



Volker Schmid

# **Bayesianische Raum-Zeit-Modellierung in der Epidemiologie**

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

**München 2004**



---

# **Bayesianische Raum-Zeit-Modellierung in der Epidemiologie**

**Volker Schmid**

---

Dissertation

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

vorgelegt von

Volker Schmid

aus Roding

München, den 30. August 2004

Erstgutachter: Prof. Dr. Leonhard Held  
Zweitgutachter: Prof. Dr. Ludwig Fahrmeir  
Drittgutachter: Prof. Dr. Dankmar Böhning  
Rigorosum: 30. November 2004

# Danksagung

Bei den Vorbereitungen zur Drucklegung (..) hat mir eigentlich niemand geholfen; diejenigen, die mich dabei behinderten, will ich nicht aufzählen, denn dies würde zu viel Platz erfordern.

*aus: Stanislav Lem, Die Sterntagebücher des Weltraumfahrers Ijon Tichy*

In Wirklichkeit wäre diese Arbeit ohne die Unterstützung unzähliger Personen niemals zustande gekommen. Besonderer Dank geht dabei an meinen Doktorvater Prof. Leonhard Held, der mir die Möglichkeit der Promotion schmackhaft gemacht und mit sanftem Druck für deren Fertigstellung gesorgt hat. Die hervorragende Betreuung dazwischen darf natürlich nicht unerwähnt bleiben. Spezieller Dank geht auch an Prof. Ludwig Fahrmeir, der für die Anfangszeit meine Betreuung übernahm. Er sorgte auch größtenteils für meine finanzielle Ausstattung.

Diese Arbeit entstand während meiner Tätigkeit am Sonderforschungsbereich 386 “Statistische Analyse diskreter Strukturen” am Institut für Statistik der Ludwig-Maximilians-Universität München. Die wissenschaftliche Arbeit wurde somit von der deutschen Forschungsgemeinschaft finanziert, der ich dafür danken möchte.

Große Teile dieser Arbeit entstanden in Kooperation mit Anwendern. Ein herzlicher Dank für die Zusammenarbeit, Daten und Anregungen geht an Dr. Christoph Staubach (Kapitel 3), Dr. Anna Hansell (Kapitel 5), Mathias Hofmann (Kapitel 7) und all jene, die sonst noch Daten zur Verfügung stellten oder anderweitig halfen. Vielen Dank auch an alle BAMP-Benutzer, speziell natürlich an diejenigen mit hilfreichen Kommentaren.

Ein herzliches Dankeschön geht an alle Mitarbeiter des Lehrstuhls Fahrmeir und der Arbeitsgruppe Held für das gute Arbeitsklima. Spezieller Dank geht an Dr. Günter Raßer für viele interessante fachliche und private Gespräche, an Dr. Leyre Estíbaliz Osuna Echavarría, nicht nur für Kaffee und Zigaretten, und an Renata Zambrzycka für ein gutes Büroklima.

Ein großer Dank geht auch an meine Familie, die mich auf meinem Weg immer unterstützt hat und weiter unterstützt, wo immer er auch hinget – und die mir immer ein Zuhause bietet.

Danken möchte ich auch all jenen, die mir das Leben in München angenehmer gemacht haben, speziell Klaus, Holger und Flo, und allen Freunden, die trotz räumlicher Distanz Kontakt zu mir gehalten haben. Danke allen Dionern. Lieben Dank an Eva für alles. Vielen lieben Dank an die, die mir das letzte Jahr so viel gegeben hat.

*München, Dezember 2004*

*Volker Schmid*

## Zusammenfassung

In dieser Arbeit werden räumliche und zeitliche Strukturen epidemiologischer Daten mittels moderner Bayes-Verfahren analysiert. Als Glättungsprioris finden hauptsächlich autoregressive Verteilungen wie Gauss-Markov-Zufallsfelder und Random Walks Verwendung. Derartige komplexe Modelle können nur mit MCMC-Methoden geschätzt werden. Es werden effiziente Algorithmen vorgestellt, welche die Schätzung der Parameter in annehmbarem Zeitbedarf zulassen. Insbesondere für die Modellierung von Raum-Zeit-Interaktionen sind diese Algorithmen wichtig.

Als Anwendung räumlicher Bayesianischer Modelle wird die Analyse von Daten zur Inzidenz von Wildtierkrankheiten vorgestellt. An einem Datensatz zur Kindersterblichkeit werden diskreter und stetiger Ansatz räumlicher Analyse verglichen. Alters-Perioden-Kohorten-Modelle werden ausführlich beschrieben und auf räumliche Probleme erweitert. Schließlich wird für infektiöse Krankheiten ein stochastisches Modell mit räumlich-zeitlichen Elementen beschrieben.

---

## Abstract

This thesis is concerned with the analysis of spatial and temporal structures of epidemiological data using modern Bayes techniques. Mainly autoregressive distributions as Gaussian Markov random fields or random walks are used as smoothing priors.

Such extensive models can be estimated using MCMC methods only. Some effective algorithms are introduced to get estimates in acceptable time. Especially for space time interactions such algorithms are essential.

As example spatial Bayesian models are applied for wildlife disease incidence data. Discrete and continuous frameworks for spatial analysis are compared on a data set on infant mortality cases. Age-period-cohort models are discussed in detail and an extension for spatial data is presented. Finally a stochastic model for space time data on infectious diseases is described.

Prognosen sind immer schwierig zu machen, insbesondere, wenn sie die Zukunft betreffen.

*Pascal Couchepin, Schweizer Bundespräsident*

# Inhaltsverzeichnis

<b>1</b>	<b>Einleitung</b>	<b>1</b>
1.1	Räumliche Statistik . . . . .	2
1.2	Inhaltsübersicht . . . . .	4
<b>2</b>	<b>Methoden und Techniken</b>	<b>7</b>
2.1	Hierarchische Bayesianische Modelle für epidemiologische Fragestellungen .	7
2.2	Prioriverteilungen . . . . .	10
2.2.1	Gauss-Markov-Zufallsfelder . . . . .	10
2.2.2	Random Walk Prioris . . . . .	13
2.2.3	Interaktionen . . . . .	15
2.2.4	Hyperparameter . . . . .	16
2.3	MCMC-Methoden . . . . .	17
2.3.1	Ziehen aus multivariaten Normalverteilungen . . . . .	19
2.3.2	Approximation einer Normalverteilung durch Taylorentwicklung . .	21
2.3.3	Gemeinsames Ziehen von Parametervektor und Präzisionsparameter	23
2.4	Auswertung . . . . .	24
2.4.1	Zum Umgang mit den Ergebnissen . . . . .	24
2.4.2	Devianz und DIC . . . . .	24
<b>3</b>	<b>Modelle für räumliche Prävalenzdaten von Wildtier-Krankheiten</b>	<b>27</b>
3.1	Disease Mapping in der Veterinärepidemiologie . . . . .	28
3.2	Vergleich zweier räumlicher Modelle . . . . .	29

3.2.1	Daten . . . . .	29
3.2.2	Modelle . . . . .	31
3.2.3	Anwendung . . . . .	35
3.2.4	Diskussion . . . . .	37
3.3	Ein Ansatz mit Hilfsvariablen . . . . .	42
3.3.1	Modellierung . . . . .	43
3.3.2	Vergleich . . . . .	45
3.3.3	Erweiterung um Kovariablen . . . . .	46
3.4	Fazit und Ausblick . . . . .	49
<b>4</b>	<b>Räumliche Effekte auf regulären Gittern in der Epidemiologie</b>	<b>51</b>
4.1	Fallkontrollstudie Porto Alegre . . . . .	52
4.2	Modellformulierung . . . . .	55
4.3	GMRF-Prioris . . . . .	57
4.4	Ergebnisse . . . . .	63
4.5	Diskussion . . . . .	68
<b>5</b>	<b>Alters-Perioden-Kohorten-Modelle</b>	<b>71</b>
5.1	Daten auf dem Lexis-Diagramm . . . . .	72
5.1.1	Das APC-Modell . . . . .	74
5.1.2	Identifizierbarkeitsproblem . . . . .	75
5.2	Bayesianische Alters-Perioden-Kohorten-Modelle . . . . .	77
5.2.1	Random Walks und Prognose . . . . .	77
5.2.2	Inferenz . . . . .	80
5.2.3	Modell mit zusätzlicher Heterogenität . . . . .	83
5.2.4	Unstrukturierte Zeiteffekte . . . . .	85
5.3	Anwendung auf COPD-Daten . . . . .	86
5.3.1	Schätzungen . . . . .	87
5.3.2	Prognose . . . . .	93
5.4	Zusammenfassung . . . . .	95

