (12) in die Modellgleichung des unrestringierten Modells (10) ein, so erhält man

$$\begin{aligned} \log(\mu_t) \quad = \quad & \ldots + x_t\gamma_0 + x_{t-1}(\gamma_0 + \gamma_1 + \gamma_2 + \ldots + \gamma_d) + \ldots \\ & + x_{t-L}(\gamma_0 + L \cdot \gamma_1 + L^2 \cdot \gamma_2 + \ldots + L^d \cdot \gamma_d) + \ldots \end{aligned}$$

Durch Ausmultiplizieren lässt sich die Regressionsgleichung umformen zu

$$\begin{aligned}\log(\mu_t) &= \ldots + \underbrace{(x_t + x_{t-1} + \ldots + x_{t-L})}_{w_0}\gamma_0 + \ldots \\ &\quad + \underbrace{(x_{t-1} + 2^d \cdot x_{t-2} + \ldots + L^d \cdot x_{t-L})}_{w_d}\gamma_d + \ldots\end{aligned}$$

Es ergibt sich also wiederum ein linearer Prädiktor in den Parametern $\gamma_0, \ldots, \gamma_d$ mit den neuen aus $x_t, \ldots, x_{t-L}$ konstruierten Kovariablen $w_0, \ldots, w_d$. Anstelle von $L+1$ Parametern müssen pro Kovariable nur noch $d+1$ Parameter geschätzt werden, was bei der vorliegenden großen Zahl an Kovariablen von Vorteil ist. Eine Schätzung der DLF-Koeffizienten β_{t-l} erhält man leicht durch Einsetzen der $\hat{\gamma}_0, \ldots, \hat{\gamma}_d$ in die Restriktionsgleichung (12).

Das Almon-Lag-Modell bietet relativ große Flexiblität bei der Schätzung der DLF. Durch die Annahme eines polynomialen Verlaufs sind im Vergleich zum Solow-Modell etwa Vorzeichenwechsel und bei ausreichend großem Polynomgrad d auch mehrmodale Formen möglich. Um eine realistische, glatte DLF zu erhalten, wählt man d in der Praxis jedoch eher niedrig. In dieser Arbeit wurde der Polynomgrad für die Meteorologie-Parameter mit maximalem Lag $L = 3$ auf $d = 2$ festgelegt und für die Luftschadstoffe mit maximalem Lag $L = 14$ auf $d = 4$.

Ein wesentlicher Nachteil der Almon-Methode ist das Verhalten der DLF im Bereich der größeren Lags. Polynome in l tendieren generell mit wachsendem l gegen $+\infty$ oder $-\infty$, so dass die erwartete Konvergenz der DLF gegen 0 nicht gewährleistet ist. Die Varianz der DLF-Schätzung beruht auf

$$\begin{aligned}\widehat{\mathrm{Var}}(\hat{\beta}_{t-l}) &= \widehat{\mathrm{Var}}\left(\sum_{c=0}^{d} l^c \cdot \hat{\gamma}_c\right) \\ &= \sum_{c=0}^{d} l^{2c} \cdot \widehat{\mathrm{Var}}(\hat{\gamma}_c) + 2\sum_{c=0}^{d}\sum_{c\neq\tilde{c}} l^{c+\tilde{c}} \cdot \widehat{\mathrm{Cov}}(\hat{\gamma}_c, \hat{\gamma}_{\tilde{c}}), \quad l = 0, \ldots, L.\end{aligned}$$

Daraus lassen sich punktweise 95%-Konfidenzbänder für die DLF konstruieren. Man kann feststellen, dass die minimale Varianz im Zentrum des Lag-Bereichs auftritt und an den Rändern deutlich anwächst (vgl. Abbildung G.15). Dies ist vor allem deswegen problematisch, da man theoretisch eine Abnahme der Varianz im mittleren bis hohen Lag-Bereich erwartet.

Ein Ansatz, welcher das Problem des fehlenden Shrinkage der DLF gegen 0 beheben könnte, wäre eine vom Lag abhängige Gewichtung der Polynomfunktion. So könnte man wie beim in Abschnitt 4.4 vorgestellten PDLF-Ansatz die Restriktion (12) erweitern zu

$$\beta_{t-l} = w_l \cdot (\gamma_0 + l \cdot \gamma_1 + l^2 \cdot \gamma_2 + \ldots + l^d \cdot \gamma_d),$$

wobei w_l eine Folge von abnehmenden Gewichten ist.

Abbildung G.15 zeigt die mit der Almon-Methode geschätzten DLFs für die Luftschadstoffe und ausgewählte Meteorologie-Parameter basierend auf den KVB-Call-Center-Daten in der Pilotregion München. Die geschätzten DLFs zeigen bei allen betrachteten Parametern einen hinreichend glatten Verlauf. Bei den Luftschadstoffen ergibt sich beispielsweise ein signifikanter verzögerter Kovariableneffekt im mittleren Lag-Bereich (Lag 5 bis 8). Außerdem führt ein Anstieg der Luftfeuchtigkeit am heutigen Tag zu einer signifikanten Abnahme der Fallzahl an den beiden Folgetagen. Vor allem die DLFs für die Luftschadstoffe sind aufgrund des fehlenden Shrinkage der Lag-Koeffizienten gegen 0 im mittleren bis höheren Lag-Bereich jedoch mit Vorsicht zu genießen. Alle weiteren nach dem Almon-Verfahren geschätzten DLFs für die nach Schema a) und b) reduzierten Datensätze sind auf der CD in Appendix I im Ordner „Almon“ zu finden. Generelle Aussagen über die Existenz und Richtung von verzögerten Kovariableneffekten zu treffen, fällt aufgrund der uneinheitlichen Ergebnisse für die verschiedenen Datensätze schwer. Häufiger zu beobachten ist beispielsweise ein signifikant negativer Effekt in den Lags 1 und 2 für Feuchtigkeit und Luftdruck.

Insgesamt betrachtet liefert die Almon-Methode brauchbare DLFs bei kleinem maximalen Lag. Da sich die Almon-Terme außerdem leicht und ohne wesentliche Erhöhung der Rechenzeit in den linearen Prädiktor der betrachteten longitudinalen Regressionsmodelle einbauen lassen, wurden mögliche zeitverzögerte Effekte der Meteorologie-Parameter auf diese Weise in die Prognosemodelle integriert. Um die Anwendung der Almon-Methode zu vereinfachen, wurden die Konstruktion der neuen Kovariablen $w_0, \ldots, w_d$, die Schätzung der $\gamma_0, \ldots, \gamma_d$ und die Darstellung der resultierenden DLF in der R-Funktion „lag_regress“ (vgl. Abschnitt 4.5) implementiert. Parallel kann darin auch die neu entwickelte PDLF-Methode (vgl. Abschnitt 4.4) zur Schätzung von DLFs angewendet werden.

Bei größerem L erweist es sich als problematisch, dass die Flexibilität der DLF nicht über den Lag-Bereich hinweg angepasst wird. Gerade im vordersten Bereich wäre eine bessere Anpassung an die Daten nötig, im hinteren Lag-Bereich dagegen eine stärkere Glättung. Vergleicht man beispielsweise die Almon-DLF für Kohlenstoffmonoxid bei den Call-Center-Daten in der Pilotregion München (vgl. Abbildung G.15) mit der im BDLM geschätzten Funktion (vgl. Abbildung G.14), wird das auch im unrestringierten Modell (vgl. Abbildung 4.3) sichtbare Maximum bei Lag 1 vom Almon-Modell nicht erkannt. Umgekehrt ergibt sich etwa bei der Almon-DLF von Ozon ein signifikant negativer Effekt bei Lag 11 bis 12, der bei der entsprechenden Funktion des BDLM nicht erkennbar ist.

Für die Berücksichtigung potentieller Lag-Effekte der Luftschadstoffe im Prognosemodell, bedarf es also einer Methode, die ähnlich realistische Ergebnisse liefert wie die BDLMs, jedoch eine parallele Betrachtung mehrerer Kovariablen mit zeit-

verzögertem Effekt ermöglicht. Zudem soll sich die Rechenzeit in einem vertretbaren Rahmen bewegen, um ein fortlaufendes Prognosemodell realisieren zu können. Um diese Probleme zu lösen, wurde ein neuer Ansatz entwickelt, der im folgenden Abschnitt detailliert erläutert wird.

4.4 Penalized Distributed Lag Function (PDLF)

Zanobetti et al. (2000) stellen in ihrem Artikel einen P-Spline-basierten Ansatz vor, die DLFs bei großem maximalen Lag zu schätzen (im Anwendungsbeispiel des Artikels: $L = 45$). Statt wie beim Almon-Ansatz ein Polynom über den gesamten Lag-Bereich anzupassen, konstruieren sie darauf eine TP-Basis (vgl. Abschnitt 3.2.3) mit äquidistanten Knoten zwischen 0 und L. Um die Glattheit der Funktion zu kontrollieren wird eine Penalisierung großer Spline-Koeffizienten vorgenommen. Der hier vorgestellte PDLF-Ansatz (Penalized Distributed Lag Function) beruht auf einem ähnlichen Prinzip, jedoch erfolgt hier, ähnlich wie bei den BDLMs, eine zunehmende Glättung der DLF im hinteren Lag-Bereich durch geeignete Knotenwahl und ein zusätzliches Shrinkage gegen 0 durch Gewichtung der Basisfunktionen.

Anstelle einer TP-Basis wird hier eine numerisch stabilere, kubische B-Spline-Basis (vgl. Abschnitt 3.2.3) verwendet. Um mehr Flexibilität im vorderen Lag-Bereich zu erhalten, werden dort mehr Knoten platziert. Genauer gesagt werden die Knoten $\kappa_1, \ldots, \kappa_K$ so zwischen 0 und L angeordnet, dass ihr Abstand quadratisch zunimmt. Um eine ausreichende Anpassung an die Daten zu gewährleisten, werden insgesamt $K = \lfloor L/2 \rfloor$ Knoten innerhalb des Lag-Bereich gesetzt. Damit ergeben sich folgende Lokalisationen:

$$\kappa_k = \sum_{i=1}^{k} i^2 \cdot \left(\frac{1}{L} \sum_{k=1}^{K} k^2 \right), \quad k = 1, \ldots, K. \tag{13}$$

Die Randknoten κ_0 und κ_{K+1} werden außerhalb des Lag-Bereichs in $\kappa_0 = -1$ und $\kappa_{K+1} = L + 1$ platziert, um auch eine stabile Schätzung der DLF an den Rändern des Lag-Bereichs zu erhalten. Durch Verwendung einer kubischen Basis (kubische Polynome zwischen den Knoten) erhält man insgesamt $m = K + 3$ Basisfunktionen $\Phi_1, \ldots, \Phi_m$. Etwa für die Luftschadstoffe mit maximalem Lag $L = 14$ ergeben sich $K = 6$ Knoten und $m = 9$ Basisfunktionen. Abbildung 4.8 zeigt die Lokalisation von inneren (rot) und Randknoten (grün) sowie eine Darstellung der Basisfunktionen $\Phi_1, \ldots, \Phi_9$, ausgewertet auf einem äquidistanten Gitter von 200 Punkten, für $L = 14$.

Die Lag-Koeffizienten β_{t-l} setzen sich zusammen aus einer gewichteten Summe der Basisfunktionen, ausgewertet im jeweiligen Lag, das heißt

$$\beta_{t-l} = \Phi_1(l)\alpha_1 + \ldots + \Phi_m(l)\alpha_m, \quad l = 0, \ldots, L. \tag{14}$$

Anstelle von $L+1$ Koeffizienten müssen bei dieser Methode also m Koeffizienten pro Kovariable $(\alpha_1, \ldots, \alpha_m)$ geschätzt werden. Dies bedeutet allerdings erst ab $L \geq 7$

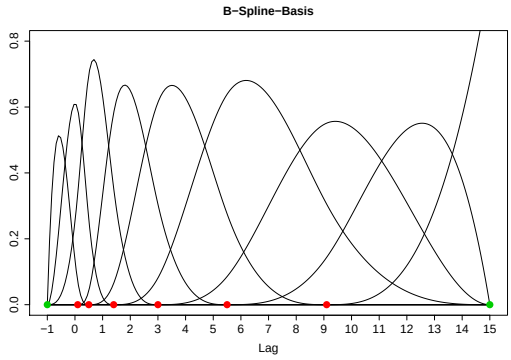


Abbildung 4.8: B-Spline-Basisfunktionen $\Phi_1, \ldots, \Phi_9$ bei $L = 14$ mit inneren (rot) und Randknoten (grün)

eine Ersparnis an Parametern gegenüber dem unrestringierten Modell und zeigt, dass die Methode eher für große maximale Lags geeignet ist.

Um das Shrinkage der Lag-Koeffizienten β_{t-l} mit wachsendem Lag l gegen 0 zu realisieren, verwendet man in Gleichung (14) statt der ursprünglichen Basisfunktionen $\Phi_j(l)$ die mit einem vom Lag abhängigen Gewicht w_l versehenen Basisfunktionen $\tilde{\Phi}_j(l)$:

$$\tilde{\Phi}_j(l) = \Phi_j(l) \cdot \underbrace{l^{-a}}_{w_l}, \quad a \geq 1, \quad j = 1 \ldots, m.$$

a kann als Tuningparameter aufgefasst werden, der die Konvergenzgeschwindigkeit der Gewichte w_l gegen 0 und damit das Shrinkage der β_{t-l} kontrolliert. In Abbildung 4.9 sind die gewichteten Basisfunktionen $\tilde{\Phi}_1, \ldots, \tilde{\Phi}_m$, wiederum ausgewertet auf einem Gitter von 200 Punkten, für $a = 1.5$ (links) und $a = 2.3$ (rechts) dargestellt. Je größer der Wert von a gewählt wird, desto stärker wird die DLF mit wachsendem l gegen 0 geshrinkt, wobei $a = 1$ kein Shrinkage bedeutet.

Setzt man wie bei der Almon-Methode die Restriktion (14) mit gewichteten Basisfunktionen in den linearen Prädiktor des unrestringierten Modells (10) ein, erhält man:

$$\begin{aligned} \log(\mu_t) &= \ldots + \sum_{l=0}^{L} x_{t-l} \underbrace{\left(\sum_{j=1}^{m} \tilde{\Phi}_j(l) \cdot \alpha_j \right)}_{\beta_{t-l}} + \ldots \\ &= \ldots + \sum_{j=1}^{m} \underbrace{\left(\sum_{l=0}^{L} x_{t-l} \cdot \tilde{\Phi}_j(l) \right)}_{z_l} \alpha_j + \ldots \end{aligned}$$

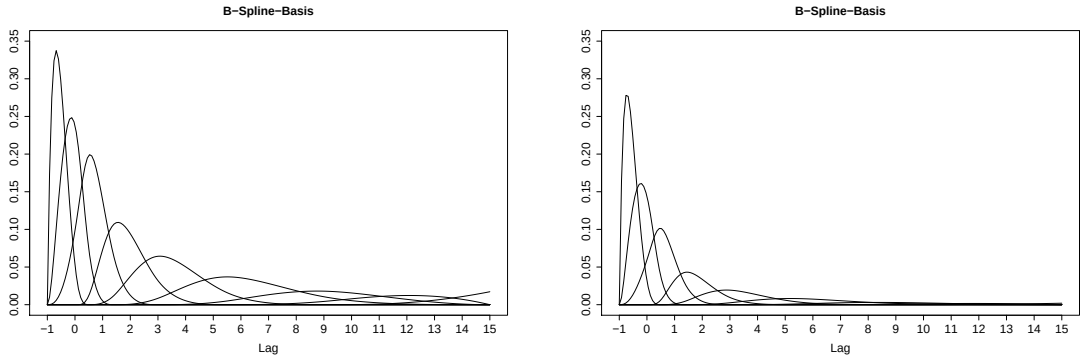


Abbildung 4.9: Gewichtete B-Spline-Basisfunktionen $\tilde{\Phi}_1, \ldots, \tilde{\Phi}_9$ für $a = 1.5$ (links) und $a = 2.3$ (rechts)

Anhand der zweiten Darstellung ist gut zu erkennen, dass sich die Methode bis hierhin problemlos in ein gewöhnliches GLM einbetten lässt. Im linearen Prädiktor stehen lediglich anstelle der gelagten Kovariablen $x_t, \ldots, x_{t-L}$ die durch Multiplikation mit gewichteten Basisfunktionen entstandenen neuen Kovariablen $z_1, \ldots, z_m$ und anstelle der ursprünglichen Parameter $\beta_t, \ldots, \beta_{t-L}$ die neuen Parameter $\alpha_1, \ldots, \alpha_m$. Die B-Spline-Basisfunktionen können manuell, beispielsweise mit der R-Funktion „bs“ im Paket „splines“ konstruiert werden. Eine Schätzung der Lag-Koeffizienten erhält man durch Einsetzen von $\hat{\alpha}_1, \ldots, \hat{\alpha}_m$ in Restriktion (14) mit gewichteten Basisfunktionen.

Der Name der PDLF-Methode ist darauf zurückzuführen, dass wie im Artikel von Zanobetti et al. (2000) zusätzlich ein Penalisierungsansatz verwendet wird, um die Rauheit der geschätzten DLF zu kontrollieren. Gemäß dem in Abschnitt 3.2.3 beschriebenen P-Spline-Ansatz erfolgt die Glättung der DLF durch die Bestrafung großer Differenzen in den Parametern α_j. Technisch wird die Penalisierung der Splinekoeffizienten mithilfe der R-Funktion „gam“ im Paket „mgcv“ (vgl. Wood (2006)) umgesetzt. Dafür werden der Funktion die manuell konstruierten Differenzenmatrizen $\boldsymbol{K}_d$ (vgl. Abschnitt 3.2.3) der Ordnung d übergeben, wobei die Rauheit der resultierenden Funktion bei konstantem Penalisierungsparameter λ mit d anwächst. Bei der Analyse der vorliegenden Daten wurden ausschließlich zweite Differenzen verwendet. λ wird mithilfe der ebenfalls in Abschnitt 3.2.3 definierten Kriterien zur Glättungsparameterwahl geschätzt. Bei bekanntem Dispersionsparameter, wie etwa im Poisson-Modell, wird das UBRE-Kriterium verwendet, bei unbekanntem Dispersionsparameter, wie etwa im loglinearen oder Quasi-Poisson-Modell, das GCV-Kriterium. Um ein Mindestmaß an Glattheit zu gewährleisten, wurde der mögliche Wertebereich von λ je nach analysiertem Datensatz durch Vorgabe eines $\lambda_{\min}$ nach unten beschränkt. Auf diese Weise können problemlos mehrere Kovariablen mit zeitverzögertem Effekt parallel betrachtet werden. Abgesehen von der

Wahl des Tuningparameters a und des minimalen Penalisierungsparameters $\lambda_{\min}$ erfolgt die Schätzung datengesteuert. a und $\lambda_{\min}$ können theoretisch für jede Kovariable separat festgelegt werden. Genau wie die Almon-Methode ist das komplette PDLF-Verfahren in der in Abschnitt 4.5 vorgestellten R-Funktion „lag_regress“ implementiert.

Es ist zu beachten, dass trotz der Verwendung der Funktion „gam“ rein technisch betrachtet kein GAM im eigentlichen Sinn gefittet wird, da der Prädiktor η nach Transformation der Kovariablen mit den von Hand konstruierten Basisfunktionen vollständig linear in den Parametern ist. Die „gam“-Funktion wird ausschließlich zur Penalisierung eingesetzt, so dass man eher von einem „penalisierten“ GLM sprechen kann. Die Schätzung ist dadurch sehr stabil und bei ausreichendem Arbeitsspeicher auch für große Datensätze ($n \approx 500000$, vgl. Abschnitt 5.2) durchführbar. Die Rechenzeit ist beispielsweise im Vergleich zum BDLM, in dem nur eine Kovariable mit zeitverzögertem Effekt analysiert werden kann, sehr gering (bei $n \approx 5000$ unter einer Minute) und kann durch Fixierung des Penalisierungsparameters λ nochmals um ein Vielfaches reduziert werden.

Eine Schätzung für die Varianz der Lag-Koeffizienten β_{t-l} zur Konstruktion von punktweisen 95%-Konfidenzbändern für die DLF erhält man wie folgt:

$$\widehat{\mathrm{Var}}(\hat{\beta}_{t-l}) = \sum_{j=1}^{m} (\tilde{\Phi}_j(l))^2 \cdot \widehat{\mathrm{Var}}(\hat{\alpha}_j) + 2 \sum_{j=1}^{m} \sum_{j \neq \tilde{j}} \tilde{\Phi}_j(l) \tilde{\Phi}_{\tilde{j}}(l) \cdot \widehat{\mathrm{Cov}}(\hat{\alpha}_j, \hat{\alpha}_{\tilde{j}}).$$

Dazu ist anzumerken, dass die R-Funktion „gam“ im Falle einer Penalisierung der Regressionskoeffizienten eine frequentistische und eine bayesianische Schätzung der Kovarianzmatrix von $\hat{\boldsymbol{\alpha}}$ zurückgibt. Hier wird letztere verwendet, da der bayesianische Ansatz besser zur Varianzschätzung nichtlinearer Funktionen geeignet ist (vgl. Wood (2006, Kap. 4.8)). Streng genommen handelt es sich bei den konstruierten Konfidenzbändern also um Kredibilitätsbänder.

Abbildung G.16 zeigt die geschätzten DLFs für die Luftschadstoffe basierend auf den KVB-Call-Center-Daten in der Pilotregion München zusammen mit 95%-Konfidenzbändern. Für den Shrinkage-Parameter a wurde durchgängig der Wert 2.3 angenommen, der minimale Penalisierungsparameter $\lambda_{\min}$ wurde hier auf 10 festgelegt. Die folgende Tabelle beinhaltet die durch Minimierung des UBRE geschätzten Penalisierungsparameter $\hat{\lambda}$ für die einzelnen Luftschadstoffe:

Schadstoff	SO2.q95	PM10.q95	O3.q95	NO2.q95	CO.q95
$\hat{\lambda}$	$2.74 \cdot 10^2$	$1.04 \cdot 10^4$	$6.06 \cdot 10^6$	$7.01 \cdot 10^7$	$1.04 \cdot 10^4$

Die stärkste Penalisierung wurde für Stickstoffdioxid und Ozon vorgenommen. Die resultierenden PDLFs (vgl. Abbildung G.16) weisen alle ein vergleichbares Maß an Glattheit auf. Abgesehen von der DLF für Feinstaub verlaufen die Funktionen monoton, das heißt der stärkste Effekt tritt im Lag 0 auf. Diese maximalen Effekte

fallen bei Schwefel- und Stickstoffdioxid positiv, bei den übrigen Schadstoffen positiv aus. Ein signifikant positiver Effekt auf die Anruferzahl am gleichen Tag konnte für Ozon nachgewiesen werden (+0.16% Anrufe pro Zunahme um $1\mu g/m^3$), ein signifikant negativer Effekt dagegen für Schwefeldioxid (−0.94% Anrufe pro Zunahme um $1\mu g/m^3$). Die PDLFs weisen grundsätzlich einen ähnlichen Verlauf wie die DLFs der BDLMs auf (vgl. Abbildung G.14). Langfristige Lag-Effekte konnten auch hier nicht beobachtet werden.

Alle gefitteten Luftschadstoff-PDLFs für die Call-Center- und Abrechnungsdaten in der Pilotregion München und aggregiert über die bayerischen Landkreise sind auf der CD in Appendix I im Ordner „PDLF“ gespeichert. Über alle reduzierten Datensätze hinweg wurde für a der Wert 2.3 gewählt. $\lambda_{\min}$ wurde bei den Call-Center-Daten auf 10 gesetzt, bei den Abrechnungsdaten dagegen auf 100. Im folgenden Absatz werden auftretende gemeinsame Tendenzen in den Modellen für die verschiedenen reduzierten Datensätze zusammengefasst.

Schwefeldioxid weist einen signifikanten mittel- bis langfristigen Effekt auf, der sowohl bei den bayernweit aggregierten Call-Center-Daten als auch bei den Abrechnungsdaten in der Pilotregion München negativ ausfällt. Der stärkste Effekt tritt am Tag der Beobachtung des Messwerts oder am folgenden Tag auf und ist tendenziell eher negativ. Feinstaub besitzt einen weitgehend übereinstimmend zu beobachtenden positiven Effekt bei Lag 0. Danach fällt die DLF vor allem bei den Abrechnungsdaten in den (signifikant) negativen Bereich. Dieser Rebound-Effekt (siehe oben) könnte darauf zurückzuführen sein, dass sich die Wahrscheinlichkeit eines Arztbesuchs reduziert, falls der Patient bereits in den Vortagen einen Arzt aufgesucht hat. Klar ausgeprägte Langzeiteffekte sind nicht zu erkennen. Eine Zunahme der Ozonkonzentration verursacht offensichlich kurzfristig sowohl eine Zunahme der Arztbesuche als auch eine Zunahme der Anrufe beim KVB-Call-Center. Der größte positive Effekt tritt bei den Anrufen noch am gleichen Tag, bei den Arztbesuchen erst am Folgetag auf. Bei den Abrechnungsdaten ergibt sich ein signifikant positiver Effekt von Stickstoffdioxid im Lag 0, ein langfristiger (positiver) Effekt ergibt sich nur bei den aggregierten Call-Center-Daten in der Pilotregion München (ab dem 5. Lag). Bei Kohlenstoffmonoxid ergibt sich ein ähnliches Bild wie bei Ozon. Es konnten kurzfristige positive Effekte beobachtet werden, die bei den Call-Center-Daten in der Pilotregion München am gleichen Tag und bei beiden bayernweit aggregierten Datensätzen am Folgetag maximal ausgeprägt waren. Mittel- und langfristige Effekte traten auch hier nicht auf.

Insgesamt betrachtet, stellt die PDLF-Methode ein weitgehend datengesteuertes Werkzeug zur Analyse verzögerter Kovariableneffekte dar, durch das ein Großteil der beschriebenen Schwierigkeiten bei der Modellierung solcher Effekte behoben werden kann. Außerdem lässt sich das Schätzverfahren problemlos und ohne unverhältnismäßige Erhöhung des Rechenaufwands in die etablierten Standardmodelle integrieren. Bei Kovariablen, denen nur ein kurzfristig verzögerter Effekt unterstellt

wird, wie z. B. den Meteorologie-Parametern mit maximalem Lag $L = 3$, bringt die PDLF-Methode keinen wesentlichen Vorteil gegenüber dem Almon-Modell, benötigt aber aufgrund der Penalisierung mehr Rechenzeit. Der folgende Abschnitt befasst sich detailliert mit der Implementierung der R-Funktion „lag_regress“, die beide Verfahren umfasst.

4.5 Programmierung eines Regressionstools für verzögerte Kovariableneffekte in R

Mit der R-Funktion „lag_regress“ können longitudinale Regressionsmodelle mit zeitverzögerten Effekten gefittet werden. Für Kovariablen mit kurzfristig verzögertem Effekt kann entweder die Almon-Methode oder die neu entwickelte PDLF-Methode verwendet werden. Bei größerem maximalen Lag $L > 4$ empfiehlt sich aus den in Abschnitt 4.4 erläuterten Gründen die Verwendung des PDLF-Verfahrens. Die Modellgleichung des zu fittenden Modells stellt sich in Matrixnotation wie folgt dar:

$$\log(\mathbb{E}(\boldsymbol{y}|\boldsymbol{X})) = \boldsymbol{U}\boldsymbol{\delta} + \boldsymbol{W}\boldsymbol{\gamma} + \boldsymbol{Z}\boldsymbol{\alpha}.$$

Die Designmatrix $\boldsymbol{X}$ setzt sich zusammen aus den Matrizen $\boldsymbol{U}$, $\boldsymbol{W}$ und $\boldsymbol{Z}$. $\boldsymbol{U}$ beinhaltet alle Kovariablen ohne zeitverzögerten Effekt inklusive der manuell konstruierten Kovariablen für gruppenspezifische Zeittrends und Bruchpunkteffekte (vgl. die Abschnitte 3.3 und 3.4). $\boldsymbol{W}$ und $\boldsymbol{Z}$ beinhalten jeweils die aus den Lags neu berechneten Kovariablen der Almon- und PDLF-Methode. Der Parametervektor $\boldsymbol{\beta}$ besteht dementsprechend aus den zu den Komponenten der Designmatrix gehörigen Teilvektoren $\boldsymbol{\delta}$, $\boldsymbol{\gamma}$ und $\boldsymbol{\alpha}$. Der Code der R-Funktion sowie ein Anwendungsbeispiel für die Call-Center-Daten in der Pilotregion München kann Appendix H.3 entnommen werden. Alle Ein- und Ausgabewerte der Funktion werden in Tabelle G.17 definiert. Im Folgenden werden die einzelnen Schritte, aus denen sich die Funktion zusammensetzt, kurz erläutert.

Zunächst werden die Hilfsfunktionen „mylag“ und „mylag_sep“ definiert, die zur Berechnung der gelagten Kovariablen dienen. Voraussetzung für deren Anwendung ist, dass der Datensatz nach den Faktoren sortiert ist, nach denen die Aufteilung der Fallzahlen vorgenommen wird. Bei den vorliegenden Daten sind das die designbedingten Variablen Alter und Geschlecht bzw. nur das Alter bei den Call-Center-Daten. Innerhalb der Gruppen müssen die Beobachtungen nach zeitlicher Abfolge angeordnet sein. Nach entsprechender Sortierung wird der Datensatz auf die Kovariablen reduziert, die zum Fitten des Regressionsmodells nötig sind. Als nächster Schritt müssen die Komponenten der Designmatrix $\boldsymbol{X}$, insbesondere die Almon-Designmatrix $\boldsymbol{W}$ und die PDLF-Designmatrix $\boldsymbol{Z}$ spezifiziert werden. Beide Designmatrizen können mehrere Kovariablen mit zeitverzögertem Effekt beinhalten und haben dann die Form $\boldsymbol{W} = (\boldsymbol{W}_1, \ldots, \boldsymbol{W}_{p_{\text{Almon}}})$ bzw. $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{p_{\text{PDLF}}})$.

Zunächst wird die Berechnung einer Komponente der Almon-Designmatrix $\boldsymbol{W}$ für die Kovariable x betrachtet. Es gilt:

$$\boldsymbol{W} = \boldsymbol{X}_{\text{Lag}} \boldsymbol{A}.$$

Dabei besteht die $n \times (L+1)$-Matrix $\boldsymbol{X}_{\text{Lag}}$ aus allen Lags der Kovariable x, das heißt $\boldsymbol{X}_{\text{Lag}} = (\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \ldots, \boldsymbol{x}_{t-L})$. Die Transformationsmatrix $\boldsymbol{A}$ hat folgende Einträge (vgl. die Restriktion (12)):

$$\boldsymbol{A} = \begin{pmatrix} 0^0 & 0^1 & \cdots & 0^d \\ 1^0 & 1^1 & \cdots & 1^d \\ \vdots & \vdots & \vdots & \vdots \\ L^0 & L^1 & \cdots & L^d \end{pmatrix}_{(L+1)\times d}.$$

Für die PDLF-Methode werden zunächst die Knoten $\kappa_1, \ldots, \kappa_K$ gemäß Formel (13) festgelegt. Die Komponenten der PDLF-Designmatrix $\boldsymbol{Z}$ entstehen durch:

$$\boldsymbol{Z} = \boldsymbol{X}_{\text{Lag}} \operatorname{diag}(0^{-a}, \ldots, L^{-a})\, \boldsymbol{\Phi} = \boldsymbol{X}_{\text{Lag}} \tilde{\boldsymbol{\Phi}}.$$

$\boldsymbol{X}_{\text{Lag}}$ hat die gleiche Gestalt wie bei der Almon-Methode und $\boldsymbol{\Phi}$ ist die Matrix der ungewichteten Basisfunktionen $\Phi_1, \ldots, \Phi_m$, ausgewertet in den Lags 0 bis L. Die Diagonalmatrix in der Mitte enthält die vom Shrinkage-Parameter a abhängigen Gewichte. Die Transformationsmatrix $\tilde{\boldsymbol{\Phi}}$ besteht dementsprechend aus den gewichteten Basisfunktionen $\tilde{\Phi}_1, \ldots, \tilde{\Phi}_m$ (vgl. Restriktion (14) mit gewichteten Basisfunktionen):

$$\tilde{\boldsymbol{\Phi}} = \begin{pmatrix} \tilde{\Phi}_1(0) & \cdots & \tilde{\Phi}_m(0) \\ \vdots & \ddots & \vdots \\ \tilde{\Phi}_1(L) & \cdots & \tilde{\Phi}_m(L) \end{pmatrix}_{(L+1)\times m}.$$

Zusätzlich müssen bei der PDLF-Methode noch die Differenzenmatrizen $\boldsymbol{K}_d$ für die Penalisierung definiert werden.

Nun kann die R-Funktion „gam“ zum Fitten des Modells aufgerufen werden. Aus dem resultierenden Modell-Objekt können anschließend die Schätzungen $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\delta}}^\top, \hat{\boldsymbol{\gamma}}^\top, \hat{\boldsymbol{\alpha}}^\top)^\top$ und $\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}})$ extrahiert werden. Um die DLFs darstellen zu können, müssen aus den Parametern $\hat{\boldsymbol{\gamma}}$ und $\hat{\boldsymbol{\alpha}}$ die Lag-Koeffizienten $\hat{\boldsymbol{\beta}}_{\text{Lag}} = (\hat{\beta}_t, \hat{\beta}_{t-1}, \ldots, \hat{\beta}_{t-L})$ und aus den geschätzten Kovarianzmatrizen $\widehat{\text{Cov}}(\hat{\boldsymbol{\gamma}})$ und $\widehat{\text{Cov}}(\hat{\boldsymbol{\alpha}})$ die Kovarianzmatrix $\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}_{\text{Lag}})$ berechnet werden. Bei der Almon-Methode erfolgt dies durch

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\text{Lag}} &= \boldsymbol{A}\hat{\boldsymbol{\gamma}} \quad \text{und} \\ \widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}_{\text{Lag}}) &= \boldsymbol{A}\, \widehat{\text{Cov}}(\hat{\boldsymbol{\gamma}})\, \boldsymbol{A}^\top, \end{aligned}$$

bei der PDLF-Methode analog durch

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\text{Lag}} &= \tilde{\boldsymbol{\Phi}}\hat{\boldsymbol{\alpha}} \quad \text{und} \\ \widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}_{\text{Lag}}) &= \tilde{\boldsymbol{\Phi}}\, \widehat{\text{Cov}}(\hat{\boldsymbol{\alpha}})\, \tilde{\boldsymbol{\Phi}}^\top. \end{aligned}$$

Schließlich werden die DLFs geplottet und die wichtigsten Schätzwerte ($\hat{\boldsymbol{\delta}}$ und $\hat{\boldsymbol{\beta}}_{\text{Lag}}$), das „gam“-Modell-Objekt sowie die verwendeten Penalisierungsparameter $\lambda_1, \ldots, \lambda_{p_{\text{PDLF}}}$ zurückgegeben.

Im folgenden Abschnitt findet die R-Funktion „lag_regress“ Anwendung bei der Prognose der Fallzahlen.

5 Zeitlich-räumliches Prognosemodell

In diesem Abschnitt wird ein fortlaufend lernendes Prognosemodell zur Vorhersage der Fallzahlen unter Berücksichtigung der zeitlichen und räumlichen Korrelationsstruktur in den vorliegenden Daten entwickelt. Darin werden auch Bruchpunkteffekte und verzögerte Kovariableneffekte berücksichtigt. Finales Ziel ist der Einsatz des Modells im Tagesbetrieb. Dazu soll das Modell zunächst anhand von über einen langen Zeitraum erfassten Ziel- und Kovariablenwerten trainiert werden. Basierend darauf und auf einer Vorhersage der im Modell enthaltenen Meteorologie- und Luftqualitätsparameter können dann zukünftige Fallzahlen prädiktiert werden. Der angestrebte Prognosezeitraum für die Fallzahlen beträgt zunächst drei Tage, da Wetter- und Luftqualitätsmodelle eine zuverlässige Prognose innerhalb dieser Zeitspanne ermöglichen. Um das Modell ständig zu verbessern, fließen die tatsächlich realisierten Fallzahlen nach ihrer Beobachtung in die Menge der Trainingsdaten ein. Die Beurteilung der Prognosequalität des Modells ist dann retrospektiv durch den Vergleich von prädiktierten und wahren Fallzahlen möglich. Die prognostizierten Fallzahlen sollen schließlich in einen kategorialen Gesundheitsindex (Health Index, HI) umgesetzt werden, der beispielsweise der Vorabinformation von Risikopatienten oder Ärzten dienen kann.

Um solch einen fortlaufenden Prozess zu simulieren, wurden hier die Beobachtungen des Jahres 2006 als grundlegender Trainingsdatensatz festgelegt und verschiedene Trainingsmodelle gefittet (vgl. Abschnitt 5.1). Unter Hinzunahme der Kovariableninformation für die ersten 3 Tage des Jahres 2007 wurden anschließend die Fallzahlen an diesen 3 Tagen prädiktiert. Im nächsten Schritt wurde der Trainingsdatensatz um den 1. Januar 2007 erweitert, um erneut die Fallzahlen der 3 Folgetage vorherzusagen. Nach diesem Schema wurde der Prozess dann über den gesamten Beobachtungszeitraum hinweg fortgeführt. Abschnitt 5.2 befasst sich sowohl mit der technischen Umsetzung als auch mit den Ergebnissen der fortlaufenden Prognosemodelle. Durch die retrospektive Betrachtungsweise ist es zum einen möglich, verschiedene Kandidatenmodelle auf Basis von prädiktiven Maßen hinsichtlich ihrer Prognosequalität zu vergleichen. Zum anderen kann durch den Vergleich von 1-Tages-, 2-Tages- und 3-Tagesprognosen mit der tatsächlich beobachteten Fallzahl beurteilt werden, wie stark die Prognosequalität abnimmt, je weiter die zu prädiktierende Fallzahl in der Zukunft liegt (vgl. Abschnitt 5.3). Schließlich wird in Abschnitt 5.4 ein quantilbasiertes Verfahren zur Einteilung der prognostizierten Fallzahlen in Gefährdungskategorien vorgestellt.

5.1 Konstruktion verschiedener Trainingsmodelle basierend auf den bisherigen Ergebnissen

Die Vorgehensweise bei der Prognose wird durch Abbildung G.17 veranschaulicht. Ziel ist es sowohl für die Call-Center- als auch für die Abrechnungsdaten diejenige Kovariablenkonfiguration zu finden, welche die größte Prognosequalität besitzt. Die Kandidatenmodelle werden aus den Ergebnissen der schrittweisen AIC-Selektion (stepaic) und der bayesianischen Shrinkage-Verfahren (shrinkage) gebildet. Wie bereits am Ende von Abschnitt 3.8 erwähnt, beruht die Kovariablenauswahl aufgrund der höheren Kovariablenqualität und des potentiellen Informationsverlusts durch die Aggregation über die Landkreise ausschließlich auf den Selektionsmodellen für die Pilotregion München. Um zu überprüfen, ob die ausgewählten Meteorologie- und Luftqualitätsparameter überhaupt eine Verbesserung der Prognose bewirken, wurde zum Vergleich auch ein Prognosemodell (admin) nur basierend auf den designbedingten und administrativen Kovariablen sowie dem Deprivationsscore betrachtet. Umgekehrt wurde die Prognose auch mit allen zur Verfügung stehenden Kovariablen durchgeführt (full), um zu testen, ob das Entfernen von Kovariablen ohne nachweisbaren Einfluss auf die Fallzahl tatsächlich zu einer Verbesserung der Vorhersage führt. Um eine krankheitsspezifische Prognose zu erhalten, wurden bei den Abrechnungsdaten die Gesamtfallzahlen sowie die Asthma- und COPD-spezifischen Anzahlen jeweils separat betrachtet, die drei Modelle besitzen jedoch jeweils dieselben Kovariablen. Tabelle G.18 fasst die Kovariablenkonfiguration aller betrachteten Trainings- und Prognosemodelle auf einen Blick zusammen. Es ist zu erkennen, dass die Shrinkage-Modelle tendenziell etwas sparsamer parametrisiert sind, vor allem hinsichtlich der Luftschadstoffe.

Die Trainingsmodelle basieren auf den Daten aus dem Jahr 2006 und werden in 2 Stufen angepasst. Auf der ersten Stufe wird zunächst der über alle (Geschlechts- und) Altersgruppen aggregierte Datensatz beschränkt auf das Jahr 2006 verwendet, um die räumliche Datenstruktur mithilfe von GLMMs (vgl. Abschnitt 3.5) zu erfassen. Aufgrund der kleinen landkreisspezifischen Fallzahlen bei den Call-Center-Daten wird hier wiederum ein Poisson-Modell verwendet, bei den Abrechnungsdaten dagegen ein loglineares Modell. Der geschätzte strukturelle räumliche Effekt fließt dann als zusätzlicher Modelloffset in den zweiten Modellierungsschritt mit ein. Der Versuch, die GLMMs basierend auf den vollen Datensätzen für das Jahr 2006 zu fitten, scheiterte an der großen Beobachtungszahl ($n = 175200$ bei den Call-Center-Daten bzw. $n = 280320$ bei den Abrechnungsdaten). Verzögerte Kovariableneffekte sollten bereits auf der ersten Stufe berücksichtigt werden, jedoch besteht bei der verwendeten BayesX-Funktion „bayesreg“ nicht die Möglichkeit, die für die PDLF erforderliche Penalisierung durchzuführen. Aus diesem Grund wurde die Almon-Methode (vgl. Abschnitt 4.3) bei allen Meteorologie- und Luftqualitätsparametern angewandt, bei denen ein Hinweis auf die Existenz eines solchen Effekts bestand.

Insgesamt ergibt sich folgende Modellgleichung für die GLMMs:

$$\begin{aligned}\log(\mu_{st}) &= \text{offset}(\log(x_{\text{inhabitants},s})) + \text{offset}(\log(x_{\text{doctors},s})) + \beta_{\text{Intercept}} \\ &\quad + x_{\text{deprivation},s}\ \beta_{\text{deprivation}} + \boldsymbol{x}_{\text{admin},t}^{\top}\ \boldsymbol{\beta}_{\text{admin}} \\ &\quad + \boldsymbol{x}_{\text{select},st}^{\top}\ \boldsymbol{\beta}_{\text{select}} + f(t) + \alpha_s + \gamma_s.\end{aligned}$$

Darin beinhaltet $\boldsymbol{x}_{\text{admin}}$ alle administrativen Kovariablen und $\boldsymbol{x}_{\text{select}}$ alle Almon-Terme bzw. Haupteffekte der im jeweiligen Modell enthaltenen Meteorologie- und Luftqualitätsparameter. Beim administrativen Modell entfällt $\boldsymbol{x}_{\text{select}}$ komplett. α_s und γ_s bezeichnen wie in Abschnitt 3.5 die zufälligen und strukturellen räumlichen Effekte. Der grau markierte Offset-Term für die Arztdichte wird nur bei den Abrechnungsmodellen verwendet. Die den Prognosemodellen zugrunde liegenden $\hat{\gamma}_s$ finden sich auf der CD in Appendix I im Ordner „Datensätze“. Vergleicht man die unterschiedlichen Kandidatenmodelle hinsichtlich der geschätzten strukturellen räumlichen Effekte, fallen keine schwerwiegenden Unterschiede auf.

Auf der zweiten Stufe wurde ein (penalisiertes) GLM, basierend auf den vollen Daten des Jahres 2006, also aufgegliedert nach Landkreis und den designbedingten Kovariablen, mithilfe der neu entwickelten R-Funktion „lag_regress“ angepasst. Verzögerte Kovariableneffekte bei den Luftschadstoffen werden hier mithilfe der PDLF-Methode modelliert (maximales Lag $L = 14$), Lag-Effekte bei den Meteorologie-Parametern dagegen weiterhin mit der Almon-Methode (maximales Lag $L = 3$), um vor allem im Hinblick auf das fortlaufende Prognosemodell die Rechenzeit zu reduzieren. Der Tuningparameter a, der die Stärke des Shrinkage der DLF-Koeffizienten steuert wurde für die Trainingsmodelle basierend auf den Call-Center-Daten auf $a = 3.5$ und für die Trainingsmodelle basierend auf den Abrechnungsdaten auf $a = 2.5$ festgelegt, um ein angemessenes Shrinkage zu erzielen. Für die Penalisierung wurden weiterhin Differenzen zweiter Ordnung verwendet. Als Polynomgrad für die Almon-DLFs wurde wiederum $d = 2$ gewählt. Die untere Grenze des Wertebereichs der Penalisierungsparameter wurde bei den Modellen für beide Datenquellen auf $\lambda_{\min} = 1000$ gesetzt. Die Modellgleichung der finalen Trainingsmodelle lautet

$$\begin{aligned}\log(\mu_{ijst}) &= \text{offset}(\log(x_{\text{inhabitants},ijs})) + \text{offset}(\log(x_{\text{doctors},s})) + \text{offset}(\log(\hat{\gamma}_s)) \\ &\quad + \beta_{\text{Intercept}} + x_{\text{deprivation},s}\ \beta_{\text{deprivation}} + \boldsymbol{x}_{\text{admin},t}^{\top}\ \boldsymbol{\beta}_{\text{admin}} + \boldsymbol{x}_{\text{design},ij}{}^{\top}\ \boldsymbol{\beta}_{\text{design}} \\ &\quad + \boldsymbol{x}_{\text{select},st}^{\top}\ \boldsymbol{\beta}_{\text{select}} + f(t) + \sum_{i=2}^{I} f_i(t). \qquad (15)\end{aligned}$$

Die grau markierten Geschlechtsindizes sowie der Offset-Term für die Arztdichte entfallen für die Call-Center-Daten. Um mögliche Unter- oder Überdispersion zu berücksichtigen, wurde hier bei den Call-Center-Daten anstelle eines gewöhnlichen

Poisson-Modells ein Quasi-Poisson-Modell gefittet. Die auf den Trainingsdaten basierenden (penalisierten) GLMs stellen zugleich den Ausgangspunkt des im folgenden Abschnitt beschrieben iterativen Modellierungsprozesses dar.