---

<b>6</b>	<b>Extrapolation von Raum-Zeit-Trends bei Krebsregisterdaten</b>	<b>97</b>
6.1	Einführung . . . . .	98
6.1.1	Datensatz zur Magenkrebsmortalität . . . . .	98
6.1.2	Literaturüberblick . . . . .	101
6.2	Modell . . . . .	103
6.2.1	Das Binomialmodell . . . . .	103
6.2.2	Prioriannahmen . . . . .	104
6.2.3	MCMC Simulation . . . . .	106
6.2.4	Vorhersage von Mortalitätsraten . . . . .	108
6.3	Anwendung auf Magenkrebsdaten . . . . .	110
6.3.1	Modellwahl . . . . .	110
6.3.2	Ergebnisse . . . . .	111
6.4	Diskussion und Ausblick . . . . .	117
<b>7</b>	<b>Ein stochastisches Modell für Surveillance-Daten</b>	<b>119</b>
7.1	Ein Verzweigungsprozeß als Modell für infektiöse Krankheiten . . . . .	120
7.1.1	Daten . . . . .	120
7.1.2	Modell . . . . .	121
7.1.3	Überdispersion . . . . .	122
7.2	Modell mit festem $\nu$ . . . . .	123
7.2.1	Priori-Annahmen . . . . .	123
7.2.2	Inferenz . . . . .	123
7.3	Modell mit variierendem $\nu$ . . . . .	125
7.3.1	Raum-Zeit-Modellierung für $\nu$ . . . . .	125
7.3.2	Inferenz . . . . .	127
7.4	Ergebnisse . . . . .	134
7.5	Diskussion und Ausblick . . . . .	140
<b>8</b>	<b>Schlußbemerkungen</b>	<b>143</b>

<b>A Beweis</b>	<b>145</b>
A.1 Gemeinsame Verteilung bei zwölf nächsten Nachbarn . . . . .	145
<b>B BAMP</b>	<b>147</b>
B.1 Allgemeines . . . . .	149
B.2 Erste Schritte . . . . .	150
B.3 Eingabedateien . . . . .	151
B.3.1 *.ini-Datei . . . . .	151
B.3.2 Mögliche Modellierungen der APC-Effekte . . . . .	153
B.3.3 Daten . . . . .	154
B.4 Ausgabedateien . . . . .	154
B.5 Devianz und Prognose . . . . .	155
B.6 S-Plus- / R- Skript . . . . .	156

# Kapitel 1

## Einleitung

Epidemiologische Probleme rücken in jüngster Zeit immer mehr in den Vordergrund. Die steigenden Gesundheitskosten erfordern mehr Wissen über Krankheiten und die Prognose zukünftiger Krankheitsraten um vorbeugende Maßnahmen zu veranlassen und den zu erwartenden Finanzbedarf abzuschätzen. Bayesianische Methoden haben in den letzten Jahren immer größere Verwendung in der Epidemiologie gefunden. Diese Arbeit behandelt die Modellierung räumlicher und zeitlicher (und räumlich-zeitlicher) Strukturen im Bayesianischen Kontext zum Zwecke epidemiologischer Untersuchungen.

Vor allem das sogenannte *disease mapping*, also die deskriptive Darstellung räumlicher Strukturen, ist eine wichtige Anwendung in der Epidemiologie. Durch Verwendung von zum Beispiel autoregressiven Prioris wie Gauss-Markov-Zufallsfeldern oder mit Clusterverfahren (Raßer 2003) erreicht man eine geeignete Glättung der räumlichen Struktur der Daten oder kann Kanten in der räumlichen Verteilung erkennen. Diese Verfahren lassen sich auch leicht auf eindimensionale, vor allem also zeitliche Probleme anwenden oder für räumlich-zeitliche Probleme verallgemeinern. Die Untersuchung von Raum-Zeit-Interaktionen kann auch speziell bei der Frage nach der Ausbreitung bzw. Ansteckung einer Krankheit behilflich sein.

Die deskriptive Darstellung der Krankheit einer Verteilung wird immer nur der er-

ste Schritt zu einer genaueren Untersuchung der Ursachen einer Krankheit sein. Daher versucht man Kovariablen in epidemiologische Untersuchungen zu integrieren, um diejenigen herauszufinden, die bei der Verbreitung einer Krankheit mitwirken oder diese sogar auslösen. Auch der Erfolg von Gegenmaßnahmen, wie zum Beispiel Impfungen, läßt sich über Kovariablen überprüfen.

Ein weiterer Gesichtspunkt epidemiologischer Untersuchungen ist die Erstellung von Prognosen zukünftig zu erwartender Krankheitsraten und -fälle. Neben der Abschätzung zu erwartender Gesundheitskosten ist hier vor allem die Möglichkeit frühzeitigen Gegensteuerns Ziel der Untersuchung.

## 1.1 Räumliche Statistik

Erste Forschungsarbeiten in räumlicher Statistik entstanden in der Geologie. Die dort vorliegenden sogenannten *geostatistischen* Daten haben räumliche Information in Form steti-

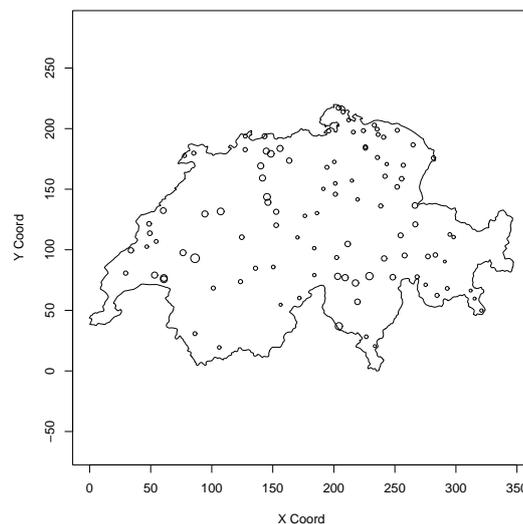


Abbildung 1.1: Schweizer Regenfall-Daten. Abgebildet sind Wetterstationen in der Schweiz. Die Größe der Kreise entspricht der Regenmenge der entsprechenden Station.

ger Koordinaten und jeder Datenpunkt hat einen stetigen Response. Ein Beispiel hierfür zeigt Abbildung 1.1, die Messung des Regenfalls eines Jahres an den Wetterstationen in der Schweiz.

Die Standardmethode für geostatistische Daten ist das sogenannte *Kriging* (Krige 1951). Dabei faßt man die Beobachtungen  $Y_1, \dots, Y_n$  als Realisationen einer Zufallsvariablen in Abhängigkeit der Koordinaten  $s_1, \dots, s_n$  auf:

$$Y(s_i) = \mu + S(s_i) + \epsilon(s_i) \quad \text{für } i = 1, \dots, n.$$

Dabei ist  $\mu$  der deterministische Trend und  $\epsilon$  weißes Rauschen. Die glatte Komponente  $S$  steht für den räumlichen Trend, wobei eine vorgegebene Korrelationsfunktion die räumliche Abhängigkeit beschreibt, meist in Abhängigkeit von der Entfernung zweier Daten. Daraus läßt sich für alle Koordinaten des Beobachtungsraums eine glatte räumliche Komponente schätzen.

Nach Cressie (1993) unterscheidet man von den genannten geostatistischen Daten noch räumliche Punktprozesse und diskrete räumliche Daten. In dieser Arbeit werden (mit Ausnahme von Kapitel 4) nur diskrete Daten behandelt.

Bei Punktprozessen liegt die räumliche Information in stetigen Koordinaten vor, wobei jeder Datenpunkt ein Fall ist. Abbildung 1.2 zeigt als Beispiel die räumlichen Daten einer Fall-Kontrollstudie zur Kindersterblichkeit. Die Glättung der räumlichen Komponente erfolgt dann zum Beispiel durch zweidimensionale Splines oder räumliche Kernschätzer (Kelsall & Diggle 1998).

Dagegen werden bei Daten mit diskreter räumlicher Information die Fälle oft aggregiert. Die Unterteilung der geographischen Einheiten beruht dabei meist auf administrativen Grenzen (z.B. Gemeinden, Regierungsbezirke, Länder), kann unter Umständen aber auch künstlich gewählt werden (vgl. Kapitel 4). Im Regelfall sind dabei geographische Unterteilung und zu untersuchende Krankheit voneinander unabhängig.

Im Vergleich haben Punktdaten die genaueren räumlichen Informationen, die bei Aggregation verloren gehen. Aus Plausibilitätsgründen ist jedoch die Genauigkeit einer räum-

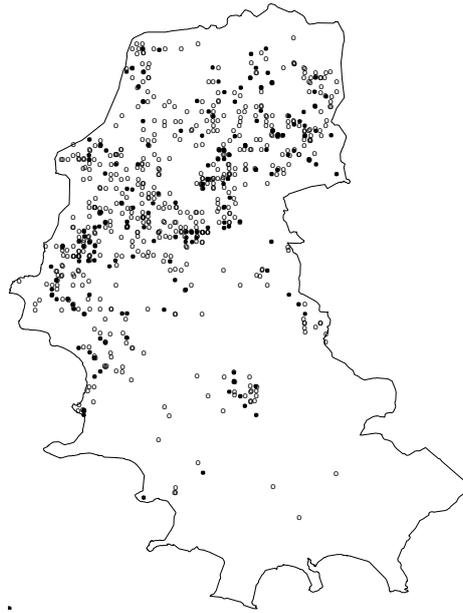


Abbildung 1.2: Daten einer Fall-Kontrollstudie über Kindersterblichkeit in Porto Alegre, Brasilien. Abgebildet sind die Wohnorte der Mütter der untersuchten Kinder. Punkte sind Fälle, Kreise sind Kontrollfälle. Vgl. Kapitel 4.

lichen Punktinformation oft fraglich, so wird z.B. bei der Wildtieruntersuchung in Kapitel 3 nicht der genaue Schußort der untersuchten Tiere als Information verwendet, weil die Tiere keinen punktförmigen Lebensraum haben; zudem ist der Schußort meist speziell ausgezeichnet (Fütterungsplätze etc.). Aus praktischen Gründen ist es meist auch schwer die genaue räumliche Punktinformation zu erhalten. Insbesondere Daten über Krankheiten liegen aus Datenschutzgründen oft auch nur aggregiert vor.

## 1.2 Inhaltsübersicht

Kapitel 2 stellt die grundlegenden Methoden und Techniken vor, mit denen in dieser Arbeit epidemiologische Daten analysiert werden. Die dafür eingesetzten hierarchischen Bayesianischen Modelle können nur mit effizienten Markovkern-Monte-Carlo-Methoden geschätzt werden. In diesen Modellen finden autoregressive Prioris Anwendung, wie Gauss-Markov-Zufallsfelder, Random Walks oder daraus abgeleitete Raum-Zeit-Interaktionen. Die weite-

ren Kapitel zeigen unterschiedliche Modelle jeweils an ihren Anwendungen.

Kapitel 3 behandelt die Anwendung von Gauss-Markov-Zufallsfeldern im Kontext von Wildtierkrankheiten. Diese Modellierung wird mit dem in der Veterinärepidemiologie verbreiteten Beta-Binomial-Modell verglichen. Mit einer von Holmes & Held (2003) vorgestellten Modellformulierung mit Hilfsvariablen wird das Modell um die Schätzung von Kovariableneffekten erweitert. Die Daten für dieses Kapitel wurden vom Bundesforschungsinstitut für Tiergesundheit zur Verfügung gestellt. Die Modellerstellung erfolgte in Zusammenarbeit mit dem Bundesforschungsinstitut, welches die im Rahmen dieser Arbeit entstandene Software auch für eigene Auswertungen einsetzt. Teile dieses Kapitels basieren auf Staubach, Schmid, Knorr-Held & Ziller (2002).

In Kapitel 4 liegt die räumliche Information der Daten stetig vor. Wir diskretisieren die Daten und wenden die für diskrete räumliche Daten übliche Methodik an. Da die Diskretisierung über ein reguläres Gitter erfolgt, sind hier verschiedenste Nachbarschaftsdefinitionen möglich, während bei den sonst in dieser Arbeit vorliegenden irregulären Gittern die Nachbarschaft immer als Besitz einer gemeinsamen Grenze definiert ist. Das Modell wird mit der Modellierung als räumlicher Punktprozeß von Shimakura, Carvalho, Aerts & Flores (2001) verglichen, die freundlicherweise die Daten zur Verfügung stellten.

Ein gern angewandter Ansatz speziell in der Krebsepidemiologie sind Alters-Perioden-Kohorten(APC)-Modelle. Kapitel 5 zeigt die Bayesianische Analyse von APC-Modellen. Neben dem von Knorr-Held & Rainer (2001) vorgestellten Modell mit zusätzlicher Heterogenität wird auch die technisch aufwendigere Modellierung ohne zusätzliche Heterogenität und die Aufnahme weiterer unstrukturierter Zeiteffekte gezeigt. Besonderes Augenmerk wird auch auf die Erstellung von Prognosen gelegt. Die verschiedenen Modelle werden anhand eines Datensatzes zu COPD in England und Wales verglichen — die Daten wurden vom Office for National Statistics, Großbritannien zur Verfügung gestellt. Die in diesem Kapitel vorgestellten Analysen können alle mit der Software BAMP durchgeführt werden, die im Rahmen dieser Arbeit entstand und in Anhang B beschrieben ist.

Kapitel 6 baut auf Kapitel 5 auf. Das APC-Modell mit zusätzlicher Heterogenität wird

hier auf räumliche Daten angewandt. Dazu wird das Modell um eine räumliche Komponente mit Gauss-Markov-Zufallsfelder-Priori erweitert und verschiedene Ansätze zur Modellierung von Raum-Zeit-Interaktionen vorgestellt. Die Modelle werden anhand eines Datensatzes über Magenkrebssterblichkeit bei Frauen in Westdeutschland verglichen. Die Daten wurden vom Deutschen Krebsforschungsinstitut zur Verfügung gestellt (Becker & Wahrendorf 1997). Diese Kapitel basiert zu großen Teilen auf Schmid & Held (2004).

Kapitel 7 erweitert ein von Hofmann, Hoehle & Held (2004) vorgestelltes stochastisches Modell für Surveillance-Daten infektiöser Krankheiten um eine räumlich-zeitliche Komponente. Ziel des Modells ist es, ein Alarmsystem für Ausbrüche infektiöser Krankheiten zu erstellen. Dieses System soll am Robert-Koch-Institut Anwendung finden, welches auch die Daten der Campylobacter-Infektionen in Deutschland zur Verfügung stellte. Der Vergleich mit dem Modell von Hofmann *et al.* (2004) zeigt, daß die räumlich-zeitliche Modellierung zumindest für die Campylobacter-Daten eine sinnvolle Verbesserung ist. Kapitel 8 rundet die Arbeit mit einigen Schlußbemerkungen ab.

# Kapitel 2

## Methoden und Techniken

Hierarchische Bayesianische Modelle sind eine Möglichkeit, räumliche epidemiologische Daten zu modellieren. Zweck dieser Modelle ist es, räumliche und bzw. oder zeitliche Effekte zu diskriminieren und gleichzeitig mittels Glättungsprioris glatt zu schätzen. Die Schätzung der Glättungsparameter erfolgt dabei ebenfalls im Modell.

Inferenz ist in diesen Modellen nur mit MCMC-Methoden möglich. Die von Rue (2001) vorgestellten Methoden ermöglichen dabei eine effiziente Ziehung aus der Posteriori. Mit dem von Spiegelhalter, Best, Carlin & van der Linde (2002) vorgeschlagenen DIC lassen sich verschiedene hierarchische Modelle vergleichen.

### 2.1 Hierarchische Bayesianische Modelle für epidemiologische Fragestellungen

Zumeist wird in der Epidemiologie mit einer von zwei Grundmodellannahmen gearbeitet, entweder mit dem Log-Poisson-Modell oder mit dem Logit-Binomial-Modell. Gehen wir im Folgenden von Fallzahlen einer Krankheit aus, welche für  $I$  Regionen vorliegen.

## Binomial-Modell

Beim Binomial-Modell geht man davon aus, daß die Anzahl der Fälle in einer Region  $i$  binomialverteilt ist mit unbekannter Wahrscheinlichkeit  $p_i$  und der Anzahl der unter Risiko stehenden Personen  $n_i$ . Auf der nächsten Ebene des hierarchischen Modell spezifiziert man den Logit der Wahrscheinlichkeit

$$\eta_i = \text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$$

für jede Region  $i$ . Den *linearen Prädiktor*  $\eta_i$  kann man weiter zerlegen, z.B. in einen strukturierten und einen unstrukturierten räumlichen Effekt (Besag, York & Mollié 1991). Die Festlegung der „Struktur“ des strukturierten Effekts wird dabei durch die Priori spezifiziert, z.B. durch ein Gauss-Markov-Zufallsfeld, siehe Abschnitt 2.2.1. Entsprechend wird auch der unstrukturierte Effekt durch unabhängige Prioris für jede Region modelliert.

Auf einer weiteren Ebene des hierarchischen Modells werden Prioris für die Varianzparameter festgelegt. Damit wird auch die Glättung der Effekte im Modell mitgeschätzt, siehe auch Abschnitt 2.2.4.

In Fällen, wo man für jede Beobachtung weitere individuelle Kovariableninformationen hat, arbeitet man statt dessen gern mit dem Bernoulli-Modell. Hier ist jede Beobachtung Bernoulliverteilt

$$y_j \sim B(1, p_j).$$

Auch hier modelliert man wieder auf einer weiteren hierarchischen Ebene den Logit der Wahrscheinlichkeit  $p_j$ , wobei nun neben der räumlichen Information  $i(j)$ , dem Ort des Individuums  $j$ , auch weitere Kovariablen eingehen können. Beispiele hierfür sind in den Kapiteln 3 und 4 zu finden.