Es wurde auch erwogen, die R-Funktion „gamm“ (Paket „mgcv“) anstelle der in die „lag_regress“-Methode integrierte „gam“-Funktion zu verwenden, um wie in den bayesianischen GLMMs zusätzlich zum Modelloffset für den strukturellen räumlichen Effekt einen zufälligen Landkreiseffekt spezifizieren zu können. Aus zeitlichen Gründen war dies für die Simulation der fortlaufenden Modellierung allerdings nicht realisierbar, da die Schätzung eines Modells bei Verwendung der „gam“-Funktion mit vorgebenen Penalisierungsparametern bereits ca. 30 Minuten benötigt. Bei impliziter Bestimmung der Penalisierungsparameter ergibt sich eine Rechenzeit von ca. 2 Stunden. Die Schätzung mit der „gamm“-Funktion dagegen benötigt bei den vollen Trainingsdaten etwa 20 Stunden und erwies sich bei größerer Datenmenge als instabil.

5.2 Fortlaufend „lernendes“ retrospektives Prognosemodell

In diesem Abschnitt wird zunächst der Ablauf des täglich aktualisierten Prognosemodells erläutert. Anschließend wird kurz auf den zeitlichen Verlauf der geschätzten Parameter sowie der gefitteten DLFs in den Prognosemodellen mit allen Kovariablen eingegangen. Um die Analyse der Kovariableneffekte zu komplettieren, werden ausgewählte Ergebnisse der finalen Modelle des Prognoseprozesses präsentiert. Schließlich werden beispielhaft erste Prognoseergebnisse dargestellt.

Der R-Code zur Realisierung des fortlaufenden Prognoseprozesses kann Appendix H.4 entnommen werden. Zunächst werden die Knoten für die altersspezifischen Zeittrends festgelegt, wobei die Anzahl der Knoten an die Länge des Beobachtungszeitraums angepasst wird, der durch die Hinzunahme neuer Beobachtungen permanent erweitert wird. Wie bereits in den GLMs für die reduzierten Datensätze (vgl. Abschnitt 3) werden die Knoten immer im Abstand von 3 Monaten, das heißt zu Beginn/Ende eines neuen Quartals gesetzt. Dementsprechend ergeben sich 5 Knoten für die Zeittrends im Trainingsmodell. Sobald der Beobachtungszeitraum den ersten Tag eines neuen Quartals umfasst, wird ein weiterer Knoten am Ende des angebrochenen Quartals ergänzt. Da bei den Call-Center-Daten die Simulation über 2 Jahre (731 Tage) und bei den Abrechnungsdaten über 1 Jahr (365 Tage) hinweg erfolgt, muss die gleiche Anzahl an Modellen gefittet werden. Um die gesamte Rechenzeit zu begrenzen wurde daher darauf verzichtet, die Penalisierungsparameter für die PDLF-Terme in jedem Schritt neu zu bestimmen. Stattdessen werden die Penalisierungsparameter immer dann neu gewählt, wenn ein neuer Knoten für die altersspezifischen Zeittrends hinzugenommen wird, also zu Beginn eines neuen Quartals. Für die restlichen Modelle innerhalb des Quartals werden diese Schätzun-

gen der Penalisierungsparameter vorgegeben. Vor Beginn der Iteration werden leere Datensätze erzeugt, in denen später die Parameterschätzer und Standardfehler der einzelnen Modelle sowie die prädiktierten Werte und Prognosefehler gespeichert werden können. Zudem wird festgelegt, in welchem Abstand Trace-Plots erstellt werden sollen, welche die Almon-DLFs und PDLFs der jeweils aktuellen Modelle darstellen. Dies erfolgt etwa einmal pro Monat.

Innerhalb der Iteration werden zunächst die Eingabeparameter für die Funktion „lag_regress“ festgelegt und die Modellgleichung definiert, die sich aufgrund der fortlaufenden Aktualisierung der Zeitsplines verändert. Die Eingabeparameter zur Steuerung der Form der DLFs a, $\lambda_{\min}$ und d werden analog zu den Trainingsmodellen gewählt. Anschließend wird die „lag_regress“-Funktion aufgerufen und das penalisierte GLM gefittet. Je nach Update-Rhythmus werden die Penalisierungsparameter $\hat{\boldsymbol{\lambda}}_c$ für die PDLFs entweder neu bestimmt oder es werden die Werte aus der vorangegangenen Iteration $\hat{\boldsymbol{\lambda}}_{c-1}$ verwendet. c stellt dabei den Iterationsindex dar, der von 0 (ursprüngliches Trainingsmodell) bis $C = 730$ bei den Call-Center-Daten und bis $C = 364$ bei den Abrechnungsdaten läuft. In der Folge werden die Schätzwerte $\hat{\boldsymbol{\delta}}_c$ und $\text{se}(\hat{\boldsymbol{\delta}}_c)$ aus dem zurückgegebenen „gam“-Objekt extrahiert und abgespeichert. $\hat{\boldsymbol{\delta}}_c$ enthält Schätzwerte für alle administrativen und designbedingten Kovariablen, die Deprivation, die Splinekoeffizienten der Zeittrends, die ausgewählten Cutpoint-Variablen und diejenigen Meteorologie- und Luftqualitätsparameter, die auf den Haupteffekt reduziert wurden. Ebenso werden aus den Schätzungen $\hat{\boldsymbol{\gamma}}_c$ und $\hat{\boldsymbol{\alpha}}_c$ die Lag-Koeffizienten für Almon-DLFs und PDLFs bestimmt. Die Darstellung erfolgt abhängig vom zuvor festgelegten Rhythmus für die Trace-Plots. Schließlich werden die Penalty-Parameter $\hat{\boldsymbol{\lambda}}_c$ für die nächste Iteration gespeichert.

Für die Prädiktion der Fallzahlen an den 3 Tagen nach Ende des jeweiligen Beobachtungszeitraums wird die zur Funktion „gam“ zugehörige „predict.gam“-Methode (vgl. Wood (2006)) verwendet. Dieser muss sowohl das Modellobjekt basierend auf dem Beobachtungszeitraum 1 bis t als auch eine Datenmatrix $\boldsymbol{X}^*$ mit allen beobachteten Kovariablenwerten des relevanten Prognosezeitraums $t+1$ bis $t+3$ übergeben werden. In der tagesaktuellen Anwendung müssten hier Vorhersagen für die Meteorologie- und Luftqualitätsparameter herangezogen werden. Mithilfe der Funktion „predict.gam“ können daraus die prognostizierten Fallzahlen $\hat{\boldsymbol{y}}^* = (\hat{\boldsymbol{y}}_{t+1}^{*\top}, \hat{\boldsymbol{y}}_{t+2}^{*\top}, \hat{\boldsymbol{y}}_{t+3}^{*\top})^\top$ für jede (Geschlechts- und) Altersgruppe in allen Landkreisen berechnet werden. Zurückgegeben wird der mit der Response-Funktion transformierte lineare Prädiktor, ausgewertet in $\hat{\boldsymbol{\beta}}_c$ ($c = t - 365$), also $\hat{\boldsymbol{y}}^* = \exp(\boldsymbol{X}^* \hat{\boldsymbol{\beta}}_c)$ bei den Quasi-Poisson-Modellen für die Call-Center-Daten (durch Verwendung des log-Links) und $\log(\hat{\boldsymbol{y}}^* + 1) = \boldsymbol{X}^* \hat{\boldsymbol{\beta}}_c$ bei den loglinearen Modellen (durch Verwendung des identischen Links nach Logarithmieren der Fallzahlen). Bei den Abrechnungsdaten erhält man die prognostizierten Fallzahlen $\hat{\boldsymbol{y}}_c^*$ dementsprechend durch die Retransformation $\hat{\boldsymbol{y}}_c^* = \exp(\boldsymbol{X}^* \hat{\boldsymbol{\beta}}_c) - 1$. Aufgrund dieser Umformung können die prädiktierten Fallzahlen theoretisch auch negative Werte annehmen ($\hat{\boldsymbol{y}}_c^* > -1$).

Neben den Prognosewerten können mit der „predict.gam“-Funktion auch Standardfehler für die vorhergesagten Werte geschätzt werden. Bei der zugrunde liegenden Varianzschätzung gilt es generell zu beachten, dass die Varianz des Prognosefehlers $y_i^* - y_i$ für eine Beobachtung i größer ist als die Varianz des Schätzfehlers $\hat{y}_i - y_i$, da bei der Prognose nicht beobachteter y-Werte zusätzliche Unsicherheit ins Spiel kommt. Zudem nimmt die Varianz des Prognosefehlers zu, je weiter der neue Wert x_i^* der Kovariable x vom Zentrum der bisher beobachteten Kovariablenwerte $\bar{x}$ entfernt liegt. Das heißt bei bis auf den Zeitfaktor gleichen Kovariablenwerten nimmt die Varianz der Prognose zu, je weiter die zu prädiktierende Fallzahl in der Zukunft liegt. Bei den Quasi-Poisson-Modellen erhält man automatisch approximative Standardfehler für y_i^* beruhend auf einer Taylor-Entwicklung (vgl. Wood (2006)). Bei den loglinearen Modellen erhält man zunächst Standardfehler für $g(y_i^*) = \log(y_i^* + 1)$. Da die prognostizierte Zielvariable $g(y_i^*)$ im linearen Modell asymptotisch einer Normalverteilung folgt, kann hier die Delta-Methode (vgl. Held (2008)) angewandt werden, um aus diesen Standardfehlern näherungsweise eine Schätzung der Varianz von y_i^* zu berechnen:

$$\begin{aligned}
\widehat{\mathrm{Var}}(y_i^*) &= \left[(g^{-1})'(y_i^*)\right]^2 \cdot \widehat{\mathrm{Var}}(g(y_i^*)) \\
&= \left[(\exp(y_i^*) - 1)'\right]^2 \cdot \left[\mathrm{se}(g(y_i^*))\right]^2 \\
&= \exp(2y_i^*) \cdot \left[\mathrm{se}(g(y_i^*))\right]^2 .
\end{aligned}$$

Bevor eine nähere Betrachtung der Prognoseergebnisse erfolgt, wird an dieser Stelle ein kurzer Blick auf die Entwicklung der Parameterschätzer im Verlauf des Prognoseprozesses ($c = 0$ bis $C = 364$ bzw. $C = 730$) geworfen. Die folgenden Analysen basieren auf den Modellen mit vollem Kovariablensatz (full). Für die designbedingten und administrativen Kovariablen sowie für die Deprivation, die Cutpoint-Variablen, die Windrichtung und die Tagesranges von Temperatur, Luftdruck und Luftfeuchtigkeit wurde der jeweils aktuelle Schätzwert $\hat{\delta}_c$ gegen c aufgetragen. Abbildung 5.1 zeigt exemplarisch den Verlauf des Parameterschätzers für die Deprivation im Prognosemodell für die KVB-Call-Center-Daten. Grau eingezeichnet ist wiederum ein punktweises 95%-Konfidenzband, konstruiert aus den gespeicherten Standardfehlern $\mathrm{se}(\hat{\delta}_c)$. Die vertikalen gestrichelten Linien markieren diejenigen Modelle, bei denen ein Update der Penalisierungsparameter vorgenommen wurde, also den Beginn/das Ende des Quartals.

In diesem Beispiel fällt auf, dass der Schätzwert der Prognosemodelle in einem ähnlichen Bereich liegt wie der Schätzwert der bayesianischen GLMMs basierend auf dem über alle Altersgruppen aggregierten Datensatz (vgl. Tabelle G.11). Die Standardfehler sind im Prognosemodell dagegen im Durchschnitt etwa um den Faktor 10 kleiner. Dies begründet sich in der größeren Beobachtungszahl der Prognosemodelle (n zwischen 350400 und 525600) im Vergleich zu den in Abschnitt 3.5 vorgestellten GLMMs ($n = 105216$). Die Parametertests der Prognosemodelle besitzen also eine größere Power und führen tendenziell zu mehr signifikanten Ergebnissen. Diese

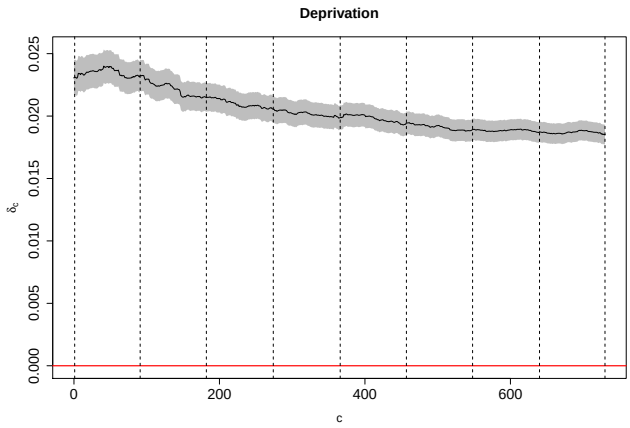


Abbildung 5.1: Verlauf eines Schätzwerts $\hat{\beta}_c$ im Prognoseprozess mit punktweisem 95%-Konfidenzband (grau) am Beispiel des Deprivationseffekts auf die Anrufe beim KVB-Call-Center (Modell mit allen Kovariablen)

Signifikanzen sollten jedoch nicht überbewertet werden, da die große Beobachtungszahl nicht durch die Hinzunahme neuer Beobachtungen, sondern lediglich durch die Aufsplittung der Fallzahlen in die Altersgruppen zustandekommt.

Auf der CD in Appendix I im Ordner „Trace-Plots“ befinden sich die Trace-Plots der Komponenten von $\boldsymbol{\delta}$ für alle Prognosemodelle mit vollem Kovariablensatz (full). Grundsätzlich stimmen die Effekte der administrativen und designbedingten Kovariablen hinsichtlich Stärke und Richtung weitestgehend mit den vorherigen Ergebnissen überein. Beispielsweise pendelt der Schulferieneffekt bei den Call-Center-Daten im Simulationszeitraum (2007 bis 2008) zwischen 0.1 und 0.2. Der mittlere Effekt der Call-Center-Modelle basierend auf den reduzierten Datensätzen ergab sich zu 0.1382 (vgl. Tabelle G.14). Bis auf wenige Ausnahmen zeigen die Trace-Plots einen relativ stabilen Zeitverlauf. Auch die Signifikanzbewertung bleibt trotz der kleineren Standardfehler der Prognosemodelle zum großen Teil bestehen. Eine erwähnenswerte Änderung ergibt sich bei der Betrachtung des Deprivationseffekts auf die Anzahl der Arztbesuche. Hier kann über alle Prognosemodelle hinweg ein stabiler, signifikant positiver Einfluss der Deprivation nachgewiesen werden, während im loglinearen GLMM noch ein nichtsignifikant negativer Effekt festgestellt wurde. Bei den Tagesranges und den Cutpoint-Variablen zeigen die Trace-Funktionen durchweg einen sehr unruhigen Verlauf. Die Schätzwerte hängen offenbar stark von einzelnen Beobachtungen ab, so dass es schwer fällt, verallgemeinerbare Aussagen zu treffen. Häufig alternieren die Funktionen auch mehrfach zwischen signifikant negativen und positiven Effekten. Angesichts dessen, dass bereits in Abschnitt 3 je nach Art der Datenreduktion stark variierende Ergebnisse auftraten, überrascht die Instabilität der Schätzungen in den Prognosemodellen allerdings wenig. Des Weiteren wurden

Trace-Plots für die altersspezifischen Zeittrends der Prädiktionsmodelle mit allen Kovariablen nach dem Schema der Abbildungen G.10 und G.11 erstellt. Diese sind in den gleichen Dateien zu finden wie die Trace-Plots für $\boldsymbol{\delta}$. Die Zeitsplines wurden bis 3 Tage nach Ende des jeweiligen Quartals eingezeichnet, um die für die Prognose an diesen Tagen verwendeten Werte zu visualisieren. Bei starker Vergrößerung der Ansicht ist hier die lineare Fortsetzung der kubischen B-Splines außerhalb des Knotenbereichs zu erkennen. Ein starker Anstieg der Zeitfunktionen fällt bei den Call-Center-Daten im letzten Quartal des Jahres 2008 auf, der bei der Analyse der reduzierten Datensätze nicht zu beobachten war. Es gilt zu beachten, dass sich diese Instabilität der Schätzung am Rand des Simulationszeitraums möglicherweise problematisch auf die angestrebte Prognose der Fallzahlen in diesem Bereich auswirken kann.

Um die Entwicklung der geschätzten Almon-DLFs und PDLFs während des Prognoseprozesses beobachten zu können, wurden auch davon mithilfe der entwickelten Funktion „lag_regress“ Trace-Plots erzeugt. Die CD in Appendix I (Ordner „Trace-Plots“) beinhaltet die Verlaufsgrafiken für alle Meteorologie- und Luftqualitätsparameter. Auch hier darf den sehr engen Konfidenzbändern keine allzu große Bedeutung beigemessen werden. Bei vielen Meteorologie-Parametern weist die geschätzte Almon-DLF erhebliche Schwankungen im zeitlichen Verlauf auf und es ergeben sich wie bereits in Abschnitt 4.3 wenig konsistente Ergebnisse. Anders stellt sich das Bild bei den Luftqualitätsparametern dar. Die Form der DLF verändert sich hier im Laufe des Prognoseprozesses nur leicht, was sicherlich zum Teil auf das verhältnismäßig starke Shrinkage ($a = 2.5$ bei den Abrechnungsdaten und $a = 3.5$ bei den Call-Center-Daten) zurückzuführen ist.

An dieser Stelle soll nun die Analyse der Kovariableneffekte durch die Betrachtung der finalen Modelle des Prognoseprozesses mit vollem Kovariablensatz (full) vervollständigt werden. Diese Modelle basieren auf den nicht reduzierten Datensätzen im jeweiligen gesamten Beobachtungszeitraum und beinhalten somit die komplette zur Verfügung stehende Kovariableninformation. Zunächst konnte im finalen Quasi-Poisson-Modell eine deutliche Unterdispersion wohl aufgrund der zahlreichen Nullen im Datensatz festgestellt werden. Der Dispersionsparameter φ wurde auf 0.5537 geschätzt. Tabelle G.19 beinhaltet die Schätzwerte $\hat{\boldsymbol{\delta}}$ (roh und exponentiell transformiert) aller finalen Modelle des Prognoseprozesses und die zugehörigen p-Werte. Bei den Cutpoints ist wiederum kein exponentiell transformierter Schätzwert angegeben, da die Interpretation nur in Kombination mit dem Haupteffekt sinnvoll ist. Dieser muss hier aus der entsprechenden Almon-DLF extrahiert werden. Im Vergleich zu den in Abschnitt 3 vorgestellten Modellen fallen die Standardfehler und p-Werte durch die erwähnte „künstliche“ Erhöhung der Fallzahl tendenziell kleiner aus, was bei der Interpretation unbedingt zu beachten ist.

Bei der Betrachtung der administrativen Kovariablen fallen insbesondere die positiven Koeffizienten für Feier- und Brückentage bei allen Abrechnungsmodellen auf.

Diese lassen sich wenn überhaupt durch die Instabilität der Schätzung aufgrund der geringen Anzahl solcher Tage im Beobachtungszeitraum erklären (Feiertage: 26, Brückentage: 10). Unterschiede zwischen Asthma und COPD ergeben sich erwartungsgemäß in der Altersverteilung der Fälle. Während COPD eher in den höheren Altersgruppen auftritt ($\exp(\beta_{\text{age=1}}) = 0.52$ im Vergleich zu $\exp(\beta_{\text{age=4}}) = 2.98$) verteilen sich die Asthmafälle gleichmäßiger auf die Altersgruppen ($\exp(\beta_{\text{age=1}}) = 1.06$ im Vergleich zu $\exp(\beta_{\text{age=4}}) = 1.33$). Unter Berücksichtigung der Interaktion zwischen Alter und Geschlecht lässt sich feststellen, dass Frauen in Altersgruppe 4 ($\exp(\beta_{\text{sex=2}}+\beta_{\text{age=4}}+\beta_{\text{sex=2,age=4}}) = 4.71$ bzw. 2.05) stärker von COPD bzw. Asthma betroffen sind als Männer ($\exp(\beta_{\text{age=4}}) = 2.97$ bzw. 1.32). In der jüngsten Altersgruppe sind Jungen ($\exp(\beta_{\text{age=1}}) = 1.06$) stärker von Asthma betroffen als Mädchen ($\exp(\beta_{\text{sex=2}} + \beta_{\text{age=1}} + \beta_{\text{sex=2,age=1}}) = 0.79$). In den Abrechnungsmodellen ergaben sich keine signifikanten Effekte der Cutpoints und Tagesranges für Luftdruck und Luftfeuchtigkeit. Dagegen konnten in allen Modellen signifikante Änderungen des Temperatureffekts im unteren und oberen Cutpoint sowie ein signifikant negativer Einfluss der Temperatur-Tagesrange nachgewiesen werden. Bei den Abrechnungsdaten fällt grundsätzlich eine signifikante Zunahme der Fallzahl bei nördlichen Windrichtungen auf, bei Windrichtungen zwischen Südost und West nimmt die Fallzahl dagegen ab.

Die geschätzten Almon-DLFs und PDLFs der finalen Modelle können der CD in Appendix I (Ordner „Prädiktion“) entnommen werden. Gemeinsame Tendenzen sind bei den meteorologischen Parametern schwerlich auszumachen. Bei beiden Datenquellen zeichnet sich beispielsweise ein positiver Temperatureffekt in den Lags 1 und 2 ab. Die Windgeschwindigkeit besitzt ebenfalls einen positiv signifikanten Effekt am Tag der Beobachtung und am dritten Folgetag. Vor allem bei den Abrechnungsdaten lässt sich immer wieder beobachten, dass der Effekt von Kovariablen mit dem Lag anwächst. Dies ist z. B. bei konvektivem Niederschlag, Luftfeuchtigkeit und Luftdruck zu beobachten. Bei den Call-Center-Daten besitzen Feinstaub, Ozon und Stickstoffdioxid einen signifikant positiven Effekt am gleichen Tag. Hinweise auf einen längerfristigen positiven Effekt ergaben sich darunter allerdings nur bei Stickstoffdioxid. Eine Zunahme von Kohlenstoffmonoxid und Schwefeldioxid bewirkt hingegen eine Abnahme der Fallzahl am gleichen Tag. Längerfristige Effekte traten hier nicht auf. Bei den Abrechnungsdaten konnte übereinstimmend ein positiver Effekt von Stickstoffdioxid und ein negativer Effekt von Kohlenstoffmonoxid am gleichen Tag nachgewiesen werden. Feinstaub und Ozon besitzen hier einen mittelbis langfristigen positiven Effekt, der allerdings erst am Tag nach der Messung der Werte einsetzt. Die PDLF von Schwefeldioxid ist maximal in Lag 0 und fällt danach monoton. Ein signifikanter Effekt kann etwa bis Lag 5 beobachtet werden.

Im Folgenden werden die Prognoseergebnisse detailliert untersucht. Dazu werden exemplarisch die Modelle basierend auf der schrittweisen Variablenselektion (stepaic) herangezogen. Bei den Abrechnungsdaten werden ausschließlich die separaten

Prognosemodelle für COPD und Asthma betrachtet. Ein Vergleich aller Prognosemodelle erfolgt im folgenden Abschnitt 5.3. Tabelle G.20 beinhaltet die prognostizierten Fallzahlen in den Landkreisen München (Landeshauptstadt, 9162) und Berchtesgadener Land (9172) in den Altersgruppen 2 und 4 für den 3. und 13. Januar 2007 mit entsprechenden Prognoseintervallen. Dabei basiert y^*_{t+1} auf der Beobachtung der Fallzahlen bis einschließlich zum Vortag. Für y^*_{t+2} bzw. y^*_{t+3} werden analog nur die Fallzahlen bis 2 bzw. 3 Tage vor dem betrachteten Tag verwendet. Zum Vergleich sind die tatsächlich beobachteten Werte tabelliert. Der 3. Januar 2007 war ein Mittwoch in den Schulferien und in der ersten Woche eines neuen Quartals, allerdings kein Feier- oder Brückentag. Der 13. Januar war ein Samstag außerhalb der Schulferien, kein Feier- oder Brückentag und lag auch nicht in der ersten Woche eines neuen Quartals.

Vergleicht man die wahren Werte mit den prognostizierten Werten fällt auf, dass zum Teil deutliche Abweichungen nach oben und nach unten auftreten, wobei die Abweichungen tendenziell mit der Fallzahl anwachsen. Beispielsweise wurden in München am 3. Januar 2007 640 Arztbesuche wg. COPD von Frauen über 60 Jahren registriert. Die 1-Tagesprognose des Modells dagegen liegt nur bei 394, die 2-Tages- (377) und 3-Tagesprognosen (343) sind sogar noch weiter vom wahren Wert entfernt. Es treten jedoch vereinzelt auch größere Abweichungen bei kleinen Fallzahlen auf, z. B. wurden am 13. Januar in München nur 2 Arztbesuche wg. COPD von Männern zwischen 21 und 40 Jahren gezählt, prognostiziert wurden allerdings 24 (1-Tagesprognose). Bei den Call-Center-Daten werden verhältnismäßig große Fallzahlen aufgrund der vielen Nullen im Datensatz häufig unterschätzt. Etwa in Altersgruppe 4 traten am 13. Januar 9 Fälle in München auf, prognostiziert wurden dagegen nur ca. 6. Die Ursache dafür können nicht ins Modell einbezogene Kovariablen sein, aber auch eine natürliche, rein zufällige Variation der Fallzahlen.

Wie bereits bei der Betrachtung der Standardfehler für die Modellparameter fällt auf, dass die Standardfehler der vorhergesagten Fallzahlen relativ klein sind. Dies begründet sich wiederum zum großen Teil durch die künstlich hohe Beobachtungszahl der Prognosemodelle. Dementsprechend überdecken die berechneten Prognoseintervalle die wahren Fallzahlen sehr selten. Insgesamt liegt die geschätzte Überdeckungswahrscheinlichkeit bei ca. 1% bis 2%. Eine genaue Prognose der Fallzahlen erscheint generell, zumindest basierend auf der vorliegenden Kovariableninformation, kaum möglich. Insgesamt lässt sich aber feststellen, dass hohe prognostizierte Werte grundsätzlich mit hohen wahren Fallzahlen einhergehen und umgekehrt. Aus diesem Grund kann die Konstruktion eines Gesundsheitsindex mit wenigen Kategorien, basierend auf den betrachteten Prognosemodellen, durchaus als realistisches Ziel betrachtet werden (vgl. dazu Abschnitt 5.4).

Um bei den Call-Center-Daten zu überprüfen, ob die Fallzahlen genauer prädiktiert werden können, wenn weniger nicht besetzte Zellen in der Häufigkeitstabelle der Anrufe auftreten, wurde ein zusätzlicher Prognoseprozess simuliert, der auf nach De-

privationsscore gebildeten Landkreis-Clustern beruht. Insgesamt wurden 8 Cluster bestehend aus jeweils 12 Landkreisen gebildet, wobei Cluster 1 aus den Landkreisen mit der höchsten Deprivation besteht und Cluster 8 aus den Landkreisen mit den niedrigsten Deprivationsscores. Abbildung 5.2 zeigt die auf diese Weise konstruierten Cluster. Durch die Aggregation der Fallzahlen innerhalb der Cluster reduziert sich der Prozentsatz der Nullen im Zielvariablenvektor von 86.90% auf 32.78%.

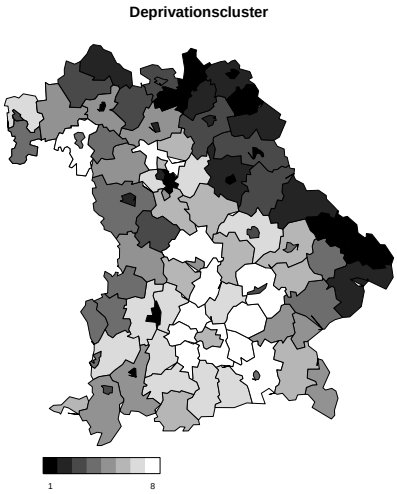


Abbildung 5.2: Deprivationscluster für den Prognoseprozess mit aggregierten Fallzahlen (1: Cluster mit höchster Deprivation, ..., 8: Cluster mit niedrigster Deprivation)

Aufgrund der ungleichmäßigen räumlichen Verteilung der Cluster wird davon ausgegangen, dass keine strukturellen räumlichen Effekte auftreten. Der entsprechende Offset-Term entfällt somit in Modellgleichung (15). Um die Deprivation weiterhin als Kovariable im Modell zu behalten, wurden für den aggregierten Datensatz clusterweise Mittelwerte $\bar{x}_{\text{Deprivation}}$ gebildet. Ansonsten wurden alle zur Verfügung stehenden Kovariablen (full) verwendet.

Tabelle 5.1 zeigt die prognostizierten Fallzahlen mit zugehörigen Prognoseintervallen in Cluster 5 (enthält den Landkreis Berchtesgadener Land) und 6 (enthält die Landeshauptstadt München) wiederum am Beispiel des 3. und 13. Januar sowie der Altersgruppen 2 und 4. Durch die kleinere Beobachtungszahl der Cluster-Modelle ($n/12$) im Vergleich zu den nicht aggregierten Daten nehmen die Standardfehler für die vorhergesagten Werte und damit die Breite der Prognoseintervalle zu. Die Wahrscheinlichkeit, mit den Prognoseintervallen die wahren Werte zu überdecken, liegt nun bei ca. 13%. Größere Fallzahlen werden jedoch tendenziell immer noch unterschätzt. Beispielsweise am 13. Januar wurden in Altersgruppe 4 13 Anrufe in Cluster 6 erfasst, jedoch nur etwa 8 prädiktiert. Betrachtet man die Gesamtheit

x_{cluster}	x_{time}	x_{age}	y	y^*_{t+1}	y^*_{t+2}	y^*_{t+3}
3.1.07	5	2	1	0.89 (0.66, 1.12)	0.90 (0.66, 1.14)	0.95 (0.75, 1.16)
3.1.07	5	4	3	3.71 (3.01, 4.42)	3.71 (2.95, 4.46)	3.17 (2.58, 3.75)
3.1.07	6	2	0	1.94 (1.43, 2.44)	1.96 (1.42, 2.49)	2.07 (1.60, 2.53)
3.1.07	6	4	11	6.53 (5.27, 7.80)	6.50 (5.14, 7.86)	5.53 (4.47, 6.59)
13.1.07	5	2	2	1.03 (0.74, 1.31)	0.98 (0.70, 1.26)	1.01 (0.72, 1.30)
13.1.07	5	4	6	4.81 (3.93, 5.68)	4.65 (3.77, 5.53)	4.96 (4.00, 5.91)
13.1.07	6	2	2	2.14 (1.55, 2.73)	2.05 (1.47, 2.64)	2.10 (1.49, 2.71)
13.1.07	6	4	13	8.10 (6.64, 9.55)	7.83 (6.37, 9.29)	8.33 (6.75, 9.92)

Tabelle 5.1: 1-Tages-, 2-Tages- und 3-Tagesprognosen für das Cluster-Prognosemodell bei den Call-Center-Daten für die Cluster 5 und 6 in den Altersgruppen 2 und 4 am 3. und 13. Januar 2007 mit zugehörigen Prognoseintervallen (in Klammern) im Vergleich zu den tatsächlich beobachteten Werten

der prädiktierten Werte, kann im Vergleich zur landkreisspezifischen Prognose keine nennenswerte Zunahme der Prognosequalität festgestellt werden. Zudem sind die Prognoseergebnisse aufgrund der nicht zusammenhängenden geographischen Lage der Cluster schwer zu interpretieren.

5.3 Predictive Model Checking und Vergleich der verschiedenen Kandidatenmodelle

Dieser Abschnitt beschäftigt sich mit der Bewertung der Prognoseergebnisse anhand von prädiktiven Maßen. Zunächst wird die Prognosequalität in einzelnen Subgruppen der Fallzahlen untersucht. In der Folge werden die 1-Tages-, 2-Tages- und 3-Tagesprognosen gegenübergestellt, um festzustellen, ob und wie stark die Prognosegüte abnimmt, wenn weiter in der Zukunft liegende Fallzahlen prädiktiert werden sollen. Ein besonderes Augenmerk liegt schließlich auf dem Vergleich der verschiedenen Kovariablenkombinationen der Prognosemodelle hinsichtlich ihrer Prognosequalität mit dem Ziel, das bestmögliche Prognosemodell für jede Zielvariable zu finden.

Da die Qualität der Prognose anhand von einzelnen prädiktierten Fallzahlen schwierig zu beurteilen ist, bedarf es eines prädiktiven Maßes, dass die Abweichungen zwischen prädiktierten und wahren Werten in ihrer Gesamtheit quantifiziert. Dazu wird im Folgenden der sogenannte Mean Predictive Squared Error (MPSE) als prädiktives Maß herangezogen:

$$\text{MPSE}(\boldsymbol{y}^*) = \frac{1}{N}\sum_{i=1}^{N}(y_i^* - y_i)^2.$$

N bezeichnet darin die Anzahl der prognostizierten Werte und nicht die den Modellen zugrunde liegende Beobachtungszahl. Der MPSE ist gut geeignet, um die

Unsicherheit der Prädiktion im Ganzen oder in einzelnen Subgruppen zu quantifizieren. Ein Nachteil des MPSE ist allerdings, dass er die Richtung der Abweichungen nicht berücksichtigt. Systematische Über- oder Unterschätzungen können deswegen basierend auf dem MPSE nicht aufgedeckt werden. Da solche Informationen, vor allem im Hinblick auf die Anwendung der Prognoseergebnisse sehr wichtig sind, wird als zusätzliches prädiktives Maß die mittlere Differenz zwischen prädiktieren und beobachteten Werten $\bar{\Delta}\boldsymbol{y}^*$ herangezogen.

Um mögliche Probleme bei der Prognose zu erkennen, werden im Folgenden die prädiktierten Fallzahlen in einzelnen Subgruppen analysiert. Als Beispiel dienen weiterhin die Prognosemodelle basierend auf der schrittweisen Variablenselektion. Grundsätzlich ist bei der Betrachtung gruppenspezifischer MPSEs zu erwarten, dass diejenigen Gruppen mit der größten Fallzahl auch die größten MPSEs besitzen, da die Variabilität der Prognose tendenziell mit der zu prädiktierenden Fallzahl wächst. Dennoch ist das Verhältnis der MPSEs im Hinblick auf die Interpretation der Prognoseergebnisse von Interesse.

Tabelle G.21 zeigt die MPSEs der 1-Tagesprognosen $\boldsymbol{y}^*_{t+1}$ sowie die mittleren Differenzen $\bar{\Delta}\boldsymbol{y}^*_{t+1}$ zwischen prädiktierten und beobachteten Werten auf den Faktorstufen der designbedingten Variablen. In den höheren Altersklassen mit durchschnittlich größerer Fallzahl (vgl. Tabellen G.1 und G.2) können erwartungsgemäß die größten MPSE-Werte beobachtet werden. Besonders auffällig sind die großen MPSEs für Männer in Altersgruppe 4 bei den Arztbesuchen wg. COPD und für Frauen in Altersgruppe 4 bei den Arztbesuchen wg. Asthma. Bei den Abrechnungsdaten wird die Anzahl der Arztbesuche insbesondere in den Altersgruppen 3 und 4 durchschnittlich unterschätzt. Etwa in der Gruppe der über 60-jährigen Männer werden durchschnittlich etwa 44 Arztbesuche wg. COPD zu wenig vorhergesagt. Bei den Call-Center-Daten werden die Fallzahlen über alle Altersgruppen hinweg im Durchschnitt minimal überschätzt, ohne dass eine einzelne Kategorie besonders hervorsticht.

Im Folgenden soll untersucht werden, ob es einzelne Tage gibt, an denen besonders große Abweichungen auftreten. Abbildung G.18 zeigt den Zeitverlauf der Differenzen zwischen prädiktierten und beobachteten Werten. Für die Abbildung wurden die prognostizierten und tatsächlich beobachteten Fallzahlen separat über alle designbedingten Kovariablen und Landkreise aufsummiert. Dargestellt ist die Differenz der jeweiligen Tagessummen. Bei den Call-Center-Daten treten sehr große positive Differenzen am 6. Januar 2007, am 6. Januar 2008 und am 1. November 2008 auf. Die starke Überschätzung der Fallzahlen lässt sich dadurch erklären, dass es sich bei allen 3 Daten gleichzeitig um Wochenend- und Feiertage handelt. Der 6. Januar liegt jeweils noch zusätzlich in den Schulferien und in der ersten Quartalswoche. Die große negative Differenz am 24. Dezember 2007 begründet sich wohl darin, dass der Heilige Abend nicht als Feiertag gerechnet wird, so dass die Fallzahl deutlich unterschätzt wird. Bei den Arztbesuchen wg. COPD und wg. Asthma werden die Fallzahlen im Allgemeinen relativ deutlich unterschätzt, besonders jedoch an einigen

Tagen im April und am 1. Oktober 2007, wobei diese extremen Beobachtungen nicht durch administrative Gegebenheiten zu erklären sind. Große positive Differenzen ergaben sich vermehrt an Feiertagen, etwa am 1., 17. und 28. Mai 2007. Zudem fällt ein periodischer Verlauf der Differenzen im Wochenrhythmus auf. An Wochenenden wird die Fallzahl grundsätzlich leicht überschätzt, unter der Woche dagegen relativ deutlich unterschätzt. Die Berücksichtigung der Kovariablen x_{dow} und x_{holiday} im Prognosemodell reicht offenbar nicht aus, um die auftretenden Unterschiede in den Fallzahlen gänzlich zu erklären. Für die Wochentage könnte man eventuell zusätzlich zu den Dummy-Variablen einen periodischen Trend in den linearen Prädiktor der Prognosemodelle integrieren.

Um noch konkreter zu überprüfen, ob generell an einzelnen Wochentagen, Feiertagen, Schulferientagen, Brückentagen oder in der ersten/letzten Woche eines Quartals systematische Abweichungen von den wahren Fallzahlen auftreten, werden wiederum gruppenweise MPSEs und mittlere Differenzen zwischen prädiktierten und wahren Werten betrachtet. Tabelle G.22 beinhaltet die entsprechenden Maße auf allen Faktorstufen der administrativen Kovariablen für die Anzahl der Anrufe beim Call-Center und der Arztbesuche wg. Asthma bzw. COPD. Die Angaben beruhen weiterhin auf 1-Tagesprognosen aus den durch schrittweise AIC-Selektion konstruierten Prognosemodellen. Obwohl beispielsweise an Wochenenden oder Feiertagen deutlich mehr Anrufe beim Call-Center eingehen (vgl. Abbildung G.6), nehmen die entsprechenden gruppenspezifischen MPSEs nur in geringem Umfang zu. An Brückentagen wird die Anzahl der Anrufe tendenziell stärker überschätzt als an den übrigen Tagen. Bei der Anzahl der Arztbesuche bestätigt sich der optische Eindruck, dass die Fallzahlen unter der Woche zu klein vorhergesagt werden, insbesondere an Montagen, Dienstagen und Donnerstagen, an denen die Arztpraxen gewöhnlich am längsten geöffnet sind. Wie schon in Abbildung G.18 zu erkennen, tritt insgesamt gesehen eine Unterschätzung der Fallzahl auf. Die erste und letzte Quartalswoche, Schulferien- und Brückentage spielen dabei keine wesentliche Rolle. An Feiertagen dagegen wird die Anzahl der Arztbesuche im Durchschnitt deutlich überschätzt, etwa um ca. 21 Fälle bei den Arztbesuchen wg. Asthma.

Schließlich soll ergründet werden, in welchen Landkreisen sich die Vorhersage der Fallzahlen als besonders schwierig erweist. Tabelle G.23 fasst, separat für die Anrufe beim KVB-Call-Center und für die Arztbesuche wg. COPD bzw. Asthma, diejenigen Landkreise mit den höchsten MPSEs zusammen. Zusätzlich sind wieder die zugehörigen mittleren Differenzen zwischen prädiktierten und wahren Werten tabelliert. Die mit Abstand größten MPSEs ergeben sich erwartungsgemäß für die Landkreise mit den größten absoluten Fallzahlen (Landeshauptstadt München und Stadt Nürnberg). Bei den Call-Center-Daten befinden sich mit Augsburg und Passau zwei weitere größere Städte unter den Landkreisen mit maximalem MPSE. In Augsburg und Nürnberg wird die Fallzahl im Durchschnitt überschätzt, in München und Passau werden dagegen tendenziell zu wenig Anrufe vorhergesagt. Bei der Anzahl der

Arztbesuche resultieren die 5 größten landkreisspezifischen MPSEs aus einer durchschnittlichen Unterschätzung der Fallzahlen. Bei den Abrechnungsdaten sind auch ländlichere Regionen unter den Landkreisen mit größtem MPSE, etwa der Landkreis Main-Spessart oder der Landkreis Ansbach. Auch die räumliche Variation der Fallzahlen kann also nur zum Teil durch die Verwendung des geschätzten strukturellen räumlichen Effekts als Offset erklärt werden.

Als nächstes sollen die 1-Tages- mit den 2-Tages- und 3-Tagesprognosen verglichen werden. Auf den ersten Blick (vgl. Tabelle G.20) lässt sich schwer beurteilen, ob die 1-Tagesprognose wirklich am besten abschneidet. Teilweise sind die Abweichungen hier sogar am größten, obwohl am meisten Dateninformation im entsprechenden Modell enthalten ist. Betrachtet man die Gesamtheit aller prognostizierten Werte eines Modells, so besitzen die Prognoseintervalle der 3-Tagesprognosen die größte Wahrscheinlichkeit, den wahren Wert zu überdecken. Dies sagt allerdings wenig über die Güte der 3-Tagesprognosen aus, da diese im Mittel auch die höchste Varianz besitzen.

Tabelle 5.2 zeigt die MPSEs von 1-Tages-, 2-Tages- und 3-Tagesprognosen zusammen mit mittleren Differenzen zwischen prädiktieren und beobachteten Werten für alle berechneten Prognosemodelle. Sowohl die entsprechenden MPSEs als auch die mittleren Differenzen liegen grundsätzlich sehr nahe beisammen. Hinsichtlich des MPSE schneiden die 1-Tagesprognosen bei den administrativen Prognosemodellen (admin) und den auf schrittweiser Variablenselektion basierenden Modellen (stepaic) meist am besten ab. Bei den Shrinkage-Prognosemodellen (shrinkage) und den Prognosemodellen mit vollem Kovariablensatz (full) dagegen besitzt die 3-Tagesprognose häufig den kleinsten MPSE. Im Bezug auf die mittleren Differenzen weisen die 1-Tagesprognosen in der Mehrzahl der Fälle die geringsten Abweichungen auf. Die Qualität der Prognose hängt also offenbar nicht unmittelbar davon ab, wie weit die zu prädiktierende Fallzahl in der Zukunft liegt. Höchstwahrscheinlich könnte man sogar weiter in der Zukunft liegende Fallzahlen ohne wesentlichen Verlust an Genauigkeit prädiktieren. Jedoch gilt es zu bedenken, dass man im tagesaktuellen Prognosebetrieb von der Qualität der Kovariablenvorhersage abhängig ist, die abnimmt, je weiter man in die Zukunft geht. Bei der retrospektiven Simulation dagegen kann man auf tatsächliche gemessene Werte zurückgreifen. Aus diesem Grund sollte man, sofern technisch realisierbar, die aktuellen Fallzahlen in die Prognose einbeziehen, um die bestmögliche Prognose für den nächsten Tag zu erhalten.