## Poisson-Modell

Beim häufiger verwendeten Poisson-Modell geht man davon aus, daß die Anzahl der Fälle in Region  $i$  Poissonverteilt mit Parameter  $\lambda_i e_i$  ist. Dabei bezeichnet  $\lambda_i$  einen unbekanntem

## 2.1 Hierarchische Bayesianische Modelle für epidemiologische Fragestellungen

Parameter und  $e_i$  entweder die (bekannte) Anzahl der unter Risiko stehenden Personen oder die (bekannte) Anzahl erwarteter Fälle. Die Anzahl erwarteter Fälle erhält man zum Beispiel durch Standardisierung bezüglich der Altersstruktur.

Auf der nächsten hierarchischen Ebene modelliert man nun den Logarithmus des unbekanntes Parameters  $\log(\lambda_i)$  analog zu oben. Für kleine Werte von  $p_i$  und große Werte von  $n_i$ , die in der Epidemiologie praktisch immer vorliegen, ist das Log-Poisson-Modell approximativ gleich dem exakten Logit-Binomial-Modell.

Vorteil des Poisson-Modells gegenüber dem Binomial-Modell ist die Möglichkeit der Standardisierung der Daten z.B. bezüglich der Altersstruktur, aber auch bezüglich anderer Kovariablen. In dieser Arbeit werden Kovariablen aber meist in das Modell mit aufgenommen, deswegen arbeiten wir hauptsächlich mit dem Binomial-Modell.

### Bayes-Prinzip

Die Parameter des hierarchischen Modells werden durch Anwendung der Bayes'schen Formel geschätzt. Aus der Likelihood des gewählten Modells  $l(\mathbf{y}|\mathbf{h})$  und den Prioris  $p(\mathbf{h})$  ergibt sich die Posteriori  $p(\mathbf{h}|\mathbf{y})$  durch

$$p(\mathbf{h}|\mathbf{y}) = \frac{p(\mathbf{h})l(\mathbf{y}|\mathbf{h})}{\int p(\mathbf{H})l(\mathbf{y}|\mathbf{H})d\mathbf{H}}, \quad (2.1)$$

wobei  $\mathbf{h}$  die Menge der Parameter des Modells ist und  $\mathbf{y}$  die beobachteten Daten.

Schlüsse werden nach dem Bayes-Prinzip nur aus der Posteriori gezogen (Rüger 1999, Kapitel 2.4). Da diese in den meisten Fällen nur bis auf die Normalisierungskonstante bekannt ist, müssen Markovkern-Monte-Carlo-Methoden (engl. *Markov chain Monte Carlo*, MCMC) angewendet werden, um die Parameter zu schätzen (siehe Abschnitt 2.3).

## 2.2 Prioriverteilungen

Die Prioriverteilung drückt das Vorwissen aus, welches wir über die Parameter haben. Geht man davon aus, daß die Effekte in irgendeiner Weise „glatt“ sind, so verwendet man *Glättungsprioris*. So werden Prioris bezeichnet, deren Zweck es ist, eine glatte Schätzung eines Effekts zu erhalten. In dieser Arbeit werden fast nur Prioris verwendet, die keine Vorinformation über den Mittelwert eines Effekts enthalten, sondern nur eine Priori auf die Differenzen von einzelnen Parametern legen, sogenannte *autoregressive* Prioris.

### 2.2.1 Gauss-Markov-Zufallsfelder

Die wohl am häufigsten verwendete Glättungspriori für diskrete räumliche Daten sind Markov-Zufallsfelder (Besag *et al.* 1991). Diese beschreiben eine gemeinsame Verteilung eines Zufallsvektor, um die Abhängigkeiten zwischen räumlichen Einheiten zu beschreiben. Am bekanntesten sind Gauss-Markov-Zufallsfelder (engl. *Gaussian Markov random fields*, GMRF), bei denen die gemeinsame Verteilung eine multivariate Normalverteilung ist. Auch andere Verteilungen sind zur Beschreibung von Markov-Zufallsfeldern möglich (siehe z.B. Besag 1974). Diese werden in jüngster Zeit im Zusammenhang mit robusten GMRF-Modellen genauer untersucht (Gelfand & Vounatsou 2003).

Grundprinzip der GMRF's ist die (Priori-)Annahme, daß benachbarte räumliche Einheiten nicht zu sehr voneinander abweichen. Zwei Regionen sollen als benachbart gelten, wenn sie eine gemeinsame Grenze haben. Dies ist die für irreguläre Gitter wie Landkarten übliche Vorgehensweise. Nachbarschaft und damit auch GMRFs lassen sich aber auch anders definieren. Eine allgemeinere Definition von GMRFs und verschiedene Arten von Nachbarschaft werden Kapitel 4 ausführlicher behandelt. Durch die Berücksichtigung der Nachbarschaftsstruktur werden zusätzlich für jeden Pixel auch die Informationen der Daten der benachbarten Pixel ausgenutzt, man spricht hierbei von *borrowing strength*.

Gehen wir im Folgenden von einem interessierenden Parametervektor  $\boldsymbol{\eta} = (\eta_1, \dots, \eta_n)'$  auf  $n$  räumlichen Einheiten aus. Man formuliert also eine bedingte Verteilung auf  $\eta_i$  gegeben

alle anderen Parameter nur in Abhängigkeit der Menge der Nachbarn  $\partial(i)$ ,

$$\eta_i | \boldsymbol{\eta}_{-i} = \eta_i | \boldsymbol{\eta}_{\{j:j \in \partial(i)\}} \sim N \left( \left( \frac{1}{|\partial(i)|} \sum_{\{j:j \in \partial(i)\}} \eta_j \right), \frac{1}{|\partial(i)|\kappa} \right). \quad (2.2)$$

Dabei bezeichnet  $\boldsymbol{\eta}_{-i}$  den Vektor  $(\eta_1, \dots, \eta_{i-1}, \eta_{i+1}, \dots, \eta_n)'$  und  $|\partial(i)|$  die Anzahl der Nachbarn von  $i$ . In dieser Arbeit wird durchgehend mit der Präzision statt mit der Varianz von Normalverteilungen gearbeitet ( $\sigma^2 = \kappa^{-1}$ ), weil sich vieles dadurch kompakter schreiben läßt. Eine alternative Notation für die Nachbarschaft von  $i$  und  $j$  ist  $i \sim j$ . Mit der zusätzlichen Abkürzung  $\boldsymbol{\eta}_{\partial(i)} = \boldsymbol{\eta}_{\{j:j \in \partial(i)\}}$  läßt sich (2.2) dann schreiben als

$$\eta_i | \boldsymbol{\eta}_{\partial(i)} \sim N \left( \frac{1}{n_i} \sum_{i \sim j} \eta_j, \frac{1}{n_i \kappa} \right),$$

wobei  $n_i = |\partial(i)|$ .

Über das Hammersley-Clifford-Theorem (Hammersley & Clifford 1971; Besag 1974) läßt sich aus der bedingten lokalen Verteilung die gemeinsame Verteilung berechnen. Diese lautet

$$p(\boldsymbol{\eta}) \propto |\mathbf{Q}|^{1/2} \kappa^{rg(\mathbf{Q})/2} \exp \left( -\frac{\kappa}{2} \sum_{i \sim j} (\eta_i - \eta_j)^2 \right) = |\mathbf{Q}|^{1/2} \kappa^{rg(\mathbf{Q})/2} \exp \left( -\frac{\kappa}{2} \boldsymbol{\eta}' \mathbf{Q} \boldsymbol{\eta} \right). \quad (2.3)$$

Dabei ist  $|\mathbf{Q}|$  nach Besag & Higdon (1999) das Produkt der Nicht-Null-Eigenwerte von  $\mathbf{Q}$ . Der Parametervektor  $\boldsymbol{\eta}$  ist damit (uneigentlich) multivariat normalverteilt mit Mittelwertsvektor  $\mathbf{0}$  und Präzision  $\mathbf{Q}$ .

Bei (2.3) handelt es sich daher streng gesehen nicht um eine Dichte. Die Priori auf  $\boldsymbol{\eta}$  ist keine Verteilung (auf  $\mathbb{R}^n$ , wohl aber auf  $\mathbb{R}^{rg(\mathbf{Q})}$ ). Man spricht von einer uneigentlichen Verteilung, in diesem Fall auch von einer singulären multivariaten Normalverteilung. Zur Übersichtlichkeit werden im Folgenden trotzdem die Begriffe Dichte und Verteilung verwendet. Die Priori wird hier nur auf die Differenzen der Effekte gelegt, nicht aber auf den globalen Mittelwert.

Die Präzisionsmatrix  $\mathbf{Q}$  wird auch als *Nachbarschaftsmatrix* bezeichnet. Sie beschreibt die Nachbarschaftsstruktur der zu untersuchenden Regionen. Für die Elemente  $q_{ij}$  von  $\mathbf{Q}$

gilt

$$\begin{aligned} q_{ij} &= -1 && \text{für } i \sim j \\ q_{ij} &= 0 && \text{für } i \not\sim j \text{ und } i \neq j \\ q_{ii} &= |\partial(i)| && \text{für alle } i. \end{aligned}$$

Zum Beispiel zeigt Abbildung 2.1 die Nachbarschaftsmatrix für die deutschen Bundesländer.

Anders als Besag & Higdon (1999) verwenden wir nicht  $\kappa^n$  als ersten Term der Dichte der gemeinsamen Verteilung (2.3), sondern nach Knorr-Held (2003)  $\kappa^{rg(\mathbf{Q})}$ . Die Normalisierungskonstante läßt sich auch als  $|\kappa\mathbf{Q}|$ , also als Produkt der Nicht-Null-Eigenwerte von  $\kappa\mathbf{Q}$  schreiben. Davon gibt es  $rg(\kappa\mathbf{Q}) = rg(\mathbf{Q})$  Stück,  $\kappa$  läßt sich also  $rg(\mathbf{Q})$ -mal vorziehen. In der Regel hat  $\mathbf{Q}$  dabei den Rang  $n - 1$ , wenn das zu untersuchende Gebiet nicht in zwei oder mehr Teilgebiete zerfällt (dann fällt der Rang entsprechend weiter ab). Hodges, Carlin & Fan (2003) zeigen, daß die richtige Wahl der Normalisierungskonstante wichtig für eine unverzerrte Schätzung ist und empfehlen ebenfalls, den Term  $\kappa^{rg(\mathbf{Q})}$  zu verwenden.

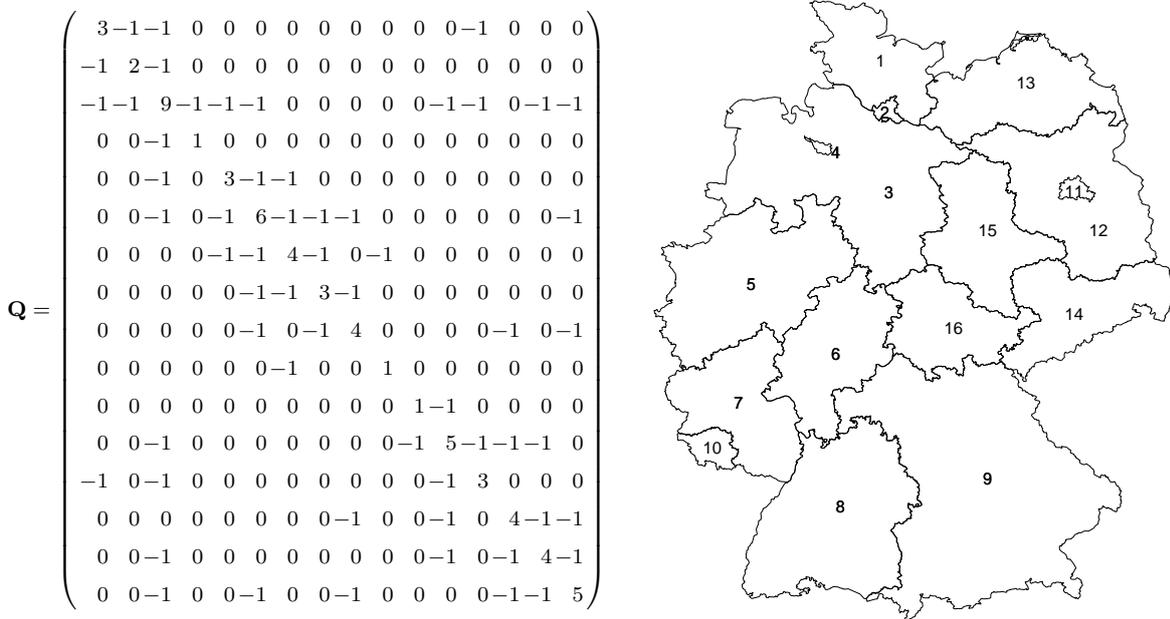


Abbildung 2.1: Nachbarschaftsmatrix der deutschen Bundesländer. Rechts die Karte der Bundesländer mit der Angabe der Zeilenindizes (bzw. Spaltenindizes).

### 2.2.2 Random Walk Prioris

Der eindimensionale Spezialfall von Gauss-Markov-Zufallsfeldern sind *Random Walk Prioris*. Diese finden vor allem bei der Modellierung von Parametern Anwendung, welche eine zeitliche Komponente beschreiben, aber z.B. auch bei Alterseffekten. Sei im Folgenden  $\boldsymbol{\phi} = (\phi_1, \dots, \phi_J)$  ein interessierender Parametervektor.

Grundlegende Idee der Random Walk Prioris ist die Annahme, daß der Effekt zum Zeitpunkt  $j$  nicht sehr stark vom Effekt zum Zeitpunkt  $j - 1$  abweicht. Genauer geht man davon aus, daß für jeden Zeitpunkt  $j = 2, \dots, J$  die bedingte Verteilung von  $\phi_j$  gegeben die Effekte aller anderen Zeitpunkte  $\boldsymbol{\phi}_{-j}$  nur vom jeweils vorherigen Zeitpunkt abhängt. Es gilt also

$$\phi_j | \boldsymbol{\phi}_{-j} = \phi_j | \phi_{j-1} \sim N(\phi_{j-1}, \kappa^{-1}) \text{ für } j = 2, \dots, J. \quad (2.4)$$

Setzt man nun für den ersten Zeitpunkt

$$p(\phi_1) \propto \text{const.}$$

an, so ergibt sich die gemeinsame (Priori-)Verteilung als

$$\begin{aligned} p(\boldsymbol{\phi}) &= p(\phi_1, \phi_2, \dots, \phi_J) \\ &= p(\phi_1) p(\phi_2 | \phi_1) \cdots p(\phi_J | \phi_{J-1}) \\ &\propto \prod_{j=2}^J \sqrt{\frac{\kappa}{2\pi}} \exp \left[ -\frac{\kappa}{2} (\phi_j - \phi_{j-1})^2 \right] \\ &= \left( \frac{\kappa}{2\pi} \right)^{(J-1)/2} \exp \left[ -\frac{\kappa}{2} \sum_{j=2}^J (\phi_j - \phi_{j-1})^2 \right] \\ &= \left( \frac{\kappa}{2\pi} \right)^{(J-1)/2} \exp \left[ -\frac{\kappa}{2} \boldsymbol{\phi}' \mathbf{R}^{(1)} \boldsymbol{\phi} \right]. \end{aligned} \quad (2.5)$$





Dann ist  $\mathbf{K} = \mathbf{A} \otimes \mathbf{B}$  eine  $(mp) \times (nq)$ -Matrix mit den Einträgen

$$\mathbf{K} = \begin{pmatrix} a_{11}b_{11} & \cdots & a_{11}b_{1q} & \cdots & a_{1n}b_{11} & \cdots & a_{1n}b_{1q} \\ \vdots & & \vdots & & \vdots & & \vdots \\ a_{11}b_{p1} & \cdots & a_{11}b_{pq} & \cdots & a_{1n}b_{p1} & \cdots & a_{1n}b_{pq} \\ \vdots & & \vdots & & \vdots & & \vdots \\ a_{m1}b_{11} & \cdots & a_{m1}b_{1q} & \cdots & a_{mn}b_{11} & \cdots & a_{mn}b_{1q} \\ \vdots & & \vdots & & \vdots & & \vdots \\ a_{m1}b_{p1} & \cdots & a_{m1}b_{pq} & \cdots & a_{mn}b_{p1} & \cdots & a_{mn}b_{pq} \end{pmatrix}.$$

Sei  $\mathbf{R}$  die Präzisionsmatrix der Dimension  $T \times T$  eines zeitlichen Random Walks und  $\mathbf{Q}$  die Präzisionsmatrix der Dimension  $N \times N$  eines räumlichen GMRFs, sowie  $\mathbf{I}_T$  bzw.  $\mathbf{I}_N$  Einheitsmatrizen der Dimension  $T \times T$  bzw.  $N \times N$ . Dann lassen sich nach Knorr-Held (2000) vier Typen von Raum-Zeit-Interaktionen formulieren:

Typ I	unabhängige Interaktionen	$\mathbf{P} = \mathbf{I}_N \otimes \mathbf{I}_T$
Typ II	unabhängige Random Walks pro Region	$\mathbf{P} = \mathbf{I}_N \otimes \mathbf{R}$
Typ III	unabhängige GMRFs pro Zeitpunkt	$\mathbf{P} = \mathbf{Q} \otimes \mathbf{I}_T$
Typ IV	abhängige Interaktion über Raum und Zeit	$\mathbf{P} = \mathbf{Q} \otimes \mathbf{R}$

Bei der komplexesten Form Typ IV ist also die lokale bedingte Verteilung eines Raum-Zeit-Interaktionsparameter abhängig von den zeitlichen und räumlichen Nachbarn, aber auch von den zeitlichen Nachbarn der räumlichen Nachbarn (oder den räumlichen Nachbarn der zeitlichen Nachbarn). Die vorgestellte Definition von Interaktionen wird zum Beispiel in Kapitel 6 angewandt, aber auch die Definition der Modelle für GMRFs auf regulären Gittern in Kapitel 4 können als Interaktionen gesehen werden.