Anhand der gleichen Tabelle kann auch ein erster Vergleich der Prognosequalität der unterschiedlichen Kovariablenkombinationen angestellt werden. Eindeutig am besten im Modellvergleich basierend auf MPSE und mittleren Differenzen sind die administrativen Prognosemodelle, die auf die verfügbaren Informationen zu Wetter und Luftqualität gänzlich verzichten. Während die Prognosequalität der Modelle für die Call-Center-Daten noch auf einem vergleichbaren Niveau liegt, bestehen bei allen Abrechnungsmodellen deutliche Unterschiede hinsichtlich der verwendeten

Datensatz	Zeit	admin	stepaic	shrinkage	full
CC	$t+1$	0.2781 (0.0034)	0.2848 (0.0043)	0.2860 (0.0038)	0.2848 (0.0046)
	$t+2$	0.2811 (0.0040)	0.2885 (0.0048)	0.2894 (0.0043)	0.2891 (0.0051)
	$t+3$	0.2829 (0.0042)	0.2899 (0.0050)	0.2908 (0.0044)	0.2901 (0.0052)
AB (gesamt)	$t+1$	667.3 (-3.7044)	1801.2 (-9.3824)	1572.3 (-8.7123)	1887.8 (-9.9569)
	$t+2$	685.9 (-3.7558)	1814.2 (-9.5404)	1545.5 (-8.8661)	1884.8 (-10.0499)
	$t+3$	683.1 (-3.6706)	1806.5 (-9.4646)	1519.1 (-8.6916)	1882.0 (-9.9306)
AB (COPD)	$t+1$	502.5 (-5.2537)	1059.7 (-8.3310)	995.0 (-8.1605)	1061.7 (-8.4011)
	$t+2$	509.2 (-5.3113)	1066.9 (-8.4180)	991.7 (-8.2474)	1063.4 (-8.4603)
	$t+3$	490.8 (-5.1663)	1064.3 (-8.3946)	982.2 (-8.1805)	1058.8 (-8.4104)
AB (Asthma)	$t+1$	239.0 (-1.8587)	451.3 (-4.3253)	396.1 (-4.0921)	477.4 (-4.8436)
	$t+2$	244.5 (-1.8761)	452.0 (-4.3902)	388.5 (-4.1708)	477.5 (-4.8983)
	$t+3$	244.9 (-1.8417)	449.9 (-4.3458)	382.5 (-4.0722)	475.6 (-4.8579)

Tabelle 5.2: Vergleich der 1-Tages-, 2-Tages- und 3-Tagesprognosen aus den Modellen mit unterschiedlicher Kovariablenkombination (admin: Modell ohne Meteorologie- und Luftqualitätsvariablen, stepaic: Modell basierend auf schrittweiser AIC-Selektion, shrinkage: Modell basierend auf Variablenselektion durch Shrinkage-Verfahren, full: Modell mit allen Kovariablen): MPSE (gelb: bestes Modell) und mittlere Differenz zwischen prädiktierten und beobachteten Werten (in Klammern)

prädiktiven Maße. Beispielsweise ist der MPSE des administrativen Modells für die Anzahl der Arztbesuche wg. COPD etwa halb so groß wie der MPSE des zweitbesten Shrinkage-Prognosemodells. Auch die durchschnittliche Unterschätzung der Fallzahlen fällt bei den administrativen Modellen deutlich geringer aus als bei den übrigen Kandidatenmodellen. Etwa bei der Anzahl der Arztbesuche wg. Asthma liegen die Prognosen der administrativen Modelle im Durchschnitt um ca. 2 Fälle unter der tatsächlich beobachteten Fallzahl, die Prognosen der übrigen Modelle dagegen weisen im Mittel eine Unterschätzung zwischen 4 und 5 Fällen auf. Die sparsameren Shrinkage-Modelle schneiden bei den Abrechnungsdaten generell besser ab als die Modelle basierend auf schrittweiser Variablenselektion. Die Modelle mit vollem Kovariablensatz weisen hier die größte Prognoseunsicherheit auf.

Um diese Ergebnisse grafisch zu unterstreichen, wurden Boxplots der MPSEs sowie der Differenzen zwischen prädiktieren und beobachteten Werten für die 1-Tagesprognosen aller Prognosemodelle erstellt (vgl. Abbildung G.19). Aufgrund der besseren Darstellbarkeit wurden dazu datumspezifische MPSE-Werte bzw. Differenzen der Tagessummen von prädiktierten und beobachteten Werten verwendet. Analoge Darstellungen für die 2-Tages- und 3-Tagesprognosen können der CD in Appendix I (Ordner „Prädiktion“) entnommen werden. Anhand der Breite der Boxen (25%- bis 75%-Quantil) sowie der unteren und oberen Zäune (5%- bis 95%-Quantil) lassen sich die deutlich geringeren Abweichungen der vom administrativen Prognosemodell prädiktieren Fallzahlen von den wahren Werten gut erkennen. Auch die datumspezifischen MPSEs der administrativen Modelle weisen eine deutlich geringere Streubreite auf als die entsprechenden Werte der übrigen Modelle. Bei den verschiedenen Prädiktionsmodellen für die Anrufe beim KVB-Call-Center lassen sich auch optisch kaum Unterschiede erkennen.

Des Weiteren wurden die Modelle hinsichtlich ihrer Erklärungsqualität im Bezug auf die Zielvariable verglichen. Im linearen Modell kann dazu das Bestimmtheitsmaß R^2 (vgl. Draper und Smith (1998)) verwendet werden:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

Im Zähler des Bruchs steht die Residuenquadratsumme, also der Anteil der Gesamtvarianz, der nicht durch die Kovariablen im Modell erklärt werden kann, im Nenner die Gesamtstreuung der Zielvariable. Ein Nachteil des Bestimmtheitsmaßes ist allerdings, dass es generell anwächst, wenn neue Kovariablen ins Modell aufgenommen werden, selbst wenn diese nicht zur Verbesserung des Modells beitragen. Das adjustierte Bestimmtheitsmaß R^2_{adj} dagegen berücksichtigt auch die Modellkomplexität in Form der Parameterzahl p und kann bei der Aufnahme zusätzlicher Terme sinken. R^2_{adj} ist wie folgt definiert:

$$R^2_{\text{adj}} = 1 - (1 - R^2)\frac{n-1}{n-p-1}.$$

Bei nicht normalverteilten Fehlern ist das (adjustierte) Bestimmtheitsmaß allerdings weniger geeignet. Hier kann stattdessen der Anteil der erklärten Devianz D_{expl} verwendet werden, der sich ergibt aus

$$D_{\text{expl}} = 1 - \frac{D_{\text{res}}}{D_0}.$$

Darin ist D_{res} die Residualdevianz (vgl. Definition (6) in Abschnitt 3.2.1) und D_0 die Devianz eines Modells nur mit Intercept. Bei den für die Call-Center-Daten verwendeten Quasi-Poisson-Modellen wird in der R-Funktion „gam“ anstelle der regulären Likelihood-basierten Devianz eine Quasi-Devianz berechnet (vgl. Abschnitt 3.2.1 und Wood (2006)).

Tabelle 5.3 zeigt den Anteil der erklärten Devianz D_{expl} aller Prognosemodelle und das adjustierte Bestimmtsheitsmaß R^2_{adj} der loglinearen Prognosmodelle für die Anzahl der Arztbesuche. Bei den Abrechnungsdaten bestätigen sich die Ergebnisse der MPSE-Betrachtung. Der Anteil der erklärten Devianz und das adjustierte Bestimmtsheitsmaß liegen bei den administrativen Prognosemodellen um etwa 10% höher als bei den übrigen Modellen, die vergleichbare Werte aufweisen. Bei den Call-Center-Daten stellte sich im Bezug auf den Anteil der erklärten Devianz das Modell mit vollem Kovariablensatz als bestes Modell heraus. Die Hinzunahme der gesamten Meteorologie- und Luftqualitätsparameter bewirkt allerdings nur eine Steigerung des Erklärungswerts um 0.42%.

Als Fazit des Modellvergleichs lässt sich festhalten, dass die Berücksichtigung der Meteorologie- und Luftqualitätskovariablen kaum zur Verbesserung der Prognose-

Datensatz	admin	stepaic	shrinkage	full
CC	0.3607/–	0.3639/–	0.3635/–	0.3649/–
AB (gesamt)	0.8925/0.8999	0.7739/0.7895	0.7693/0.7853	0.7742/0.7898
AB (COPD)	0.8201/0.8269	0.7372/0.7472	0.7327/0.7429	0.7378/0.7478
AB (Asthma)	0.8674/0.8786	0.7482/0.7696	0.7430/0.7648	0.7486/0.7699

Tabelle 5.3: Vergleich der Modelle mit unterschiedlicher Kovariablenkombination (admin: Modell ohne Meteorologie- und Luftqualitätsvariablen, stepaic: Modell basierend auf schrittweiser AIC-Selektion, shrinkage: Modell basierend auf Variablenselektion durch Shrinkage-Verfahren, full: Modell mit allen Kovariablen) hinsichtlich der erklärten Streuung der Zielvariable: $D_{\text{expl}}/R^2_{\text{adj}}$ (gelb: bestes Modell)

qualität beiträgt. Vor allem bei den Abrechnungsdaten wirken sich die Umweltbedingungen störend auf die Prädiktion der Fallzahlen aus. Aus diesem Grund werden die administrativen Modelle im folgenden Abschnitt zur Konstruktion des angestrebten kategorialen Gesundheitsindex verwendet. Angesichts der dominanten Effekte der administrativen Kovariablen und der Inkonsistenz der Ergebnisse für die Meteorologie- und Luftqualitätsparameter, ist dieses Resultat im Grunde wenig überraschend. Um die Genauigkeit der Prognose zu steigern, ist offensichtlich eine weitere Verbesserung der Kovariablenqualität etwa in Form einer kleinräumigeren und individuelleren Erfassung erforderlich.

5.4 Umsetzung der Prognoseergebnisse in einen kategorialen Gefährdungsindex

In diesem Abschnitt wird ein experimentelles Konzept präsentiert, welches die Umsetzung der prädiktierten Fallzahlen in einen kategorialen, landkreisspezifischen Gesundheitsindex (Health Index, HI) ermöglicht. Entsprechend der Zielgrößen, die dem entwickelten Index zugrunde liegen, kann die Gefährdungsstufe für Atembeschwerden im Allgemeinen (basierend auf den Call-Center-Daten) oder separat für Asthmatiker und COPD-Patienten (basierend auf den Abrechnungsdaten) vorhergesagt werden. Des Weiteren kann der regionale Prognoseindex alters- und geschlechtsspezifisch berechnet werden oder gemeinsam für alle Bevölkerungsgruppen. Die in diesem Abschnitt exemplarisch dargestellten, ausgewählten Indizes beruhen aufgrund der festgestellten Verschlechterung der Prognosequalität durch die Umweltparameter ausschließlich auf den administrativen Modellen.

Ziel des Index ist es, die prädiktierten Fallzahlen in 5 Kategorien einzuteilen, welche die Gefährdungsstufe betroffener Personen beschreiben. Die Kategorien könnten beispielsweise wie folgt bezeichnet werden:

- sehr geringe Gefährdung (-2)
- geringe Gefährdung (-1)

- normale Gefährdung (0)
- hohe Gefährdung (+1)
- sehr hohe Gefährdung (+2)

Die Häufigkeitsverteilung der realisierten Indizes sollte dabei idealerweise symmetrisch um die am stärksten besetzte Kategorie 0 sein (vgl. Abbildung 5.3).

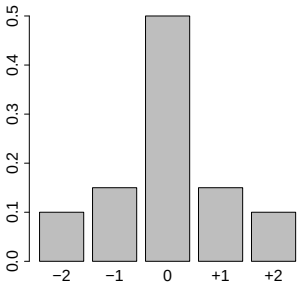


Abbildung 5.3: Angestrebte relative Häufigkeitsverteilung des Prognoseindex

Aufgrund der im vorangehenden Abschnitt beobachteten systematischen Abweichungen der prädiktierten von den wahren Fallzahlen, insbesondere bei den Abrechnungsdaten, muss die Einordnung eines neu prädiktierten Werts zwangsläufig auf Basis der bisher prädiktierten Fallzahlen erfolgen. Idee des entwickelten Index ist es, bestimmte Quantile aller im Simulationszeitraum prädiktierten, alters- und geschlechtsspezifischen Fallzahlen eines Landkreises als landkreisspezifische Grenzen für die Kategorisierung der prognostizierten Werte festzulegen. Die jeweils vorliegenden Ausprägungen der designbedingten und administrativen Kovariablen werden berücksichtigt, indem die Kategoriengrenzen mit den entsprechenden exponentiell transformierten Modellparametern multipliziert werden (vgl. das nachfolgende Rechenbeispiel). Mithilfe der daraus resultierenden transformierten Grenzen können neu prädiktierte Fallzahlen in eine der 5 Kategorien eingeteilt werden. Um obige Häufigkeitsverteilung des Index zu realisieren, wurden zur Berechung der rohen Grenzen zunächst das 10%-, 25%-, 75%- und 90%-Quantil der prädiktierten Werte im jeweiligen Landkreis verwendet. Dabei stellte sich allerdings heraus, dass die transformierten Grenzen für Subgruppen der Fallzahlen stark vom 10%-, 25%-, 75%- und 90%-Quantil der entsprechenden gruppenspezifischen Verteilung der prädiktierten Fallzahlen abwichen. In vielen Landkreisen lagen die prädiktierten Fallzahlen fast ausschließlich in den extremen Indexkategorien. Um dem Rechnung zu tragen, wurden zentralere Quantile (20%-, 40%-, 60%- und 80%-Quantil) zur Festlegung der rohen Grenzen verwendet. Damit ergaben sich die in Abbildung 5.4 dargestellten relativen Häufigkeitsverteilungen der Prognoseindizes im Simulationszeitraum. Um der angestrebten Häufigkeitsverteilung noch näher zu kommen, ist vor allem bei den Abrechnungsdaten eine weitere Anpassung der verwendeten Quantile nötig.

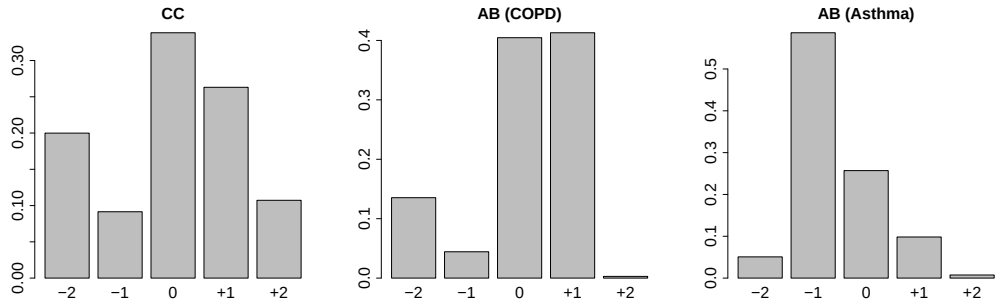


Abbildung 5.4: Häufigkeitsverteilungen der im jeweiligen Simulationszeitraum berechneten Prognoseindizes basierend auf den Call-Center- (links) und Abrechnungsdaten (Mitte: COPD, rechts: Asthma)

Rechenbeispiel zum Prognoseindex:

An dieser Stelle soll die Berechnung eines gruppenspezifischen Gefährdungsindex am Beispiel eines männlichen Asthma-Patienten über 60 Jahre in der Landeshauptstadt München am 1. Januar 2008 demonstriert werden. Die rohen Grenzen basierend auf allen im Simulationszeitraum für die Landeshauptstadt München prognostizierten Fallzahlen liegen bei

$$q_{0.2} = 17.0, q_{0.4} = 119.0, q_{0.6} = 220.0, q_{0.8} = 391.2.$$

Der 1. Januar 2008 war ein Dienstag, Schulferientag und Feiertag und lag in der ersten Quartalswoche. Dementsprechend müssen die rohen Grenzen mit

$$\exp(\hat{\beta}_{\text{age=4}})\cdot\exp(\hat{\beta}_{\text{dow=Di}})\cdot\exp(\hat{\beta}_{\text{school=ja}})\cdot\exp(\hat{\beta}_{\text{holiday=ja}})\cdot\exp(\hat{\beta}_{\text{quartal=Anfang}}) = 0.5409$$

multipliziert werden. In diesem Fall wurden die Parameterschätzer der finalen administrativen Modelle verwendet, die auf allen bis zum 31. Dezember 2007 beobachteten Ziel- und Kovariablenwerten beruhen. Damit ergibt sich folgende Berechnungsvorschrift für den zu prognostizierenden Index HI_1, abhängig von der am 1. Januar 2008 prädiktierten Fallzahl y^*_{t+1}:

$$\text{HI}_1 = \begin{cases} 1 & \text{falls} & & & y^*_{t+1} & < & 9.1951 \\ 2 & \text{falls} & 9.1951 & \geq & y^*_{t+1} & < & 64.3659 \\ 3 & \text{falls} & 64.3659 & \geq & y^*_{t+1} & \leq & 118.9958 \\ 4 & \text{falls} & 118.9958 & > & y^*_{t+1} & \leq & 211.5962 \\ 5 & \text{falls} & & & y^*_{t+1} & > & 211.5962 \end{cases}.$$

Durch die Verwendung der exponentiell transformierten Modellparameter können ohne großen Aufwand für jede mögliche Kombination der administrativen und designbedingten Kovariablen eigene Grenzen spezifiziert werden. Nach Berücksichtigung aller designbedingter und administrativer Kovariablen hängt die Ausprägung

des prognostizierten Index theoretisch nur noch von den übrigen Modellkovariablen ab. Aus diesem Grund ist der praktische Einsatz des Index auf Basis der administrativen Prognosemodelle fragwürdig, da abgesehen vom Deprivationsscore und den altersspezifischen Zeitfunktionen keine weiteren individuellen Einflüsse auf die prädiktierten Fallzahlen wirken. Etwaige Umweltparameter, die eine Steigerung der Prognosequalität bewirken, können jedoch problemlos in die Indexberechnung integriert werden, indem man sie in die zugrunde liegenden Prognosemodelle aufnimmt.

Tabelle G.24 zeigt die aus dem administrativen Modell prädiktierten Fallzahlen y^*_{t+1}, y^*_{t+2} und y^*_{t+3} mit den zugehörigen Prognoseindizes HI_1, HI_2 und HI_3, wiederum am Beispiel der Altersklassen 2 und 4 am 3. und 13. Januar 2007 in der Landeshauptstadt München (9162) und im Landkreis Berchtesgadener Land (9172). Für HI_1 wurden die rohen landkreisspezifischen Grenzen aus den Quantilen der 1-Tagesprognosen bestimmt und die Zuordnung der Indizes auf deren Basis vorgenommen. Analog ergeben sich die Indizes HI_2 und HI_3, die eingesetzt werden können, falls die aktuell beobachteten Fallzahlen erst am nächsten oder übernächsten Tag verfügbar sind. Zur Transformation der rohen Grenzen wurden die exponentiell transformierten Koeffizienten der administrativen Prognosemodelle basierend auf den Ziel- und Kovariablenwerten bis zum Zeitpunkt t verwendet. Die negativen prädiktierten Fallzahlen bei den Arztbesuchen wg. COPD am 13. Januar 2007 entstehen, wie in Abschnitt 5.2 beschrieben, durch die Retransformation der Rückgabewerte der R-Funktion „gam.predict“ auf Response-Ebene. HI_1, HI_2 und HI_3 besitzen meist die gleiche Ausprägung. Die gruppenspezifischen Indizes an den betrachteten Tagen verhalten sich relativ homogen, mit Ausnahme der COPD-Indizes für den 13. Januar im Landkreis Berchtesgadener Land (9172). Während hier für die 20- bis 40-Jährigen jeweils eine sehr niedrige Gefährdung vorhergesagt wird, sind die über 60-Jährigen einer hohen Gefährdung ausgesetzt.

Diese gruppen- und krankheitsspezifischen Indizes können zur gezielten Information betroffener Patienten eingesetzt werden. Beispielsweise für Ärzte oder das Call-Center-Personal ist dagegen eher eine Gesamtprognose, welche das regionale Patienten- bzw. Anruferaufkommen vorhersagt, von Interesse. Um dafür keine eigenen Prognosemodelle berechnen zu müssen, wurden die Indizes aller Geschlechts- und Altersgruppen hier experimentell zu einem landkreisspezifischen Gesamtindex zusammengefasst. Dies erfolgte durch Bildung des Medians $\mathrm{HI}_{\mathrm{median}}$ der jeweiligen 5 (Call-Center-Daten) bzw. 8 (Abrechnungsdaten) Werte. Der Median ist definiert als mittlerer Wert der geordneten Stichprobe bzw. bei gerader Beobachtungszahl als arithmetisches Mittel der beiden mittleren Werte. Um eine eindeutige Zuordnung zu erhalten, die zentralere Kategorien im Zweifelsfall bevorzugt, wurde der Gesamtindex $\mathrm{HI}_{\mathrm{ges}}$ wie folgt berechnet:

$$\mathrm{HI}_{\mathrm{ges}} = \begin{cases} \lfloor \mathrm{HI}_{\mathrm{median}} \rfloor & \text{falls } \mathrm{HI}_{\mathrm{median}} \geq 0 \\ \lceil \mathrm{HI}_{\mathrm{median}} \rceil & \text{falls } \mathrm{HI}_{\mathrm{median}} < 0 \end{cases} .$$

Abbildung G.20 zeigt die auf 1-Tagesprognosen basierenden Gesamtindizes am 3. und 13. Januar 2007 für alle bayerischen Landkreise. Vergleicht man die Abrechnungsindizes für Asthma (unten) und COPD (Mitte) mit dem Call-Center-Index für Atembeschwerden im Allgemeinen (oben), ergeben sich zum Teil deutliche Unterschiede. Beispielsweise am 3. Januar besteht laut Asthma-Index eine niedrige Gefährdung des Auftretens akuter Beschwerden für betroffene Personen, der Call-Center-Index liegt hingegen in weiten Teilen Bayerns beim Wert +1. Eine mögliche Ursache für die abweichenden Ergebnisse ist mit Sicherheit die unterschiedliche Häufigkeitsverteilung der gruppenspezifischen Indizes, die sich weitgehend auf den Gesamtindex überträgt.

Insgesamt betrachtet liefern die vorgestellten Indizes einen intuitiven Ansatz, die Prognoseergebnisse für verschiedene Anwendergruppen nutzbar zu machen, wobei hinsichtlich der Konfiguration sicherlich noch Verbesserungsbedarf besteht.

6 Zusammenfassung und Ausblick

Dieser Abschnitt bietet zunächst einen schwerpunktartigen Überblick über die in dieser Arbeit verwendeten statistischen Methoden sowie die gewonnenen inhaltlichen Erkenntnisse. Abschließend folgt eine kurze Diskussion der Ergebnisse und eine Zusammenstellung von Ansatzpunkten für eine weitere wissenschaftliche Auseinandersetzung mit der bearbeiteten Thematik.

6.1 Überblick über die verwendeten Methoden

Als erstes wurden die Zielvariablen sowie die zur Verfügung stehenden Kovariablen separat betrachtet, um einen Einblick von der zeitlichen und räumlichen Korrelationsstruktur zu gewinnen. Um sich der Frage nach dem Zusammenhang zwischen den beobachteten Fallzahlen und den Umweltbedingungen als potenziellen Einflussfaktoren zu nähern, wurden paarweise Korrelationen zwischen Ziel- und Kovariablenwerten betrachtet. Insbesondere aufgrund der hohen Korrelationen innerhalb der Meteorologie- und Luftqualitätsparameter erschien es erforderlich, eine simultane Analyse der Kovariableneffekte durchzuführen. Dazu wurden verschiedene Regressionsmodelle herangezogen, die eine Quantifizierung der auftretenden Effekte sowie eine Beurteilung ihrer Signifikanz ermöglichen.

Durch die Aufsplittung der Fallzahlen nach Landkreis, Alter und Geschlecht ergab sich eine sehr große Anzahl an Beobachtungen. Da gleichzeitig eine große Menge zum Teil hochkorrelierter Kovariablen zur Verfügung stand, mussten die verwendeten Call-Center- und Abrechnungsdaten zunächst reduziert werden, um die Rechenzeit der Modelle im Hinblick auf die geplante Durchführung einer schrittweisen Variablenselektion zu begrenzen und ihre Konvergenz zu gewährleisten. Um sowohl die zeitliche als auch die räumliche Datenstruktur zu erfassen, wurden verschiedene Reduktionsverfahren verwendet (Beschränkung auf die Pilotregion Landeshauptstadt München und Aggregation über alle bayerischen Landkreise bzw. über alle Alters- und Geschlechtsgruppen).

Zur Analyse der vorliegenden Zähldaten wurden verschiedene generalisierte Regressionsansätze vorgestellt. Für die Call-Center-Daten wurden aufgrund der geringen Anruferzahlen in den einzelnen Landkreisen und Altersgruppen klassische Poisson-GLMs gefittet. Zur Berücksichtigung möglicher Über- oder Unterdispersion wurden Negativ-Binomial-, Quasi-Poisson- und Zero-Inflated-Poisson-Modelle betrachtet. Es trat allerdings nur eine sehr geringe Überdispersion auf, was den Einsatz von Poisson-Modellen für die weiteren Analysen rechtfertigte. Die Abrechnungsdaten wurden durch lineare Modelle mit logarithmisch transformierter Zielvariable analysiert.

Um die longitudinale Datenstruktur und die daraus resultierenden Autokorrelationen in den Fallzahlen aufeinanderfolgender Tage zu berücksichtigen, wurden altersspezifische Zeittrends in die GLMs integriert, die in B-Spline-Basisfunktionen mit quartalsweisen Knoten entwickelt wurden. GEEs konnten aufgrund der großen Datenmenge nicht eingesetzt werden. Die Kovariablen Luftdruck, Luftfeuchtigkeit und Temperatur wurden auf das Auftreten nichtlinearer Effekte in Form einer Comfort Range untersucht. Dies erfolgte durch Aufnahme spezieller Cutpoint-Variablen in den linearen Prädiktor der GLMs. Die Lokalisationen der optimalen Cutpoints wurden zuvor datengesteuert durch AIC-Selektion bestimmt. Um die räumliche Datenstruktur zu erfassen, wurden bayesianische gemischte Modelle herangezogen. Die landkreisspezifischen räumlichen Effekte wurden darin in eine unstrukturierte und eine systematische Komponente aufgeteilt.

Schließlich wurden die Ergebnisse aller auf den reduzierten Datensätzen beruhenden Modelle zusammengefasst, um gemeinsame Tendenzen aufzudecken. Für die Durchführung der Variablenselektion wurden zusätzliche gelagte Kovariablen in den Prädiktor aufgenommen, um einen ersten Eindruck von der Existenz verzögerter Effekte zu erhalten. Neben der standardmäßigen schrittweisen AIC-Selektion wurden bayesianische Ridge- und Lasso-Shrinkage-Modelle verwendet, um Kovariablen mit großem Einfluss auf die Zielvariable deutlicher von irrelevanten Größen abzugrenzen. Die aus der Variablenselektion basierend auf der Pilotregion München resultierenden Modelle wurden für die spätere Konstruktion der Prognosemodelle herangezogen.

Im Anschluss erfolgte eine methodische Diskussion verschiedener Verfahren zur Berücksichtigung verzögerter Kovariableneffekte in Regressionsmodellen, veranschaulicht durch Beispiele basierend auf den reduzierten Datensätzen. Ziel dieser Verfahren ist es, die Distributed Lag Function zu schätzen, welche den Effekt der betrachteten Kovariable in Abhängigkeit vom jeweiligen Lag darstellt. Neben einigen naiven Ansätzen wurden die Solow- und Almon-Lag-Modelle zur Schätzung der DLF betrachtet, die auf verschiedenen Restriktionen der Lag-Koeffizienten beruhen. Darüber hinaus wurde ein bayesianischer Ansatz nach Welty et al. (2009) erläutert, der bei den vorliegenden Daten gute Ergebnisse zeigte, sich allerdings als sehr rechenzeitaufwendig erwies. Mit der dazugehörigen Software war es zudem nicht möglich, mehrere Kovariablen mit verzögertem Effekt gleichzeitig zu betrachten. Schließlich wurde ein eigenes P-Spline-basiertes Verfahren (Penalized Distributed Lag Function) präsentiert, das bei relativ geringem Zeitaufwand gute Resultate lieferte, vor allem bei Kovariablen mit langfristigem Lag-Effekt. Die Schätzung der DLF erfolgt dabei weitgehend datengesteuert und kann gleichzeitig für mehrere Kovariablen durchgeführt werden. Außerdem wird die Glattheit der geschätzten Funktion durch eine Penalisierung von Differenzen benachbarter Lag-Koeffizienten kontrolliert. Mit wachsendem Lag erfolgt ein zunehmendes Shrinkage der DLF gegen 0, wodurch sich auch die Varianz der geschätzten Funktion reduziert, allerdings auf Kosten der Datentreue. Um die Anwendung zu vereinfachen, wurde die PDLF-Methode ge-

meinsam mit dem Almon-Verfahren in der R-Funktion „lag_regress“ implementiert. Diese ermöglicht das Fitten penalisierter GLMs für longitudinale Datenstrukturen basierend auf der R-Funktion „gam“ (vgl. Wood (2006)).

Aufbauend auf den bisherigen Ergebnissen wurden verschiedene Prognosemodelle zur Vorhersage der Fallzahlen konstruiert. Zunächst wurden dazu zweistufige Trainingsmodelle basierend auf den Daten des Jahres 2006 gerechnet. Auf der ersten Stufe wurde der strukturelle räumliche Effekt mithilfe bayesianischer GLMMs geschätzt. Dieser wurde dann als Modelloffset der Regressionsfunktion „lag_regress“ übergeben. Auf diese Weise konnten auch verzögerte Effekte, Zeittrends und Cutpoint-Effekte in die Modelle integriert werden. Um einen Einsatz der Prognosemodelle im tagesaktuellen Betrieb zu simulieren, wurde nach Bestimmung von 1-Tages-, 2-Tages- und 3-Tagesprognosen die Menge der Trainingsdaten stückweise erweitert. Neben den Punktprognosen wurden auch approximative Prognoseintervalle berechnet. Durch die retrospektive Betrachtungsweise standen auch die jeweils beobachteten wahren Fallzahlen zur Verfügung, so dass ein Vergleich der Prognosequalität zwischen 1-Tages-, 2-Tages- und 3-Tagesprognosen, aber auch zwischen den verschiedenen Kovariablenkombinationen der Modelle möglich war. Dazu wurden der Mean Predictive Squared Error (MPSE) und die mittlere Differenz zwischen prädiktierten und wahren Werten als prädiktive Maße herangezogen. Auch der Anteil der erklärten Variation der Zielvariable wurde in die Bewertung miteinbezogen. Abschließend wurde ein experimentelles Konzept zur Umsetzung der Prognoseergebnisse in einen kategorialen Gesundheitsindex präsentiert, der die Gefährdungslage betroffener Patienten vorhersagt. Die Kategoriengrenzen des Index werden dabei aus landkreisspezifischen Quantilen der prädiktieren Fallzahlen berechnet und mithilfe von geschätzten Modellparametern bezüglich der designbedingten und administrativen Kovariablen adjustiert. Der entwickelte Index ist sowohl für Patienten direkt als auch durch eine Modifikation für medizinisches Personal nutzbar.

6.2 Zusammenfassung der Ergebnisse

Bereits bei der anfänglichen Betrachtung der Korrelationen zwischen den Fallzahlen und den Meteorologie-Parametern konnten nur schwache Zusammenhänge beobachtet werden. Vergleichsweise hohe positive Korrelationen mit der Anzahl der Arztbesuche zeigten Stickstoffdioxid, Feinstaub- und Kohlenstoffmonoxid.

Bei den Regressionsergebnissen fällt im Allgemeinen auf, dass die Effekte der designbedingten und administrativen Kovariablen deutlich stärker ausgeprägt sind als die Effekte der Meteorologie- und Luftqualitätsparameter. Für letztere ergaben sich sehr unterschiedliche und teilweise widersprüchliche Aussagen, die zwar für den jeweils betrachteten Datensatz gültig, jedoch nur schwer verallgemeinerbar sind. Im Folgenden werden über mehrere Modelle hinweg übereinstimmend beobachtete Tendenzen

zusammengestellt.

In den räumlichen Modellen erwies sich die Deprivation als signifikanter Einflussfaktor auf die Anzahl der Anrufe beim KVB-Call-Center aufgrund von Lungenerkrankungen im Allgemeinen und der Arztbesuche wegen COPD und/oder Asthma. Die Fallzahlen steigen um etwa 1% bis 2% pro Zunahme des Deprivationsscores um eine Einheit. Erwartungsgemäß ergaben sich signifikante Alters- und Geschlechtseffekte. Bei den über 60-Jährigen treten generell wesentlich mehr Fälle auf als in den jüngeren Altersgruppen. Betrachtet man die Arztbesuche wegen Asthma separat, ist eine Zunahme der Fallzahl bei männlichen Patienten bis 20 Jahren zu erkennen. Der Geschlechtseffekt variiert abhängig von der Altersgruppe. Bezogen auf die Gesamtpopulation sind Frauen stärker von Asthma betroffen als Männer, bei Männern treten dagegen mehr COPD-Fälle auf. In den administrativen Effekten spiegeln sich deutlich die Öffnungszeiten der Arztpraxen wider. Eine signifikante Zunahme der Call-Center-Anrufe bei einer gleichzeitigen Abnahme der Arztbesuche konnte an Wochenenden, Feiertagen, Schulferientagen und größtenteils auch an Brückentagen und in der letzten Quartalswoche nachgewiesen werden.

Keine übereinstimmenden Hinweise auf einen signifikanten Effekt ergaben sich für großskaligen und konvektiven Niederschlag sowie für den Oberflächenwärmefluss. Der Luftfeuchtigkeit wird zum Teil ein signifikant negativer Effekt unterstellt, der sich mit wachsendem Lag verstärkt. Ebenso konnten für den Luftdruck signifikant negative verzögerte Effekte nachgewiesen werden, jedoch kein Einfluss auf die Fallzahlen am gleichen Tag. Für die Temperatur konnten vereinzelt positive Effekte, die möglicherweise mit 1 bis 2 Tagen Verspätung eintreten, festgestellt werden. Bayernweit betrachtet ergab sich durchgängig eine Zunahme der Fallzahl unterhalb der unteren Temperatur-Cutpoints, die für die Anrufe beim KVB-Call-Center bei ca. $2°C$ und für die Arztbesuche bei ca. $-4°C$ Tagesdurchschnittstemperatur liegen. Die Tagesranges von Luftfeuchtigkeit, Luftdruck und Temperatur besitzen wenn überhaupt sehr geringe Effekte. Die Effekte von niedriger und mittelhoher Bewölkung, Windgeschwindigkeit und Surface Stress sind aufgrund ihrer Inkonsistenz schwer zu beurteilen. Gleiches gilt für die Windrichtung. Einzig bei den finalen Modellen des Prognoseprozesses für die Anzahl der Arztbesuche konnte ein strukturierter Effekt nachgewiesen werden. Demzufolge steigt die Fallzahl bei nördlichen Windrichtungen und sinkt bei Windrichtungen von Südost bis West.

Schwefeldioxid weist bezogen auf die Abrechnungsdaten einen positiven Effekt bis etwa zum Lag 5 auf. Bei den Call-Center-Daten dagegen ist lediglich ein negativ signifikanter Haupteffekt zu beobachten. Feinstaub und Ozon besitzen einen signifikant positiven Effekt auf die Fallzahlen. Bei den Call-Center-Daten beschränkt sich dieser auf den gleichen Tag, während bei den Abrechnungsdaten ein kurz- bis mittelfristiger Effekt vorliegt, der in Lag 1 am stärksten ausgeprägt ist. Für Schwefeldioxid kann häufig ein positiver Effekt in Lag 0 nachgewiesen werden. Zum Teil bestehen auch Hinweise auf einen längerfristigen positiven Effekt. Eine Erhöhung

der Kohlenstoffmonoxid-Konzentration bewirkt dagegen eine Abnahme der Fallzahl am gleichen Tag. Die Existenz verzögerter Effekte ist hier eher unwahrscheinlich.

Bei der Betrachtung der Prognoseergebnisse für die Call-Center-Anrufe fiel eine Unterschätzung verhältnismäßig großer Fallzahlen, ausgelöst durch die zahlreichen Nullen im Zielvariablenvektor, auf. Diese Problematik konnte auch durch eine Zusammenfassung der Landkreise in nach Deprivationsscore gebildeten Clustern nicht behoben werden. Im Mittel wurden die Anruferzahlen leicht überschätzt. Bei den Prognosemodellen für die Abrechnungsdaten ergab sich trotz der Dummy-Variablen für die Wochentage eine deutliche Unterschätzung der Fallzahlen zwischen Montag und Freitag. Auch in den höheren Altersgruppen wurden im Durchschnitt zu wenige Arztbesuche prognostiziert. Bei beiden Datenquellen traten einzelne besonders große Abweichungen häufig an Feiertagen auf. Die größte Prognoseunsicherheit ergab sich generell in den Gruppen mit den höchsten Fallzahlen, das heißt in Bezug auf die Landkreise vor allem in den Großstädten München und Nürnberg.

Bei der Mehrzahl der Prognosemodelle wiesen die 1-Tagesprognosen die geringsten MPSE-Werte und die geringsten mittleren Differenzen zwischen prädiktierten und beobachteten Werten auf. Unter der Voraussetzung, dass aufgrund der retrospektiven Betrachtungsweise keine Vorhersage der Kovariablenwerte notwendig war, lieferten die 2-Tages- und 3-Tagesprognosen allerdings eine vergleichbare Prognosequalität. Im Vergleich der verschiedenen Kovariablenkombinationen schnitt das administrative Modell, das gänzlich auf die Berücksichtigung von Meteorologie- und Luftqualitätseffekten verzichtet, am besten ab. Während die Modelle für die Call-Center-Anrufe durchweg eine vergleichbare Prognosequalität zeigten, ergaben sich bei den Prognosemodellen für die Abrechnungsdaten deutliche Abweichungen hinsichtlich der prädiktiven Maße und des Anteils der erklärten Variation in der Zielvariable. Am zweitbesten wurde die Anzahl der Arztbesuche von den relativ sparsam parametrisierten bayesianischen Shrinkage-Modellen prädiktiert, die noch deutlich vor den Prognosemodellen basierend auf der schrittweisen AIC-Selektion und den Modellen mit vollem Kovariablensatz lagen.

6.3 Diskussion und weitere Forschungsmöglichkeiten

Obwohl in vielen vorhergehenden Studien (vgl. z. B. D'Amato et al. (2002), Atkinson et al. (2001), Harré et al. (1997) und Wen-Chao et al. (2007)) klare Zusammenhänge zwischen Meteorologie- bzw. Luftqualitätsparametern und der lungenspezifischen Morbidität nachgewiesen werden konnten, ergaben sich diesbezüglich in dieser Arbeit relativ inkonsistente Ergebnisse, die kaum zur Verbesserung der entwickelten Prognosemodelle beitrugen. Eine potenzielle Ursache dafür stellt vor allem im Bezug auf die hochdynamischen Luftschadstoffe die großräumige Erfassung der Kovariablenwerte auf Landkreisebene dar. Da etwa für die beschriebenen Call-Center-

und Rettungsdienstdaten (vgl. Abschnitt 2.2) genauere geographische Informationen verfügbar sind, brächte eine höhere Auflösung der Umweltdaten auf Gemeindeebene oder gar auf ein äquidistantes Gitter von Werten einen zusätzlichen Informationsgewinn für die Modellierung. Um dies zu realisieren, könnten validierte Luftqualitätsmodelle eingesetzt werden. Ebenso ist eine Verzerrung der Effekte durch die ausschließliche Verwendung der Messwerte an vielbefahrenen Straßen denkbar. Ein weiteres Problem besteht in den hohen Korrelationen innerhalb der Meteorologie- und Luftqualitätsparameter. Hier könnte eventuell eine Verbesserung der Modelle durch gezielte Hinzunahme von Interaktionen erreicht werden. Des Weiteren zeigen die alterspezifischen Zeitfunktionen, die zur Berücksichtigung der Autokorrelation in den Zielvariablenwerten erforderlich sind, zum Teil einen parallelen Verlauf zur zeitlichen Beobachtungskurve der Meteorologie- und Luftqualitätsparameter, was sich eventuell störend auf deren Schätzung auswirken kann.

Eine grundsätzliche Schwierigkeit der populationsbasierten Betrachtung ist, dass die individuelle Exposition von Patienten nur unzureichend erfasst wird. Beispielsweise wird nicht berücksichtigt, ob und wie lange sich die Patienten außer Haus aufhalten oder ob eine erhöhte Schadstoffbelastung am Wohnort oder Arbeitsplatz vorliegt. Ideal wäre die Durchführung einer kontrollierten Studie, für die ausgewählte Personen mit tragbaren Messstationen ausgerüstet werden. Zu einer besseren Erklär- und Prognostizierbarkeit der Asthma- und COPD-Morbidität könnte auch die Erfassung weiterer Kovariablen, wie z. B. des Rauchverhaltens oder des Vorliegens von Allergien, beitragen. Interessant wäre auch eine separate Betrachtung einzelner Subgruppen, gerade im Hinblick auf die unterschiedliche Altersstruktur von Asthma- und COPD-Patienten, sowie eine Verlängerung des Beobachtungszeitraums und/oder eine Erweiterung der Analysen auf zusätzliche Regionen. Die angewendete Methodik ist grundsätzlich auch auf andere Krankheitsfelder übertragbar.

Aus statistischer Sicht besteht in jedem Fall weiterer Entwicklungsbedarf bei der Analyse verzögerter Kovariableneffekte. Insbesondere ist es notwendig, die Anwendbarkeit der bestehenden Methoden durch Integration in gängige Software zu verbessern. Die neu konstruierte PDLF-Methode soll zukünftig anhand von simulierten und weiteren real beobachteten longitudinalen Datensätzen getestet und weiterentwickelt werden. Auch die Umsetzung der Prognoseergebnisse in einen kategorialen Gesundheitsindex bedarf weiterer Untersuchungen. Gerade im Hinblick auf eine Erweiterung der Datenmenge ist es erforderlich, die verwendeten Modellierungstechniken in statistischen Programmpaketen umzusetzen, die besser mit großen Beobachtungszahlen umgehen können.

Literatur

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

S. Almon. The distributed lag between capital appropriations and expenditures. *Econometrica*, 33:178–196, 1965.

R. W. Atkinson, H. R. Anderson, J. Sunyer, J. Ayres, M. Baccini, J. M. Vonk, A. Boumghar, F. Forastiere, B. Forsberg, G. Touloumi, J. Schwartz, und K. Katsouyanni. Acute effects of particulate air pollution on respiratory admissions. *American Journal of Respiratory and Critical Care Medicine*, 164:1860–1866, 2001.

Bayer. Landesamt für Gesundheit und Lebensmittelsicherheit, editor. *Erklärungsmodelle regionaler Gesundheitsunterschiede*. 2007. Wissenschaftliche Bearbeitung: A. Mielck.

Bayer. Landesamt für Statistik und Datenverarbeitung. Datenbank GENESIS. https://www.statistikdaten.bayern.de/genesis.

C. Belitz, A. Brezger, T. Kneib, und S. Lang. *BayesX - Software for Bayesian Inference in Structured Additive Regression Models, Version 2.0.1*, 2009. erhältlich unter: http://www.stat.uni-muenchen.de/~bayesx.

J. Besag, J. York, und A. Mollié. Bayesian image restoration with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics*, 43:1–59, 1991.

K. P. Burnham und D. R. Anderson. *Model selection and inference: a practical information-theoretic approach*. Springer Verlag, 1998.

A. C. Cameron und P. K. Trivedi. *Regression Analysis of Count Data*. Cambridge University Press, New York, 1998.

J.-M. Chiou und H.-G. Müller. Quasi-likelihood regression with unknown link and variance functions. *Journal of the American Statistical Association*, 93:1376–1387, 1998.

W. S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74:829–836, 1979.

V. Connolly, N. Unwin, P. Sherriff, R. Bilous, und W. Kelly. Diabetes prevalence and socioeconomic status: a population based study showing increased prevalence of type 2 diabetes mellitus in deprived areas. *Journal of Epidemiology and Community Health*, 54(3):173–177, 2000.

G. D'Amato, G. Liccardi, M. D'Amato, und M. Cazzola. Outdoor air pollution, climatic changes and allergic bronchial asthma. *European Respiratory Monograph*, 21:30–51, 2002.

N. R. Draper und H. Smith. *Applied Regression Analysis*. Wiley-Interscience, 1998.

D. B. Dunson, N. Pillai, und J.-H. Park. Bayesian density regression. *Journal of the Royal Statistical Society, Series B*, 69:163–183, 2007.

L. E. O. Echavarría. *Semiparametric Bayesian Count Data Models*. PhD thesis, Ludwig-Maximilians-Universität München, 2004.

P. H. C. Eilers und B. D. Marx. Flexible smoothing with B-splines and penalties. *Statitical Science*, 11(2):89–121, 1996.

European Centre for Medium Range Weather Forecast. http://www.ecmwf.int.

European Environmental Agency. http://www.eea.europa.eu/data-and-maps/data/airbase-the-european-air-quality-database-2.

T. Exner. Longitudinale Modellierung pneumologischer Erkrankungen in Abhängigkeit von meteorologischen Einflussfaktoren (Pilotregion Bayern). Master's thesis, Ludwig-Maximilians-Universität München, 2009.