## 2.2.4 Hyperparameter

Die gezeigten Prioris sind alle (singuläre) multivariate Normalverteilungen mit Präzisionsmatrix  $\kappa\mathbf{Q}$ . Während die Matrix  $\mathbf{Q}$  die Nachbarschaftsbeziehungen beschreibt und somit fest vorgegeben ist, ist der Präzisionsparameter  $\kappa$  ein globaler Glättungsparameter. Dieser sogenannte Hyperparameter kann im hierarchischen Modell parallel geschätzt werden.

In der Regel wird für den Präzisionsparameter eine uninformative Gamma-Priori  $\text{Ga}(a, b)$  verwendet. In der Praxis setzt man meist  $a = 1$  und  $b$  auf einen sehr kleinen Wert ( $10^{-2}$  bis  $10^{-5}$ ).

Der Glättungsparameter wird hier global gewählt, das heißt, daß zum Beispiel bei einem Gauss-Markov-Zufallsfeld über den gesamten Untersuchungsbereich gleich geglättet wird. Gelegentlich ist diese globale Glättung aber unangebracht. In neuerer Zeit werden daher verstärkt lokal adaptive Verfahren untersucht. Speziell für Raum-Zeit-Interaktionen ist eine globale Glättung öfter problematisch, insbesondere, da hier über Raum und Zeit gleich geglättet wird. Dafür läßt sich leicht Abhilfe schaffen, indem man die Präzisionsmatrizen der Haupteffekte bei der Bildung der Präzisionsmatrix der Interaktion (vgl. Abschnitt 2.2.3) gewichtet

$$\mathbf{P} = \kappa_1 \mathbf{Q} \otimes \kappa_2 \mathbf{R}.$$

Dieses Verfahren findet in Kapitel 4 Anwendung. Bei den Interaktionen nach Typ II oder Typ III kann ein weiterer Ansatz sein, für jedes der unabhängigen Gauss-Markov-Zufallsfelder bzw. Random Walks einen eigenen Hyperparameter zu verwenden

$$\mathbf{P} = \mathbf{Q} \otimes \text{diag}(\kappa_1, \dots, \kappa_T)$$

beziehungsweise

$$\mathbf{P} = \text{diag}(\kappa_1, \dots, \kappa_N) \otimes \mathbf{R}.$$

## 2.3 MCMC-Methoden

Bei der Berechnung der Posteriori mit der Bayes-Formel (2.1) ist bei den vorgestellten Modellen die explizite Berechnung der Normalisierungskonstante  $\int p(\mathbf{H})l(\mathbf{y}|\mathbf{H})d\mathbf{H}$  nicht möglich. Die Posteriori ist also analytisch nicht zugänglich. Man verwendet daher Markovkern-Monte-Carlo-Verfahren (*Markov chain Monte Carlo, MCMC*), bei denen es genügt, die Posteriori bis auf die Normalisierungskonstante zu kennen.

Für jeden Parameter  $\theta_i$  berechnet man die volle bedingte Posterioriverteilung (*Full*

*Conditional*)  $p(\theta_i|\mathbf{h}, \mathbf{y})$ , also die Verteilung des Parameters gegeben die Daten  $\mathbf{y}$  und die restlichen Parameter  $\mathbf{h}$ . Häufig ist es sinnvoll, die Full Conditional für ganze Parametervektoren  $\boldsymbol{\theta}$  anzugeben. Sie berechnet sich als

$$p(\boldsymbol{\theta}|\cdot) = p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{h}) \propto p(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta}).$$

Man spricht dann von Blockverfahren.

Ist die Full Conditional eine Standardverteilung, so kann man den Parameter(vektor) direkt daraus ziehen (*Gibbs-Sampler*, siehe dazu auch Abschnitt 2.3.1). Andernfalls kann man einen Metropolis-Hastings-Algorithmus anwenden. Man wählt sich dazu eine Vorschlagsdichte (*Proposal*)  $q(\theta^*|\theta)$ , welche der Full Conditional möglichst ähnlich ist (siehe auch Abschnitt 2.3.2).

### Metropolis-Hastings-Algorithmus

1. Wähle einen Startwert  $\theta^0$  sowie die Anzahl der Iterationen  $D$ .
2. Setze  $d = 1$ .
3. Ziehe eine Zufallszahl  $\theta^*$  gemäß der Vorschlagsdichte  $q(\theta^*|\theta^{(d-1)})$ .
4. Akzeptiere  $\theta^*$  mit der Akzeptanzwahrscheinlichkeit

$$\alpha = \min \left\{ \frac{p(\theta^*|\cdot)q(\theta^{(d-1)}|\theta^*)}{p(\theta^{(d-1)}|\cdot)q(\theta^*|\theta^{(d-1)})}, 1 \right\},$$

andernfalls setze  $\theta^{(d)} = \theta^{(d-1)}$ .

5. Erhöhe  $d$  um 1.
6. Wiederhole Schritte 3 bis 5, bis  $d = D$ .

In den allermeisten Fällen wird man einen Hybrid-Algorithmus verwenden, d.h. einen Teil der Parameterblöcke in Gibbs-Schritten ziehen, den anderen Teil mit dem Metropolis-Hastings-Algorithmus. Für weitere Details zu MCMC-Verfahren siehe z.B. Gilks, Richardson & Spiegelhalter (1996).

### 2.3.1 Ziehen aus multivariaten Normalverteilungen

Mit den in Abschnitt 2.2 vorgestellten GRMF- und Random Walk-Prioris sowie den analog gebildeten Raum-Zeit-Interaktionen erhält man oft eine multivariate Normalverteilung als Full Conditional der entsprechenden Parametervektoren. Die Ziehung aus einer multivariaten Normalverteilung kann mit den von Rue (2001) vorgestellten Methoden sehr effizient erfolgen.

Sei die bedingte Posterioriverteilung eines Parametervektors  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)$  eine  $n$ -dimensionale Normalverteilung mit Mittelwertsvektor  $\mathbf{m}$  und Präzisionsmatrix  $\mathbf{Q}$ , so läßt sich die Dichte schreiben als

$$\begin{aligned} p(\boldsymbol{\theta}|\cdot) &= (2\pi)^{-n/2} \det(\mathbf{Q})^{1/2} \exp \left[ -\frac{1}{2} (\boldsymbol{\theta} - \mathbf{m})' \mathbf{Q} (\boldsymbol{\theta} - \mathbf{m}) \right] \\ &\propto \exp \left[ -\frac{1}{2} (\boldsymbol{\theta}' \mathbf{Q} \boldsymbol{\theta} - 2 \boldsymbol{\theta}' \mathbf{Q} \mathbf{m} + \mathbf{m}' \mathbf{Q} \mathbf{m}) \right] \\ &\propto \exp \left[ -\frac{1}{2} \boldsymbol{\theta}' \mathbf{Q} \boldsymbol{\theta} + \boldsymbol{\theta}' \mathbf{b} \right]. \end{aligned} \quad (2.8)$$

Dabei ist  $\mathbf{b} = \mathbf{Q} \mathbf{m}$ . Man spricht hierbei von *kanonischer Repräsentation* der Normalverteilung

$$\boldsymbol{\theta}|\cdot \sim N_C(\mathbf{b}, \mathbf{Q}).$$

Wir verwenden im Folgenden zwar meist die *kanonischen Parameter*  $\mathbf{b}$  und  $\mathbf{Q}$ , verwenden aber die herkömmliche Schreibweise

$$\boldsymbol{\theta}|\cdot \sim N(\mathbf{Q}^{-1} \mathbf{b}, \mathbf{Q}^{-1}).$$

Nach Rue (2001) läßt sich eine Stichprobe von  $\boldsymbol{\theta}$  durch folgenden Algorithmus ziehen:

#### Rue-Block-Algorithmus

1. Berechne die Cholesky-Zerlegung  $\mathbf{Q} = \mathbf{L} \mathbf{L}'$
2. Löse  $\mathbf{L} \mathbf{w} = \mathbf{b}$
3. Löse  $\mathbf{L}' \boldsymbol{\mu} = \mathbf{w}$
4. Ziehe einen Zufallsvektor  $\mathbf{z} \sim N(\mathbf{0}, \mathbf{I}_n)$



### Restringierte Ziehung

In der Regel wird man auf die Parametervektoren Zentrierungsrestriktionen legen. Sei allgemein eine Restriktion auf den Parametervektor  $\boldsymbol{\theta}$  der Art

$$\mathbf{A}\boldsymbol{\theta} = \mathbf{e}$$

gegeben. Dann läßt sich nach Rue (2001) aus der unrestringierten Stichprobe  $\boldsymbol{\theta}$  aus (2.8) eine restringierte Stichprobe  $\boldsymbol{\theta}^*$  erhalten, indem man

$$\boldsymbol{\theta}^* = \boldsymbol{\theta} - \mathbf{Q}^{-1}\mathbf{A}'(\mathbf{A}\mathbf{Q}^{-1}\mathbf{A}')^{-1}(\mathbf{A}\boldsymbol{\theta} - \mathbf{e})$$

berechnet.

### 2.3.2 Approximation einer Normalverteilung durch Taylorentwicklung

Nicht immer hat die bedingte Verteilung eines räumlichen Parameters die Form (2.8). Sei der Parametervektor  $\boldsymbol{\theta}$  a priori nach einem GMRF mit Präzisionsmatrix  $\mathbf{Q}$  verteilt und  $l(\boldsymbol{\theta})$  der von  $\boldsymbol{\theta}$  abhängige Teil der Log-Likelihood des Modells. Dann läßt sich die bedingte Posterioriverteilung allgemein schreiben als

$$p(\boldsymbol{\theta}|\cdot) \propto \exp\left(-\frac{\kappa}{2}\boldsymbol{\theta}'\mathbf{Q}\boldsymbol{\theta} + l(\boldsymbol{\theta})\right). \quad (2.9)$$

Meist läßt sich die Log-Likelihood weiter zerlegen

$$l(\boldsymbol{\theta}) = \boldsymbol{\theta}'\mathbf{P}\boldsymbol{\theta} + \boldsymbol{\theta}'\mathbf{m} + f(\boldsymbol{\theta})$$

Ist  $f(\boldsymbol{\theta}) = 0$ , so ist (2.9) die Dichte einer multivariaten Normalverteilung und der obige Gibbs-Sampler läßt sich anwenden. Andernfalls ist ein Metropolis-Hastings-Algorithmus notwendig, um aus der bedingten Verteilung zu ziehen.

Als Vorschlagsverteilung im Metropolis-Hastings-Algorithmus kann man häufig eine multivariate Normalverteilung verwenden, deren Dichte approximativ (2.9) entspricht. Sei

$\boldsymbol{\theta}^0$  der Zustand, in dem der Parametervektor momentan ist. Wir gehen davon aus, daß  $f$  an der Stelle  $\boldsymbol{\theta}^0$  mindestens zweimal stetig differenzierbar ist (in den behandelten Modellen ist dies auch der Fall). Dann berechnet man eine (mehrdimensionale) Taylorentwicklung zweiten Grades von  $f(\boldsymbol{\theta})$  um  $\boldsymbol{\theta}^0$ . Sie lautet

$$\begin{aligned} f(\boldsymbol{\theta}) &= f(\boldsymbol{\theta}^0) + \sum_{k=1}^n (\theta_k - \theta_k^0) \frac{\partial}{\partial \theta_k} f(\boldsymbol{\theta}^0) + \frac{1}{2} \sum_{k=1}^n \sum_{l=1}^n (\theta_k - \theta_k^0) (\theta_l - \theta_l^0) \frac{\partial^2}{\partial \theta_k \partial \theta_l} f(\boldsymbol{\theta}^0) \\ &\quad + \text{Rest}(\boldsymbol{\theta}, \boldsymbol{\theta}^0) \\ &= \sum_{k=1}^n \theta_k \left( \frac{\partial}{\partial \theta_k} f(\boldsymbol{\theta}^0) - \sum_{l=1}^n \theta_l^0 \frac{\partial^2}{\partial \theta_k \partial \theta_l} f(\boldsymbol{\theta}^0) \right) + \frac{1}{2} \sum_{k=1}^n \sum_{l=1}^n \theta_k \theta_l \frac{\partial^2}{\partial \theta_k \partial \theta_l} f(\boldsymbol{\theta}^0) \\ &\quad + \text{Rest}(\boldsymbol{\theta}, \boldsymbol{\theta}^0) + C. \end{aligned}$$

Dabei ist  $C$  eine nicht von  $\boldsymbol{\theta}$  abhängige Konstante. Den  $\text{Rest}(\boldsymbol{\theta}, \boldsymbol{\theta}^0)$  vernachlässigen wir für die Approximation. Dann lautet unsere Proposal

$$q(\boldsymbol{\theta}|\boldsymbol{\theta}^0, \cdot) \propto \exp \left( -\frac{1}{2} \boldsymbol{\theta}' (\mathbf{Q} + \mathbf{P} + \mathbf{H}_f(\boldsymbol{\theta}^0)) \boldsymbol{\theta} + \boldsymbol{\theta}' (\mathbf{m} + \mathbf{f}'(\boldsymbol{\theta}^0)) \right).$$

Dabei ist  $\mathbf{H}_f(\boldsymbol{\theta}^0)$  die Hessematrix von  $f$  ausgewertet an der Stelle  $\boldsymbol{\theta}^0$  und

$$\mathbf{f}'(\boldsymbol{\theta}^0) = \left( \frac{\partial}{\partial \theta_1} f(\boldsymbol{\theta}^0), \dots, \frac{\partial}{\partial \theta_n} f(\boldsymbol{\theta}^0) \right)'.$$

Der Parametervektor  $\boldsymbol{\theta}$  wird dann gemäß dem Algorithmus von Rue (2001, siehe Abschnitt 2.3.1) gezogen und mit Akzeptanzwahrscheinlichkeit  $\alpha$  angenommen. Diese berechnet sich als

$$\alpha = \min \left( \frac{p(\boldsymbol{\theta}|\cdot)q(\boldsymbol{\theta}^0|\boldsymbol{\theta})}{p(\boldsymbol{\theta}^0|\cdot)q(\boldsymbol{\theta}|\boldsymbol{\theta}^0)}, 1 \right).$$

Bei der Berechnung der Akzeptanzwahrscheinlichkeit ist es notwendig, die Normalisierungskonstante des Proposals  $q(\boldsymbol{\theta}^0|\boldsymbol{\theta})$  bzw. von  $q(\boldsymbol{\theta}|\boldsymbol{\theta}^0)$  explizit auszurechnen. Die Normalisierungskonstante von  $q(\boldsymbol{\theta}|\boldsymbol{\theta}^0)$  ergibt sich dabei als Nebenprodukt beim Ziehen aus der multivariaten Normalverteilung im Rue-Algorithmus. Insbesondere gilt  $|\mathbf{Q}|^{1/2} = \prod \mathbf{L}_{ii}$  für die Cholesky-Zerlegung  $\mathbf{Q} = \mathbf{L}\mathbf{L}'$ . Für weitere Details siehe Rue (2001).

Wie man leicht sieht, gilt für dummykodierte Variablen, also insbesondere für die Raum- und Zeiteffekte sowie die Raum-Zeit-Interaktionen

$$\frac{\partial^2}{\partial \theta_k \partial \theta_l} f(\boldsymbol{\theta}^0) = 0.$$

Damit ist die Hessematrix nur auf der Hauptdiagonalen besetzt. Da auch  $\mathbf{P}$  in der Regel eine Diagonalmatrix ist, bleibt die Bandstruktur der Präzisionsmatrix in der approximativen Normalverteilung erhalten. Für einen Block nicht dummykodierter Variablen ist die Präzisionsmatrix dagegen voll besetzt (siehe die Ziehung der Kovariableneffekte in Abschnitt 3.3.3 und der Saisonkomponente in Abschnitt 7.3.2).

### 2.3.3 Gemeinsames Ziehen von Parametervektor und Präzisionsparameter

Hat ein Parametervektor  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)'$  eine GMRF-Priori mit Nachbarschaftsmatrix  $\mathbf{Q}$  und Präzisionsparameter  $\kappa$ , so erhält man beim getrennten Ziehen von  $\boldsymbol{\theta}$  und  $\kappa$  oft eine schlechte Mischung des Samples (vgl. Abschnitt 2.4). Knorr-Held & Rue (2002) schlagen daher vor, den Parametervektor  $\boldsymbol{\theta}$  und den Präzisionsparameter  $\kappa$  gemeinsam zu ziehen. Diese Technik wird beispielsweise in Abschnitt 4.3 verwendet.