L. Fahrmeir und C. Heumann. Skript zur Vorlesung Testen und Schätzen I im WiSe 2009/10. erhältlich unter http://www.statistik.lmu.de/~semwiso/schaetzentesten1-ws0910/skript/ST1-ws0910-skript.pdf, 2009.

L. Fahrmeir und T. Kneib. *Bayesian Smoothing and Regression for Longitudinal, Spatial and Event History Data*. Oxford University Press, 2010. to appear.

L. Fahrmeir und G. Tutz. *Multivariate statistical modelling based on generalized linear models*. Springer Verlag, 2001.

L. Fahrmeir, R. Künstler, I. Pigeot, und G. Tutz. *Der Weg zur Datenanalyse, 5. Auflage*. Springer Verlag, 2004.

GENESIS. GENeric European Sustainable Information Space for Environment. http://www.genesis-fp7.eu/.

S. K. Ghosh, P. Mukhopadhyay, und J.-C. Lu. Bayesian analysis of zero-inflated regression models. *Journal of Statistical Planning and Inference*, 136(4):1360–1375, 2006.

J. E. Griffin und P. J. Brown. Alternative prior distributions for variable selection with very many more variables than observations. Technical report, University of Warwick, Department of Statistics, 2005.

G. Grimmett und D. Stirzaker. *Probability and Random Processes, Third Edition*. Oxford University Press, 2001.

U. Halekoh, S. Højsgaard, und J. Yan. The R Package geepack for Generalized Estimating Equations. *Journal of Statistical Software*, 15:1040–1044, 2006.

E. S. M. Harré, P. D. Price, R. B. Ayrey, L. J. Toop, I. R. Martin, und G. I. Town. Respiratory effects of air pollution in chronic obstructive pulmonary disease: a three month prospective study. *Thorax*, 52:1040–1044, 1997.

L. Held. *Methoden der statistischen Inferenz: Likelihood und Bayes*. Spektrum Verlag, 2008.

G. Hoek, B. Brunekreef, S. Goldbohm, P. Fischer, und P. A. van den Brandt. Association between mortality and indicators of traffic-related air pollution in the netherlands: a cohort study. *Lancet*, 360:1203–1209, 2002.

A. E. Hoerl und R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 42(1):80–86, 2000.

A. Ishigami, S. Hajat, R. S. Kovats, L. Bisanti, M. Rognoni, A. Russo, und A. Paldy. An ecological time-series study of heat-related mortality in three european cities. *Environmental Health*, 7(5), 2008.

P. A. Kassomenos, A. Gryparis, und K. Katsouyanni. On the association between daily mortality and air mass types in athens, greece during winter and summer. *International Journal of Biometeorology*, 51:315–322, 2007.

R. Koenker. *Quantile regression*. Cambridge University Press, 2005.

L. M. Koyck. *Distributed Lags and Investment Analysis*. North-Holland, Amsterdam, 1954.

K. Y. Liang und S. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

H. Lingner, K. Schultz, und F.-W. Schwartz. *Volkskrankheit Asthma/COPD*. Springer Medizin Verlag Heidelberg, 2007.

S. Macintyre, A. Ellaway, und S. Cummins. Place effects on health: how can we conceptualise, operationalise and measure them? *Social Science & Medicine*, 55 (1):125–139, 2002.

W. Maier, J. Fairburn, und A. Mielck. Deprivation und Mortalität in Bayern: Entwicklung eines ‚Index Multipler Deprivation' auf Gemeindebasis. In Bearbeitung, 2010.

C. L. Mallows. Some comments on C_p. *Technometrics*, 15:661–675, 1973.

P. McCullagh und J. A. Nelder. *Generalized Linear Models*. Chapman & Hall / CRC, 1989.

P. Michelozzi, G. Accetta, M. D. Sario, D. D'Ippoliti, C. Marino, M. Baccini, H.R. Biggeri, A. amd Anderson, K. Katsouyanni, F. Ballester, L. Bisanti, E. Cadum, B. Forsberg, F. Forastiere, P. G. Goodman, A. Hojs, U. Kirchmayer, S. Medina, A. Paldy, C. Schindler, J. Sunyer, und C. A. Perucci. High temperature and hospitalizations for cardiovascular and respiratory causes in 12 european cities. *American Journal of Respiratory and Critical Care Medicine*, 179:383–389, 2009.

V. M. R. Muggeo. Modeling temperature effects on mortality: Multiple segmented relationships with common breakpoints. *Biostatistics*, 9:613–620, 2008a.

V. M. R. Muggeo. segmented: An R package to fit Regression Models with Broken-Line Relationships. *R News*, 8(1):20–25, 2008b.

V. M. R. Muggeo. Analyzing temperature effects on mortality within the R environment: The constrained segmented distributed lag parametrization. *Biostatistics*, 9:613–620, 2008c.

J. A. Nelder und R. W. M. Wedderburn. Generalized Linear Models. *Journal of the Royal Statistical Society, Series A*, 135:370–384, 1972.

M. Noble, G. Wright, G. Smith, und C. Dibben. Measuring multiple deprivation at the small-area level. *Environment and Planning A*, 38:169–185, 2006.

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

R. R. Ravines, A. M. Schmidt, und H. S. Migon. Revisiting distributed lag models through a bayesian perspective. *Applied Stochastic Models in Business and Industry*, 22:193–210, 2006.

C. Ren und S. Tong. Temperature modifies the health effect of particulate matter in brisbane, australia. *International Journal of Biometeorology*, 51:87–96, 2006.

L. G. Shack, B. Rachet, D. H. Brewster, und M. P. Coleman. Socioeconomic inequalities in cancer survival in scotland 1986–2000. *British Journal of Cancer*, 97 (7):999–1004, 2007.

J. S. Simonoff. *Analyzing Categorical Data*. Springer Verlag, 2003.

R. Solow. On a family of lag distributions. *Econometrica*, 28(2):393–406, 1960.

R. Tibshirani. Regression Shrinkage and Selection via the LASSO. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

P. Townsend. *Poverty in the United Kingdom: a survey of household resources and standards of living*. University of California Press Berkeley and Los Angeles, 1979.

P. Townsend, P. Phillimore, und A. Beattie. *Health and deprivation: Inequality and the North*. Routledge, London and New York, 1988.

G. Tutz. *Analyse kategorialer Daten*. Oldenbourg Verlag, 2000.

W. N. Venables und B. D. Ripley. *Modern Applied Statistics with S, Fourth Edition*. Springer Verlag, 2002.

J. M. Ver Hoef und P. L. Boveng. Quasi-Poisson vs. Negative Binomial Regression: How should we model overdispersed count data? *Ecology*, 66(11):2766–2772, 2007.

E. Wanka. Weather and air pollutants have an impact on patients with respiratory diseases and breathing difficulties in Munich, Germany. Master's thesis, Ludwig-Maximilians-Universität München, 2010.

R. W. M. Wedderburn. Quasi-Likelihood Functions, Generalized Linear Models and the Gauss-Newton method. *Environmental Research*, 104:402–409, 1974.

L. J. Welty, R. D. Peng, S. L. Zeger, und F. Dominici. Bayesian distributed lag models: Estimating effects of particulate matter air pollution on daily mortality. *Biometrics*, 65:282–291, 2009.

L. J. Wen-Chao, R. D. Peng, S. L. Zeger, und F. Dominici. Air pollution, weather and associated risk factors related to asthma prevalence and attack rate. *Environmental Research*, 104:402–409, 2007.

S. Wild, F. Macleod, J. McKnight, G. Watt, C. Mackenzie, I. Ford, A. McConnachie, und R. S. Lindsay. Impact of deprivation on cardiovascular risk factors in people with diabetes: an observational study. *Diabetic Medicine*, 25(2):194–199, 2008.

M. Wildner, H. Zöllner, W. H. Caselmann, und G. Kerscher. Strategy for a public health initiative at regional level – the example of bavaria. *Journal of Public Health*, 13(6):318–324, 2005.

R. Winkelmann. *Econometric Analysis of Count Data, Fifth Edition*. Springer Verlag, 2008.

S. N. Wood. *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC, Boca Raton, Florida, 2006.

A. Zanobetti, M. P. Wand, J. Schwartz, und L. M. Ryan. Generalized additive distributed lag models: quantifying mortality displacement. *Biostatistics*, 1(3): 279–292, 2000.

G Tabellen und Grafiken

Zu Abschnitt 2.3:

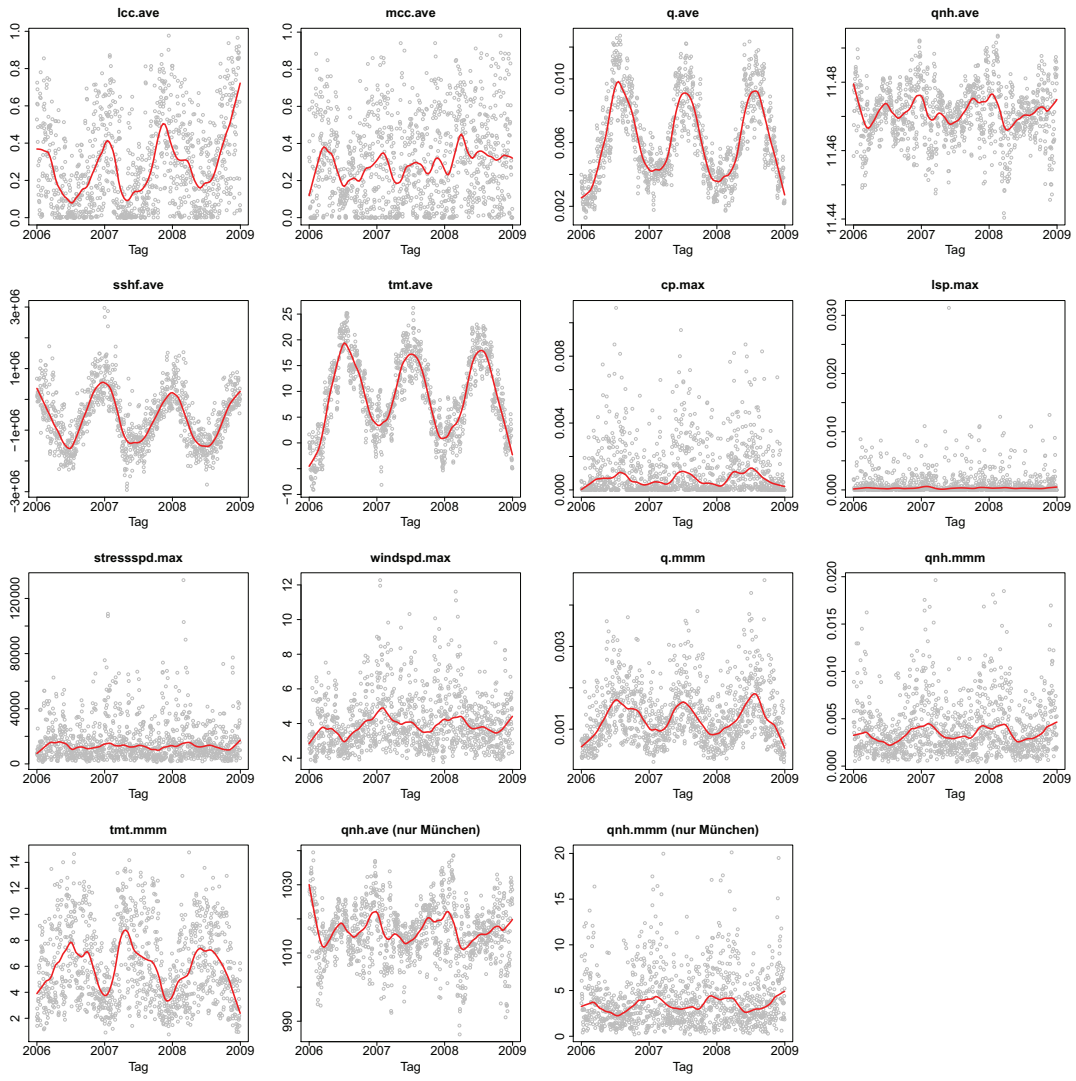


Abbildung G.1: Zeitlicher Verlauf der Meteorologie-Parameter in Bayern im Beobachtungszeitraum 2006 bis 2008: Mittelwerte über alle landkreisspezifischen Tageswerte (grau) mit lowess-Trend (rot) (vgl. Tabelle 2.1 für die zugehörigen Einheiten)

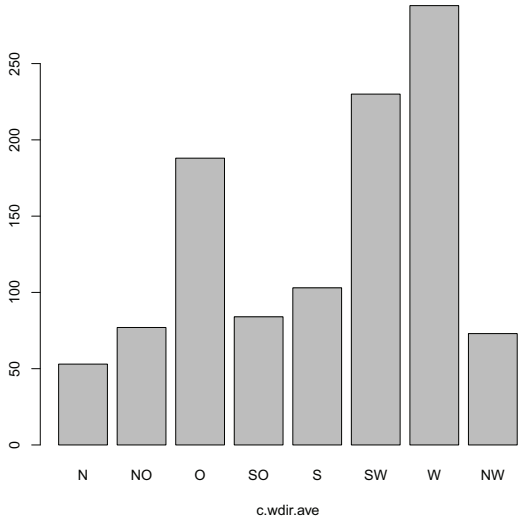


Abbildung G.2: Häufigkeitsverteilung der täglich vorherrschenden Windrichtung in Bayern im Beobachtungszeitraum 2006 bis 2008

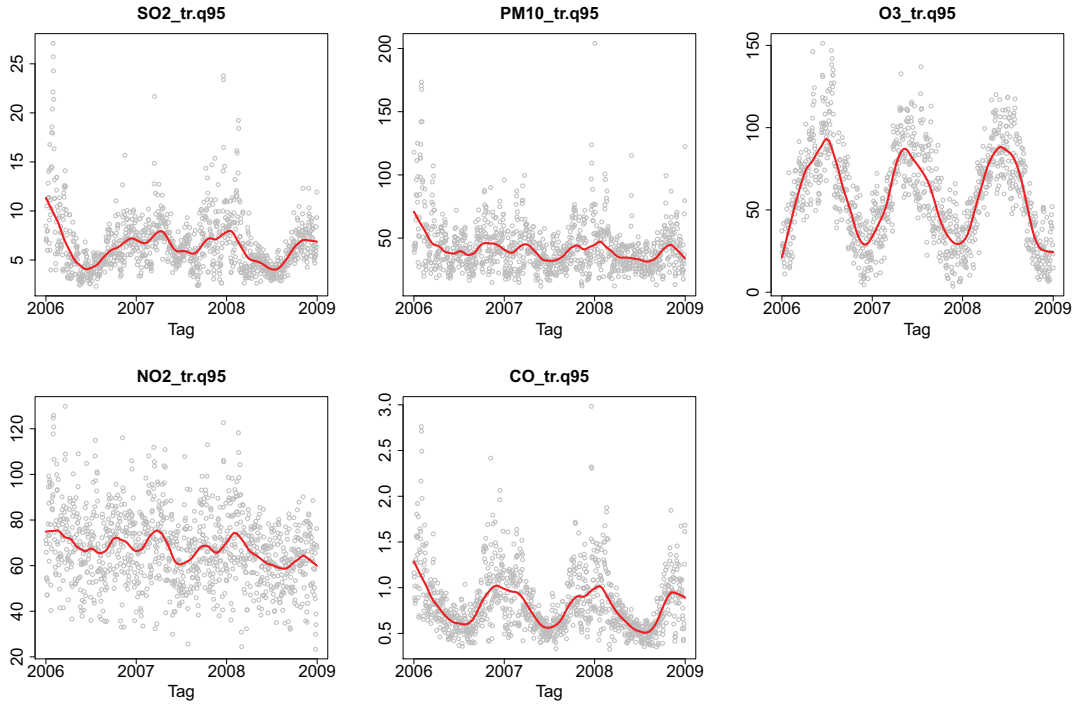


Abbildung G.3: Zeitlicher Verlauf der Luftqualitätsparameter in Bayern im Beobachtungszeitraum 2006 bis 2008: Mittelwerte über alle landkreisspezifischen Tageswerte (grau) mit lowess-Trend (rot) (vgl. Tabelle 2.2 für die zugehörigen Einheiten)

Zu Abschnitt 2.4:

Alter	Min.	1st Qu.	Median	3rd Qu.	Max.
<= 20 J.	1	3	5	8	59
21 - 40 J.	1	5	8	15	109
41 - 60 J.	1	7	11	20	65
61 - 80 J.	4	16	23	52	148
>= 81 J.	4	14	21	49	105

Tabelle G.1: 5-Punkte-Zusammenfassung der pro Tag in Bayern erfassten Anrufe bei den KVB-Call-Centern getrennt nach Altersgruppen

Zielgröße	Geschlecht	Alter	Min.	1st Qu.	Median	3rd Qu.	Max.
gesamt	männlich	<= 20 J.	40	204	2234	3358	5232
	männlich	21 - 40 J.	40	130	1520	2067	4542
	männlich	41 - 60 J.	78	256	3580	4764	11460
	männlich	>= 61 J.	251	692	10500	13370	30210
	weiblich	<= 20 J.	29	131	1460	2190	3451
	weiblich	21 - 40 J.	54	161	2241	3056	6819
	weiblich	41 - 60 J.	90	296	4988	6672	16210
	weiblich	>= 61 J.	248	711	11890	15410	34640
COPD	männlich	<= 20 J.	3	41	410	637	1185
	männlich	21 - 40 J.	8	30	268	368	884
	männlich	41 - 60 J.	38	135	1815	2361	5848
	männlich	>= 61 J.	190	550	8064	10190	23180
	weiblich	<= 20 J.	20	30	277	440	8640
	weiblich	21 - 40 J.	11	34	349	473	1158
	weiblich	41 - 60 J.	24	114	1806	2384	5971
	weiblich	>= 61 J.	146	473	7554	9764	22030
Asthma	männlich	<= 20 J.	37	170	1909	2862	4532
	männlich	21 - 40 J.	31	107	1325	1788	4070
	männlich	41 - 60 J.	44	144	2044	2794	6687
	männlich	>= 61 J.	68	210	3465	4396	10160
	weiblich	<= 20 J.	23	109	1226	1843	2834
	weiblich	21 - 40 J.	47	137	1977	2703	5948
	weiblich	41 - 60 J.	60	201	3564	4782	11620
	weiblich	>= 61 J.	104	297	5416	6998	16110

Tabelle G.2: 5-Punkte-Zusammenfassung der pro Tag in Bayern erfassten Arztbesuche getrennt nach Alters- und Geschlechtsgruppen sowie nach Diagnose

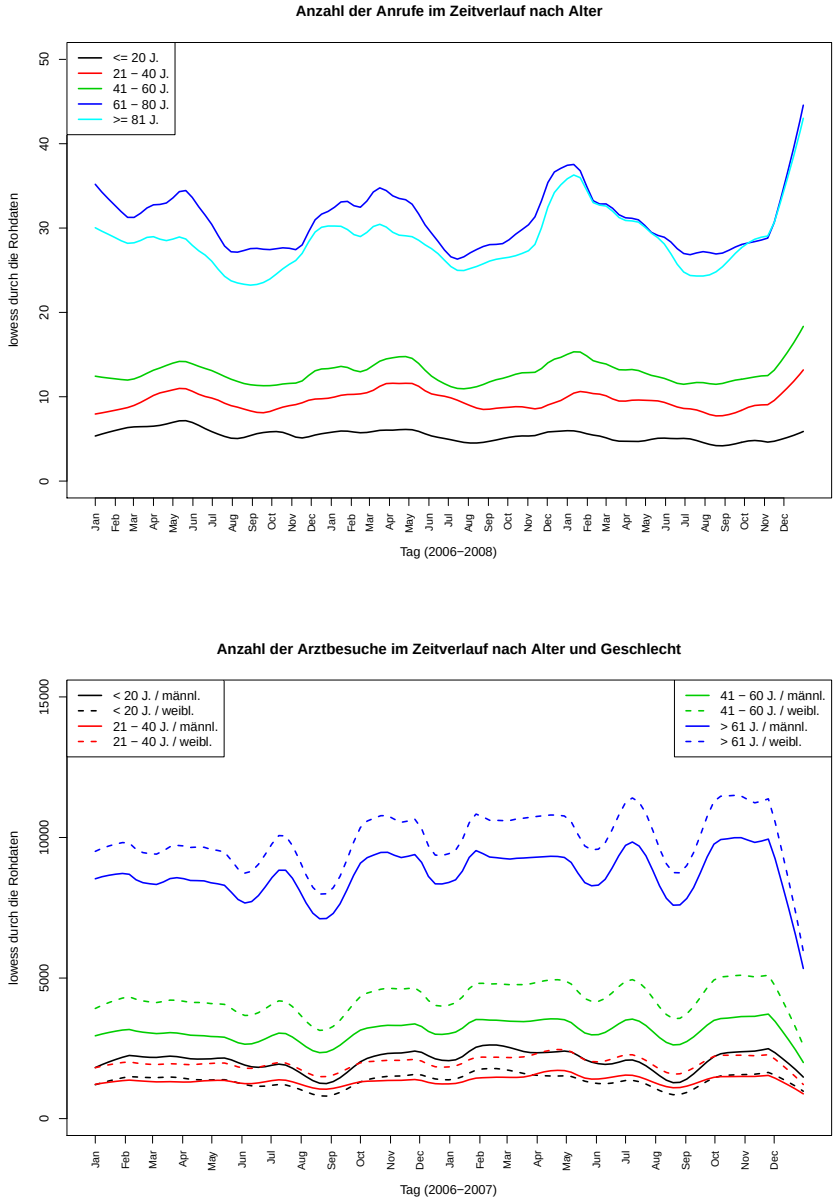


Abbildung G.4: Zeitlicher Verlauf der Anruferzahlen von 2006 bis 2008 getrennt nach Alter (oben) und der Arztbesuche gesamt getrennt nach Geschlecht und Alter von 2006 bis 2007 (unten)

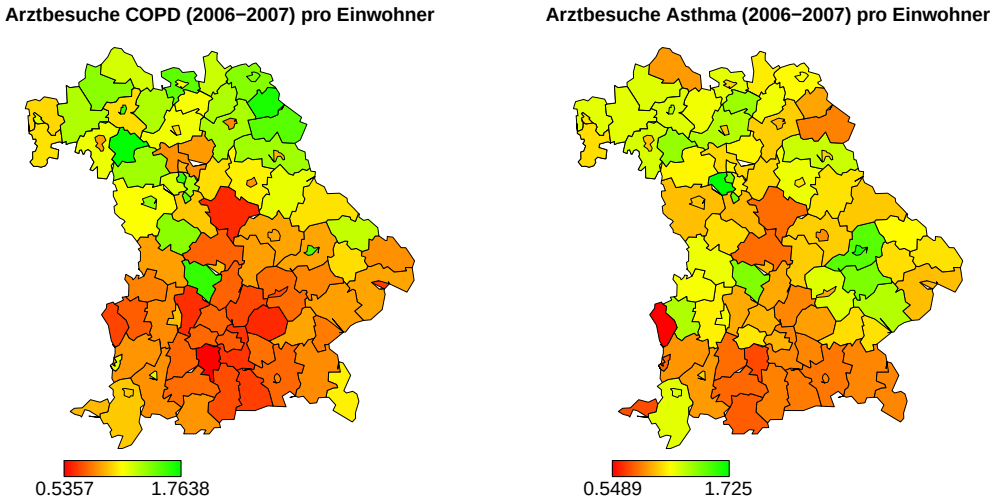


Abbildung G.5: Räumliche Verteilung der Arztbesuche aufgrund von COPD (links) und Asthma (rechts) von 2006 bis 2007 jeweils standardisiert durch die Einwohnerzahl

	lcc	mcc	q	qnh	sshf	tmt	cp	lsp	stressspd	windspd
lcc	1.00	0.48	-0.34	0.00	0.39	-0.46	0.08	0.37	0.24	0.24
mcc	0.48	1.00	-0.09	-0.48	0.28	-0.15	0.40	0.58	0.42	0.45
q	-0.34	-0.09	1.00	-0.10	-0.42	0.94	0.37	-0.01	-0.14	-0.14
qnh	0.00	-0.48	-0.10	1.00	-0.07	-0.08	-0.33	-0.32	-0.25	-0.34
sshf	0.39	0.28	-0.42	-0.07	1.00	-0.53	-0.01	0.33	0.38	0.47
tmt	-0.46	-0.15	0.94	-0.08	-0.53	1.00	0.25	-0.09	-0.10	-0.12
cp	0.08	0.40	0.37	-0.33	-0.01	0.25	1.00	0.41	0.30	0.30
lsp	0.37	0.58	-0.01	-0.32	0.33	-0.09	0.41	1.00	0.44	0.44
stressspd	0.24	0.42	-0.14	-0.25	0.38	-0.10	0.30	0.44	1.00	0.91
windspd	0.24	0.45	-0.14	-0.34	0.47	-0.12	0.30	0.44	0.91	1.00

Tabelle G.3: Korrelationsmatrix der meteorologischen Parameter

	SO2	PM10	O3	NO2	CO
SO2	1.00	0.73	-0.37	0.58	0.78
PM10	0.73	1.00	-0.24	0.62	0.74
O3	-0.37	-0.24	1.00	-0.01	-0.55
NO2	0.58	0.62	-0.01	1.00	0.72
CO	0.78	0.74	-0.55	0.72	1.00

Tabelle G.4: Korrelationsmatrix der Luftqualitätsparameter

	lcc	mcc	q	qnh	sshf	tmt	cp	lsp	stressspd	windspd
SO2	-0.09	-0.33	-0.47	0.25	0.20	-0.50	-0.40	-0.22	-0.29	-0.28
PM10	-0.15	-0.40	-0.22	0.27	0.11	-0.29	-0.32	-0.28	-0.41	-0.42
O3	-0.54	-0.18	0.53	-0.12	-0.65	0.68	0.16	-0.11	0.01	-0.06
NO2	-0.33	-0.36	-0.14	0.14	0.02	-0.10	-0.28	-0.23	-0.44	-0.39
CO	-0.02	-0.30	-0.47	0.24	0.38	-0.54	-0.38	-0.18	-0.36	-0.31

Tabelle G.5: Korrelationen zwischen Meteorologie- und Luftqualitätsparametern

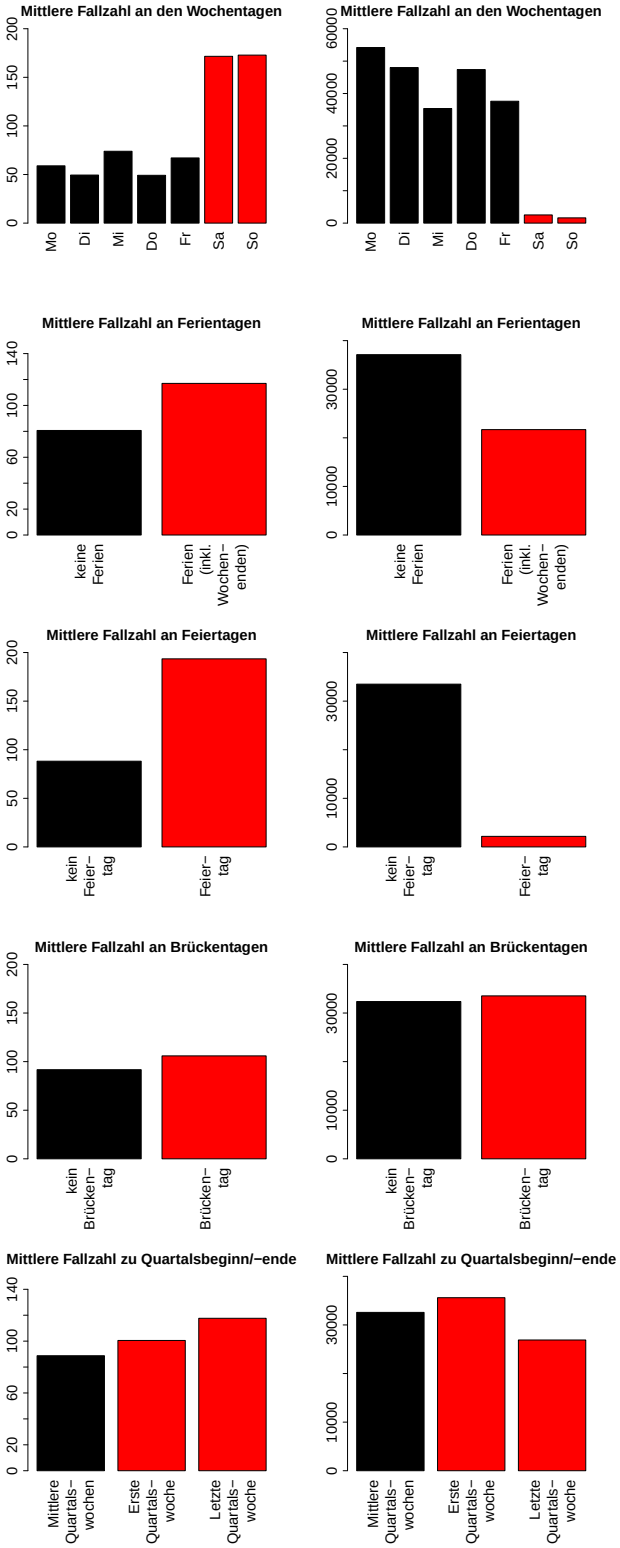


Abbildung G.6: Mittelwerte der Anruferzahlen (2006 bis 2008, links) bzw. der Anzahl der Arztbesuche gesamt (2006 bis 2007, rechts) auf den Faktorstufen der jeweiligen administrativen Kovariable

Zu Abschnitt 3.2:

$$
\begin{aligned}
\log(\mu_{it}) \;=\; & \text{offset}(\log(x_{\text{inhabitants},i})) + \beta_{\text{Intercept}} + x_{\text{dow=Di},t}\ \beta_{\text{dow=Di}} + \ldots + x_{\text{dow=So},t}\ \beta_{\text{dow=So}} \\
& + x_{\text{school=ja},t}\ \beta_{\text{school=ja}} + x_{\text{holiday=ja},t}\ \beta_{\text{holiday=ja}} + x_{\text{bridge=ja},t}\ \beta_{\text{bridge=ja}} \\
& + x_{\text{quartal=Anfang},t}\ \beta_{\text{quartal=Anfang}} + x_{\text{quartal=Ende},t}\ \beta_{\text{quartal=Ende}} \\
& + x_{\text{age=2},i}\ \beta_{\text{age=2}} + \ldots + x_{\text{age=5},i}\ \beta_{\text{age=5}} + x_{\text{cp.max},t}\ \beta_{\text{cp.max}} + x_{\text{lsp.max},t}\ \beta_{\text{lsp.max}} \\
& + x_{\text{lcc.ave},t}\ \beta_{\text{lcc.ave}} + x_{\text{mcc.ave},t}\ \beta_{\text{mcc.ave}} + x_{\text{q.ave},t}\ \beta_{\text{q.ave}} \\
& + x_{\text{q.mmm},t}\ \beta_{\text{q.mmm}} + x_{\text{q.ave\_lcp},t}\ \beta_{\text{q.ave\_lcp}} + x_{\text{q.ave\_ucp},t}\ \beta_{\text{q.ave\_ucp}} \\
& + x_{\text{qnh.ave},t}\ \beta_{\text{qnh.ave}} + x_{\text{qnh.mmm},t}\ \beta_{\text{qnh.mmm}} + x_{\text{qnh.ave\_lcp},t}\ \beta_{\text{qnh.ave\_lcp}} \\
& + x_{\text{qnh.ave\_lcp},t}\ \beta_{\text{qnh.ave\_lcp}} + x_{\text{qnh.ave\_ucp},t}\ \beta_{\text{qnh.ave\_ucp}} + x_{\text{sshf.ave},t}\ \beta_{\text{sshf.ave}} \\
& + x_{\text{stressspd.max},t}\ \beta_{\text{stressspd.max}} + x_{\text{tmt.ave},t}\ \beta_{\text{tmt.ave}} + x_{\text{tmt.mmm},t}\ \beta_{\text{tmt.mmm}} \\
& + x_{\text{tmt.ave\_lcp},t}\ \beta_{\text{tmt.ave\_lcp}} + x_{\text{tmt.ave\_ucp},t}\ \beta_{\text{tmt.ave\_ucp}} + x_{\text{windspd.max},t}\ \beta_{\text{windspd.max}} \\
& + x_{\text{c.wdir=NO},t}\ \beta_{\text{c.wdir=NO}} + \ldots + x_{\text{c.wdir=W},t}\ \beta_{\text{c.wdir=W}} \\
& + x_{\text{SO2.q95},t}\ \beta_{\text{SO2.q95}} + x_{\text{PM10.q95},t}\ \beta_{\text{PM10.q95}} + x_{\text{O3.q95},t}\ \beta_{\text{O3.q95}} \\
& + x_{\text{NO2.q95},t}\ \beta_{\text{NO2.q95}} + x_{\text{CO.q95},t}\ \beta_{\text{CO.q95}} + f(t) + f_2(t) + \ldots + f_5(t)
\end{aligned}
$$

Abbildung G.7: Modellgleichung für die KVB-Call-Center-Modelle basierend auf Reduktionsverfahren a) und b)

$$
\begin{aligned}
\log(\mu_{ijt}) \;=\; & \text{offset}(\log(x_{\text{inhabitants},i})) + \beta_{\text{Intercept}} + x_{\text{dow=Di},t}\ \beta_{\text{dow=Di}} + \ldots + x_{\text{dow=So},t}\ \beta_{\text{dow=So}} \\
& + x_{\text{school=ja},t}\ \beta_{\text{school=ja}} + x_{\text{holiday=ja},t}\ \beta_{\text{holiday=ja}} + x_{\text{bridge=ja},t}\ \beta_{\text{bridge=ja}} \\
& + x_{\text{quartal=Anfang},t}\ \beta_{\text{quartal=Anfang}} + x_{\text{quartal=Ende},t}\ \beta_{\text{quartal=Ende}} \\
& + x_{\text{sex=2},j}\ \beta_{\text{sex=2}} + \ x_{\text{age=2},i}\ \beta_{\text{age=2}} + \ldots + x_{\text{age=4},i}\ \beta_{\text{age=4}} \\
& + x_{\text{sex=2},j}\ x_{\text{age=2},i}\ \beta_{\text{sex=2,age=2}} + \ldots + x_{\text{sex=2},j}\ x_{\text{age=4},i}\ \beta_{\text{sex=2,age=4}} \\
& + x_{\text{cp.max},t}\ \beta_{\text{cp.max}} + x_{\text{lsp.max},t}\ \beta_{\text{lsp.max}} + x_{\text{lcc.ave},t}\ \beta_{\text{lcc.ave}} \\
& + x_{\text{mcc.ave},t}\ \beta_{\text{mcc.ave}} + x_{\text{q.ave},t}\ \beta_{\text{q.ave}} + x_{\text{q.mmm},t}\ \beta_{\text{q.mmm}} \\
& + x_{\text{q.ave\_lcp},t}\ \beta_{\text{q.ave\_lcp}} + x_{\text{q.ave\_ucp},t}\ \beta_{\text{q.ave\_ucp}} + x_{\text{qnh.ave},t}\ \beta_{\text{qnh.ave}} \\
& + x_{\text{qnh.mmm},t}\ \beta_{\text{qnh.mmm}} + x_{\text{qnh.ave\_lcp},t}\ \beta_{\text{qnh.ave\_lcp}} + x_{\text{qnh.ave\_lcp},t}\ \beta_{\text{qnh.ave\_lcp}} \\
& + x_{\text{qnh.ave\_ucp},t}\ \beta_{\text{qnh.ave\_ucp}} + x_{\text{sshf.ave},t}\ \beta_{\text{sshf.ave}} + x_{\text{stressspd.max},t}\ \beta_{\text{stressspd.max}} \\
& + x_{\text{tmt.ave},t}\ \beta_{\text{tmt.ave}} + x_{\text{tmt.mmm},t}\ \beta_{\text{tmt.mmm}} + x_{\text{tmt.ave\_lcp},t}\ \beta_{\text{tmt.ave\_lcp}} \\
& + x_{\text{tmt.ave\_ucp},t}\ \beta_{\text{tmt.ave\_ucp}} + x_{\text{windspd.max},t}\ \beta_{\text{windspd.max}} \\
& + x_{\text{c.wdir=NO},t}\ \beta_{\text{c.wdir=NO}} + \ldots + x_{\text{c.wdir=W},t}\ \beta_{\text{c.wdir=W}} \\
& + x_{\text{SO2.q95},t}\ \beta_{\text{SO2.q95}} + x_{\text{PM10.q95},t}\ \beta_{\text{PM10.q95}} + x_{\text{O3.q95},t}\ \beta_{\text{O3.q95}} \\
& + x_{\text{NO2.q95},t}\ \beta_{\text{NO2.q95}} + x_{\text{CO.q95},t}\ \beta_{\text{CO.q95}} + f(t) + f_2(t) + \ldots + f_4(t)
\end{aligned}
$$

Abbildung G.8: Modellgleichung für die KVB-Abrechnungsmodelle basierend auf Reduktionsverfahren a) und b)

Kovariable	$\hat{\beta}$	exp($\hat{\beta}$)	se($\hat{\beta}$)	T	p-Wert	$\hat{l}_\beta$	$\hat{u}_\beta$
Intercept	-36.0014	0.0000	15.1512	-2.3761	0.0175	-65.6973	-6.3056
dow=Mo	(-0.3906)	0.6767	(0.0140)	-27.9125	0.0000	-0.4180	-0.3632
dow=Di	-0.4688	0.6257	0.0142	-33.0846	0.0000	-0.4966	-0.4411
dow=Mi	-0.0250	0.9753	0.0143	-1.7432	0.0813	-0.0531	0.0031
dow=Do	-0.4574	0.6329	0.0146	-31.4300	0.0000	-0.4859	-0.4289
dow=Fr	-0.1503	0.8605	0.0142	-10.6202	0.0000	-0.1780	-0.1226
dow=Sa	0.7544	2.1264	0.0153	49.1688	0.0000	0.7243	0.7845
dow=So	0.7377	2.0911	0.0192	38.3639	0.0000	0.7000	0.7754
school=ja	0.0999	1.1051	0.0133	7.5063	0.0000	0.0738	0.1260
holiday=ja	0.9610	2.6144	0.0367	26.2185	0.0000	0.8892	1.0329
bridge=ja	0.4139	1.5128	0.0543	7.6200	0.0000	0.3075	0.5204
quartal=Anfang	0.0135	0.0220	1.0136	0.6128	0.5400	-0.0296	0.0566
quartal=Ende	0.1308	1.1398	0.0212	6.1690	0.0000	0.0893	0.1724
age=1	(-1.0008)	0.3676	(0.0756)	-13.2331	0.0000	-1.1490	-0.8526
age=2	-1.0086	0.3647	0.0756	-13.3362	0.0000	-1.1568	-0.8604
age=3	-0.7630	0.4663	0.0756	-10.0886	0.0000	-0.9112	-0.6148
age=4	0.6431	1.9024	0.0756	8.5035	0.0000	0.4949	0.7914
age=5	2.1293	8.4090	0.0756	28.1543	0.0000	1.9811	2.2775
cp.max (0.001)	0.0019	1.0019	0.0053	0.3698	0.7115	-0.0084	0.0122
lsp.max (0.001)	0.0041	1.0041	0.0040	1.0148	0.3102	-0.0038	0.0119
lcc.ave (0.1)	-0.0097	0.9904	0.0045	-2.1466	0.0318	-0.0185	-0.0008
mcc.ave (0.1)	0.0018	1.0018	0.0043	0.4122	0.6802	-0.0066	0.0101
q.ave (0.001)	0.0016	1.0016	0.0149	0.1094	0.9129	-0.0276	0.0309
q.mmm (0.001)	0.0038	1.0038	0.0114	0.3345	0.7380	-0.0185	0.0261
qnh.ave (0.001)	0.0021	1.0021	0.0013	1.5855	0.1129	-0.0005	0.0047
qnh.mmm (0.001)	-0.0057	0.9943	0.0023	-2.4821	0.0131	-0.0102	-0.0012
sshf.ave (100000)	0.0015	1.0015	0.0018	0.8175	0.4136	-0.0021	0.0051
stressspd.max (10000)	-0.0147	0.9854	0.0115	-1.2789	0.2009	-0.0372	0.0078
tmt.ave	0.0004	1.0004	0.0061	0.0706	0.9437	-0.0115	0.0124
tmt.mmm	0.0076	1.0076	0.0041	1.8502	0.0643	-0.0005	0.0157
windspd.max	-0.0071	0.9930	0.0118	-0.6003	0.5483	-0.0302	0.0160
c.wdir=N	(0.0064)	1.0064	(0.0244)	0.2618	0.7935	-0.0414	0.0542
c.wdir=NO	-0.0523	0.9490	0.0202	-2.5845	0.0098	-0.0920	-0.0126
c.wdir=NW	0.0238	1.0241	0.0212	1.1235	0.2612	-0.0178	0.0654
c.wdir=O	-0.0512	0.9501	0.0154	-3.3344	0.0009	-0.0813	-0.0211
c.wdir=S	-0.0009	0.9991	0.0195	-0.0472	0.9624	-0.0392	0.0374
c.wdir=SO	0.0600	1.0618	0.0211	2.8410	0.0045	0.0186	0.1014
c.wdir=SW	0.0053	1.0053	0.0141	0.3730	0.7091	-0.0224	0.0330
c.wdir=W	0.0090	1.0090	0.0142	0.6330	0.5267	-0.0188	0.0368
SO2.q95	-0.0049	0.9951	0.0040	-1.2260	0.2202	-0.0128	0.0029
PM10.q95	0.0022	1.0022	0.0005	4.0267	0.0001	0.0011	0.0033
O3.q95	0.0010	1.0010	0.0007	1.5728	0.1158	-0.0003	0.0023
NO2.q95	-0.0031	0.9969	0.0011	-2.8539	0.0043	-0.0052	-0.0010
CO.q95	-0.0442	0.9568	0.0583	-0.7589	0.4479	-0.1584	0.0700

Tabelle G.6: Geschätzte Regressionskoeffizienten des loglinearen Modells (roh und exponentiell transformiert) für die Call-Center-Anrufe aggregiert über alle bayerischen Landkreise (Reduktionsverfahren b)) zusammen mit Standardfehlern, p-Werten und Konfidenzintervallen (gelb: signifikante Ergebnisse)

Kovariable	Poisson-Modell	Quasi-Poisson-Modell	Negativ-Binomial-Modell
Intercept	-12.7155/1.6573/0.0000	-12.7155/1.8913/0.0000	-12.6966/1.9052/0.0000
dow=Mo	(0.0015)/(0.0208)/0.9415	(0.0015)/(0.0237)/0.9488	(-0.0002)/(0.0238)/0.9943
dow=Di	-0.1312/0.0224/0.0000	-0.1312/0.0255/0.0000	-0.1299/0.0252/0.0000
dow=Mi	-0.0825/0.0219/0.0002	-0.0825/0.0250/0.0010	-0.0832/0.0249/0.0008
dow=Do	-0.2047/0.0233/0.0000	-0.2047/0.0265/0.0000	-0.2069/0.0261/0.0000
dow=Fr	-0.0738/0.0218/0.0007	-0.0738/0.0248/0.0029	-0.0704/0.0247/0.0043
dow=Sa	0.2166/0.0203/0.0000	0.2166/0.0232/0.0000	0.2143/0.0237/0.0000
dow=So	0.2741/0.0235/0.0000	0.2741/0.0269/0.0000	0.2762/0.0275/0.0000
school=ja	0.1668/0.0190/0.0000	0.1668/0.0216/0.0000	0.1731/0.0219/0.0000
holiday=ja	0.2700/0.0445/0.0000	0.2700/0.0507/0.0000	0.2647/0.0529/0.0000
bridge=ja	0.0802/0.0793/0.3124	0.0802/0.0906/0.3760	0.0724/0.0916/0.4292
quartal=Anfang	0.0138/0.0319/0.6665	0.0138/0.0364/0.7057	0.0154/0.0367/0.6754
quartal=Ende	0.0976/0.0306/0.0014	0.0976/0.0349/0.0052	0.0970/0.0354/0.0062
age=1	(-2.5647)/(0.3845)/0.0000	(-2.5647)/(0.4389)/0.0000	(-2.5622)/(0.3878)/0.0000
age=2	-0.9174/0.1719/0.0000	-0.9174/0.1961/0.0000	-0.9202/0.1800/0.0000
age=3	-0.4158/0.1572/0.0082	-0.4158/0.1794/0.0204	-0.4189/0.1660/0.0116
age=4	1.2521/0.1250/0.0000	1.2521/0.1427/0.0000	1.2438/0.1359/0.0000
age=5	2.6458/0.1280/0.0000	2.6458/0.1461/0.0000	2.6576/0.1385/0.0000
cp.max (0.001)	0.0027/0.0048/0.5745	0.0027/0.0054/0.6227	0.0040/0.0054/0.4579
lsp.max (0.001)	-0.0005/0.0029/0.8637	-0.0005/0.0034/0.8804	-0.0004/0.0033/0.9110
lcc.ave (0.1)	0.0049/0.0046/0.2891	0.0049/0.0053/0.3529	0.0050/0.0053/0.3439
mcc.ave (0.1)	-0.0044/0.0047/0.3489	-0.0044/0.0053/0.4117	-0.0055/0.0054/0.3036
q.ave (0.001)	-0.0022/0.0183/0.9039	-0.0022/0.0208/0.9158	-0.0072/0.0209/0.7293
q.mmm (0.001)	0.0164/0.0125/0.1900	0.0164/0.0142/0.2508	0.0193/0.0142/0.1741
qnh.ave (München)	0.0008/0.0016/0.5994	0.0008/0.0018/0.6453	0.0008/0.0018/0.6556
qnh.mmm (München)	-0.0027/0.0032/0.3938	-0.0027/0.0036/0.4550	-0.0020/0.0036/0.5818
sshf.ave (100000)	0.0002/0.0016/0.9176	0.0002/0.0018/0.9278	0.0007/0.0018/0.7172
stressspd.max (10000)	-0.0071/0.0081/0.3820	-0.0071/0.0092/0.4437	-0.0085/0.0092/0.3554
tmt.ave	0.0083/0.0074/0.2605	0.0083/0.0084/0.3242	0.0100/0.0085/0.2374
tmt.mmm	0.0067/0.0045/0.1408	0.0067/0.0052/0.1968	0.0059/0.0052/0.2537
windspd.max	-0.0027/0.0081/0.7427	-0.0027/0.0092/0.7736	-0.0025/0.0091/0.7805
c.wdir=N	(-0.0360)/(0.0342)/0.2929	(-0.0360)/(0.0390)/0.3567	(-0.0305)/(0.0388)/0.4318
c.wdir=NO	-0.0301/0.0463/0.5157	-0.0301/0.0528/0.5690	-0.0291/0.0527/0.5816
c.wdir=NW	0.0345/0.0249/0.1662	0.0345/0.0284/0.2251	0.0391/0.0285/0.1697
c.wdir=O	-0.0062/0.0406/0.8777	-0.0062/0.0463/0.8927	-0.0084/0.0470/0.8587
c.wdir=S	-0.0416/0.0269/0.1218	-0.0416/0.0307/0.1752	-0.0422/0.0308/0.1716
c.wdir=SO	0.0188/0.0361/0.6034	0.0188/0.0412/0.6489	0.0175/0.0416/0.6750
c.wdir=SW	0.0321/0.0215/0.1347	0.0321/0.0245/0.1900	0.0272/0.0247/0.2698
c.wdir=W	0.0286/0.0188/0.1280	0.0286/0.0214/0.1824	0.0264/0.0216/0.2216
SO2.q95	-0.0079/0.0044/0.0726	-0.0079/0.0050/0.1156	-0.0085/0.0051/0.0944
PM10.q95	0.0005/0.0005/0.2572	0.0005/0.0005/0.3208	0.0005/0.0006/0.3864
O3.q95	0.0011/0.0008/0.1729	0.0011/0.0009/0.2323	0.0011/0.0009/0.2441
NO2.q95	-0.0013/0.0007/0.0710	-0.0013/0.0008/0.1137	-0.0013/0.0008/0.1170
CO.q95	0.0746/0.0507/0.1409	0.0746/0.0578/0.1970	0.0811/0.0584/0.1645

Tabelle G.7: Geschätzte Regressionskoeffizienten (roh)/Standardfehler/p-Werte des Poisson-, Quasi-Poisson- und Negativ-Binomial-Modells für die Call-Center-Anrufe in der Pilotregion München (gelb: signifikante Ergebnisse)

Kovariable	Pilotregion München	Aggregation über alle bayer. LKs
Intercept	-9.2449/0.0001/1.5306/0.0000	-48.5926/0.0000/10.8298/0.0000
dow=Mo	(1.0381)/2.8238/(0.0117)/0.0000	(1.0951)/2.9894/(0.0088)/0.0000
dow=Di	0.8615/2.3667/0.0119/0.0000	0.9461/2.5757/0.0090/0.0000
dow=Mi	0.5899/1.8039/0.0119/0.0000	0.6095/1.8395/0.0090/0.0000
dow=Do	0.8849/2.4227/0.0121/0.0000	0.9401/2.5601/0.0091/0.0000
dow=Fr	0.6088/1.8382/0.0119/0.0000	0.7536/2.1247/0.0089/0.0000
dow=Sa	-1.7669/0.1709/0.0123/0.0000	-1.8601/0.1557/0.0097/0.0000
dow=So	-2.2163/0.1090/0.0148/0.0000	-2.4843/0.0834/0.0124/0.0000
school=ja	-0.2786/0.7569/0.0112/0.0000	-0.1961/0.8219/0.0087/0.0000
holiday=ja	-2.6509/0.0706/0.0287/0.0000	-2.9788/0.0509/0.0229/0.0000
bridge=ja	-0.5621/0.5700/0.0420/0.0000	-0.4646/0.6284/0.0309/0.0000
quartal=Anfang	0.2301/1.2587/0.0187/0.0000	0.2568/1.2928/0.0142/0.0000
quartal=Ende	-0.1262/0.8815/0.0181/0.0000	-0.0845/0.9190/0.0133/0.0000
sex=2	0.1458/1.1570/0.0092/0.0000	0.0203/1.0205/0.0068/0.0028
age=1	(-0.4066)/0.6659/(0.0459)/0.0000	(-0.2087)/0.8116/(0.0339)/0.0000
age=2	-0.8746/0.4170/0.0459/0.0000	-0.8526/0.4263/0.0339/0.0000
age=3	0.1128/1.1194/0.0459/0.0140	-0.1801/0.8352/0.0339/0.0000
age=4	1.1685/3.2171/0.0459/0.0000	1.2414/3.4605/0.0339/0.0000
sex=2,age=1	(-0.5201)/0.5944/(0.0160)/0.0000	(-0.4063)/0.6661/(0.0118)/0.0000
sex=2,age=2	0.3584/1.4310/0.0160/0.0000	0.3248/1.3838/0.0118/0.0000
sex=2,age=3	0.1941/1.2142/0.0160/0.0000	0.2632/1.3010/0.0118/0.0000
sex=2,age=4	-0.0323/0.9682/0.0160/0.0429	-0.1817/0.8339/0.0118/0.0000
cp.max (0.001)	0.0013/1.0013/0.0026/0.6062	0.0026/1.0026/0.0035/0.4541
lsp.max (0.001)	0.0025/1.0025/0.0014/0.0706	-0.0032/0.9968/0.0024/0.1761
lcc.ave (0.1)	-0.0072/0.9928/0.0028/0.0092	0.0024/1.0024/0.0029/0.4032
mcc.ave (0.1)	-0.0008/0.9992/0.0027/0.7723	0.0069/1.0069/0.0028/0.0136
q.ave (0.001)	-0.0239/0.9763/0.0107/0.0248	-0.0101/0.9899/0.0077/0.1906
q.mmm (0.001)	0.0041/1.0042/0.0071/0.5603	-0.0188/0.9814/0.0076/0.0135
qnh.ave (München)	0.0024/1.0024/0.0015/0.1032	–
qnh.mmm (München)	-0.0017/0.9983/0.0019/0.3831	–
qnh.ave (0.001)	–	0.0036/1.0036/0.0009/0.0001
qnh.mmm (0.001)	–	0.0009/1.0009/0.0015/0.5620
sshf.ave (100000)	-0.0014/0.9986/0.0009/0.1319	-0.0011/0.9989/0.0012/0.3296
stressspd.max (10000)	0.0033/1.0033/0.0043/0.4416	0.0094/1.0094/0.0075/0.2119
tmt.ave	0.0064/1.0064/0.0041/0.1208	0.0096/1.0096/0.0030/0.0015
tmt.mmm	-0.0028/0.9972/0.0025/0.2673	0.0010/1.0010/0.0026/0.6848
windspd.max	-0.0001/0.9999/0.0043/0.9811	0.0197/1.0198/0.0078/0.0119
c.wdir=N	(0.0262)/1.0266/(0.0268)/0.3269	(-0.0193)/0.9809/(0.0174)/0.2684
c.wdir=NO	0.0065/1.0065/0.0278/0.8159	0.0041/1.0041/0.0133/0.7571
c.wdir=NW	-0.0080/0.9921/0.0143/0.5775	-0.0326/0.9680/0.0134/0.0149
c.wdir=O	-0.0030/0.9970/0.0224/0.8940	0.0219/1.0221/0.0097/0.0239
c.wdir=S	0.0245/1.0248/0.0163/0.1330	0.0466/1.0477/0.0122/0.0001
c.wdir=SO	-0.0017/0.9983/0.0224/0.9408	0.0103/1.0103/0.0138/0.4571
c.wdir=SW	-0.0333/0.9672/0.0127/0.0086	0.0158/1.0160/0.0092/0.0839
c.wdir=W	-0.0113/0.9888/0.0106/0.2874	-0.0469/0.9542/0.0092/0.0000
SO2.q95	-0.0046/0.9954/0.0024/0.0576	0.0025/1.0025/0.0023/0.2924
PM10.q95	0.0009/1.0009/0.0003/0.0115	-0.0009/0.9991/0.0004/0.0203
O3.q95	-0.0020/0.9980/0.0004/0.0000	-0.0026/0.9974/0.0004/0.0000
NO2.q95	0.0032/1.0032/0.0004/0.0000	0.0039/1.0039/0.0007/0.0000
CO.q95	-0.1391/0.8701/0.0294/0.0000	-0.0511/0.9502/0.0353/0.1484

Tabelle G.8: Geschätzte Regressionskoeffizienten (roh/exponentiell transformiert)/Standardfehler/p-Werte des loglinearen Modells für die Arztbesuche in der Pilotregion München und aggregiert über die bayerischen Landkreise (Reduktionsverfahren b)) (gelb: signifikante Ergebnisse)

Kovariable	$\hat{\beta}$	se($\hat{\beta}$)	p-Wert
Intercept	-12.8673	1.6863	0.0000
dow=Mo	(0.0031)	(0.0212)	0.8850
dow=Di	-0.1306	0.0228	0.0000
dow=Mi	-0.0821	0.0224	0.0002
dow=Do	-0.2050	0.0238	0.0000
dow=Fr	-0.0759	0.0221	0.0006
dow=Sa	0.2135	0.0207	0.0000
dow=So	0.2770	0.0240	0.0000
school=ja	0.1695	0.0194	0.0000
holiday=ja	0.2596	0.0451	0.0000
bridge=ja	0.0859	0.0809	0.2878
quartal=Anfang	0.0182	0.0326	0.5779
quartal=Ende	0.0979	0.0310	0.0016
age=1	(-2.0873)	(0.4189)	0.0000
age=2	-0.9795	0.1833	0.0000
age=3	-0.5058	0.1696	0.0029
age=4	1.0999	0.1327	0.0000
age=5	2.4727	0.1358	0.0000
cp.max (0.001)	0.0027	0.0049	0.5848
lsp.max (0.001)	-0.0007	0.0030	0.8022
lcc.ave (0.1)	0.0051	0.0047	0.2783
mcc.ave (0.1)	-0.0039	0.0047	0.4102
q.ave (0.001)	0.0011	0.0186	0.9508
q.mmm (0.001)	0.0161	0.0127	0.2048
qnh.ave (München)	0.0011	0.0016	0.4840
qnh.mmm (München)	-0.0016	0.0032	0.6111
sshf.ave (100000)	0.0005	0.0016	0.7417
stressspd.max (10000)	-0.0078	0.0082	0.3462
tmt.ave	0.0059	0.0075	0.4366
tmt.mmm	0.0070	0.0046	0.1273
windspd.max	-0.0041	0.0082	0.6227
c.wdir=N	(-0.0407)	(0.0351)	0.2451
c.wdir=NO	-0.0333	0.0474	0.4820
c.wdir=NW	0.0409	0.0255	0.1080
c.wdir=O	-0.0045	0.0416	0.9135
c.wdir=S	-0.0403	0.0274	0.1417
c.wdir=SO	0.0213	0.0367	0.5617
c.wdir=SW	0.0300	0.0220	0.1727
c.wdir=W	0.0267	0.0192	0.1647
SO2.q95	-0.0086	0.0045	0.0544
PM10.q95	0.0005	0.0005	0.3275
O3.q95	0.0011	0.0008	0.1702
NO2.q95	-0.0011	0.0007	0.1150
CO.q95	0.0699	0.0517	0.1764

Tabelle G.9: Geschätzte Regressionskoeffizienten, Standardfehler und p-Werte des ZIP-Modells für die Anrufe beim KVB-Call-Center in der Pilotregion München (gelb: signifikante Ergebnisse)

Zu Abschnitt 3.3:

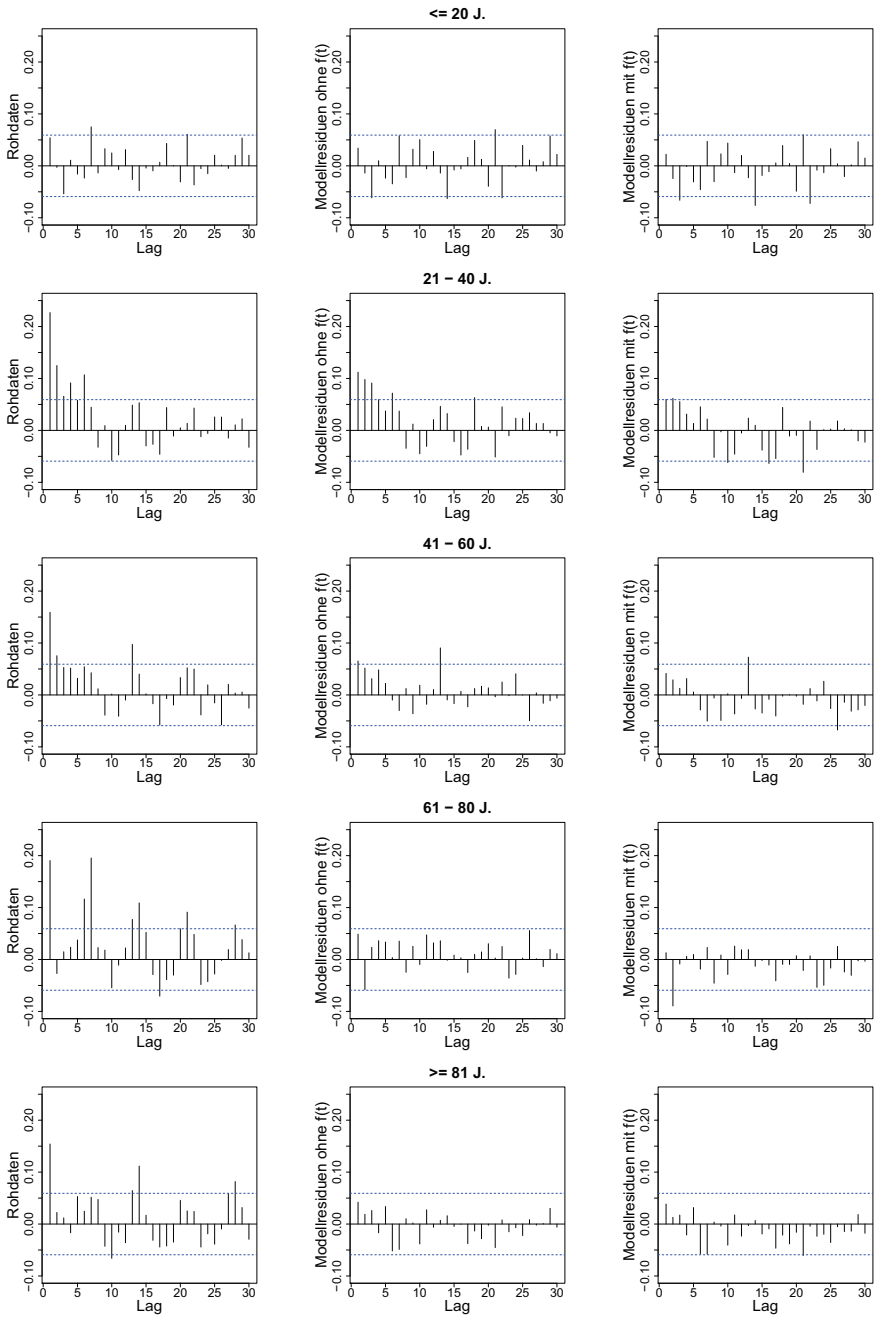


Abbildung G.9: Partielle Autokorrelationsfunktionen mit 95%-Konfidenzintervallen für die Call-Center-Anrufe in der Pilotregion München getrennt nach Altersgruppen basierend auf den rohen Anruferzahlen (links) sowie den Modellresiduen ohne (Mitte) und mit altersspezifischem Zeittrend (rechts)

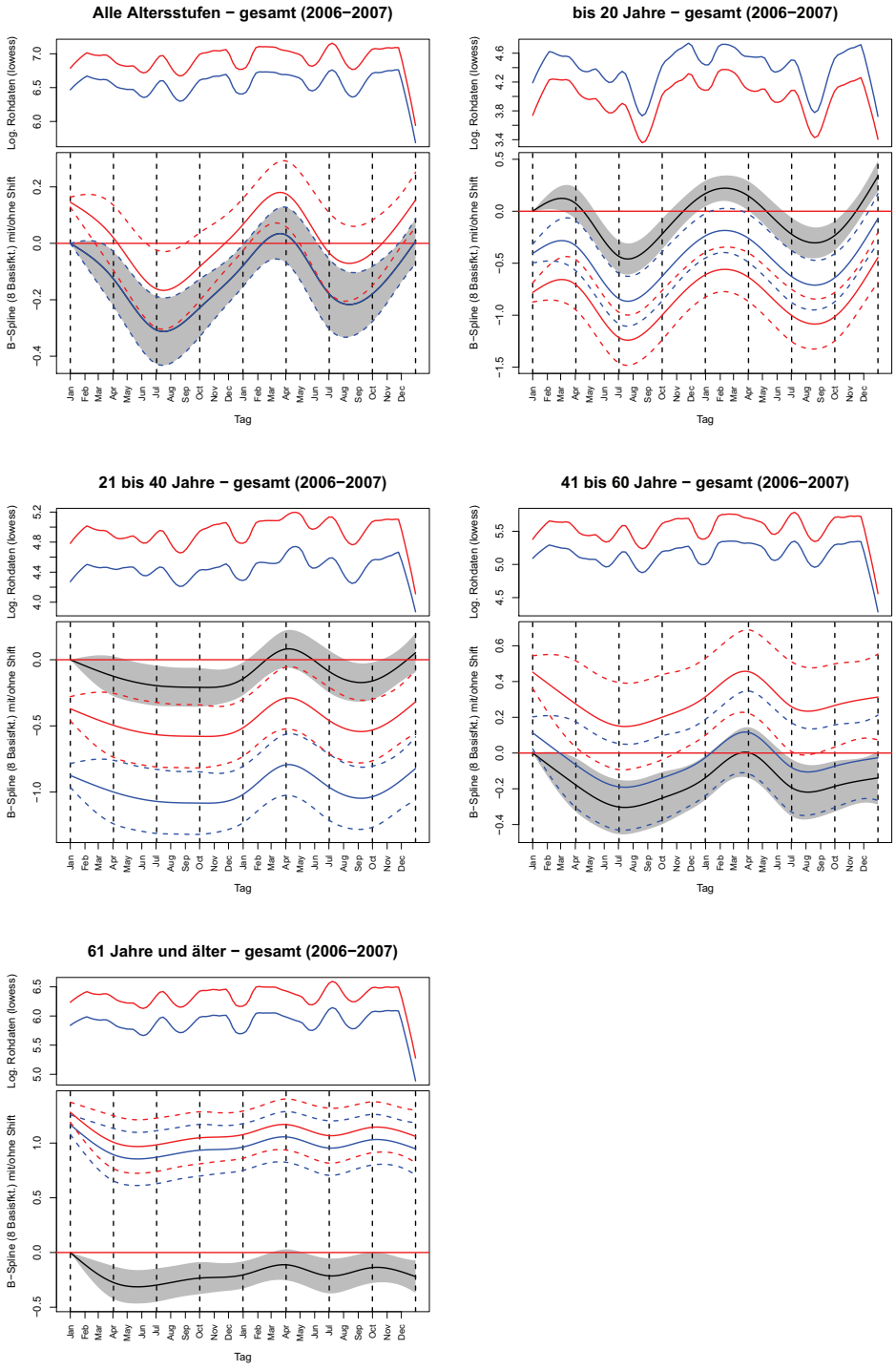


Abbildung G.10: Darstellung verschiedener Zeittrends des loglinearen Modells für die KVB-Abrechnungsdaten in der Pilotregion München zusammen mit zugehörigen punktweisen Konfidenzbändern im Vergleich zu lowess-Kurven durch die Rohdaten (jeweils oben): $\hat{f}(t)$ bzw. $\hat{f}_j(t)$ (schwarz), $\hat{f}_{\text{sex=1}}(t)$ bzw. $\hat{f}_{\text{age=i,sex=1}}(t)$ (blau), $\hat{f}_{\text{sex=2}}(t)$ bzw. $\hat{f}_{\text{age=i,sex=2}}(t)$ (rot)

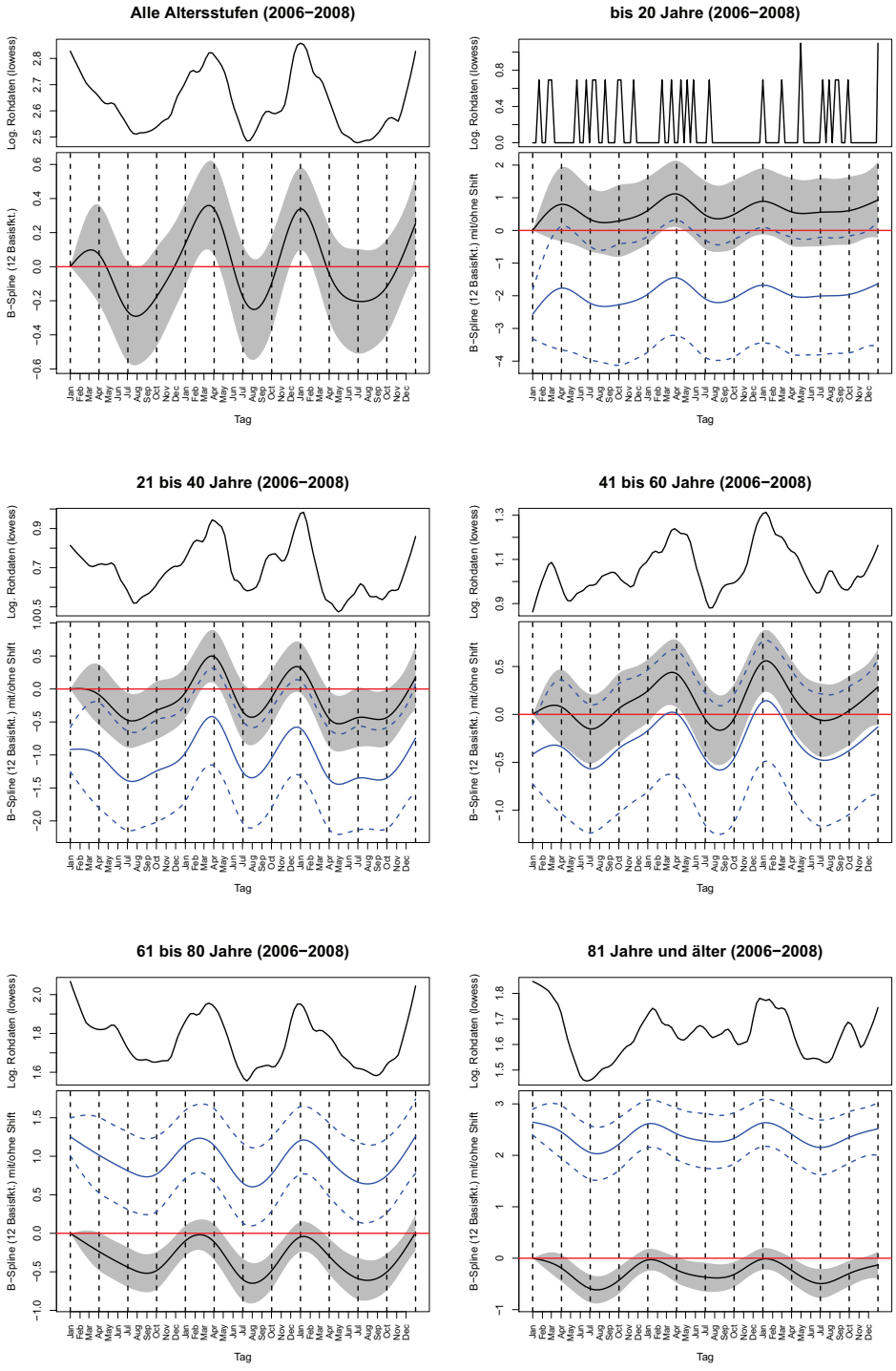


Abbildung G.11: Darstellung verschiedener Zeittrends des Poisson-Modells für die KVB-Call-Center-Daten in der Pilotregion München zusammen mit zugehörigen punktweisen Konfidenzbändern im Vergleich zu lowess-Kurven durch die Rohdaten (jeweils oben): $\hat{f}(t)$ bzw. $\hat{f}_i(t)$ (schwarz), $\hat{f}_{\text{age=i}}(t)$ (blau)

Zu Abschnitt 3.4:

KVB-Call-Center-Daten			Selektionsmodell ohne Haupteffekt	Volles GLM mit Haupteffekt
Kovariable	Datensatz	Cutpoint	$\hat{\beta}_{cp}$/exp($\mp\hat{\beta}_{cp}$)/p-Wert	$\hat{\beta}/\hat{\beta}_{cp}$/p-Wert
q.ave_lcp (0.001)	München (a))	0.0037 (20%)	-0.0309/1.0314/0.1732	-0.0022/0.0037/0.9445
	Bayern (b))	0.0038 (20%)	-0.0413/1.0422/0.0133	0.0016/0.0727/0.1262
q.ave_ucp (0.001)	München (a))	0.0090 (80%)	0.0460/1.0471/0.0022	-0.0022/0.0442/0.1220
	Bayern (b))	0.0091 (83%)	0.0342/1.0348/0.0020	0.0016/-0.0064/0.7401
qnh.ave_lcp (Mü.)	München (a))	1002.063 (5%)	-0.0181/1.0183/0.0023	0.0008/-0.0252/0.0005
qnh.ave_lcp (0.001)	Bayern (b))	11.4564 (5%)	-0.0045/1.045/0.3460	0.0021/-0.0155/0.0066
qnh.ave_ucp (Mü.)	München (a))	1030.079 (95%)	-0.0010/0.9990/0.3039	0.0008/-0.0150/0.1924
qnh.ave_ucp (0.001)	Bayern (b))	11.4776 (80%)	0.0064/1.0064/0.0113	0.0021/0.0035/0.3375
tmt.ave_lcp	München (a))	2.2540 (20%)	-0.0126/1.0127/0.0240	0.0083/-0.0232/0.0871
	Bayern (b))	1.8018 (19%)	-0.0152/1.0153/0.0005	0.0004/-0.0307/0.0114
tmt.ave_ucp	München (a))	16.8158 (80%)	0.0155/1.0156/0.0258	0.0083/-0.0200/0.0814
	Bayern (b))	19.3971 (91%)	0.0366/1.0373/0.0000	0.0004/0.0232/0.0479

KVB-Abrechnungsdaten			Selektionsmodell ohne Haupteffekt	Volles GLM mit Haupteffekt
Kovariable	Datensatz	Cutpoint	$\hat{\beta}_{cp}$/exp($\mp\hat{\beta}_{cp}$)/p-Wert	$\hat{\beta}/\hat{\beta}_{cp}$/p-Wert
q.ave_lcp (0.001)	München (a))	0.0037 (20%)	0.0719/0.9306/0.0000	-0.0239/0.0012/0.9755
	Bayern (b))	0.0025 (5%)	-0.5773/1.7812/0.0002	-0.0101/0.1726/0.0397
q.ave_ucp (0.001)	München (a))	0.0090 (80%)	-0.0063/0.9937/0.4410	-0.0239/0.0150/0.3180
	Bayern (b))	0.0099 (89%)	0.1866/1.2051/0.0000	-0.0101/0.0512/0.0006
qnh.ave_lcp (Mü.)	München (a))	1010.687 (20%)	0.0022/0.9978/0.2185	0.0024/-0.0010/0.7495
qnh.ave_lcp (0.001)	Bayern (b))	11.4606 (9%)	-0.0182/1.0184/0.0529	0.0036/-0.0061/0.0478
qnh.ave_ucp (Mü.)	München (a))	1023.284 (80%)	-0.0049/0.9951/0.0371	0.0024/-0.0082/0.0194
qnh.ave_ucp (0.001)	Bayern (b))	11.4778 (80%)	-0.0162/0.9839/0.0312	0.0036/-0.0049/0.0554
tmt.ave_lcp	München (a))	2.2315 (20%)	0.0222/0.9780/0.0000	0.0064/0.0241/0.0074
	Bayern (b))	-4.3516 (5%)	-0.1872/1.2059/0.0000	0.0096/-0.0731/0.0002
tmt.ave_ucp	München (a))	17.9320 (85%)	0.0037/1.0037/0.3866	0.0064/0.0010/0.8776
	Bayern (b))	16.5372 (81%)	0.0415/1.0424/0.0002	0.0096/0.0106/0.0184

Tabelle G.10: Luftfeuchtigkeits-, Luftdruck- und Temperatur-Cutpoints für die Call-Center- und Abrechnungsdaten in der Pilotregion München und aggregiert über die bayerischen Landkreise zusammen mit geschätzten Koeffizienten (roh und exponentiell transformiert) und p-Werten des Selektionsmodells sowie des vollen GLMs (gelb: signifikante Ergebnisse)

Zu Abschnitt 3.5:

$$
\begin{aligned}
\log(\mu_{st}) \;=\;& \text{offset}(\log(x_{\text{inhabitants},s})) + \text{offset}(\log(x_{\text{doctors},s})) + \beta_{\text{Intercept}} \\
&+ x_{\text{deprivation},s}\ \beta_{\text{deprivation}} + x_{\text{dow=Di},t}\ \beta_{\text{dow=Di}} + \ldots + x_{\text{dow=So},t}\ \beta_{\text{dow=So}} \\
&+ x_{\text{school=ja},t}\ \beta_{\text{school=ja}} + x_{\text{holiday=ja},t}\ \beta_{\text{holiday=ja}} \\
&+ x_{\text{bridge=ja},t}\ \beta_{\text{bridge=ja}} + x_{\text{quartal=Anfang},t}\ \beta_{\text{quartal=Anfang}} \\
&+ x_{\text{quartal=Ende},t}\ \beta_{\text{quartal=Ende}} + x_{\text{cp.max},st}\ \beta_{\text{cp.max}} + x_{\text{lsp.max},st}\ \beta_{\text{lsp.max}} \\
&+ x_{\text{lcc.ave},st}\ \beta_{\text{lcc.ave}} + x_{\text{mcc.ave},st}\ \beta_{\text{mcc.ave}} + x_{\text{q.ave},st}\ \beta_{\text{q.ave}} \\
&+ x_{\text{q.mmm},st}\ \beta_{\text{q.mmm}} + x_{\text{q.ave\_lcp},st}\ \beta_{\text{q.ave\_lcp}} + x_{\text{q.ave\_ucp},st}\ \beta_{\text{q.ave\_ucp}} \\
&+ x_{\text{qnh.ave},st}\ \beta_{\text{qnh.ave}} + x_{\text{qnh.mmm},st}\ \beta_{\text{qnh.mmm}} + x_{\text{qnh.ave\_lcp},st}\ \beta_{\text{qnh.ave\_lcp}} \\
&+ x_{\text{qnh.ave\_lcp},st}\ \beta_{\text{qnh.ave\_lcp}} + x_{\text{qnh.ave\_ucp},st}\ \beta_{\text{qnh.ave\_ucp}} + x_{\text{sshf.ave},st}\ \beta_{\text{sshf.ave}} \\
&+ x_{\text{stressspd.max},st}\ \beta_{\text{stressspd.max}} + x_{\text{tmt.ave},st}\ \beta_{\text{tmt.ave}} + x_{\text{tmt.mmm},st}\ \beta_{\text{tmt.mmm}} \\
&+ x_{\text{tmt.ave\_lcp},st}\ \beta_{\text{tmt.ave\_lcp}} + x_{\text{tmt.ave\_ucp},st}\ \beta_{\text{tmt.ave\_ucp}} + x_{\text{windspd.max},st}\ \beta_{\text{windspd.max}} \\
&+ x_{\text{c.wdir=NO},st}\ \beta_{\text{c.wdir=NO}} + \ldots + x_{\text{c.wdir=W},st}\ \beta_{\text{c.wdir=W}} \\
&+ x_{\text{SO2.q95},st}\ \beta_{\text{SO2.q95}} + x_{\text{PM10.q95},st}\ \beta_{\text{PM10.q95}} + x_{\text{O3.q95},st}\ \beta_{\text{O3.q95}} \\
&+ x_{\text{NO2.q95},st}\ \beta_{\text{NO2.q95}} + x_{\text{CO.q95},st}\ \beta_{\text{CO.q95}} + f(t) + \alpha_s + \underbrace{f(x_{\text{district},s})}_{\gamma_s}
\end{aligned}
$$

Abbildung G.12: Modellgleichung der räumlichen Modelle basierend auf Reduktionsverfahren c) für Call-Center- und Abrechnungsdaten

	Poisson-GLMM	Negativ-Binomial-GLMM
Kovariable	$\hat{\beta}$/exp($\hat{\beta}$) ($\hat{\beta}_{\text{lo}}$, $\hat{\beta}_{\text{up}}$)	$\hat{\beta}$/exp($\hat{\beta}$) ($\hat{\beta}_{\text{lo}}$, $\hat{\beta}_{\text{up}}$)
Intercept	-14.1876/0.0000 (-14.8742, -13.3725)	-14.0249/0.0000 (-14.6949, -13.2857)
Deprivation	0.0230/1.0233 (0.0152, 0.0318)	0.0251/1.0254 (0.0155, 0.0330)
dow=Mo	(-0.3652)/0.6941 (-0.3837, -0.3463)	(-0.4438)/0.6416 (-0.4655, -0.4228)
dow=Di	-0.4963/0.6088 (-0.5163, -0.4774)	-0.5556/0.5737 (-0.5786, -0.5316)
dow=Mi	-0.0558/0.9457 (-0.0741, -0.0389)	-0.0387/0.9621 (-0.0581, -0.0191)
dow=Do	-0.5064/0.6026 (-0.5259, -0.4856)	-0.5545/0.5744 (-0.5793, -0.5312)
dow=Fr	-0.2121/0.8088 (-0.2306, -0.1926)	-0.2259/0.7978 (-0.2478, -0.2029)
dow=Sa	0.8147/2.2584 (0.8022, 0.8290)	0.9119/2.4892 (0.8969, 0.9278)
dow=So	0.8212/2.2732 (0.8059, 0.8363)	0.9065/2.4757 (0.8876, 0.9247)
school=ja	0.1435/1.1543 (0.1286, 0.1580)	0.1478/1.1592 (0.1296, 0.1641)
holiday=ja	0.9775/2.6579 (0.9485, 1.0069)	1.1389/3.1235 (1.1019, 1.1749)
bridge=ja	0.5436/1.7222 (0.4846, 0.5987)	0.6225/1.8635 (0.5594, 0.6908)
quartal=Anfang	-0.0339/0.9667 (-0.0579, -0.0105)	-0.0096/0.9905 (-0.0385, 0.0226)
quartal=Ende	0.1463/1.1575 (0.1221, 0.1697)	0.1573/1.1703 (0.1292, 0.1879)
cp.max (0.001)	0.0009/1.0009 (-0.0039, 0.0055)	0.0003/1.0003 (-0.0050, 0.0055)
lsp.max (0.001)	0.0004/1.0004 (-0.0029, 0.0036)	0.0012/1.0012 (-0.0027, 0.0049)
lcc.ave (0.1)	0.0023/1.0023 (-0.0014, 0.0059)	-0.0002/0.9998 (-0.0044, 0.0039)
mcc.ave (0.1)	-0.0053/0.9947 (-0.0090, -0.0017)	-0.0044/0.9956 (-0.0089, 0.0002)
q.ave (0.001)	0.0008/1.0008 (-0.0119, 0.0136)	0.0034/1.0034 (-0.0118, 0.0202)
q.ave_lcp (0.001)	0.0672/1.0695 (0.0266, 0.1051)	0.0772/1.0803 (0.0312, 0.1254)
q.ave_ucp (0.001)	0.0077/1.0077 (-0.0118, 0.0288)	0.0034/1.0034 (-0.0218, 0.0262)
q.mmm (0.001)	0.0012/1.0012 (-0.0081, 0.0105)	0.0023/1.0023 (-0.0091, 0.0135)
qnh.ave (0.001)	0.0000/1.0000 (-0.0000, 0.0000)	-0.0000/1.0000 (-0.0000, 0.0000)
qnh.ave_lcp (0.001)	-0.0015/0.9985 (-0.0036, 0.0008)	-0.0001/0.9999 (-0.0025, 0.0023)
qnh.ave_ucp (0.001)	0.0020/1.0020 (0.0004, 0.0033)	0.0014/1.0014 (-0.0003, 0.0032)
qnh.mmm (0.001)	-0.0057/0.9943 (-0.0080, -0.0034)	-0.0053/0.9947 (-0.0080, -0.0026)
sshf.ave (100000)	0.0011/1.0011 (-0.0003, 0.0026)	0.0012/1.0013 (-0.0005, 0.0030)
stressspd.max (10000)	-0.0073/0.9927 (-0.0147, 0.0006)	-0.0115/0.9886 (-0.0210, -0.0020)
tmt.ave	-0.0047/0.9953 (-0.0097, 0.0010)	-0.0052/0.9948 (-0.0115, 0.0006)
tmt.ave_lcp	-0.0301/0.9703 (-0.0398, -0.0197)	-0.0291/0.9713 (-0.0415, -0.0163)
tmt.ave_ucp	0.0099/1.0099 (-0.0024, 0.0215)	0.0178/1.0179 (0.0024, 0.0329)
tmt.mmm	-0.0036/0.9964 (-0.0070, -0.0003)	-0.0042/0.9959 (-0.0082, -0.0000)
windspd.max	-0.0005/0.9995 (-0.0082, 0.0072)	-0.0025/0.9975 (-0.0120, 0.0070)
c.wdir=N	(-0.0166)/0.9835 (-0.0445, 0.0116)	(-0.0101)/0.9900 (-0.0439, 0.0247)
c.wdir=NO	-0.0250/0.9753 (-0.0475, -0.0020)	-0.0226/0.9777 (-0.0493, 0.0050)
c.wdir=NW	0.0255/1.0258 (0.0024, 0.0473)	0.0159/1.0160 (-0.0128, 0.0438)
c.wdir=O	-0.0453/0.9557 (-0.0625, -0.0281)	-0.0479/0.9533 (-0.0667, -0.0306)
c.wdir=S	0.0122/1.0123 (-0.0093, 0.0325)	0.0117/1.0117 (-0.0116, 0.0371)
c.wdir=SO	0.0402/1.0411 (0.0201, 0.0588)	0.0366/1.0373 (0.0112, 0.0611)
c.wdir=SW	-0.0105/0.9896 (-0.0256, 0.0047)	-0.0069/0.9931 (-0.0246, 0.0110)
c.wdir=W	0.0200/1.0202 (0.0044, 0.0357)	0.0258/1.0261 (0.0077, 0.0437)
SO2.q95	-0.0067/0.9933 (-0.0092, -0.0044)	-0.0033/0.9967 (-0.0060, -0.0006)
PM10.q95	0.0018/1.0018 (0.0014, 0.0021)	0.0014/1.0014 (0.0010, 0.0018)
O3.q95	0.0016/1.0016 (0.0010, 0.0021)	0.0015/1.0015 (0.0009, 0.0021)
NO2.q95	0.0003/1.0003 (-0.0002, 0.0007)	0.0001/1.0001 (-0.0005, 0.0008)
CO.q95	-0.0031/0.9969 (-0.0341, 0.0274)	-0.0222/0.9781 (-0.0649, 0.0169)

Tabelle G.11: Geschätzte Regressionskoeffizienten (Posteriori-Mittelwerte, roh/exponentiell transformiert) der räumlichen GLMMs für die Call-Center-Daten (reduziert nach Verfahren c)) zusammen mit 95%-Kredibilitätsintervallen (in Klammern) (gelb: signifikante Ergebnisse)

Kovariable	$\hat{\beta}$/exp($\hat{\beta}$) ($\hat{\beta}_{\text{lo}}$, $\hat{\beta}_{\text{up}}$)
Intercept	-10.5845/0.0000 (-12.8314, -9.3072)
Deprivation	-0.0126/0.9875 (-0.0295, 0.0160)
dow=Mo	(1.1067)/3.0242 (1.0999, 1.1134)
dow=Di	0.9794/2.6628 (0.9725, 0.9859)
dow=Mi	0.7010/2.0157 (0.6941, 0.7080)
dow=Do	0.9624/2.6181 (0.9560, 0.9692)
dow=Fr	0.7652/2.1493 (0.7586, 0.7716)
dow=Sa	-1.9556/0.1415 (-1.9628, -1.9488)
dow=So	-2.5589/0.0774 (-2.5664, -2.5513)
school=ja	-0.1823/0.8333 (-0.1892, -0.1755)
holiday=ja	-3.0036/0.0496 (-3.0194, -2.9880)
bridge=ja	-0.4943/0.6100 (-0.5195, -0.4672)
quartal=Anfang	0.3027/1.3536 (0.2917, 0.3133)
quartal=Ende	-0.0833/0.9200 (-0.0943, -0.0723)
cp.max (0.001)	-0.0013/0.9987 (-0.0034, 0.0006)
lsp.max (0.001)	-0.0012/0.9988 (-0.0025, 0.0000)
lcc.ave (0.1)	-0.0014/0.9986 (-0.0030, 0.0002)
mcc.ave (0.1)	0.0039/1.0039 (0.0023, 0.0055)
q.ave (0.001)	-0.0041/0.9959 (-0.0088, 0.0012)
q.ave_lcp (0.001)	-0.0008/0.9992 (-0.0491, 0.0480)
q.ave_ucp (0.001)	0.0296/1.0300 (0.0192, 0.0395)
q.mmm (0.001)	-0.0111/0.9889 (-0.0154, -0.0073)
qnh.ave (0.001)	-0.0000/1.0000 (-0.0000, -0.0000)
qnh.ave_lcp (0.001)	-0.0009/0.9991 (-0.0018, -0.0000)
qnh.ave_ucp (0.001)	-0.0002/0.9998 (-0.0009, 0.0006)
qnh.mmm (0.001)	0.0009/1.0009 (-0.0002, 0.0020)
sshf.ave (100000)	0.0005/1.0005 (-0.0002, 0.0011)
stressspd.max (10000)	0.0036/1.0036 (0.0001, 0.0070)
tmt.ave	0.0066/1.0066 (0.0045, 0.0085)
tmt.ave_lcp	-0.0008/0.9992 (-0.0112, 0.0097)
tmt.ave_ucp	-0.0013/0.9987 (-0.0042, 0.0016)
tmt.mmm	0.0013/1.0013 (-0.0002, 0.0028)
windspd.max	0.0039/1.0039 (0.0005, 0.0074)
c.wdir=N	(-0.0168)/0.9834 (-0.0283, -0.0053)
c.wdir=NO	-0.0047/0.9953 (-0.0143, 0.0052)
c.wdir=NW	0.0071/1.0072 (-0.0022, 0.0171)
c.wdir=O	0.0034/1.0034 (-0.0045, 0.0108)
c.wdir=S	0.0101/1.0101 (0.0016, 0.0189)
c.wdir=SO	-0.0187/0.9814 (-0.0281, -0.0101)
c.wdir=SW	0.0353/1.0359 (0.0287, 0.0422)
c.wdir=W	-0.0195/0.9807 (-0.0261, -0.0123)
SO2.q95	0.0002/1.0002 (-0.0008, 0.0012)
PM10.q95	0.0008/1.0008 (0.0007, 0.0010)
O3.q95	-0.0009/0.9991 (-0.0012, -0.0007)
NO2.q95	0.0014/1.0015 (0.0012, 0.0017)
CO.q95	-0.0324/0.9681 (-0.0448, -0.0192)

Tabelle G.12: Geschätzte Regressionskoeffizienten (Posteriori-Mittelwerte, roh/exponentiell transformiert) des loglinearen räumlichen GLMMs für die Abrechnungsdaten (reduziert nach Verfahren c)) zusammen mit 95%-Kredibilitätsintervallen (in Klammern) (gelb: signifikante Ergebnisse)

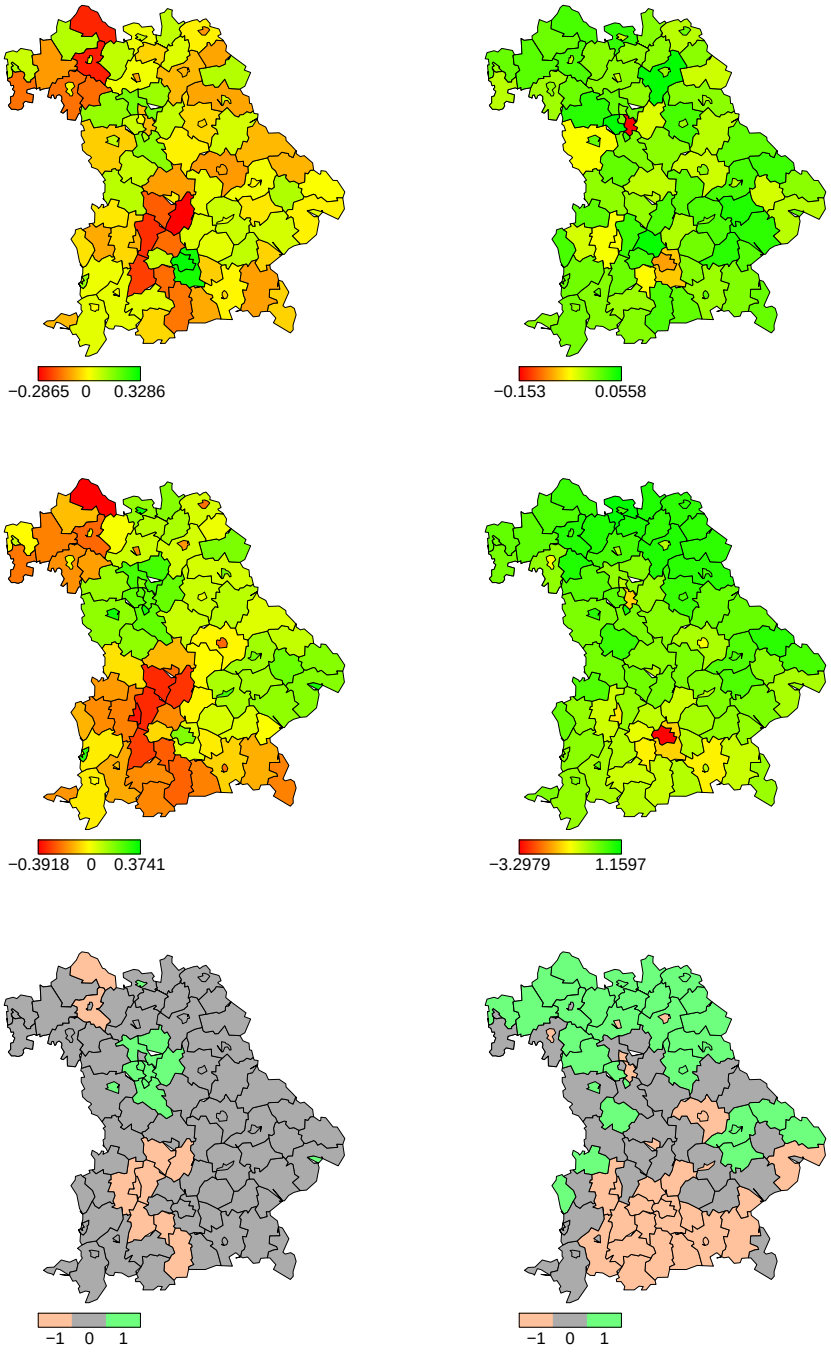


Abbildung G.13: Geschätzte räumliche Effekte des Poisson-GLMMs für die Call-Center-Daten (links) und des loglinearen GLMMs für die Abrechnungsdaten (rechts): zufällige Komponente (oben), strukturelle Komponente (Mitte) und Signifikanz-Darstellung der strukturellen Effekte (unten)

Zu Abschnitt 3.6:

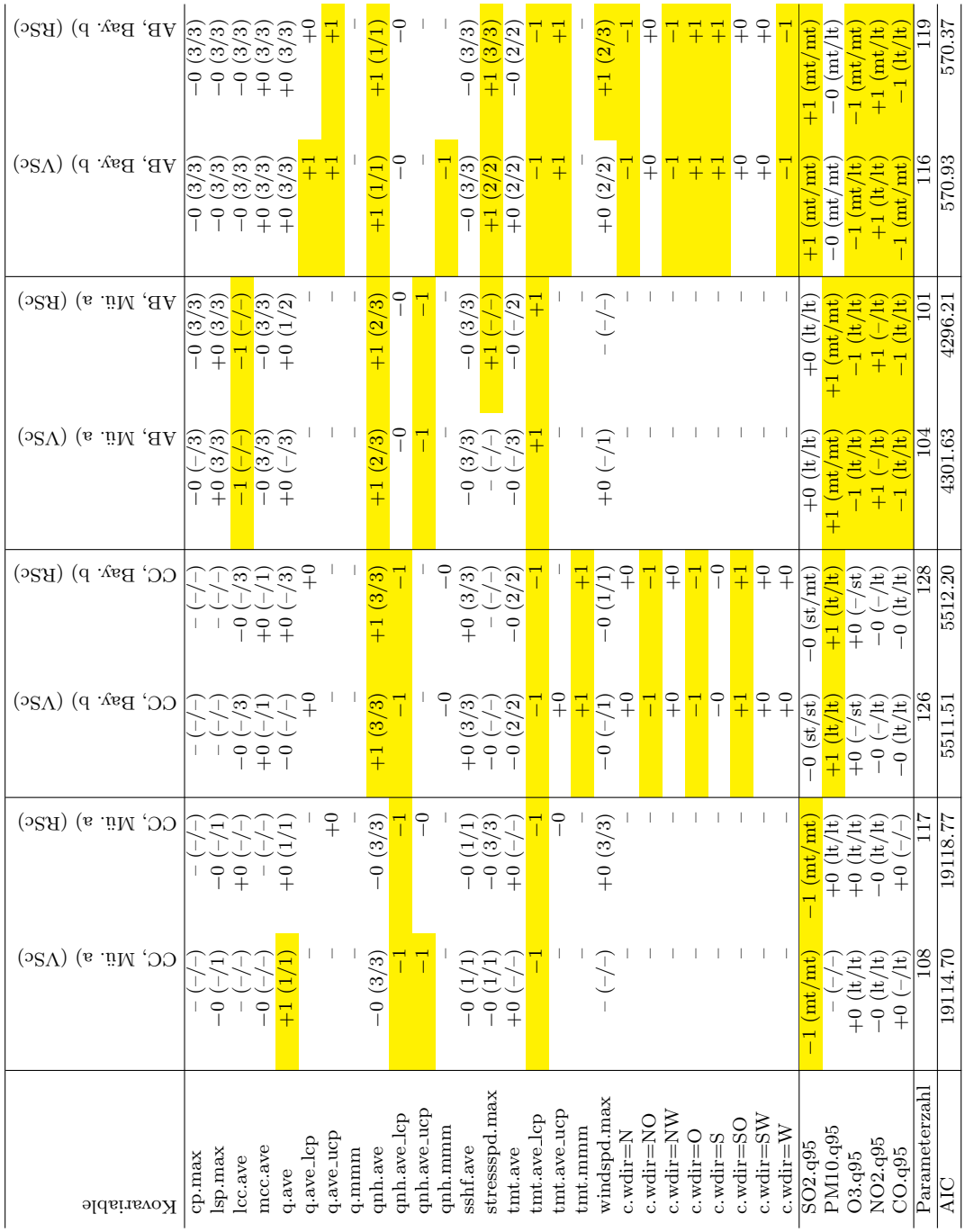
Kovariable	CC a) (Poisson-GLM)	CC a) (Quasi-Poisson-Mod.)	CC a) (Neg.-Bin.-GLM)	CC a) (ZIP-Mod.)	CC b) (loglin. GLM)	CC c) (Poisson-GLMM)	CC c) (Neg.-Bin.-GLMM)	AB a) (loglin. GLM)	AB b) (loglin. GLM)	AB c) (loglin. GLMM)
Intercept	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1
Deprivation	–	–	–	–	–	+1	+1	–	–	−0
dow=Mo	+0	+0	−0	+0	−1	−1	−1	+1	+1	+1
dow=Di	−1	−1	−1	−1	−1	−1	−1	+1	+1	+1
dow=Mi	−1	−1	−1	−1	−0	−1	−1	+1	+1	+1
dow=Do	−1	−1	−1	−1	−1	−1	−1	+1	+1	+1
dow=Fr	−1	−1	−1	−1	−1	−1	−1	+1	+1	+1
dow=Sa	+1	+1	+1	+1	+1	+1	+1	−1	−1	−1
dow=So	+1	+1	+1	+1	+1	+1	+1	−1	−1	−1
school=ja	+1	+1	+1	+1	+1	+1	+1	−1	−1	−1
holiday=ja	+1	+1	+1	+1	+1	+1	+1	−1	−1	−1
bridge=ja	+0	+0	+0	+0	+1	+1	+1	−1	−1	−1
quartal=Anfang	+0	+0	+0	+0	+0	−1	−0	+1	+1	+1
quartal=Ende	+1	+1	+1	+1	+1	+1	+1	−1	−1	−1
sex=2	–	–	–	–	–	–	–	+1	+1	–
age=1	−1	−1	−1	−1	−1	–	–	−1	−1	–
age=2	−1	−1	−1	−1	−1	–	–	−1	−1	–
age=3	−1	−1	−1	−1	−1	–	–	+1	−1	–
age=4	+1	+1	+1	+1	+1	–	–	+1	+1	–
age=5	+1	+1	+1	+1	+1	–	–	–	–	–
sex=2,age=1	–	–	–	–	–	–	–	−1	−1	–
sex=2,age=2	–	–	–	–	–	–	–	+1	+1	–
sex=2,age=3	–	–	–	–	–	–	–	+1	+1	–
sex=2,age=4	–	–	–	–	–	–	–	−1	−1	–
cp.max (0.001)	+0	+0	+0	+0	+0	+0	+0	+0	+0	−0
lsp.max (0.001)	−0	−0	−0	−0	+0	+0	+0	+0	−0	−0
lcc.ave (0.1)	+0	+0	+0	+0	−1	+0	−0	−1	+0	−0
mcc.ave (0.1)	−0	−0	−0	−0	+0	−1	−0	−0	+1	+1
q.ave (0.001)	−0	−0	−0	+0	+0	+0	+0	−1	−1	−0
q.ave_lcp (0.001)	+0	+0	+0	−0	+0	+1	+1	+0	+1	−0
q.ave_ucp (0.001)	+0	+0	+0	+0	−0	+0	+0	+0	+1	+1
q.mmm (0.001)	+0	+0	+0	+0	+0	+0	+0	+0	−1	−1
qnh.ave (Mü.)	+0	+0	+0	+0	–	–	–	+0	–	–
qnh.ave_lcp (Mü.)	−1	−1	−1	−1	–	–	–	−0	–	–
qnh.ave_ucp (Mü.)	−0	−0	−0	−0	–	–	–	−1	–	–
qnh.mmm (Mü.)	−0	−0	−0	−0	–	–	–	−0	–	–
qnh.ave (0.001)	–	–	–	–	+0	+0	−0	–	+1	−1
qnh.ave_lcp (0.001)	–	–	–	–	−1	−0	−0	–	−1	−1
qnh.ave_ucp (0.001)	–	–	–	–	+0	+1	+0	–	−0	−0
qnh.mmm (0.001)	–	–	–	–	−1	−1	−1	–	+0	+0
sshf.ave (100000)	+0	+0	+0	+0	+0	+0	+0	−0	−0	+0
stressspd.max (10000)	−0	−0	−0	−0	−0	−0	−1	+0	+0	+1
tmt.ave	+0	+0	+0	+0	+0	−0	−0	+0	+1	+1
tmt.ave_lcp	−0	−0	−0	−0	−1	−1	−1	+1	−1	−0
tmt.ave_ucp	−0	−0	−0	−0	+1	+0	+1	+0	+1	−0
tmt.mmm	+0	+0	+0	+0	+0	−1	−1	−0	+0	+0
windspd.max	−0	−0	−0	−0	−0	−0	−0	−0	+1	+1
c.wdir=N	−0	−0	−0	−0	+0	−0	−0	+0	−0	−1
c.wdir=NO	−0	−0	−0	−0	−1	−1	−0	+0	+0	−0
c.wdir=NW	+0	+0	+0	+0	+0	+1	+0	−0	−1	+0
c.wdir=O	−0	−0	−0	−0	−1	−1	−1	−0	+1	+0
c.wdir=S	−0	−0	−0	−0	−0	+0	+0	+0	+1	+1
c.wdir=SO	+0	+0	+0	+0	+1	+1	+1	−0	+0	−1
c.wdir=SW	+0	+0	+0	+0	+0	−0	−0	−1	+0	+1
c.wdir=W	+0	+0	+0	+0	+0	+1	+1	−0	−1	−1
SO2.q95	−0	−0	−0	−0	−0	−1	−1	−0	+0	+0
PM10.q95	+0	+0	+0	+0	+1	+1	+1	+1	−1	+1
O3.q95	+0	+0	+0	+0	+0	+1	+1	−1	−1	−1
NO2.q95	−0	−0	−0	−0	−1	+0	+0	+1	+1	+1
CO.q95	+0	+0	+0	+0	−0	−0	−0	−1	−0	−1

Tabelle G.13: Überblick über Richtung (+: positiv, −: negativ) und Signifikanz (1/gelb: signifikant, 0: nichtsignifikant) der Effekte aller in Abschnitt 3 gefitteten Modelle für Call-Center- (CC) und Abrechnungsdaten jeweils reduziert auf die Pilotregion München (a)), aggregiert über die bayerischen Landkreise (b)) und aggregiert über die (Geschlechts- und) Altersgruppen (c))

Kovariable	CC $\min(\hat{\beta})/\max(\hat{\beta})$	CC $\hat{\beta}_{\text{ave}}/\exp(\hat{\beta}_{\text{ave}})$	AB $\min(\hat{\beta})/\max(\hat{\beta})$	AB $\hat{\beta}_{\text{ave}}/\exp(\hat{\beta}_{\text{ave}})$
Intercept	-36.0014/-12.6966	-20.9521/0.0000	-48.5926/-9.2449	-22.8073/0.0000
Deprivation	0.0230/0.0251	0.0241/1.02439	-0.0126/-0.0126	-0.0126/0.9875
dow=Mo	-0.4438/0.0031	-0.2645/0.7676	1.0381/1.1067	1.0800/2.9446
dow=Di	-0.5556/-0.1299	-0.3752/0.6872	0.8615/0.9794	0.929/2.5320
dow=Mi	-0.0832/-0.0250	-0.1934/0.8242	0.5899/0.7010	0.6335/1.8841
dow=Do	-0.5545/-0.2047	-0.3977/0.6718	0.8849/0.9624	0.9291/2.5323
dow=Fr	-0.2259/-0.0704	-0.1476/0.8628	0.6088/0.7652	0.7092/2.0324
dow=Sa	0.2135/0.9119	0.6110/1.8423	-1.9556/-1.7669	-1.8609/0.1555
dow=So	0.2741/0.9065	0.6257/1.8695	-2.5589/-2.2163	-2.4198/0.0889
school=ja	0.0999/0.1731	0.1382/1.1482	-0.2786/-0.1823	-0.2190/0.8033
holiday=ja	0.2596/1.1389	0.7618/2.1421	-3.0036/-2.6509	-1.1105/0.3294
bridge=ja	0.0724/0.6225	0.3589/1.4318	-0.5621/-0.4646	-0.5070/0.6023
quartal=Anfang	-0.0339/0.0154	-0.0629/0.9390	0.2301/0.3027	0.2632/1.3011
quartal=Ende	0.0970/0.1573	0.1267/1.1351	-0.1262/-0.0833	-0.0980/0.9066
sex=2	–/–	–/–	0.0203/0.1458	0.0831/1.0866
age=1	-2.5647/-1.0008	-1.7228/0.1786	-0.4066/-0.2087	-0.3076/0.7352
age=2	-1.0086/-0.9174	-0.0375/0.9632	-0.8746/-0.8526	-0.8636/0.4216
age=3	-0.7630/-0.4158	-0.6010/0.5482	-0.1801/0.1128	-0.0337/0.9669
age=4	0.6431/1.2521	0.9276/2.5283	1.1685/1.2414	1.205/3.3366
age=5	2.1293/2.6576	2.3674/10.6696	–/–	–/–
sex=2,age=1	–/–	–/–	-0.5201/-0.4063	-0.4632/0.6293
sex=2,age=2	–/–	–/–	0.3248/0.3584	0.3416/1.4072
sex=2,age=3	–/–	–/–	0.1941/0.2632	0.2286/1.2569
sex=2,age=4	–/–	–/–	-0.1817/-0.0323	-0.1070/0.8985
cp.max (0.001)	0.0003/0.0040	0.0018/1.0018	-0.0013/0.0026	0.0009/1.0009
lsp.max (0.001)	-0.0007/0.0041	0.0015/1.0015	-0.0032/0.0025	-0.0006/0.9994
lcc.ave (0.1)	-0.0097/0.0051	-0.0012/0.9988	-0.0072/0.0024	-0.0021/0.9979
mcc.ave (0.1)	-0.0055/0.0018	-0.0026/0.9974	-0.0008/0.0069	0.0033/1.0033
q.ave (0.001)	-0.0072/0.0034	0.0002/1.0002	-0.0239/-0.0041	-0.0127/0.9874
q.ave_lcp (0.001)	-0.0023/0.0772	0.0492/1.0504	-0.0008/0.1726	0.0577/1.0594
q.ave_ucp (0.001)	-0.0064/0.0442	0.0143/1.0144	0.0150/0.0512	0.0319/1.0324
q.mmm (0.001)	0.0012/0.0193	0.0076/1.0076	-0.0188/0.0041	-0.0086/0.9914
qnh.ave (Mü.)	0.0008/0.0011	0.0009/1.0009	0.0024/0.0024	0.0024/1.0024
qnh.ave_lcp (Mü.)	-0.0251/-0.0252	-0.0252/0.9751	-0.0010/-0.0010	-0.0010/0.9990
qnh.ave_ucp (Mü.)	-0.0135/-0.0150	-0.0143/0.9858	-0.0082/-0.0082	-0.0082/0.9918
qnh.mmm (Mü.)	-0.0027/-0.0016	-0.0023/0.9977	-0.0017/-0.0017	-0.0017/0.9983
qnh.ave (0.001)	-0.0000/0.0021	0.0010/1.0011	-0.0000/0.0036	0.0018/1.0018
qnh.ave_lcp (0.001)	0.0155/-0.0001	-0.0081/0.9919	-0.0061/-0.0009	-0.0035/0.9965
qnh.ave_ucp (0.001)	0.0014/0.0035	0.0026/1.0026	-0.0049/-0.0002	-0.0025/0.9975
qnh.mmm (0.001)	-0.0057/-0.0053	-0.0056/0.9944	0.0009/0.0009	0.0009/1.0009
sshf.ave (100000)	0.0002/0.0015	0.0010/1.0010	-0.0014/0.0005	-0.0007/0.9993
stressspd.max (10000)	-0.0147/-0.0071	-0.0106/0.9895	0.0033/0.0094	0.0054/1.0054
tmt.ave	-0.0052/0.0100	0.0012/1.0012	0.0064/0.0096	0.0075/1.0076
tmt.ave_lcp	-0.0307/-0.0232	-0.0278/0.9726	-0.0731/0.0241	-0.0166/0.9835
tmt.ave_ucp	-0.0200/0.0232	0.0060/1.0060	-0.0013/0.0106	0.0034/1.0034
tmt.mmm	-0.0042/0.0076	0.0034/1.0034	-0.0028/0.0013	-0.0002/0.9998
windspd.max	-0.0071/-0.0005	-0.0039/0.9961	-0.0001/0.0197	0.0078/1.0079
c.wdir=N	-0.0407/0.0064	-0.0143/0.9858	-0.0193/0.0262	-0.0033/0.9967
c.wdir=NO	-0.0523/-0.0226	-0.0356/0.9650	-0.0047/0.0065	0.0020/1.0020
c.wdir=NW	0.0159/0.0409	0.0273/1.0276	-0.0326/0.0071	-0.0112/0.9889
c.wdir=O	-0.0512/-0.0045	-0.0347/0.9659	-0.0030/0.0219	0.0074/1.0075
c.wdir=S	-0.0422/0.0122	-0.0101/0.9900	0.0101/0.0466	0.0271/1.0274
c.wdir=SO	0.0175/0.0600	0.0392/1.0399	-0.0187/0.0103	-0.0034/0.9966
c.wdir=SW	-0.0105/0.0321	0.0090/1.0090	-0.0333/0.0353	0.0059/1.0060
c.wdir=W	0.0090/0.0286	0.0198/1.0200	-0.0469/-0.0113	-0.0259/0.9744
SO2.q95	-0.0086/-0.0033	-0.0060/0.9940	-0.0046/0.0025	-0.0006/0.9994
PM10.q95	0.0005/0.0022	0.0014/1.0014	-0.0009/0.0009	0.0003/1.0003
O3.q95	0.0010/0.0016	0.0012/1.0012	-0.0026/-0.0009	-0.0018/0.9982
NO2.q95	-0.0031/0.0003	-0.0014/0.9986	0.0014/0.0039	0.0028/1.0028
CO.q95	-0.0442/0.0811	0.0061/1.0061	-0.1391/-0.0324	-0.0742/0.9285

Tabelle G.14: Wertebereich der geschätzten Effekte und „mittlerer“ Effekt (roh/exponentiell transformiert) aller in Abschnitt 3 gefitteten Modelle für die Anzahl der Anrufe beim KVB-Call-Center (CC) und der Arztbesuche gesamt (AB)

Zu Abschnitt 3.7:



Kovariable	CC, Mii. a) (VSc)	CC, Mii. a) (RSc)	CC, Bay. b) (VSc)	CC, Bay. b) (RSc)	AB, Mii. a) (VSc)	AB, Mii. a) (RSc)	AB, Bay. b) (VSc)	AB, Bay. b) (RSc)
cp.max	– (–/–)	– (–/–)	– (–/–)	– (–/–)	−0 (–/3)	−0 (3/3)	−0 (3/3)	−0 (3/3)
lsp.max	−0 (–/1)	−0 (–/1)	– (–/–)	– (–/–)	+0 (3/3)	+0 (3/3)	−0 (3/3)	−0 (3/3)
lcc.ave	– (–/–)	+0 (–/–)	−0 (–/3)	−0 (–/3)	−1 (–/–)	−1 (–/–)	−0 (3/3)	−0 (3/3)
mcc.ave	−0 (–/–)	– (–/–)	+0 (–/1)	+0 (–/1)	−0 (3/3)	−0 (3/3)	+0 (3/3)	+0 (3/3)
q.ave	+1 (1/1)	+0 (1/1)	−0 (–/–)	+0 (–/3)	+0 (–/3)	+0 (1/2)	+0 (3/3)	+0 (3/3)
q.ave_lcp	–	–	+0	+0	–	–	+1	+0
q.ave_ucp	–	+0	–	–	–	–	+1	+1
q.mmm	–	–	–	–	–	–	–	–
qnh.ave	−0 (3/3)	−0 (3/3)	+1 (3/3)	+1 (3/3)	+1 (2/3)	+1 (2/3)	+1 (1/1)	+1 (1/1)
qnh.ave_lcp	−1	−1	−1	−1	−0	−0	−0	−0
qnh.ave_ucp	−1	−0	–	–	−1	−1	–	–
qnh.mmm	–	–	−0	−0	–	–	−1	–
sshf.ave	−0 (1/1)	−0 (1/1)	+0 (3/3)	+0 (3/3)	−0 (3/3)	−0 (3/3)	−0 (3/3)	−0 (3/3)
stresspd.max	−0 (1/1)	−0 (3/3)	−0 (–/–)	– (–/–)	– (–/–)	+1 (–/–)	+1 (2/2)	+1 (3/3)
tmt.ave	+0 (–/–)	+0 (–/–)	−0 (2/2)	−0 (2/2)	−0 (–/3)	−0 (–/2)	+0 (2/2)	−0 (2/2)
tmt.ave_lcp	−1	−1	−1	−1	+1	+1	−1	−1
tmt.ave_ucp	–	−0	+0	–	–	–	+1	+1
tmt.mmm	–	–	+1	–	–	–	–	–
windspd.max	– (–/–)	+0 (3/3)	−0 (–/1)	−0 (1/1)	+0 (–/1)	– (–/–)	+0 (2/2)	+1 (2/3)
c.wdir=N	–	–	+0	+0	–	–	−1	−1
c.wdir=NO	–	–	−1	−1	–	–	+0	+0
c.wdir=NW	–	–	+0	+0	–	–	−1	−1
c.wdir=O	–	–	−1	−1	–	–	+1	+1
c.wdir=S	–	–	−0	−0	–	–	+1	+1
c.wdir=SO	–	–	+1	+1	–	–	+0	+0
c.wdir=SW	–	–	+0	+0	–	–	+0	+0
c.wdir=W	–	–	+0	+0	–	–	−1	−1
SO2.q95	−1 (mt/mt)	−1 (mt/mt)	−0 (st/st)	−0 (st/mt)	+0 (lt/lt)	+0 (lt/lt)	+1 (mt/mt)	+1 (mt/mt)
PM10.q95	– (–/–)	+0 (lt/lt)	+1 (lt/lt)	+1 (lt/lt)	+1 (mt/mt)	+1 (mt/mt)	−0 (mt/mt)	−0 (mt/lt)
O3.q95	+0 (lt/lt)	+0 (lt/lt)	+0 (–/st)	+0 (–/st)	−1 (lt/lt)	−1 (lt/lt)	−1 (mt/lt)	−1 (mt/mt)
NO2.q95	−0 (lt/lt)	−0 (lt/lt)	−0 (–/lt)	−0 (–/lt)	+1 (–/lt)	+1 (–/lt)	+1 (lt/lt)	+1 (mt/lt)
CO.q95	+0 (–/lt)	+0 (–/–)	−0 (lt/lt)	−0 (lt/lt)	−1 (lt/lt)	−1 (lt/lt)	−1 (mt/mt)	−1 (lt/lt)
Parameterzahl	108	117	126	128	104	101	116	119
AIC	19114.70	19118.77	5511.51	5512.20	4301.63	4296.21	570.93	570.37

Tabelle G.15: Ergebnisse der korrigierten Vorwärts- (VSc) und Rückwärtsselektionsmodelle (RSc) für Call-Center- (CC) und Abrechnungsdaten (AB) jeweils reduziert auf die Pilotregion München (a)) und aggregiert über Gesamtbayern (b)): Richtung (+: positiv, −: negativ) und Signifikanz (1/gelb: signifikant, 0: nichtsignifikant) der Effekte (maximal signifikantes Lag/maximales Lag im Modell), Parameterzahl und AIC

Zu Abschnitt 3.8:

Kovariable	CC, Mü. a) (La)	CC, Mü. a) (Ri)	CC, Bay. b) (La)	CC, Bay. b) (Ri)	AB, Mü. a) (La)	AB, Mü. a) (Ri)	AB, Bay. b) (La)	AB, Bay. b) (Ri)
cp.max (sd)	– (–/–)	– (–/–)	– (–/–)	– (–/–)	– (–/–)	+0 (2/3)	−0 (2/2)	– (–/–)
lsp.max (sd)	– (–/–)	−0 (–/1)	– (–/–)	– (–/–)	+1 (3/3)	+1 (3/3)	−1 (3/3)	−1 (3/3)
lcc.ave (sd)	– (–/–)	– (–/–)	– (–/–)	– (–/–)	−1 (–/–)	−1 (–/–)	−0 (3/3)	−0 (3/3)
mcc.ave (sd)	– (–/–)	– (–/–)	– (–/–)	– (–/–)	−0 (3/3)	−0 (3/3)	+0 (3/3)	+0 (3/3)
q.ave (std)	– (–/–)	+1 (–/2)	– (–/–)	– (–/–)	−0 (–/–)	−0 (–/–)	+1 (2/3)	+1 (2/3)
q.ave_lcp (sd)	–	–	–	–	−1	−1	–	–
q.ave_ucp (sd)	–	–	–	–	–	–	–	–
q.mmm (sd)	–	–	–	–	–	–	–	–
qnh.ave (sd)	−1 (3/3)	−1 (3/3)	+0 (3/3)	+0 (3/3)	+1 (1/1)	+1 (1/1)	+1 (1/1)	+1 (1/1)
qnh.ave_lcp (sd)	–	–	–	–	–	–	−1	−1
qnh.ave_ucp (sd)	–	–	–	–	–	–	–	–
qnh.mmm (sd)	–	–	–	–	–	–	–	–
sshf.ave (sd)	– (–/–)	– (–/–)	– (–/–)	– (–/–)	−0 (3/3)	−0 (3/3)	−0 (3/3)	−0 (3/3)
stressspd.max (sd)	−0 (3/3)	−0 (3/3)	– (–/–)	– (–/–)	– (–/–)	– (–/–)	+0 (2/2)	+0 (2/2)
tmt.ave (sd)	– (–/–)	– (–/–)	– (–/–)	– (–/–)	+1 (–/–)	+1 (–/–)	−0 (2/2)	−0 (2/2)
tmt.ave_lcp (sd)	–	–	–	–	+1	+1	–	–
tmt.ave_ucp (sd)	–	–	–	–	–	–	+1	+1
tmt.mmm (sd)	–	–	–	–	–	–	–	–
windspd.max (sd)	−0 (3/3)	+0 (3/3)	– (–/–)	– (–/–)	– (–/–)	– (–/–)	+1 (–/–)	+1 (–/–)
c.wdir=N	–	–	−0	−0	–	–	−1	−1
c.wdir=NO	–	–	−1	−1	–	–	+0	−0
c.wdir=NW	–	–	+0	+0	–	–	−0	−0
c.wdir=O	–	–	−1	−1	–	–	+0	+0
c.wdir=S	–	–	+0	+0	–	–	+1	+1
c.wdir=SO	–	–	+1	+1	–	–	+0	+0
c.wdir=SW	–	–	+0	+0	–	–	+0	+0
c.wdir=W	–	–	−0	−0	–	–	−1	−1
SO2.q95 (sd)	– (–/–)	−0 (mt/mt)	+0 (st/st)	+0 (st/st)	– (–/–)	– (–/–)	+1 (mt/mt)	+1 (mt/mt)
PM10.q95 (sd)	– (–/–)	– (–/–)	+1 (–/–)	+1 (–/–)	+1 (mt/mt)	+1 (mt/mt)	−0 (mt/mt)	−0 (mt/mt)
O3.q95 (sd)	+0 (–/–)	+0 (–/–)	– (–/–)	– (–/–)	−1 (–/–)	−1 (–/–)	−1 (mt/mt)	−1 (mt/mt)
NO2.q95 (sd)	−0 (lt/lt)	−0 (lt/lt)	– (–/–)	+0 (–/lt)	+1 (–/–)	+1 (–/–)	+1 (mt/mt)	+1 (mt/mt)
CO.q95 (sd)	– (–/–)	– (–/–)	−1 (lt/lt)	−1 (lt/lt)	−1 (–/–)	−1 (–/–)	−1 (mt/mt)	−1 (mt/mt)

Tabelle G.16: Ergebnisse der korrigierten Lasso- (La) und Ridge-Shrinkage-Modelle (Ri) für Call-Center- (CC) und Abrechnungsdaten (AB) jeweils reduziert auf die Pilotregion München (a)) und aggregiert über Gesamtbayern (b)): Richtung (+: positiv, −: negativ) und Signifikanz (1/gelb: signifikant, 0: nichtsignifikant) der Effekte (maximal signifikantes Lag/maximales Lag im Modell)

Zu Abschnitt 4.2:

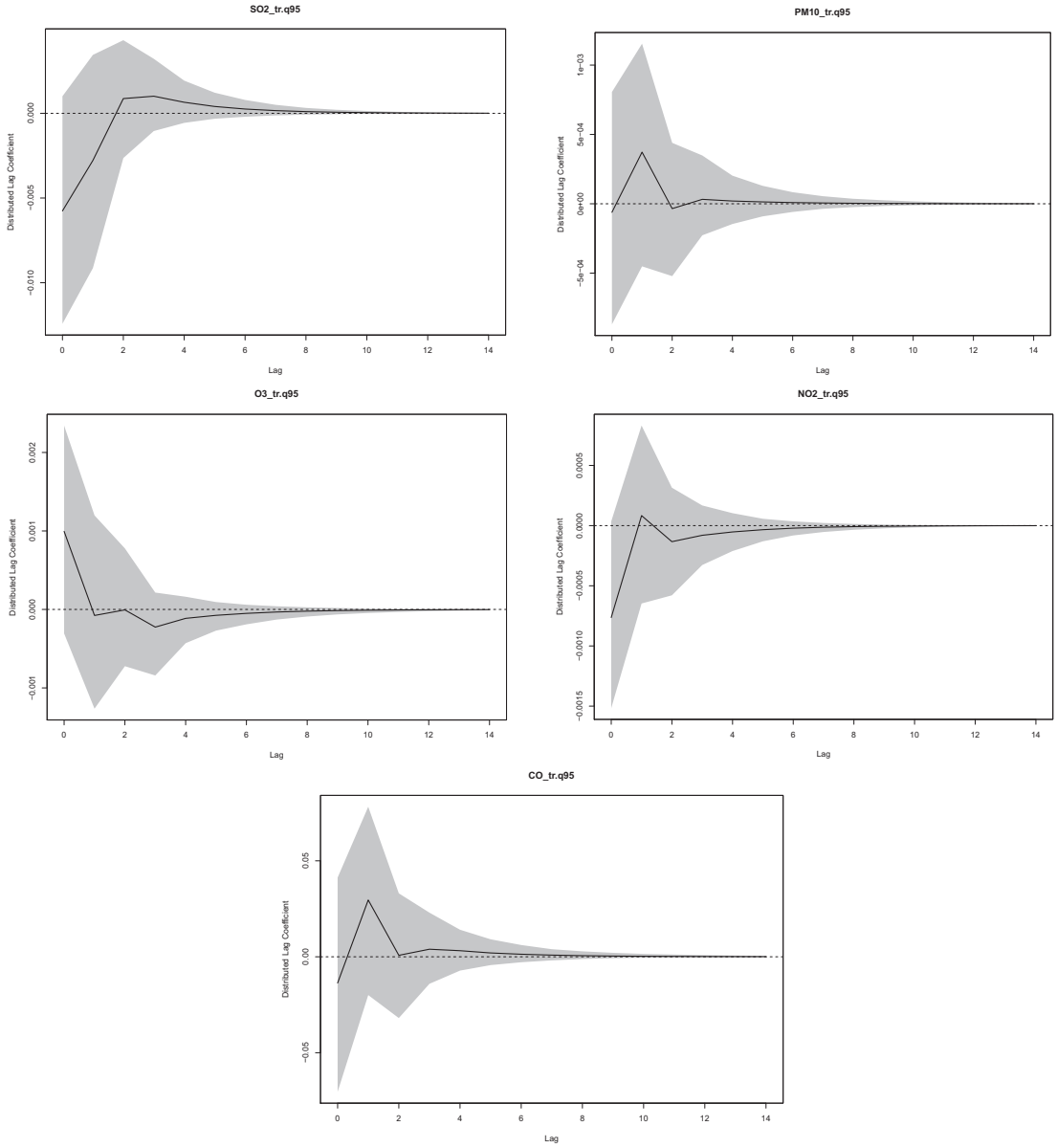


Abbildung G.14: Distributed Lag Functions von SO2.q95 (oben links), PM10.q95 (oben rechts), O3.q95 (Mitte links), NO2.q95 (Mitte rechts) und CO.q95 (unten) in den jeweils gefitteten BDLMs für die Call-Center-Daten in der Pilotregion München mit punktweisen 95%-Kredibilitätsbändern (grau)

Zu Abschnitt 4.3:

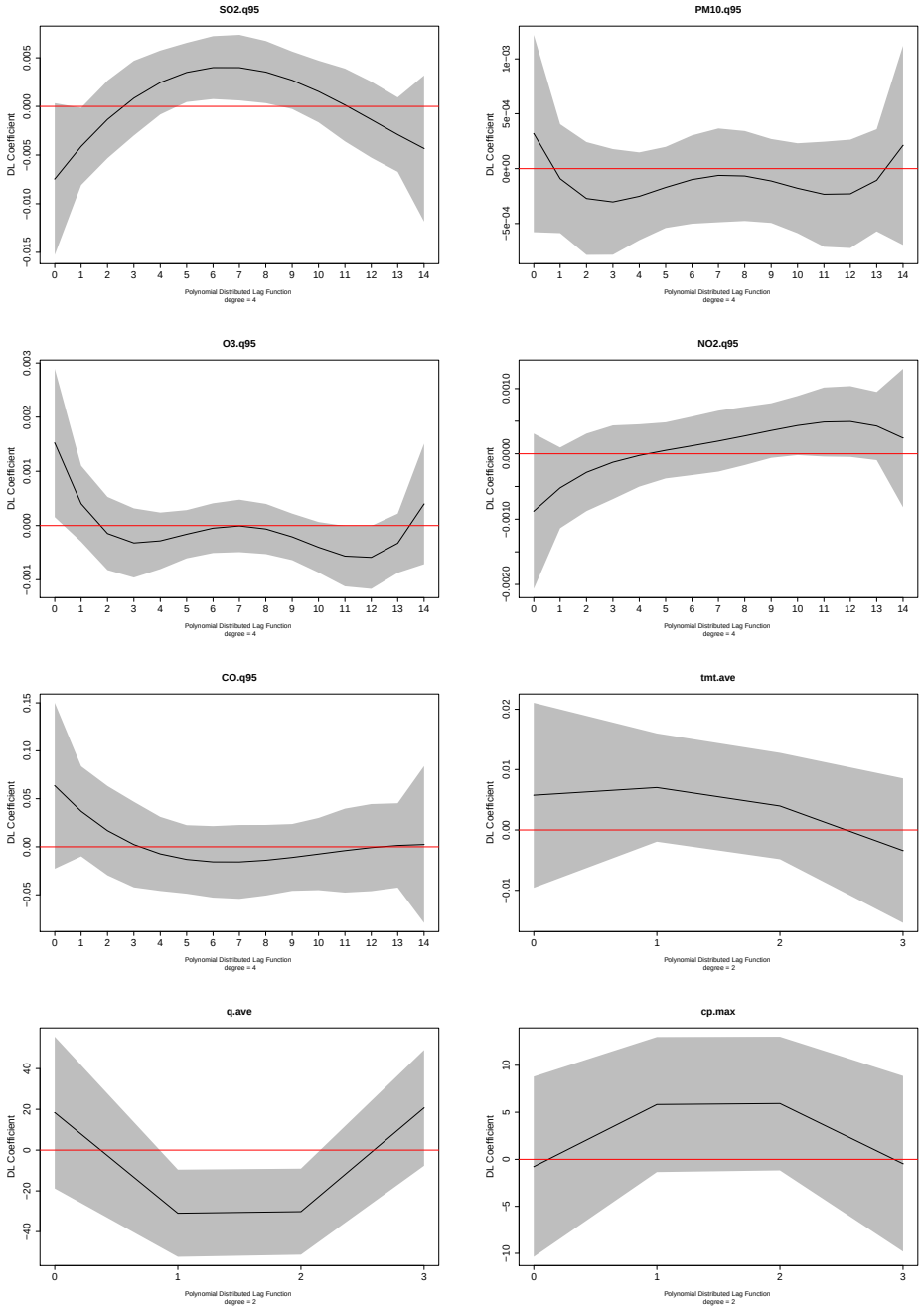


Abbildung G.15: Distributed Lag Functions von SO2.q95 (1. Reihe links), PM10.q95 (1. Reihe rechts), O3.q95 (2. Reihe links), NO2.q95 (2. Reihe rechts) und CO.q95 (3. Reihe links), tmt.ave (3. Reihe rechts), q.ave (4. Reihe links), cp.max (4.Reihe rechts) im Almon-Lag-Modell für die Call-Center-Daten in der Pilotregion München mit punktweisen 95%-Konfidenzbändern (grau)

Zu Abschnitt 4.4:

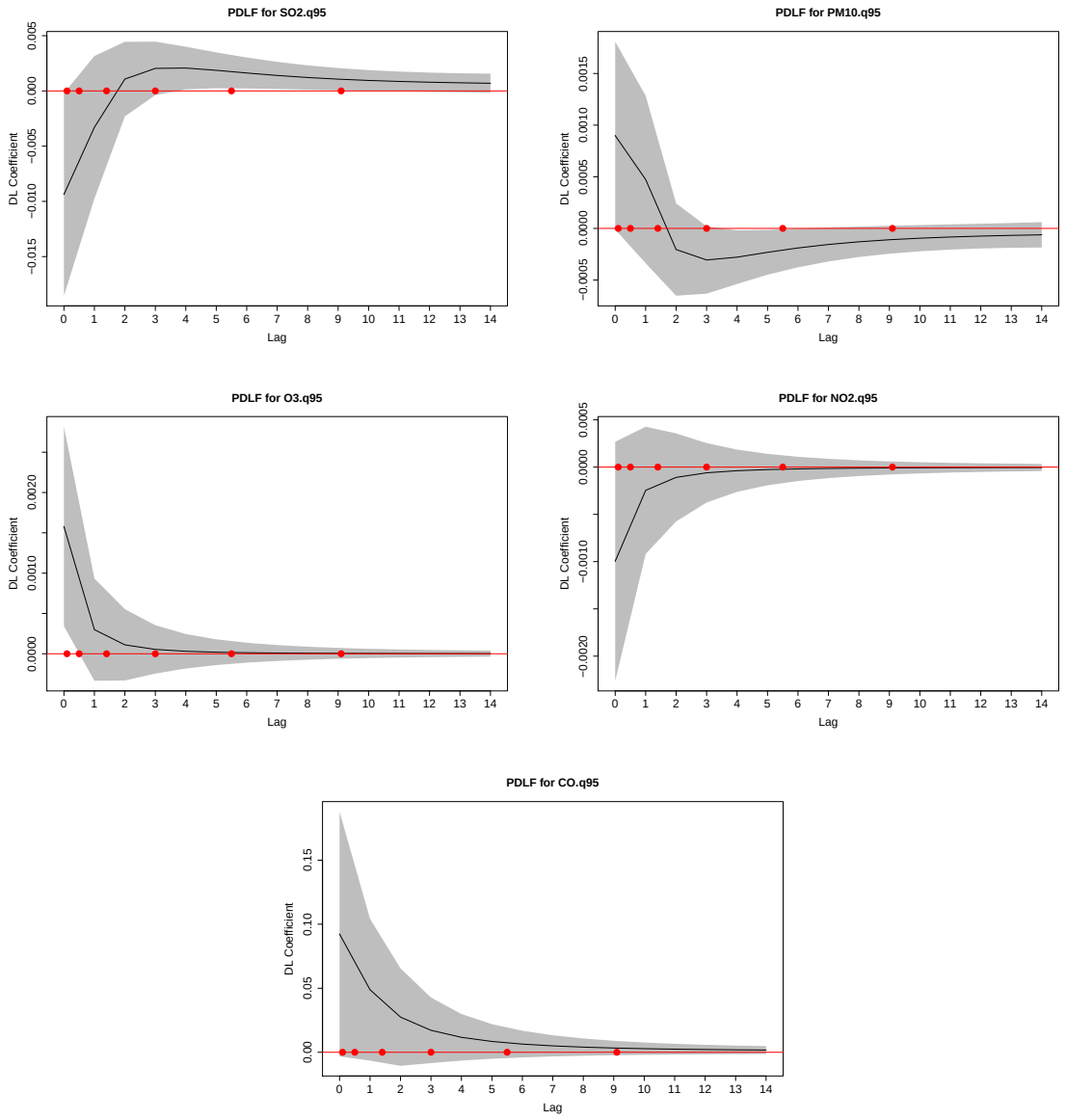


Abbildung G.16: PDLFs der Luftschadstoffe für die Call-Center-Daten in der Pilotregion München mit punktweisen 95%-Konfidenzbändern (grau) und Knoten der B-Spline-Basis (rot): SO2.q95 (oben links), PM10.q95 (oben rechts), O3.q95 (Mitte links), NO2.q95 (Mitte rechts) und CO.q95 (unten links)

Zu Abschnitt 4.5:

Parameter	Datenstruktur in R	Definition
formula	character	Formel des Regressionsmodells nach dem Schema „$y \sim u_1 + \ldots + u_p$“ mit manuell konstruierten Zeittrend- und Cutpoint-Variablen und ohne Kovariablen mit verzögertem Effekt (ohne Standardwert)
data	data.frame	Datensatz, der die Zielvariable und alle Kovariablen des Regressionsmodells (die Lags der Kovariablen müssen nicht enthalten sein) beinhaltet (ohne Standardwert)
family	family	Angenommene Verteilung für die $y_i \vert \boldsymbol{x}_i$ (für die vorliegenden Zähldaten können folgende Werte verwendet werden: “poisson” (Standardwert), “gaussian” für ein loglineares Modell (in diesem Fall muss die Zielgröße manuell logarithmiert werden), “quasipoisson”, “negbin”)
time_cov	character	Name der Variable, die die Zeit t definiert (zur Konstruktion der Lags) (ohne Standardwert)
group_cov	character (vector)	Name der Variable, die die Fallzahlgruppe definiert (im vorliegenden Fall Alter (und Geschlecht)) (ohne Standardwert)
almon	character (vector)	Name der Variable, für die eine DLF nach der Almon-Methode geschätzt werden soll (Standardwert ist NULL)
almon_L	integer (vector)	Maximales Lag für die Almon-DLF (Standardwert ist $L = 3$)
almon_d	integer (vector)	Polynomgrad für die Almon-DLF (Standardwert ist $d = \lfloor \sqrt{L} + 0.5 \rfloor$)
pdlf	character (vector)	Name der Variable, für die eine PDLF geschätzt werden soll (Standardwert ist NULL)
pdlf_L	integer (vector)	Maximales Lag für die PDLF (Standardwert ist $L = 14$)
pdlf_a	numeric (vector)	Tuningparameter für das Shrinkage der DLF gegen 0 (kein Shrinkage erfolgt für $a \leq 1$; je größer $a \geq 1$, desto stärker das Shrinkage; Standardwert ist $a = 2.3$)
pdlf_d	integer (vector)	Verwendete Differenzenordnung für die Penalisierung der DLF-Splinekoeffizienten (zulässige Werte sind $d = 1$, 2 und 3, wobei die Rauheit der DLF bei festem λ mit wachsendem d zunimmt; Standardwert ist $d = 2$)
sp	numeric (vector)	Wert des Penalisierungsparameters λ für die PDLF (je größer λ, desto glatter die resultierende DLF; wenn kein Wert angegeben wird (Standardwert ist NULL), wird λ durch Minimierung von GCV/UBRE bestimmt)
min_sp	numeric (vector)	Untere Grenze $\lambda_{\min}$ des Wertebereichs des Penalisierungsparameters λ (Standardwert ist NULL, das bedeutet $\lambda_{\min} = 0$)
ret_mod	logical	Wenn TRUE (Standardwert), wird eine Liste mit den folgenden Bestandteilen zurückgegeben: a) Regressionskoeffizienten $\hat{\delta}_0, \ldots, \hat{\delta}_p$ mit Standardfehlern und p-Werten b) Almon-DLF- und PDLF-Koeffizienten $\hat{\beta}_t, \ldots, \hat{\beta}_{t-L}$ c) erzeugtes Modellobjekt (zur Extraktion weiterer Parameter) d) Penalisierungsparameter λ für die PDLF
plot_dlf	logical	Wenn TRUE (Standardwert), werden die geschätzten DLFs mit punktweisen 95%-Konfidenzbändern geplottet
plot_ext	logical	Wenn FALSE (Standardwert), muss die Enter-Taste zur Anzeige des jeweils nächsten DLF-Plots gedrückt werden (die Option FALSE dient zur Ausgabe der Plots in externe Dokumente)
...	–	Weitere Argumente, die an die R-Funktion „gam“ übergeben werden (vgl. die R-Hilfe)

Tabelle G.17: Ein- und Ausgabewerte für die R-Funktion „lag_regress“

Zu Abschnitt 5.1:

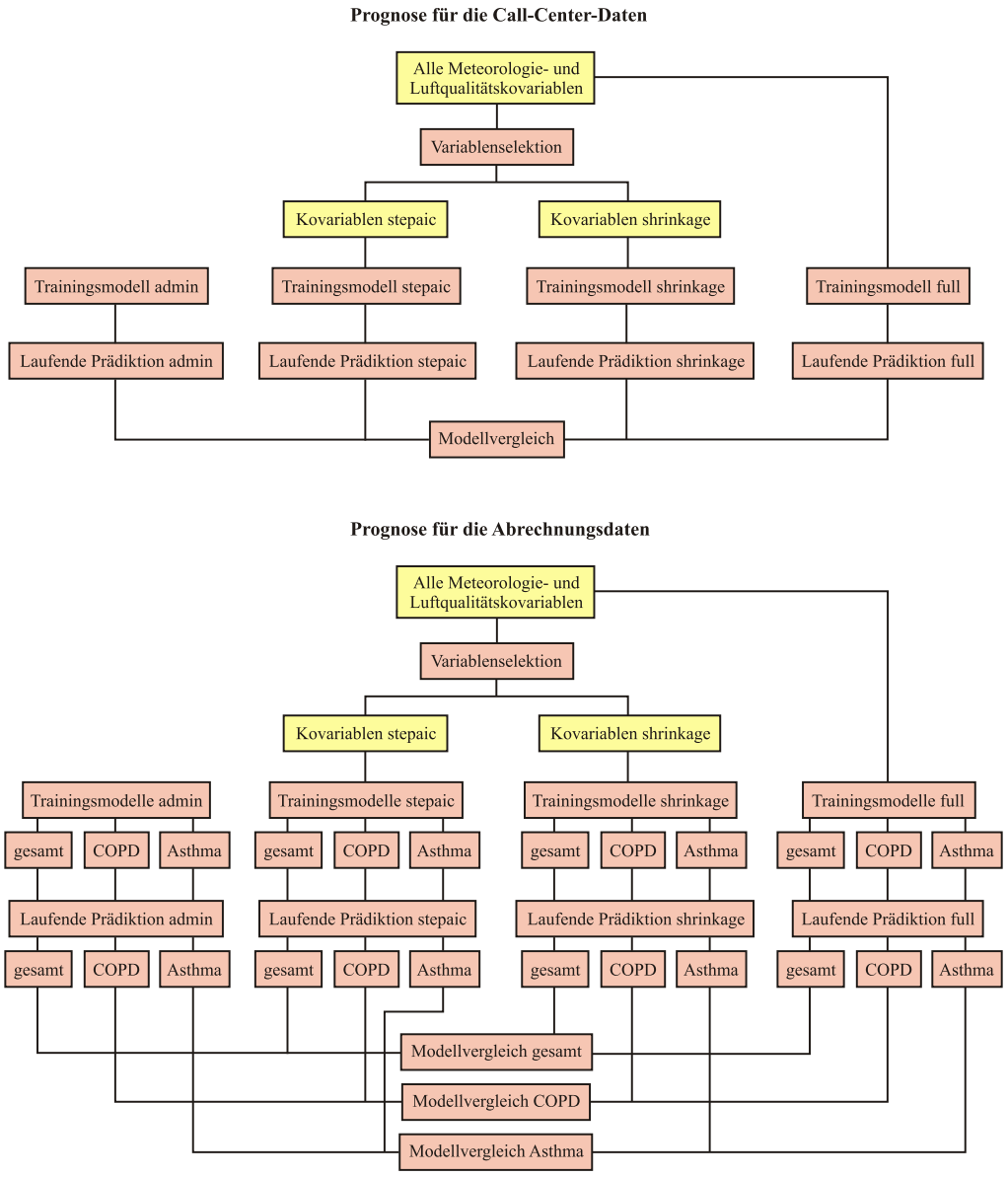


Abbildung G.17: Vorgehensweise bei der Prognose und Überblick über alle Modelle

	KVB-Call-Center-Daten			KVB-Abrechnungsdaten		
Kovariable	AIC-Selektion	Shrinkage	volles Modell	AIC-Selektion	Shrinkage	volles Modell
cp.max	–/–	–/–	H/L	H/L	H/L	H/L
lsp.max	–/–	H/L	H/L	H/L	H/L	H/L
lcc.ave	–/–	–/–	H/L	H/–	H/–	H/L
mcc.ave	–/–	–/–	H/L	H/L	H/L	H/L
q.ave	H/L	H/L	H/–	H/L	H/–	H/L
q.ave_lcp	–	–	H	–	H	H
q.ave_ucp	H	–	H	–	–	H
q.mmm	–	–	H	–	–	H
qnh.ave	H/L	H/L	H/L	H/L	H/L	H/L
qnh.ave_lcp	H	–	H	H	–	H
qnh.ave_ucp	–	–	H	H	–	H
qnh.mmm	–	–	H	–	–	H
sshf.ave	–/–	–/–	H/L	H/L	H/L	H/L
stressspd.max	H/L	H/L	H/L	H/–	–/–	H/L
tmt.ave	H/–	–/–	H/L	H/L	H/–	H/L
tmt.ave_lcp	H	–	H	H	H	H
tmt.ave_ucp	–	–	H	–	–	H
tmt.mmm	–	–	H	–	–	H
windspd.max	H/L	H/L	H/L	H/–	–/–	H/L
c.wdir	–	–	H	–	–	H
SO2.q95	H/L	H/L	H/L	H/L	–/–	H/L
PM10.q95	–/–	–/–	H/L	H/L	H/L	H/L
O3.q95	H/L	H/–	H/L	H/L	H/–	H/L
NO2.q95	H/L	H/L	H/L	H/L	H/–	H/L
CO.q95	H/–	–/–	H/L	H/L	H/–	H/L

Tabelle G.18: Kovariablenkonfiguration der Trainings- und Prognosemodelle für die Call-Center- und Abrechnungsdaten basierend auf der schrittweisen AIC-Selektion und den bayesianischen Shrinkage-Modellen: H = Haupteffekt, L = Lag-Effekt

Zu Abschnitt 5.2:

Kovariable	CC	AB (gesamt)	AB (COPD)	AB (Asthma)
Intercept	-68.5383/0.0000/0.0000	-0.0009/0.9991/0.9145	-0.0008/0.9992/0.9614	-0.0009/0.9991/0.8945
Deprivation	0.0185/1.0187/0.0000	0.0088/1.0088/0.0000	0.0172/1.0174/0.0000	0.0068/1.0068/0.0000
dow=Mo	(-0.3860)/0.6798/0.0000	(0.7696)/2.1589/0.0000	(0.5991)/1.8204/0.0000	(0.7347)/2.0848/0.0000
dow=Di	-0.5301/0.5886/0.0000	0.6259/1.8699/0.0000	0.4587/1.5821/0.0000	0.5942/1.8117/0.0000
dow=Mi	-0.0697/0.9326/0.0000	0.3867/1.4721/0.0000	0.2682/1.3076/0.0000	0.3446/1.4114/0.0000
dow=Do	-0.5125/0.5990/0.0000	0.7059/2.0256/0.0000	0.5242/1.6890/0.0000	0.6757/1.9654/0.0000
dow=Fr	-0.2108/0.8100/0.0000	0.6263/1.8707/0.0000	0.4516/1.5708/0.0000	0.5863/1.7974/0.0000
dow=Sa	0.8416/2.3201/0.0000	-1.4315/0.2390/0.0000	-1.0724/0.3422/0.0000	-1.3807/0.2514/0.0000
dow=So	0.8675/2.3809/0.0000	-1.6828/0.1858/0.0000	-1.2294/0.2925/0.0000	-1.5549/0.2112/0.0000
school=ja	0.1521/1.1643/0.0000	-0.5026/0.6050/0.0000	-0.4023/0.6688/0.0000	-0.4845/0.6160/0.0000
holiday=ja	1.0447/2.8424/0.0000	0.3418/1.4075/0.0000	0.2439/1.2763/0.0000	0.3205/1.3778/0.0000
bridge=ja	0.5776/1.7817/0.0000	0.0397/1.0405/0.0000	0.0283/1.0287/0.0000	0.0370/1.0377/0.0000
quartal=Anfang	-0.0501/0.9511/0.0000	0.0488/1.0500/0.0000	0.0341/1.0346/0.0480	0.0533/1.0547/0.0000
quartal=Ende	0.1601/1.1737/0.0000	-0.1696/0.8440/0.0000	-0.1560/0.8556/0.0000	-0.1650/0.8479/0.0000
sex=2	–	0.0087/1.0088/0.3193	-0.0970/0.9076/0.0000	0.1229/1.1307/0.0000
age=1	(-1.1516)/0.3161/0.0000	(-0.2197)/0.8027/0.0000	(-0.6497)/0.5222/0.0000	(0.0612)/1.0631/0.0000
age=2	-1.0444/0.3519/0.0000	-0.3811/0.6831/0.0000	-0.5638/0.5690/0.0000	-0.2484/0.7800/0.0000
age=3	-0.8216/0.4397/0.0000	-0.0587/0.9430/0.0000	0.1219/1.1296/0.0000	-0.0998/0.9050/0.0000
age=4	0.7375/2.0907/0.0000	0.6595/1.9338/0.0000	1.0916/2.9789/0.0000	0.2870/1.3324/0.0000
age=5	2.2801/9.7777/0.0000	–	–	–
sex=2,age=1	–	(-0.3059)/0.7364/0.0000	(-0.0587)/0.9430/0.0000	(-0.4263)/0.6529/0.0000
sex=2,age=2	–	-0.1741/0.8402/0.0000	-0.4546/0.6347/0.0000	0.0035/1.0036/0.6009
sex=2,age=3	–	0.0218/1.0220/0.0000	-0.0380/0.9628/0.0000	0.0939/1.0984/0.0000
sex=2,age=4	–	0.4582/1.5812/0.0000	0.5512/1.7354/0.0000	0.3289/1.3894/0.0000
q.ave_lcp (0.001)	0.045/–/0.0031	0.0000/–/1.0000	0.0000/–/1.0000	0.0000/–/1.0000
q.ave_ucp (0.001)	0.0063/–/0.4114	0.0000/–/0.9999	0.0000/–/0.9999	0.0000/–/0.9999
q.mmm (0.001)	-0.0119/0.9882/0.001	0.0000/1.0000/0.9998	0.0000/1.0000/0.9999	0.0000/1.0000/0.9998
qnh.ave_lcp (0.001)	0.0004/–/0.5759	0.0000/–/0.9978	0.0000/–/0.9975	0.0000/–/0.9980
qnh.ave_ucp (0.001)	-0.0035/–/0.0000	0.0000/–/0.9999	0.0000/–/1.0000	0.0000/–/0.9999
qnh.mmm (0.001)	-0.0067/0.9934/0.0000	0.0000/1.0000/1.0000	0.0000/1.0000/1.0000	0.0000/1.0000/0.9999
tmt.ave_lcp	-0.0255/–/0.0000	-0.0263/–/0.0027	-0.0178/–/0.3011	-0.0248/–/0.0003
tmt.ave_ucp	0.0044/–/0.3401	-0.0022/–/0.0000	-0.0039/–/0.0000	-0.0012/–/0.0000
tmt.mmm	-0.0061/0.9939/0.0000	0.0014/1.0014/0.0000	0.0015/1.0015/0.0000	0.0019/1.0019/0.0000
c.wdir=N	(-0.0231)/0.9772/0.0355	(0.1002)/1.1054/0.0000	(0.0841)/1.0877/0.0000	(0.0933)/1.0978/0.0000
c.wdir=NO	-0.0386/0.9621/0.0000	0.0773/1.0804/0.0000	0.0706/1.0732/0.0000	0.0727/1.0754/0.0000
c.wdir=NW	0.0151/1.0152/0.0888	0.0972/1.1020/0.0000	0.0840/1.0877/0.0000	0.0948/1.0995/0.0000
c.wdir=O	-0.0539/0.9475/0.0000	0.0132/1.0133/0.0000	-0.0064/0.9936/0.0000	0.0178/1.0179/0.0000
c.wdir=S	0.0257/1.0261/0.0011	-0.1186/0.8882/0.0000	-0.0943/0.9100/0.0000	-0.1191/0.8877/0.0000
c.wdir=SO	0.0718/1.0744/0.0000	-0.0773/0.9256/0.0000	-0.0757/0.9271/0.0000	-0.0719/0.9306/0.0000
c.wdir=SW	-0.0229/0.9773/0.0002	-0.0192/0.9810/0.0291	-0.0056/0.9944/0.7430	-0.0230/0.9772/0.0007
c.wdir=W	0.0260/1.0263/0.0000	-0.0729/0.9297/0.0000	-0.0567/0.9448/0.0000	-0.0646/0.9375/0.0000

Tabelle G.19: Parameterschätzer $\boldsymbol{\delta}$ der finalen Prognosemodelle mit vollem Kovariablensatz (full) für die Anzahl der Anrufe beim KVB-Call-Center sowie die Anzahl der Arztbesuche wg. Asthma bzw. COPD und gesamt: Schätzwert (roh/exponentiell transformiert) mit p-Wert (gelb: signifikante Ergebnisse)

Datensatz	x_{district}	x_{time}	x_{age}	x_{sex}	y	y^*_{t+1}	y^*_{t+2}	y^*_{t+3}
CC	9162	3.1.07	2	–	0	1.25 (1.06, 1.43)	1.24 (1.05, 1.43)	1.34 (1.18, 1.50)
	9162	3.1.07	4	–	5	3.67 (3.31, 4.04)	3.59 (3.22, 3.96)	3.14 (2.86, 3.43)
	9172	3.1.07	2	–	0	0.05 (0.04, 0.06)	0.05 (0.04, 0.06)	0.05 (0.04, 0.06)
	9172	3.1.07	4	–	0	0.23 (0.20, 0.25)	0.22 (0.20, 0.25)	0.19 (0.17, 0.22)
	9162	13.1.07	2	–	1	1.75 (1.46, 2.05)	1.67 (1.38, 1.97)	1.68 (1.38, 1.98)
	9162	13.1.07	4	–	9	5.76 (5.15, 6.38)	5.55 (4.93, 6.16)	5.77 (5.12, 6.42)
	9172	13.1.07	2	–	0	0.07 (0.05, 0.08)	0.07 (0.05, 0.08)	0.07 (0.05, 0.08)
	9172	13.1.07	4	–	0	0.36 (0.30, 0.41)	0.35 (0.29, 0.40)	0.36 (0.30, 0.42)
AB	9162	3.1.07	2	1	17	66.5 (64.6, 68.4)	63.9 (62.1, 65.7)	56.3 (54.4, 58.1)
(COPD)	9162	3.1.07	4	1	482	192 (187, 198)	184 (179, 189)	162 (157, 168)
	9162	3.1.07	2	2	23	39.7 (38.5, 40.8)	38.1 (37.0, 39.2)	32.9 (31.8, 34.0)
	9162	3.1.07	4	2	640	394 (383, 405)	377 (367, 388)	343 (332, 354)
	9172	3.1.07	2	1	4	4.70 (4.55, 4.85)	4.57 (4.42, 4.71)	4.41 (4.26, 4.55)
	9172	3.1.07	4	1	70	22.9 (22.3, 23.5)	22.2 (21.6, 22.8)	21.5 (20.9, 22.1)
	9172	3.1.07	2	2	2	2.73 (2.63, 2.83)	2.66 (2.56, 2.75)	2.51 (2.42, 2.61)
	9172	3.1.07	4	2	103	45.9 (44.7, 47.2)	44.4 (43.2, 45.6)	44.2 (43.0, 45.5)
	9162	13.1.07	2	1	2	24.0 (23.5, 24.4)	23.9 (23.4, 24.3)	23.8 (23.3, 24.2)
	9162	13.1.07	4	1	41	69.4 (68.1, 70.7)	68.9 (67.6, 70.2)	68.4 (67.1, 69.7)
	9162	13.1.07	2	2	10	14.3 (14.0, 14.5)	14.2 (13.9, 14.5)	14.1 (13.9, 14.4)
	9162	13.1.07	4	2	47	143 (140, 146)	142 (139, 145)	141 (138, 143)
	9172	13.1.07	2	1	1	1.85 (1.78, 1.93)	1.84 (1.76, 1.91)	1.83 (1.76, 1.90)
	9172	13.1.07	4	1	6	9.05 (8.79, 9.31)	8.94 (8.68, 9.20)	8.88 (8.62, 9.14)
	9172	13.1.07	2	2	0	1.07 (1.02, 1.13)	1.06 (1.01, 1.12)	1.06 (1.01, 1.12)
	9172	13.1.07	4	2	4	18.2 (17.7, 18.7)	18.0 (17.5, 18.5)	17.9 (17.4, 18.4)
AB	9162	3.1.2007	2	1	83	143 (139, 146)	137 (133, 140)	117 (114, 121)
(Asthma)	9162	3.1.2007	4	1	228	136 (133, 139)	130 (127, 133)	112 (109, 115)
	9162	3.1.2007	2	2	176	161 (157, 165)	153 (150, 157)	131 (128, 135)
	9162	3.1.2007	4	2	477	290 (283, 297)	277 (270, 283)	241 (234, 247)
	9172	3.1.2007	2	1	5	9.38 (9.14, 9.62)	9.06 (8.83, 9.29)	8.82 (8.59, 9.05)
	9172	3.1.2007	4	1	20	15.1 (14.7, 15.4)	14.5 (14.2, 14.9)	14.1 (13.8, 14.5)
	9172	3.1.2007	2	2	8	10.3 (10.0, 10.5)	9.9 (9.7, 10.2)	9.6 (9.4, 9.9)
	9172	3.1.2007	4	2	38	31.3 (30.6, 32.1)	30.2 (29.5, 31.0)	29.8 (29.1, 30.5)
	9162	13.1.2007	2	1	14	38.2 (37.6, 38.8)	38.0 (37.4, 38.6)	37.7 (37.1, 38.3)
	9162	13.1.2007	4	1	16	36.4 (35.8, 37.0)	36.1 (35.6, 36.7)	35.9 (35.3, 36.5)
	9162	13.1.2007	2	2	21	43.0 (42.3, 43.7)	42.7 (42.0, 43.4)	42.4 (41.7, 43.1)
	9162	13.1.2007	4	2	34	77.9 (76.6, 79.1)	77.2 (75.9, 78.5)	76.6 (75.3, 77.9)
	9172	13.1.2007	2	1	2	2.69 (2.61, 2.77)	2.65 (2.57, 2.73)	2.63 (2.55, 2.72)
	9172	13.1.2007	4	1	3	4.32 (4.20, 4.44)	4.25 (4.13, 4.37)	4.22 (4.10, 4.34)
	9172	13.1.2007	2	2	0	2.95 (2.86, 3.04)	2.90 (2.81, 2.99)	2.88 (2.79, 2.97)
	9172	13.1.2007	4	2	2	9.03 (8.80, 9.25)	8.87 (8.65, 9.10)	8.81 (8.58, 9.04)

Tabelle G.20: 1-Tages-, 2-Tages- und 3-Tagesprognosen aus den auf schrittweiser Variablenselektion basierenden Modellen (stepaic) für die Fallzahlen (Anrufe beim Call-Center, Arztbesuche wg. COPD und wg. Asthma) in den Landkreisen München (9162) und Berchtesgadener Land (9172) in den Altersgruppen 2 und 4 am 3. und 13. Januar 2007 mit zugehörigen Prognoseintervallen (in Klammern) im Vergleich zu den tatsächlich beobachteten Werten

Zu Abschnitt 5.3:

Datensatz	x_{age}	x_{sex}	MPSE($\boldsymbol{y}^*_{t+1}$)	$\Delta\boldsymbol{y}^*_{t+1}$
CC	1	–	0.0619	0.0027
	2	–	0.1582	0.0059
	3	–	0.1950	0.0035
	4	–	0.5236	0.0060
	5	–	0.4855	0.0033
AB (COPD)	1	1	19.41	1.15
	1	2	13.48	1.43
	2	1	65.48	3.96
	2	2	11.78	0.54
	3	1	111.84	-2.25
	3	2	191.10	-4.29
	4	1	5452.86	-44.48
	4	2	2611.89	-14.72
AB (Asthma)	1	1	138.09	-0.77
	1	2	69.00	0.42
	2	1	162.95	2.46
	2	2	170.23	-2.00
	3	1	178.63	0.03
	3	2	970.74	-9.59
	4	1	554.51	-9.89
	4	2	1366.13	-7.25

Tabelle G.21: MPSEs und mittlere Differenzen zwischen wahren und prädiktierten Fallzahlen (Anrufe beim Call-Center, Arztbesuche wg. COPD und wg. Asthma) in den (Geschlechts- und) Altersgruppen für die auf der schrittweisen AIC-Selektion (stepaic) beruhenden Modelle

		CC		AB (COPD)		AB (Asthma)	
Kovariable	Level	MPSE($\boldsymbol{y}^*_{t+1}$)	$\Delta\boldsymbol{y}^*_{t+1}$	MPSE($\boldsymbol{y}^*_{t+1}$)	$\Delta\boldsymbol{y}^*_{t+1}$	MPSE($\boldsymbol{y}^*_{t+1}$)	$\Delta\boldsymbol{y}^*_{t+1}$
dow	Mo	0.2908	0.0053	2325.09	-14.33	1029.60	-8.29
	Di	0.2750	0.0004	1878.24	-13.93	728.85	-7.53
	Mi	0.2798	0.0038	932.50	-8.06	379.31	-2.45
	Do	0.2720	0.0050	1529.15	-11.01	695.34	-4.79
	Fr	0.2894	0.0076	663.83	-7.19	285.30	-2.28
	Sa	0.2918	0.0026	21.03	1.89	8.68	1.32
	So	0.2951	0.0054	43.94	1.43	20.80	0.83
school	nein	0.2843	0.0044	1177.55	-8.39	471.96	-3.70
	ja	0.2860	0.0039	793.58	-4.94	404.59	-2.48
holiday	nein	0.2854	0.0045	1069.33	-8.15	430.83	-4.24
	ja	0.2682	-0.0005	799.75	14.74	1005.08	21.40
bridge	nein	0.2854	0.0042	1066.04	-7.43	451.76	-3.44
	ja	0.2373	0.0142	682.48	-1.11	422.72	3.31
quartal	Mitte	0.2889	0.0044	995.60	-7.27	417.89	-3.23
	Anfang	0.2653	0.0047	2101.94	-10.65	956.73	-6.28
	Ende	0.2598	0.0021	725.26	-4.67	314.36	-1.38

Tabelle G.22: MPSEs und mittlere Differenzen zwischen wahren und prädiktierten Fallzahlen (Anrufe beim Call-Center, Arztbesuche wg. COPD und wg. Asthma) auf den Faktorstufen der administrativen Kovariablen für die auf der schrittweisen AIC-Selektion (stepaic) beruhenden Modelle

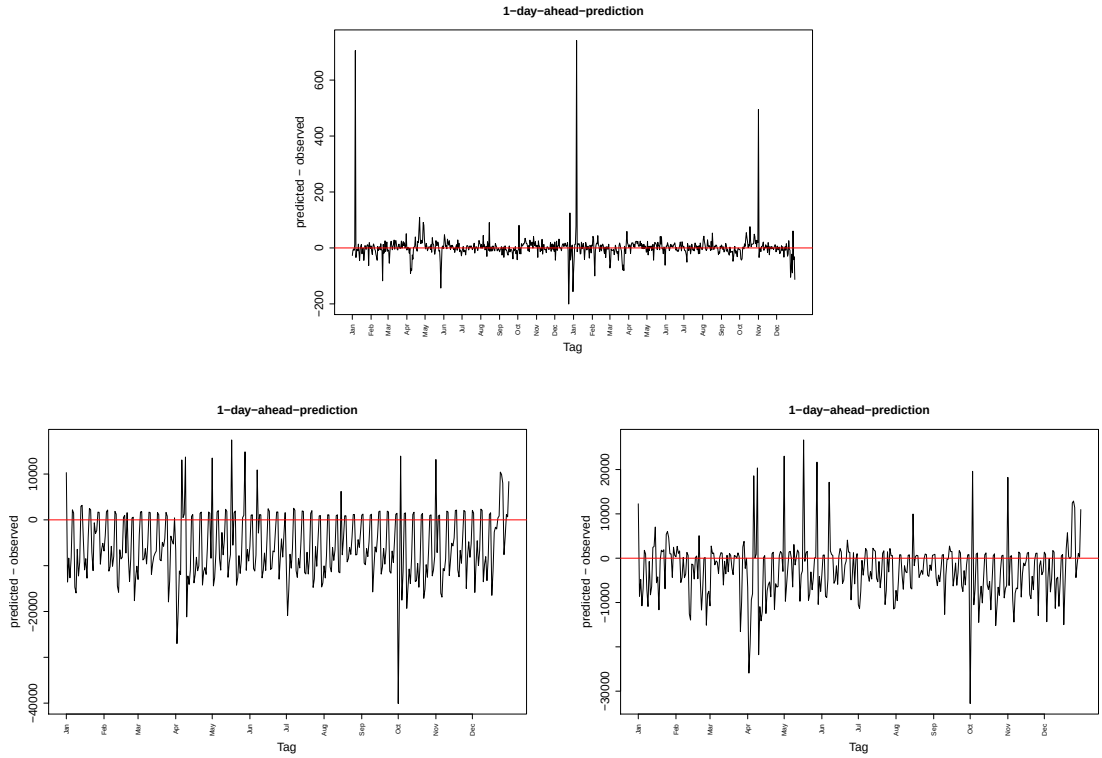


Abbildung G.18: Zeitlicher Verlauf der Differenz zwischen prädiktierten und beobachteten Werten (aufsummiert über alle Faktorstufen der designbedingten Variablen und Landkreise) für die auf schrittweiser AIC-Selektion (stepaic) beruhenden Prognosemodelle: Anrufe beim KVB-Call-Center (oben) und Arztbesuche wg. COPD (links unten) bzw. Asthma (rechts unten)

Datensatz	Landkreis	MPSE($\boldsymbol{y}^*_{t+1}$)	$\Delta\boldsymbol{y}^*_{t+1}$
CC	Landeshauptstadt München	6.9111	-0.6121
	Stadt Nürnberg	1.5807	0.1131
	Landkreis München	0.6951	-0.0374
	Stadt Augsburg	0.5758	0.2744
	Stadt Passau	0.4172	-0.0249
AB (COPD)	Landeshauptstadt München	32638.17	-55.19
	Stadt Nürnberg	15401.99	-44.02
	Landkreis Ansbach	1815.29	-13.25
	Landkreis Würzburg	1761.95	-17.47
	Landkreis Main-Spessart	1436.34	-15.12
AB (Asthma)	Landeshauptstadt München	21958.10	-43.30
	Stadt Nürnberg	3350.40	-8.57
	Landkreis Fürth	1050.05	-16.75
	Stadt Augsburg	904.68	-7.98
	Landkreis München	744.54	-4.69

Tabelle G.23: Höchste MPSEs und zugehörige mittlere Differenzen zwischen wahren und prädiktierten Fallzahlen (Anrufe beim Call-Center und Arztbesuche wg. COPD bzw. Asthma) in den Landkreisen für die auf der schrittweisen AIC-Selektion (stepaic) beruhenden Modelle

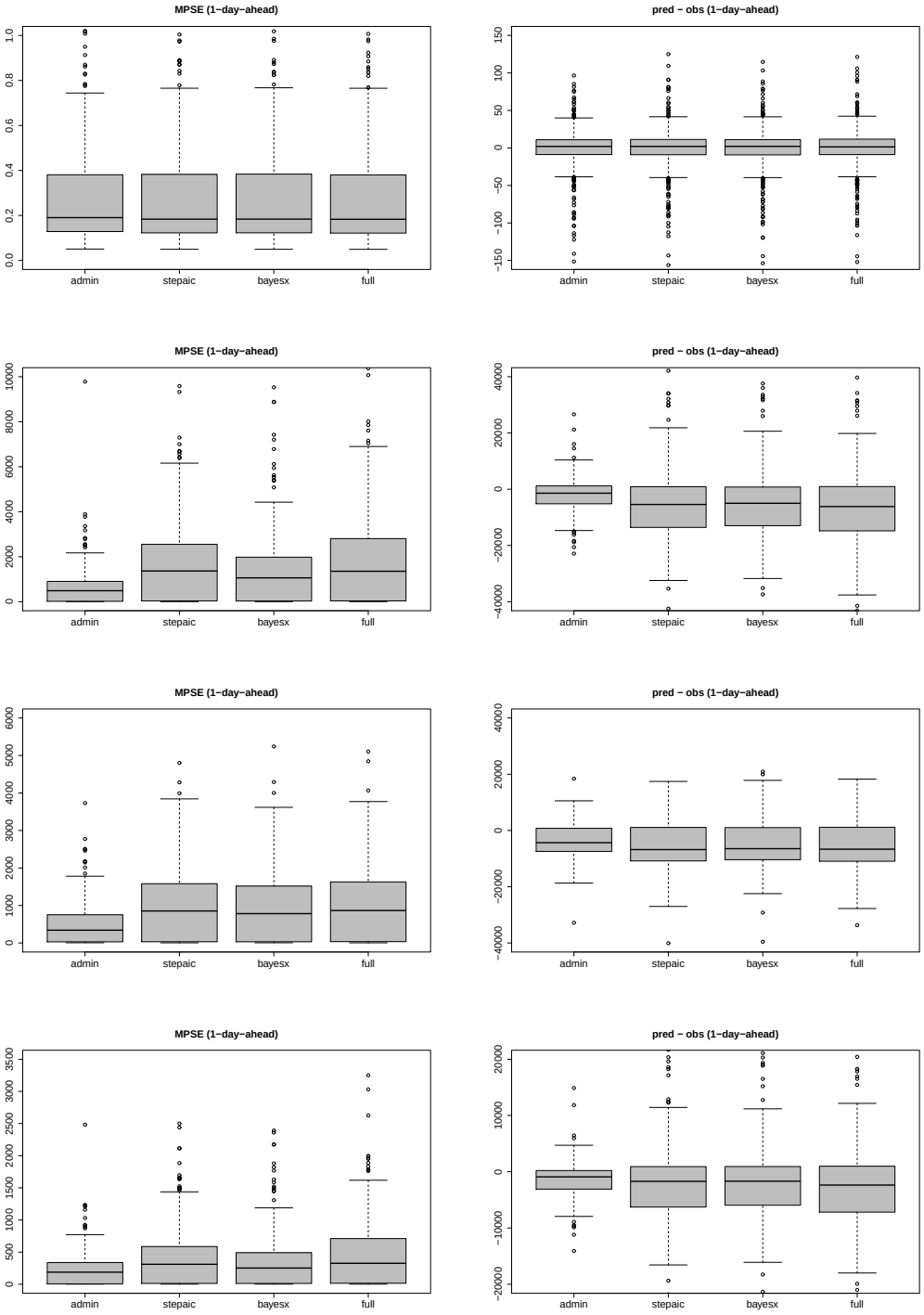


Abbildung G.19: Boxplots der datumspezifischen MPSEs (links) und der Differenzen zwischen den Tagessummen von prädiktierten und von beobachteten Fallzahlen (rechts) für die Anrufe beim KVB-Call-Center (1. Reihe) sowie für die Arztbesuche gesamt (2. Reihe) und separat wg. COPD (3. Reihe) bzw. Asthma (4. Reihe)

Zu Abschnitt 5.4:

Datensatz	x_{district}	x_{time}	x_{age}	x_{sex}	y	y^*_{t+1}	HI_1	y^*_{t+2}	HI_2	y^*_{t+3}	HI_3
CC	9162	3.1.07	2	–	0	1.16	2	1.16	2	1.23	2
	9162	3.1.07	4	–	5	3.37	0	3.34	0	2.86	0
	9172	3.1.07	2	–	0	0.07	1	0.07	1	0.07	1
	9172	3.1.07	4	–	0	0.32	0	0.32	0	0.27	0
	9162	13.1.07	2	–	1	1.59	1	1.56	1	1.57	1
	9162	13.1.07	4	–	9	5.12	0	5.08	0	5.30	0
	9172	13.1.07	2	–	0	0.10	0	0.09	0	0.09	0
	9172	13.1.07	4	–	0	0.49	0	0.49	0	0.51	0
AB	9162	3.1.07	2	1	17	59.00	1	60.32	1	61.46	1
(COPD)	9162	3.1.07	4	1	482	342.92	0	306.46	0	435.99	1
	9162	3.1.07	2	2	23	67.00	1	68.50	1	69.82	1
	9162	3.1.07	4	2	640	332.33	1	296.98	0	422.69	1
	9172	3.1.07	2	1	4	2.71	0	2.79	0	2.86	0
	9172	3.1.07	4	1	70	34.83	0	31.03	0	44.52	0
	9172	3.1.07	2	2	2	3.10	0	3.19	0	3.27	0
	9172	3.1.07	4	2	103	32.94	0	29.34	0	42.13	1
	9162	13.1.07	2	1	2	11.66	1	11.91	1	12.30	1
	9162	13.1.07	4	1	41	108.55	1	104.19	1	98.57	1
	9162	13.1.07	2	2	10	13.37	1	13.65	1	14.09	1
	9162	13.1.07	4	2	47	105.23	1	101.00	1	95.53	1
	9172	13.1.07	2	1	1	-0.22	-2	-0.20	-2	-0.18	-2
	9172	13.1.07	4	1	6	10.41	1	9.96	0	9.37	0
	9172	13.1.07	2	2	0	-0.13	-2	-0.12	-2	-0.09	-2
	9172	13.1.07	4	2	4	9.81	1	9.38	1	8.83	1
AB	9162	3.1.07	2	1	83	172.41	0	169.95	-1	197.07	0
(Asthma)	9162	3.1.07	4	1	228	209.26	-1	196.88	-1	253.93	-1
	9162	3.1.07	2	2	176	227.59	-1	224.31	-1	260.14	0
	9162	3.1.07	4	2	477	312.66	-1	294.14	-1	379.38	-1
	9172	3.1.07	2	1	5	9.39	-1	9.24	-1	10.86	-1
	9172	3.1.07	4	1	20	20.22	-1	18.97	-1	24.73	-1
	9172	3.1.07	2	2	8	12.34	-1	12.15	-1	14.24	-1
	9172	3.1.07	4	2	38	29.94	-1	28.11	-1	36.51	-1
	9162	13.1.07	2	1	14	25.78	0	25.79	0	25.76	0
	9162	13.1.07	4	1	16	36.99	-1	36.39	-1	35.84	-1
	9162	13.1.07	2	2	21	34.35	0	34.34	0	34.29	0
	9162	13.1.07	4	2	34	55.75	-1	54.85	-1	54.00	-1
	9172	13.1.07	2	1	2	0.60	-1	0.60	-1	0.60	-1
	9172	13.1.07	4	1	3	2.83	-1	2.77	-1	2.72	-1
	9172	13.1.07	2	2	0	1.06	-1	1.06	-1	1.06	-1
	9172	13.1.07	4	2	2	4.60	-1	4.51	-1	4.42	-1

Tabelle G.24: 1-Tages-, 2-Tages- und 3-Tagesprognosen aus den Modellen ohne Meteorologie- und Luftqualitätskovariablen (admin) für die Fallzahlen (Anrufe beim Call-Center, Arztbesuche wg. COPD und wg. Asthma) in den Landkreisen München (9162) und Berchtesgadener Land (9172) in den Altersgruppen 2 und 4 am 3. und 13. Januar 2007 mit zugehörigen Gesundheitsindizes (HI)

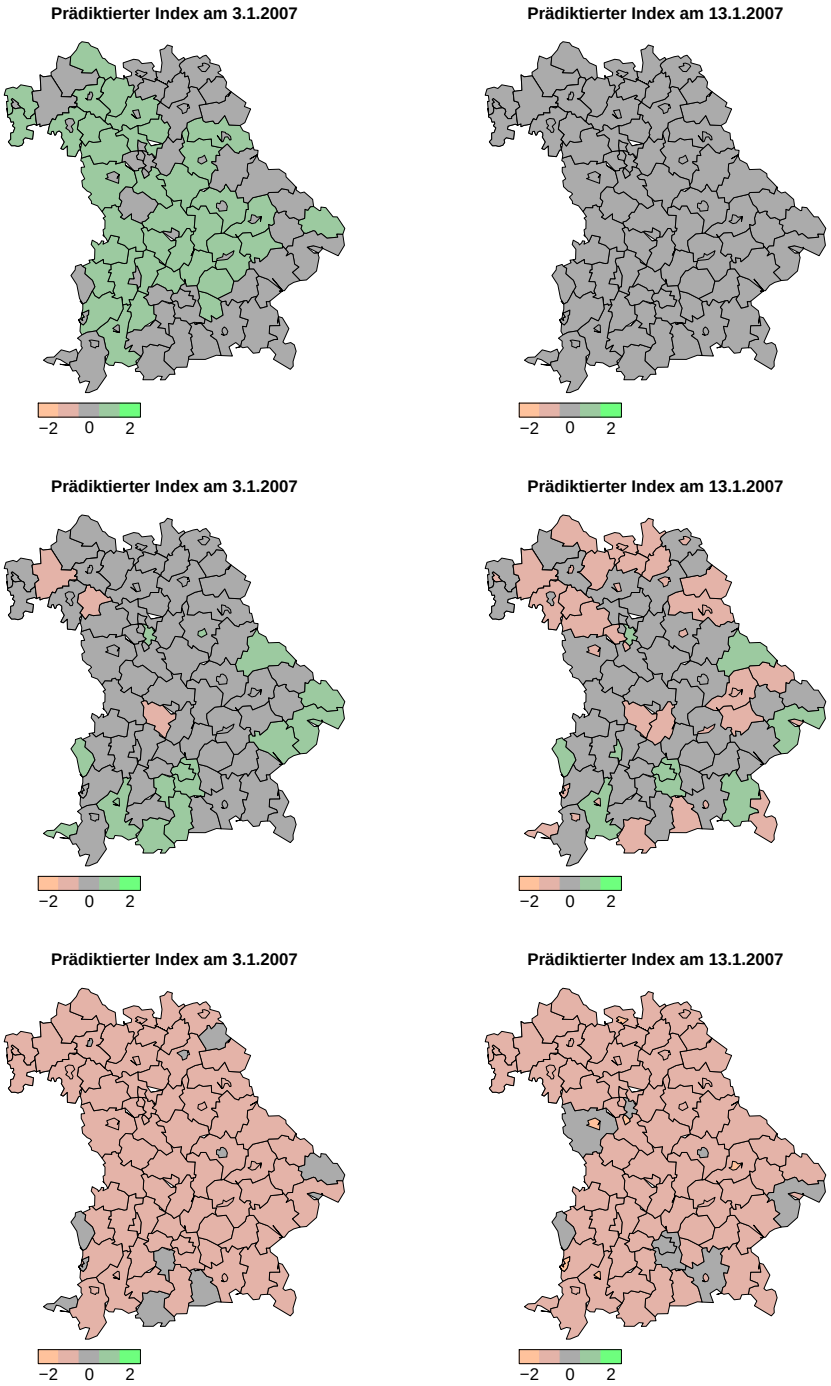


Abbildung G.20: Landkreisspezifische Gesundheitsindizes HI_{ges} (Median der Indizes aller Alters- und Geschlechtsgruppen) am 3. (links) und 13. Januar (rechts) 2007 basierend auf den Prognosemodellen nur mit administrativen Kovariablen für die Anrufe beim KVB-Call-Center (oben) und die Arztbesuche wg. COPD (Mitte) bzw. Asthma (unten) (-2: minimale Gefährdung, ..., 2: maximale Gefährdung)

H R-Code

H.1 Beispielhafter R-Code zur Darstellung der (altersspezifischen) Zeittrends

```
library(splines)

# Daten einlesen
frame <- read.table("../../Daten_Abrechnung_München_a).txt",header=T)

# Definition der Faktoren
frame$Alter <- as.factor(frame$Alter)
frame$Geschlecht <- as.factor(frame$Geschlecht)
frame$dow <- as.factor(frame$dow)
frame$school <- as.factor(frame$school)
frame$holiday <- as.factor(frame$holiday)
frame$bridge <- as.factor(frame$bridge)
frame$quartal <- as.factor(frame$quartal)

# Mittelwertcodierung
contrasts(frame$Alter)[1,] <- -1
contrasts(frame$dow)[1,] <- -1
contrasts(frame$c.wdir.ave)[1,] <- -1

# Sortieren nach Datum, Geschlecht und Alter
frame <- frame[order(frame$Datum,frame$Geschlecht,frame$Alter),]
row.names(frame) <- 1:nrow(frame)

# Modell für Anzahl_gesamt mit 8 Basisfkt.

admin <- "offset(log(Einwohner)) + dow + school + holiday + bridge +
         quartal + Geschlecht + Geschlecht:Alter + Alter*ns(time,8)"

cp <- "cp.max"
lsp <- "lsp.max"
lcc <- "lcc.ave"
mcc <- "mcc.ave"
q <- "q.ave + q.mmm + q.ave_lcp + q.ave_ucp"
qnh <- "qnh.ave + qnh.mmm + qnh.ave_lcp + qnh.ave_ucp"
sshf <- "sshf.ave"
```

```
stressspd <- "stressspd.max"
tmt <- "tmt.ave + tmt.mmm + tmt.ave_lcp + tmt.ave_ucp"
windspd <- "windspd.max"
c.wdir.ave <- "c.wdir.ave"

SO2_tr <- "SO2_tr.q95"
PM10_tr <- "PM10_tr.q95"
O3_tr <- "O3_tr.q95"
NO2_tr <- "NO2_tr.q95"
CO_tr <- "CO_tr.q95"

formula <- as.formula(paste("log(Anzahl_gesamt+1) ~",paste(admin,cp,
           lsp,lcc,mcc,q,qnh,sshf,stressspd,tmt,windspd,c.wdir.ave,
           SO2_tr,PM10_tr,O3_tr,NO2_tr,CO_tr,sep=" + "),sep=""))

model <- glm(formula,family=gaussian,data=frame,na.action=na.omit)

Sigma <- summary(model)$cov.scaled

# Vergleiche lowess durch die wahren Werte mit dem durchs Modell
# gefitteten B-Spline

mat <- matrix(0,2,4)
mat[1,] <- c(0,1,0.5,1)
mat[2,] <- c(0,1,0,0.75)

frame_agg <- aggregate(frame$Anzahl_gesamt,
             by=list(frame$Datum,frame$Geschlecht),FUN=sum)
names(frame_agg) <- c("Datum","Geschlecht","Anzahl_gesamt")

# Darstellung von f(t)

split.screen(mat)

screen(1)

male <- lowess(1:730,log(frame_agg$Anzahl_gesamt[frame_agg$
        Geschlecht==0]+1),f=0.1)
female <- lowess(1:730,log(frame_agg$Anzahl_gesamt[frame_agg$
        Geschlecht==1]+1),f=0.1)

plot(male,type="l",cex.main=1.5,lwd=2,xlab="",
```

```
     ylab="Log. Rohdaten (lowess)",
     main="Alle Altersstufen - gesamt (2006-2007)",xaxt="n",col=4,
     ylim=c(min(male$y,female$y),max(male$y,female$y)))
lines(lwd=2,female,col=2)

screen(2)

mean <- ns(frame$time[frame$Alter==1&frame$Geschlecht==0],8)%*%
        coef(model)[17:24]
se <- sqrt(diag(ns(frame$time[frame$Alter==1&frame$
      Geschlecht==0],8)%*%summary(model)$cov.scaled[17:24,17:24]%*%
      t(ns(frame$time[frame$Alter==1&frame$Geschlecht==0],8))))
lower <- mean - qnorm(0.975)*se
upper <- mean + qnorm(0.975)*se

mean2 <- mean
lower2 <- lower
upper2 <- upper

mean3 <- mean + coef(model)[13]
lower3 <- lower + coef(model)[13] - qnorm(0.975)*sqrt(Sigma[13,13])
upper3 <- upper + coef(model)[13] + qnorm(0.975)*sqrt(Sigma[13,13])

plot(1:730,mean,type="l",cex.main=1.5,lwd=2,xlab="Tag",
     ylab="B-Spline (8 Basisfkt.) mit/ohne Shift",main="",
     ylim=c(min(lower,lower2,lower3),max(upper,upper2,upper3)))
polygon(x=c(1:730,730:1,1),y=c(upper,lower[730:1],upper[1]),
        col="grey",border=NA)
lines(lwd=2,1:730,mean)
abline(lwd=2,v=seq(1,730,len=9),lty=2)
abline(lwd=2,h=0,col=2)
lines(lwd=2,1:730,mean2,col=4)
lines(lwd=2,1:730,lower2,col=4,lty=2)
lines(lwd=2,1:730,upper2,col=4,lty=2)
lines(lwd=2,1:730,mean3,col=2)
lines(lwd=2,1:730,lower3,col=2,lty=2)
lines(lwd=2,1:730,upper3,col=2,lty=2)

close.screen(all = TRUE)

# Darstellung von f_2(t)
```

```
split.screen(mat)

screen(1)

male <- lowess(1:730,log(frame$Anzahl_gesamt[frame$Alter==2&frame$
        Geschlecht==0]+1),f=0.1)
female <- lowess(1:730,log(frame$Anzahl_gesamt[frame$Alter==2&frame$
          Geschlecht==1]+1),f=0.1)
plot(male,type="l",cex.main=1.5,lwd=2,xlab="",
     ylab="Log. Rohdaten (lowess)",
     main="21 bis 40 Jahre - gesamt (2006-2007)",xaxt="n",col=4,
     ylim=c(min(male$y,female$y),max(male$y,female$y)))
lines(lwd=2,female,col=2)

screen(2)

mean <- ns(frame$time[frame$Alter==2&frame$Geschlecht==0],8)%*%
        (coef(model)[17:24]+coef(model)[59+(0:7*3)])
cov_betasum <- Sigma[17:24,17:24] + Sigma[59+(0:7*3),59+(0:7*3)] -
               2*(Sigma[17:24,59+(0:7*3)])
se <- sqrt(diag(ns(frame$time[frame$Alter==2&frame$
      Geschlecht==0],8)%*%cov_betasum%*%
      t(ns(frame$time[frame$Alter==2&frame$Geschlecht==0],8))))
lower <- mean - qnorm(0.975)*se
upper <- mean + qnorm(0.975)*se

mean2 <- mean + coef(model)[14]
lower2 <- lower + coef(model)[14] - qnorm(0.975)*sqrt(Sigma[14,14])
upper2 <- upper + coef(model)[14] + qnorm(0.975)*sqrt(Sigma[14,14])

var_sex_1_age_2 <- calc_varsum(Sigma[c(14,13,56),c(14,13,56)])
mean3 <- mean + coef(model)[14] + coef(model)[13] + coef(model)[56]
lower3 <- lower + coef(model)[14] + coef(model)[13] +
          coef(model)[56] - qnorm(0.975)*sqrt(var_sex_1_age_2)
upper3 <- upper + coef(model)[14] + coef(model)[13] +
          coef(model)[56] + qnorm(0.975)*sqrt(var_sex_1_age_2)

plot(1:730,mean,type="l",cex.main=1.5,lwd=2,xlab="Tag",
     ylab="B-Spline (8 Basisfkt.) mit/ohne Shift",main="",
     ylim=c(min(lower,lower2,lower3),max(upper,upper2,upper3)))
polygon(x=c(1:730,730:1,1),y=c(upper,lower[730:1],upper[1]),
        col="grey",border=NA)
```

```
lines(lwd=2,1:730,mean)
abline(lwd=2,v=seq(1,730,len=9),lty=2)
abline(lwd=2,h=0,col=2)
lines(lwd=2,1:730,mean2,col=4)
lines(lwd=2,1:730,lower2,col=4,lty=2)
lines(lwd=2,1:730,upper2,col=4,lty=2)
lines(lwd=2,1:730,mean3,col=2)
lines(lwd=2,1:730,lower3,col=2,lty=2)
lines(lwd=2,1:730,upper3,col=2,lty=2)

close.screen(all = TRUE)
```

H.2 Beispielhafter R-Code zur AIC-Selektion von Bruchpunkten

```
# Identifikation von Bruchpunkten durch AIC-Selektion aus
# verschiedenen Kandidaten-Werten

lower_list <- list()
upper_list <- list()

unit <- c(0.001,1,1)

for (k in 11:13){
lower_candidate <- quantile(model_frame[,k],
                   probs=seq(0.05,0.20,len=16))
upper_candidate <- quantile(model_frame[,k],
                   probs=seq(0.80,0.95,len=16))

lower_frame <- data.frame(quantile=names(lower_candidate))
upper_frame <- data.frame(quantile=names(upper_candidate))

lower_frame$value <- lower_candidate
upper_frame$value <- upper_candidate

lower_frame$AIC <- lower_frame$pval <- lower_frame$beta <-
numeric(16)
upper_frame$AIC <- upper_frame$pval <- upper_frame$beta <-
numeric(16)

model_frame$V32 <- model_frame$V31 <- model_frame$V30 <-
model_frame$V29 <- model_frame$V28 <- model_frame$V27 <-
model_frame$V26 <- model_frame$V25 <- model_frame$V24 <-
model_frame$V23 <- model_frame$V22 <- model_frame$V21 <-
model_frame$V20 <- model_frame$V19 <- model_frame$V18 <-
model_frame$V17 <- model_frame$V16 <- model_frame$V15 <-
model_frame$V14 <- model_frame$V13 <- model_frame$V12 <-
model_frame$V11 <- model_frame$V10 <- model_frame$V9 <-
model_frame$V8 <- model_frame$V7 <- model_frame$V6 <-
model_frame$V5 <- model_frame$V4 <- model_frame$V3 <-
model_frame$V2 <- model_frame$V1 <- numeric(nrow(model_frame))

for (l in 1:16){
```

```
model_frame[,ncol(model_frame)-32+l] <- I(model_frame[,k]<
lower_candidate[l])*(model_frame[,k]-lower_candidate[l])
model_frame[,ncol(model_frame)-16+l] <- I(model_frame[,k]>
upper_candidate[l])* (model_frame[,k]-upper_candidate[l])

lower_formula <- as.formula(paste("Anzahl~offset(log(Einwohner))+
dow+school+holiday+bridge+quartal+Alter*ns(time,12)+",
names(model_frame)[ncol(model_frame)-32+l],sep=""))
upper_formula <- as.formula(paste("Anzahl~offset(log(Einwohner))+
dow+school+holiday+bridge+quartal+Alter*ns(time,12)+",
names(model_frame)[ncol(model_frame)-16+l],sep=""))

lower_model <- glm(lower_formula,data=model_frame,na.action=na.omit,
               family=poisson)
upper_model <- glm(upper_formula,data=model_frame,na.action=na.omit,
               family=poisson)

lower_frame$beta[l] <- coef(lower_model)[29]*unit[k-10]
upper_frame$beta[l] <- coef(upper_model)[29]*unit[k-10]

lower_zvalue <- coef(lower_model)[29]/sqrt(diag(summary(lower_model)$
                cov.scaled)[29])
upper_zvalue <- coef(upper_model)[29]/sqrt(diag(summary(upper_model)$
                cov.scaled)[29])

lower_frame$pval[l] <- 2*(1-pnorm(abs(lower_zvalue)))
upper_frame$pval[l] <- 2*(1-pnorm(abs(upper_zvalue)))

lower_frame$AIC[l] <- extractAIC(lower_model)[2]
upper_frame$AIC[l] <- extractAIC(upper_model)[2]
}

lower_list[[k-10]] <- lower_frame; names(lower_list)[k-10] <-
names(model_frame)[k]
upper_list[[k-10]] <- upper_frame; names(upper_list)[k-10] <-
names(model_frame)[k]

model_frame <- model_frame[,1:18]
}

# Darstellung der besten Cutpoint-Modelle
```

```
for (k in 11:13){
lower_index <- which(lower_list[[k-10]]$AIC==
               min(lower_list[[k-10]]$AIC))
lower_beta <- lower_list[[k-10]]$beta[lower_index]
lower_cp <- lower_list[[k-10]]$value[lower_index]
lower_quantile <- lower_list[[k-10]]$quantile[lower_index]
lower_x <- c(min(model_frame[,k]),lower_cp)
lower_y <- c(-lower_beta*(lower_x[2]-lower_x[1]),0)
lower_perc <- round(-(exp(unit[k-10]*lower_beta)-1)*100,2)
lower_p <- round(lower_list[[k-10]]$pval[lower_index],4)

upper_index <- which(upper_list[[k-10]]$AIC==
               min(upper_list[[k-10]]$AIC))
upper_beta <- upper_list[[k-10]]$beta[upper_index]
upper_cp <- upper_list[[k-10]]$value[upper_index]
upper_quantile <- upper_list[[k-10]]$quantile[upper_index]
upper_x <- c(upper_cp,max(model_frame[,k]))
upper_y <- c(0,upper_beta*(upper_x[2]-upper_x[1]))
upper_perc <- round((exp(unit[k-10]*upper_beta)-1)*100,2)
upper_p <- round(upper_list[[k-10]]$pval[upper_index],4)

plot(model_frame[,k],log(model_frame$Anzahl+1),
     xlab=names(model_frame)[k],ylab="log(Anzahl+1)",
     sub=paste("Comfort range (quantiles): ",lower_quantile," - ",
         upper_quantile,sep=""),col="grey",ylim=c(-1,3.5),
     main=paste("Abs. increase/decrease per ",unit[k-10],
          "unit(s) below lower quantile: ",lower_perc,"% (p-value:",
          lower_p,") \n","Abs. increase/decrease per ",unit[k-10],
          " unit(s) above upper quantile: ",upper_perc,"% (p-value: ",
          upper_p,")",sep=""),cex.main=0.8,cex.lab=1.5)
lines(x=lower_x,y=lower_y,col=(lower_p<0.05)+1,lwd=2)
lines(x=c(lower_cp,upper_cp),y=c(0,0),col=1,lwd=2)
lines(x=upper_x,y=upper_y,col=(upper_p<0.05)+1,lwd=2)
points(x=lower_cp,y=0,col=(lower_p<0.05)+1,pch=16,cex=1.5)
points(x=upper_cp,y=0,col=(upper_p<0.05)+1,pch=16,cex=1.5)
}
```

H.3 R-Funktion lag_regress für verzögerte Kovariableneffekte und Anwendungsbeispiel

```
lag_regress <- function(formula,data,family=poisson,time_cov,
                        group_cov,almon=NULL,almon_L=3,almon_d=
                        round(sqrt(almon_L)),pdlf=NULL,pdlf_L=14,
                        pdlf_a=2.3,pdlf_d=2,sp=NULL,min_sp=NULL,
                        ret_mod=T,plot_dlf=T,plot_ext=F,...){

library(splines)
library(mgcv)

if(is.null(almon) == F & is.null(pdlf) == F){
if(length(intersect(almon,pdlf))!=0){
warning("Lag Terms shouldn't be specified in both almon and pdlf!")
}
}

mylag <- function(vec,lag=1,past=T){
temp <- vec
if(past){
vec[1:lag] <- NA
for (i in (lag+1):length(vec)){vec[i] <- temp[i-lag]}
}
else{
vec[(length(vec)-lag+1):length(vec)] <- NA
for (i in 1:(length(vec)-lag)){vec[i] <- temp[i+lag]}
}
return(vec)
}

mylag_sep <- function(vec_long,len,lag=1,past=T){
vec_sep <- mylag(vec=vec_long[1:len],lag=lag,past=past)
for (j in 2:(length(vec_long)/len)){
new_vec <- mylag(vec=vec_long[(j-1)*len+(1:len)],lag=lag,past=past)
vec_sep <- append(vec_sep,new_vec)
}
return(vec_sep)
}

# Sortieren des Datensatzes
```

```
message("Sorting data frame ... please wait!")

group_frame <- data.frame(data[,group_cov])
group_vector <- character(nrow(data))
for (k in 1:nrow(data)){
group_vector[k] <- as.character(group_frame[k,1])
if(ncol(group_frame)>=2){
for(l in 2:ncol(group_frame)){
group_vector[k] <- paste(group_vector[k],
                   as.character(group_frame[k,l]),sep="-")
}}}
group_vector <- as.factor(group_vector)

workdata <- data[order(group_vector,data[,time_cov]),]
row.names(workdata) <- 1:nrow(workdata)

# Reduzieren des Datensatzes auf notwendige Variablen
# und Erzeugen der verwendeten Variablenliste

gp <- interpret.gam(as.formula(formula))
vars <- all.vars(gp$fake.formula[-2])
vars <- append(vars,gp$response)
vars <- append(vars,almon)
vars <- append(vars,pdlf)
vars <- append(vars,time_cov)
vars <- append(vars,group_cov)
vars <- unique(vars)

workdata <- workdata[,vars]

varlist <- list()

for(i in 1:ncol(workdata)){
varlist[[i]] <- workdata[,i]
names(varlist)[i] <- names(workdata)[i]
}

d <- length(varlist)

# Berechnen der W-Matrizen für Almon-Terme
```

```
if(is.null(almon)==F){

message("Calculate W matrices for Almon-Terms ... please wait!")

calc_W_almon <- function(name,deg,maxlag){
Wmat <- matrix(nrow=nrow(lagframe),ncol=deg+1)
# Berechne W0
Wmat[,1] <- lagframe[,name]
for(l in 1:maxlag){
Wmat[,1] <- Wmat[,1] + lagframe[,paste(name,"_",l,sep="")]
}
# Berechne W1, W2, ...
for(d in 1:deg){
Wmat[,d+1] <- lagframe[,paste(name,"_1",sep="")]
for(l in 2:maxlag){
Wmat[,d+1] <- Wmat[,d+1] + l^d * lagframe[,paste(name,"_",l,sep="")]
}
}
return(Wmat)
}

if(length(almon)!=length(almon_d)){almon_d <- rep(almon_d[1],
                                                  times=length(almon))}
if(length(almon)!=length(almon_L)){almon_L <- rep(almon_L[1],
                                                  times=length(almon))}

for (i in 1:length(almon)){

# Berechnen der Lags
lagframe <- data.frame(workdata[,almon[i]])
names(lagframe) <- almon[i]
for (j in 1:almon_L[i]){
lagframe <- cbind(lagframe,mylag_sep(workdata[,almon[i]],
            len=nlevels(as.factor(workdata[,time_cov])),lag=j))
names(lagframe)[ncol(lagframe)] <- paste(almon[i],j,sep="_")
}

varlist[[d+i]] <- calc_W_almon(almon[i],deg=almon_d[i],
                  maxlag=almon_L[i])
names(varlist)[d+i] <- paste("Wlist_almon",almon[i],sep="_")

formula <- paste(formula,names(varlist)[d+i],sep=" + ")
```

```
}
}

d <- length(varlist)

# Berechnen der Z-Matrizen für PDLF-Terme

if(is.null(pdlf)==F){

message("Calculate Z matrices for PDLF-Terms ... please wait!")

calc_Phi_tilde <- function(maxlag,shrink){
# Knotenabstände mit quadr. wachsender Abstandsfunktion berechnen
q <- maxlag/sum((1:floor(maxlag/2))^2)
kappa <- cumsum(q*(1:floor(maxlag/2))^2)
kappa <- kappa[-length(kappa)]
# Matrix Phi der Basisfunktionen berechnen
Phi <- bs(0:(maxlag),knots=kappa,Boundary.knots=c(-1,(maxlag+1)),
       degree=3)
# Berechnen der Matrix Phi_tilde (Shrinkage gegen 0)
W_shrink <- matrix(0,maxlag+1,maxlag+1)
for (s in 1:(maxlag+1)){
W_shrink[s,s] <- 1/(s^shrink)
}
Phi_tilde <- W_shrink%*%Phi
return(Phi_tilde)
}

if(length(pdlf)!=length(pdlf_L)){pdlf_L <- rep(pdlf_L[1],
                                          times=length(pdlf))}
if(length(pdlf)!=length(pdlf_a)){pdlf_a <- rep(pdlf_a[1],
                                          times=length(pdlf))}
for(i in 1:length(pdlf)){if(pdlf[i]<1){pdlf[i]=1}}
if(length(pdlf)!=length(pdlf_d)){pdlf_d <- rep(pdlf_d[1],
                                          times=length(pdlf))}
if(length(pdlf)!=length(sp)){sp <- rep(sp[1],times=length(pdlf))}
if(length(pdlf)!=length(min_sp)){min_sp <- rep(min_sp[1],
                                          times=length(pdlf))}

penalty <- list()

for (i in 1:length(pdlf)){
```

```
# Berechnen der Lags
lagframe <- data.frame(workdata[,pdlf[i]])
names(lagframe) <- pdlf[i]
for (j in 1:pdlf_L[i]){
lagframe <- cbind(lagframe,mylag_sep(workdata[,pdlf[i]],
            len=nlevels(as.factor(workdata[,time_cov])),lag=j))
names(lagframe)[ncol(lagframe)] <- paste(pdlf[i],j,sep="_")
}

Phi_tilde <- calc_Phi_tilde(maxlag=pdlf_L[i],shrink=pdlf_a[i])

varlist[[d+i]] <- as.matrix(lagframe)%*%Phi_tilde
names(varlist)[d+i] <- paste("Wlist_pdlf",pdlf[i],sep="_")

formula <- paste(formula,names(varlist)[d+i],sep=" + ")

# Festlegen der verwendeten Differenzenordnung für die Penalisierung

if(pdlf_d[i]==1){
D1 <- matrix(0,ncol(Phi_tilde)-1,ncol(Phi_tilde))
for(t in 1:(ncol(Phi_tilde)-1)){
D1[t,t] <- -1
D1[t,t+1] <- 1
}
K1 <- t(D1)%*%D1
penalty[[i]] <- list(K1)
names(penalty)[i] <- paste("Wlist_pdlf",pdlf[i],sep="_")
}

if(pdlf_d[i]==2){
D2 <- matrix(0,ncol(Phi_tilde)-2,ncol(Phi_tilde))
for(t in 1:(ncol(Phi_tilde)-2)){
D2[t,t] <- 1
D2[t,t+1] <- -2
D2[t,t+2] <- 1
}
K2 <- t(D2)%*%D2
penalty[[i]] <- list(K2)
names(penalty)[i] <- paste("Wlist_pdlf",pdlf[i],sep="_")
}
```

```
if(pdlf_d[i]==3){
D3 <- matrix(0,ncol(Phi_tilde)-3,ncol(Phi_tilde))
for(t in 1:(ncol(Phi_tilde)-3)){
D3[t,t] <- 1
D3[t,t+1] <- -3
D3[t,t+2] <- 3
D3[t,t+3] <- -1
}
K3 <- t(D3)%*%D3
penalty[[i]] <- list(K3)
names(penalty)[i] <- paste("Wlist_pdlf",pdlf[i],sep="_")
}

if(pdlf_d[i]!=1 & pdlf_d[i]!=2 & pdlf_d[i]!=3){
penalty[[i]] <- NULL
warning(paste("pdlf_d must be in c(1,2,3)! No penalty for",pdlf[i],
        "assumed!"))
}

}
}

else{penalty <- NULL}

message("Fit the model ... please wait!")

# Rechnen des GAMs
model <- gam(as.formula(formula),paraPen=penalty,sp=sp,min.sp=min_sp,
         data=varlist,family=family,na.action=na.omit,...)

w <- 1
return_list <- list()

# Berechnung der Schätzertabelle für Non-Lag-Effects
table_other <- summary(model)$p.table[substr(row.names(
               summary(model)$p.table),start=1,stop=5)!="Wlist",]

return_list[[w]] <- round(table_other,6)
names(return_list)[w] <- "other"
w <- w+1

# Darstellung der DLFs
```

```
plot_almon <- function(name,deg,maxlag,plot){

index <- paste("Wlist_almon",name,sep="_")==substr(names(coef(model)),
         start=1,stop=12+nchar(name))

gamma <- coef(model)[index]
beta <- numeric(maxlag+1)

beta[1] <- gamma[1]

for(l in 2:(maxlag+1)){
beta[l] <- gamma[1]
for(d in 2:(deg+1)){
beta[l] <- beta[l] + (l-1)^(d-1) * gamma[d]
}
}

beta.var <- numeric(maxlag+1)
covmat.gamma <- model$Vp[index,index]

beta.var[1] <- covmat.gamma[1,1]

for(l in 2:(maxlag+1)){
weight.mat <- matrix(0,deg+1,deg+1)
for(i in 1:(deg+1)){
for(j in 1:(deg+1)){
weight.mat[i,j] <- (l-1)^(i+j-2)
}
}
beta.var[l] <- sum(covmat.gamma*weight.mat)
}

beta.sd <- sqrt(beta.var)

lower <- beta - qnorm(0.975) * beta.sd
upper <- beta + qnorm(0.975) * beta.sd

# Plotte die Distributed Lag Function
if(plot){
plot(0:maxlag,beta,type="l",xlab="",ylab=expression(paste(beta,
     "(Lag)")),main=name,sub=paste("Almon DLF \n deg = ",deg,sep=""),
```

```
     xaxt="n",ylim=c(min(lower,0),max(upper,0)))
axis(side=1,at=0:maxlag)
polygon(x=c(0:maxlag,maxlag:0,0),y=c(upper,lower[(maxlag+1):1],
        upper[1]),col="grey",border=F)
lines(0:maxlag,beta)
abline(h=0,col=2,lwd=2)
}

rframe <- data.frame(beta=beta,lower=lower,upper=upper)
rframe <- round(rframe,6)

return(rframe)
}

plot_pdlf <- function(name,maxlag,diff,shrink,plot){

index <- paste("Wlist_pdlf",name,sep="_")==substr(names(coef(model)),
         start=1,stop=11+nchar(name))

alpha <- coef(model)[index]
Phi_tilde <- calc_Phi_tilde(maxlag=maxlag,shrink=shrink)
beta <- Phi_tilde%*%alpha

covalpha <- model$Vp[index,index]
covbeta <- Phi_tilde%*%covalpha%*%t(Phi_tilde)

lower <- beta - qnorm(0.975)*sqrt(diag(covbeta))
upper <- beta + qnorm(0.975)*sqrt(diag(covbeta))

# Plotte die Distributed Lag Function
if(plot){
plot(0:maxlag,beta,type="l",ylim=c(min(lower,0),max(upper,0)),xlab="",
     ylab=expression(paste(beta,"(Lag)")),main=name,sub=paste(
     "PDLF \n diff = ",diff,", shrink = ",shrink,sep=""),xaxt="n")
axis(side=1,at=0:maxlag)
polygon(x=c(0:maxlag,maxlag:0,0),y=c(upper,lower[(maxlag+1):1],
        upper[1]),col="grey",border=F)
lines(0:maxlag,beta)
abline(h=0,col=2,lwd=2)

q <- maxlag/sum((1:floor(maxlag/2))^2)
kappa <- cumsum(q*(1:floor(maxlag/2))^2)
```

```
kappa <- kappa[-length(kappa)]

points(x=kappa,y=rep(0,length(kappa)),col=2,pch=15)
}

rframe <- data.frame(beta=beta,lower=lower,upper=upper)
rframe <- round(rframe,6)

return(rframe)
}

if(plot_dlf==F){plot_ext <- T}

if(is.null(almon)==F){
table_almon <- list()
# Darstellung der Almon-Terme
for (i in 1:length(almon)){
table_almon[[i]] <- plot_almon(name=almon[i],deg=almon_d[i],
                    maxlag=almon_L[i],plot=plot_dlf)
row.names(table_almon[[i]]) <- paste(almon[i],0:almon_L[i],sep="_")
names(table_almon)[i] <- almon[i]
if(plot_ext==F){devAskNewPage(T)}
}
return_list[[w]] <- table_almon
names(return_list)[w] <- "almon"
w <- w+1
}

if(is.null(pdlf)==F){
table_pdlf <- list()
# Darstellung der PDLF-Terme
for(i in 1:length(pdlf)){
table_pdlf[[i]] <- plot_pdlf(name=pdlf[i],maxlag=pdlf_L[i],
                   diff=pdlf_d[i],shrink=pdlf_a[i],plot=plot_dlf)
row.names(table_pdlf[[i]]) <- paste(pdlf[i],0:pdlf_L[i],sep="_")
names(table_pdlf)[i] <- pdlf[i]
if(plot_ext==F){devAskNewPage(T)}
}
return_list[[w]] <- table_pdlf
names(return_list)[w] <- "pdlf"
w <- w+1
}
```

```
return_list[[w]] <- model
names(return_list)[w] <- "model"
w <- w+1

if(is.null(min_sp)){min_sp <- rep(0,length(pdlf))}

if(is.null(sp)){
return_list[[w]] <- min_sp + model$sp
names(return_list[[w]]) <- pdlf
}
else{
return_list[[w]] <- min_sp + model$full.sp
names(return_list[[w]]) <- pdlf
}
names(return_list)[w] <- "penalty"

if(plot_ext==F){devAskNewPage(F)}

if(ret_mod){return(return_list)}

}
```

Anwendungsbeispiel für die KVB-Call-Center-Daten in der Pilot-Region München:

```
# Pilotregion München - Call-Center-Daten

# Daten einlesen
frame <- read.table("../../Daten_Call-Center_München_a).txt",header=T)

# Definition der Faktoren
frame$Alter <- as.factor(frame$Alter)
frame$dow <- as.factor(frame$dow)
frame$school <- as.factor(frame$school)
frame$holiday <- as.factor(frame$holiday)
frame$bridge <- as.factor(frame$bridge)
frame$quartal <- as.factor(frame$quartal)

# Mittelwertcodierung
contrasts(frame$Alter)[1,] <- -1
```

```
contrasts(frame$dow)[1,] <- -1
contrasts(frame$c.wdir.ave)[1,] <- -1

# Sortieren nach Datum und Alter
frame <- frame[order(frame$Datum,frame$Alter),]
row.names(frame) <- 1:nrow(frame)

formula = "Anzahl ~ offset(log(Einwohner)) + dow + school + holiday +
          bridge + quartal + Alter*ns(time,12) + q.mmm + q.ave_lcp +
          q.ave_ucp + qnh.mmm + qnh.ave_lcp + qnh.ave_ucp + tmt.mmm +
          tmt.ave_lcp + tmt.ave_ucp + c.wdir.ave"

almon = c("cp.max","lsp.max","lcc.ave","mcc.ave","q.ave","qnh.ave",
        "sshf.ave","stressspd.max","tmt.ave","windspd.max")

pdlf = c("SO2_tr.q95","PM10_tr.q95","O3_tr.q95","NO2_tr.q95",
       "CO_tr.q95")

source("../../lag_regress.r")

lag_regress(formula=formula,data=frame,family=poisson,
            time_cov="time",group_cov="Alter",almon=almon,
            almon_L=3,almon_d=2,pdlf=pdlf,pdlf_L=14,
            pdlf_a=2.3,pdlf_d=2,sp=NULL,min_sp=10,
            ret_mod=T,plot_dlf=T,plot_ext=F)
```

H.4 Beispielhafter R-Code zur Umsetzung des fortlaufenden Prognosemodells

```
source("../../lag_regress.r")

# Räumlicher Effekt

spatial_table <- read.table("../../Spatial_Call-Center_stepaic.res",
                 header=T)[,2:3]
spatial_table[,1] <- as.factor(spatial_table[,1])
names(spatial_table)[2] <- "spatial"

# Einlesen des Datensatzes

med_frame <- read.table("../../Daten_Abrechnung_gesamt.txt",header=T)
med_frame$Datum <- as.factor(med_frame$Datum)
med_frame$Kreis <- as.factor(med_frame$Kreis)

meteo_frame <- read.table(
               "../../Daten_Meteorologie_Bayern_komplett.txt",
               header=T)
meteo_frame$Datum <- as.factor(meteo_frame$Datum)
meteo_frame$Kreis <- as.factor(meteo_frame$Kreis)

aq_frame <- read.table(
            "../../Daten_Luftqualität_Bayern_komplett.txt",
            header=T)
aq_frame$Datum <- as.factor(aq_frame$Datum)
aq_frame$Kreis <- as.factor(aq_frame$Kreis)

frame <- merge(med_frame,meteo_frame,by=c("Datum","Kreis"),all.x=T)
frame <- merge(frame,aq_frame,by=c("Datum","Kreis"),all.x=T)
frame <- merge(frame,spatial_table,by="Kreis",all.x=T)

# Definition der Faktoren

frame$Alter <- as.factor(frame$Alter)
frame$dow <- as.factor(frame$dow)
frame$school <- as.factor(frame$school)
frame$holiday <- as.factor(frame$holiday)
frame$bridge <- as.factor(frame$bridge)
```

```
frame$quartal <- as.factor(frame$quartal)

# Sortieren nach Alter, Datum und Landkreis
frame <- frame[order(frame$Datum,frame$Kreis,frame$Alter),]
row.names(frame) <- 1:nrow(frame)

# Mittelwertcodierung

contrasts(frame$Alter)[1,] <- -1
contrasts(frame$dow)[1,] <- -1
contrasts(frame$c.wdir.ave)[1,] <- -1

# Fortlaufend lernendes Prädiktionsmodell

knot_list <- list()
B.knot_list <- list()

knot_list[[1]] <- quantile(-547.5:-183.5,probs=seq(0,1,len=5)[2:4])
B.knot_list[[1]] <- quantile(-547.5:-183.5,
                    probs=seq(0,1,len=5)[-(2:4)])
for(i in 2:91){
knot_list[[i]] <- quantile(-547.5:-93.5,probs=seq(0,1,len=6)[2:5]);
B.knot_list[[i]] <- quantile(-547.5:-93.5,
                    probs=seq(0,1,len=6)[-(2:5)])
}
for(i in 92:182){
knot_list[[i]] <- quantile(-547.5:-2.5,probs=seq(0,1,len=7)[2:6])
B.knot_list[[i]] <- quantile(-547.5:-2.5,
                    probs=seq(0,1,len=7)[-(2:6)])
}
for(i in 183:274){
knot_list[[i]] <- quantile(-547.5:89.5,probs=seq(0,1,len=8)[2:7])
B.knot_list[[i]] <- quantile(-547.5:89.5,
                    probs=seq(0,1,len=8)[-(2:7)])
}
for(i in 275:366){
knot_list[[i]] <- quantile(-547.5:181.5,probs=seq(0,1,len=9)[2:8])
B.knot_list[[i]] <- quantile(-547.5:181.5,
                    probs=seq(0,1,len=9)[-(2:8)])
}
for(i in 367:457){
knot_list[[i]] <- quantile(-547.5:272.5,probs=seq(0,1,len=10)[2:9])
```

```
B.knot_list[[i]] <- quantile(-547.5:272.5,
                    probs=seq(0,1,len=10)[-(2:9)])
}
for(i in 458:548){
knot_list[[i]] <- quantile(-547.5:363.5,probs=seq(0,1,len=11)[2:10])
B.knot_list[[i]] <- quantile(-547.5:363.5,
                    probs=seq(0,1,len=11)[-(2:10)])
}
for(i in 549:640){
knot_list[[i]] <- quantile(-547.5:455.5,probs=seq(0,1,len=12)[2:11])
B.knot_list[[i]] <- quantile(-547.5:455.5,
                    probs=seq(0,1,len=12)[-(2:11)])
}
for(i in 641:731){
knot_list[[i]] <- quantile(-547.5:547.5,probs=seq(0,1,len=13)[2:12])
B.knot_list[[i]] <- quantile(-547.5:547.5,
                    probs=seq(0,1,len=13)[-(2:12)])
}
names(knot_list) <- 0:730
names(B.knot_list) <- 0:730

penalty_update <- rep(F,731)
penalty_update[c(1,2,92,183,275,367,458,549,641)] <- T

plot_trace <- rep(F,731)
plot_trace[floor(seq(1,731,len=25))] <- T

predict_frame <- frame[175201:nrow(frame),c(1:3,5)]
predict_frame$pred_1_ahead <- NA
predict_frame$pred_1_error <- NA
predict_frame$pred_1_lower <- NA
predict_frame$pred_1_upper <- NA
predict_frame$pred_1_index <- NA
predict_frame$pred_1_sqdif <- NA
predict_frame$pred_2_ahead <- NA
predict_frame$pred_2_error <- NA
predict_frame$pred_2_lower <- NA
predict_frame$pred_2_upper <- NA
predict_frame$pred_2_index <- NA
predict_frame$pred_2_sqdif <- NA
predict_frame$pred_3_ahead <- NA
predict_frame$pred_3_error <- NA
```

```
predict_frame$pred_3_lower <- NA
predict_frame$pred_3_upper <- NA
predict_frame$pred_3_index <- NA
predict_frame$pred_3_sqdif <- NA

fixed <- c("(Intercept)","Deprivation","dow2","dow3","dow4","dow5",
         "dow6","dow7","school1","holiday1","bridge1","quartal1",
         "quartal2","Alter2","Alter3","Alter4","Alter5","q.ave_ucp",
         "qnh.ave_lcp","tmt.ave_lcp","tmt.ave","CO_tr.q95")

coef_frame <- data.frame(fixed=fixed)
se_frame <- data.frame(fixed=fixed)

time_list <- list()
k <- 1

postscript("../../kvbcc_stepaic_predict_trace_lag_plots.ps")

# Iteration

for(i in 0:730){

f_knots <- paste("c(",knot_list[[i+1]][1],sep="")
for(j in 2:length(knot_list[[i+1]])){
f_knots <- paste(f_knots,paste(", ",knot_list[[i+1]][j],sep=""),
           sep="")
}
f_knots <- paste(f_knots,")",sep="")

f_B.knots <- paste("c(",B.knot_list[[i+1]][1],", ",
             B.knot_list[[i+1]][2],")",sep="")

formula <- paste("Anzahl ~ offset(log(Einwohner)) +
           offset(spatial) + Deprivation + dow + school + holiday +
           bridge + quartal + Alter * ns(time, ",
           length(knot_list[[i+1]])+1,",knots = ",f_knots,",
           Boundary.knots = ",f_B.knots,") + q.ave_ucp +
           qnh.ave_lcp + tmt.ave_lcp + tmt.ave + CO_tr.q95",sep="")

almon <- c("q.ave","qnh.ave","stressspd.max","windspd.max")
pdlf <- c("SO2_tr.q95","O3_tr.q95","NO2_tr.q95")
```

```
almon_mlag <- rep(3,4)
almon_deg <- rep(2,4)

pdlf_mlag <- rep(14,3)
pdlf_shrink <- rep(3.5,3)

if(penalty_update[i+1]){sp <- NULL; min.sp <- rep(1000,3)}
else{sp <- model$penalty; min.sp <- NULL}

par(mfrow=c(3,4))

model <- lag_regress(formula=formula,data=frame[1:(175200+i*480),],
         family=quasipoisson,time_cov="time",group_cov=c("Kreis",
         "Alter"),almon=almon,almon_mlag=almon_mlag,
         almon_deg=almon_deg,pdlf=pdlf,pdlf_mlag=pdlf_mlag,
         pdlf_shrink=pdlf_shrink,pdlf_diff=rep(2,3),sp=sp,
         min.sp=min.sp,return_model=T,plot_dlf=plot_trace[i+1],
         plot_to_pdf=T)

par(mfrow=c(1,1))

if(penalty_update[i+1]){
time_list[[k]] <- list()

index1 <- substr(rownames(model$other),start=1,stop=7)=="ns(time"
time_list[[k]][[1]] <- model$other[,1][index1]
names(time_list[[k]])[1] <- "nstime"

index2 <- substr(names(coef(model$model)),start=1,stop=7)==
          "ns(time"
time_list[[k]][[2]] <- model$model$Vp[index2,index2]
names(time_list[[k]])[2] <- "nstime_cov"

index3 <- substr(rownames(model$other),start=8,stop=14)=="ns(time"
time_list[[k]][[3]] <- model$other[,1][index3]
names(time_list[[k]])[3] <- "nstimeage"

index4 <- substr(names(coef(model$model)),start=8,stop=14)==
          "ns(time"
time_list[[k]][[4]] <- model$model$Vp[index4,index4]
names(time_list[[k]])[4] <- "nstimeage_cov"
```

```
time_list[[k]][[5]] <- model$model$Vp[index2,index4]
names(time_list[[k]])[5] <- "nstime_nstimeage_cov"

index5 <- substr(rownames(model$other),start=1,stop=5)=="Alter"
time_list[[k]][[6]] <- model$other[,1][index5]
names(time_list[[k]])[6] <- "age"

index6 <- substr(names(coef(model$model)),start=1,stop=5)=="Alter"
time_list[[k]][[7]] <- model$model$Vp[index6,index6]
names(time_list[[k]])[7] <- "age_cov"

k <- k+1
}

# Liste mit Kovariablen für die Prädiktion

# Administrative Variablen und sonstige Terme

newdata <- list()

if(i<=728){newdata[[1]] <- rep(1,1440)}
if(i==729){newdata[[1]] <- rep(1,960)}
if(i==730){newdata[[1]] <- rep(1,480)}

names(newdata)[1] <- "(Intercept)"

predict_lines <- (175200+i*480+1):min(175200+i*480+1440,nrow(frame))

newdata[[2]] <- frame$Deprivation[predict_lines]
names(newdata)[2] <- "Deprivation"
newdata[[3]] <- frame$Alter[predict_lines]
names(newdata)[3] <- "Alter"
newdata[[4]] <- frame$dow[predict_lines]
names(newdata)[4] <- "dow"
newdata[[5]] <- frame$school[predict_lines]
names(newdata)[5] <- "school"
newdata[[6]] <- frame$holiday[predict_lines]
names(newdata)[6] <- "holiday"
newdata[[7]] <- frame$bridge[predict_lines]
names(newdata)[7] <- "bridge"
newdata[[8]] <- frame$quartal[predict_lines]
names(newdata)[8] <- "quartal"
```

```
newdata[[9]] <- frame$q.ave_ucp[predict_lines]
names(newdata)[9] <- "q.ave_ucp"
newdata[[10]] <- frame$qnh.ave_lcp[predict_lines]
names(newdata)[10] <- "qnh.ave_lcp"
newdata[[11]] <- frame$tmt.ave_lcp[predict_lines]
names(newdata)[11] <- "tmt.ave_lcp"

newdata[[12]] <- frame$tmt.ave[predict_lines]
names(newdata)[12] <- "tmt.ave"
newdata[[13]] <- frame$CO_tr.q95[predict_lines]
names(newdata)[13] <- "CO_tr.q95"

newdata[[14]] <- frame$time[predict_lines]
names(newdata)[14] <- "time"

newdata[[15]] <- frame$Einwohner[predict_lines]
names(newdata)[15] <- "Einwohner"
newdata[[16]] <- frame$spatial[predict_lines]
names(newdata)[16] <- "spatial"

# Almon-Lag-Terme

calc_W_almon <- function(name,deg,maxlag){
Wmat <- matrix(nrow=nrow(lagframe),ncol=deg+1)
# Berechne W0
Wmat[,1] <- lagframe[,name]
for(l in 1:maxlag){
Wmat[,1] <- Wmat[,1] + lagframe[,paste(name,"_",l,sep="")]
}
# Berechne W1, W2, ...
for(d in 1:deg){
Wmat[,d+1] <- lagframe[,paste(name,"_1",sep="")]
for(l in 2:maxlag){
Wmat[,d+1] <- Wmat[,d+1] + l^d * lagframe[,paste(name,"_",l,
              sep="")]
}
}
return(Wmat)
}

for (j in 1:length(almon)){
```

```
almon_cols <- almon[j]==substr(names(frame),start=1,
              stop=nchar(almon[j])) & substr(names(frame),start=
              nchar(names(frame))-2,stop=nchar(names(frame))) !=
              "mmm" & substr(names(frame),start=nchar(names(
              frame))-1,stop=nchar(names(frame))) != "cp"
lagframe <- frame[predict_lines,almon_cols]
almon_frame <- calc_W_almon(name=almon[j],deg=almon_deg[j],
               maxlag=almon_mlag[j])
newdata[[16+j]] <- almon_frame
names(newdata)[16+j] <- paste("Wlist_almon_",almon[j],sep="")
}

# PDLF-Terme

calc_Phi_tilde <- function(maxlag,shrink){
# Knotenabstände mit quadr. wachsender Abstandsfunktion berechnen
q <- maxlag/sum((1:floor(maxlag/2))^2)
kappa <- cumsum(q*(1:floor(maxlag/2))^2)
kappa <- kappa[-length(kappa)]
# Matrix Phi der Basisfunktionen berechnen
Phi <- bs(0:(maxlag),knots=kappa,Boundary.knots=c(-1,(maxlag+1)),
       degree=3)
# Berechnen der Matrix Phi_tilde (Shrinkage gegen 0)
W_shrink <- matrix(0,maxlag+1,maxlag+1)
for (s in 1:(maxlag+1)){
W_shrink[s,s] <- 1/(s^shrink)
}
Phi_tilde <- W_shrink%*%Phi
return(Phi_tilde)
}

for (j in 1:length(pdlf)){
pdlf_cols <- pdlf[j]==substr(names(frame),start=1,
             stop=nchar(pdlf[j]))
lagframe <- frame[predict_lines,pdlf_cols]
Phi_tilde <- calc_Phi_tilde(maxlag=pdlf_mlag[j],shrink=
             pdlf_shrink[j])
pdlf_frame <- as.matrix(lagframe)%*%Phi_tilde
newdata[[16+length(almon)+j]] <- pdlf_frame
names(newdata)[16+length(almon)+j] <- paste("Wlist_pdlf_",
                                      pdlf[j],sep="")
}
```

```
pred <- predict(model$model,newdata=newdata,type="response",
        se.fit=T)

predict_frame$pred_1_ahead[i*480+1:480] <- pred$fit[1:480]
predict_frame$pred_1_error[i*480+1:480] <- pred$se.fit[1:480]

if(i<=729){
predict_frame$pred_2_ahead[(i+1)*480+1:480] <- pred$fit[481:960]
predict_frame$pred_2_error[(i+1)*480+1:480] <- pred$se.fit[481:960]
}
if(i<=728){
predict_frame$pred_3_ahead[(i+2)*480+1:480] <- pred$fit[961:1440]
predict_frame$pred_3_error[(i+2)*480+1:480] <- pred$se.fit[961:1440]
}

coef_frame$V1 <- numeric(nrow(coef_frame))
se_frame$V1 <- numeric(nrow(se_frame))

for (j in 1:nrow(coef_frame)){
coef_frame$V1[j] <- model$other[,1][as.character(
                      coef_frame$fixed)[j]==rownames(model$other)]
se_frame$V1[j] <- model$other[,2][as.character(
                    se_frame$fixed)[j]==rownames(model$other)]
}

names(coef_frame)[ncol(coef_frame)] <- paste("model",i,sep="_")
names(se_frame)[ncol(se_frame)] <- paste("model",i,sep="_")

if(plot_trace[i+1]){
save(list=c("newdata","time_list","coef_frame","se_frame","pred",
    "predict_frame","knot_list","B.knot_list"),
    file="../../kvbcc_stepaic_postpred.RData")
}

}

dev.off()
```

I CD mit Datensätzen, weiterem R-Code und Grafiken