Man geht dazu in jeder Iteration wie folgt vor:

#### Joint-Update-Algorithmus

1. Ziehe  $\kappa''$  bei gegebenem aktuellem Zustand  $\kappa'$  aus einer Proposalverteilung  $q(\kappa''|\kappa')$ .
2. Ziehe  $\boldsymbol{\theta}''$ , eventuell gegeben dem aktuellen Zustand  $\boldsymbol{\theta}'$ , aus einer Proposalverteilung  $q(\boldsymbol{\theta}''|\kappa'', \boldsymbol{\theta}')$ . (Falls möglich wird man  $q(\boldsymbol{\theta}''|\kappa'', \boldsymbol{\theta}') = p(\boldsymbol{\theta}''|\kappa'', \cdot)$  setzen, also  $\boldsymbol{\theta}''$  unabhängig von  $\boldsymbol{\theta}'$  direkt ziehen können.)
3. Akzeptiere  $(\boldsymbol{\theta}'', \kappa'')$  **gemeinsam** mit Akzeptanzwahrscheinlichkeit

$$\alpha = \frac{p(\boldsymbol{\theta}'', \kappa''|\cdot)q(\boldsymbol{\theta}'|\kappa', \boldsymbol{\theta}'')q(\kappa'|\kappa'')}{p(\boldsymbol{\theta}', \kappa'|\cdot)q(\boldsymbol{\theta}''|\kappa'', \boldsymbol{\theta}')q(\kappa''|\kappa')}$$

andernfalls übernahm den alten Zustand  $(\boldsymbol{\theta}', \kappa')$  als Neuen.

## 2.4 Auswertung

### 2.4.1 Zum Umgang mit den Ergebnissen

Aussagen über die Parameter werden nach dem Bayes-Prinzip nur aus der Posteriori gezogen. Bei Anwendung von MCMC-Methodik liegt die Posteriori nicht direkt vor, sondern nur Zufallsziehungen (*Samples*) aus der Posteriori der einzelnen Parameter.

Als Punktschätzer dient daher der Mittelwert oder der Median der gezogenen Parameter, wobei der Median die robusteren Schätzungen liefert. Als Grundlage für Intervallschätzungen dienen meist die empirischen Quantile der *Samples*. Auch andere Funktionen der Posterioriverteilung lassen sich aus den *Samples* schätzen (siehe z.B. Abschnitt 2.4.2).

Die theoretischen Grundlagen der MCMC-Methoden erfordern, daß die ersten Iterationen des *Samples* weggelassen werden, weil sie noch nicht aus der gewünschten stationären Verteilung stammen. Man nennt dies *Burn-In-Phase*.

Die *Samples* sollen möglichst gut mischen (gutes *Mixing*), das heißt die einzelnen Ziehungen sollten möglichst unabhängig sein. Meist dünnt man die *Samples* aus, sprich man nimmt zum Beispiel nur jede 100. Ziehung. Dadurch erhält man mehr oder weniger unabhängige *Samples*. Für weitere Details siehe Gilks *et al.* (1996).

### 2.4.2 Devianz und DIC

Um die Güte der Anpassung eines Modells an die Daten zu messen, berechnet man die saturierte Posteriori-Devianz

$$D_i = -2 \left( l(\pi_i) - l(\hat{\pi}_i) \right),$$

wobei  $l(\pi_i)$  die Log-Likelihood bezeichnet,  $\pi_i$  die im Modell geschätzte Wahrscheinlichkeit und  $\hat{\pi}_i$  die relative Häufigkeit  $y_i/n_i$ . Als zusammenfassender Wert wird die Summe

$$D = \sum_i D(\pi_i)$$

verwendet.

Die Devianz ist eine Funktion der unbekannt Parameter, welche in  $\pi_i$  eingehen und hat folglich eine Posterioriverteilung. Als Punktschätzer berechnet man den Mittelwert oder den Median der Posteriori-Devianz  $\bar{D}$ . Eine geringe Devianz ist dabei ein Hinweis auf gute Anpassung der geschätzten an die beobachteten Daten.

Die Devianz sinkt allerdings natürlicherweise mit der Anzahl der Parameter. Um verschiedene Modelle trotzdem vergleichen zu können, schlagen Spiegelhalter *et al.* (2002) das *deviance information criteria* DIC vor. Es ist definiert als

$$DIC = \bar{D} + p_D,$$

wobei  $p_D$  die effektive Anzahl der Parameter bezeichnet. Die effektive Anzahl der Parameter penalisiert größere Modellkomplexität und wird über

$$p_D = \bar{D} - D(\bar{\boldsymbol{\pi}})$$

berechnet, wobei  $\bar{\boldsymbol{\pi}}$  den Posteriori-Mittelwert bzw. Posteriori-Median von  $\boldsymbol{\pi}$  bezeichnet. Die effektive Anzahl der Parameter ist in hierarchischen Modellen in der Regel geringer als die Anzahl der tatsächlichen Parameter, da die Schätzung der effektiven Zahl der Parameter die Abhängigkeiten in den Parametern, z.B. durch eine GMRF-Priori, berücksichtigt.



# Kapitel 3

## Modelle für räumliche Prävalenzdaten von Wildtier-Krankheiten

Die Analyse der räumlichen Verteilung von Krankheiten auf Ebene geographischer Regionen wie z.B. Verwaltungseinheiten spielt eine wichtige Rolle in der Veterinärepidemiologie. Bei Wildtierpopulationen basieren Prävalenzschätzungen häufig auf den regionalen Zähl-  
daten der von Jägern zufällig geschossenen Tiere. Die beobachtete Krankheitsrate pro Region ist dabei aufgrund sehr unterschiedlicher Stichprobengrößen und räumlicher Abhängigkeiten zwischen benachbarten Regionen keine gute Schätzung für die Prävalenz der Krankheit. Es ist daher notwendig, Überdispersion und räumliche Nachbarschaftsbeziehungen in den Daten zu berücksichtigen, um bessere Karten der Verteilung der Krankheit zu erhalten. Die Schätzung des räumlichen Musters ist dabei durch fehlende Daten in vielen der geographischen Regionen erschwert. Die komplette Abdeckung aller Regionen ist in Wildtieruntersuchungen praktisch unmöglich.

Um trotzdem die räumliche Verteilungsstruktur von Wildtierkrankheiten schätzen zu können, verwenden wir ein hierarchisches Bayesianisches Modell, in dem ein strukturierter räumlicher Effekt mit Gauss-Markov-Zufallsfeldern und zusätzliche unstrukturierte Effekte

geschätzt werden. Dabei finden die Algorithmen von Rue (2001) Anwendung. Das Modell wird empirisch mit den Resultaten eines in der Veterinärepidemiologie üblichen, nicht-räumlichen Beta-Binomial-Modells mit Surveillance-Daten der Aujeszky'schen Krankheit, einer Virusinfektion von Wildschweinen (*Sus scrofa scrofa* L.) im Bundesland Brandenburg, Deutschland, verglichen. Im weiteren zeigen wir eine kürzlich von Holmes & Held (2003) vorgestellte Methode, mit der sich das Modell effizienter schätzen läßt und Erweiterungen leicht möglich sind. An einem Datensatz zur Echinokokkose bei Füchsen wird das Potential dieses Ansatzes gezeigt.

Teile dieses Kapitels, insbesondere die Abschnitte 3.1 und 3.2, basieren auf einem Artikel, der in englischer Sprache in *Preventive Veterinary Medicine* erschien (Staubach, Schmid, Knorr-Held & Ziller 2002).

### 3.1 Disease Mapping in der Veterinärepidemiologie

In der geographischen Epidemiologie werden räumliche Daten häufig auf Basis künstlicher Grenzen wie Verwaltungsbezirken aggregiert. Bei Wildtieruntersuchungen ist dies häufig die einzige Möglichkeit, Stichproben abzubilden. Nur in wenigen Studien liegen für jeden Fall und für jede Kontrolle die exakten Koordinaten vor (Staubach, Tackmann, Thulke, Hugh-Jones & Conraths 2001). Zur Visualisierung werden die Gebiete gemäß den Werten einer beschreibenden Statistik eingefärbt, zum Beispiel dem Mittelwert. Die dabei häufig angewendete simple Klassifikation in unterschiedliche Prävalenzklassen hat verschiedene Probleme:

- (i) Geographische Grenzen sind künstlich und nicht relevant zur Krankheitsverbreitung gewählt.
- (ii) Die Stichprobengröße wird häufig nicht in Erwägung gezogen, wenn räumliche Daten dargestellt werden. Die Konfidenzintervalle benachbarter Regionen können sich beispielsweise überdecken und die Prävalenzunterschiede zwischen den Regionen wären damit rein zufällig.

- (iii) Bei geschichteten Daten können die Stichprobengrößen in einigen Regionen oder in einzelnen Schichten zu niedrig sein um zuverlässige Prävalenzschätzungen zu erhalten.
- (iv) Derartige Visualisierung von Surveillance-Daten kann auch zu fälschlicher oder Nicht-Erkennung von Krankheitsclustern führen (Pfeiffer & Morris 1994; Smans & Estéve 1996; Haining 1998).

Weiterhin ist in Wildtieruntersuchungen die Identifikation räumlicher Variation von Krankheiten durch das Problem fehlender Daten in vielen der geographischen Einheiten erschwert. Trotz vieler Bemühungen, den Untersuchungsbereich so komplett wie möglich zu erfassen, liegen Stichproben kranker und gesunder Tiere aus logistischen und finanziellen Gründen nur in einem Teil der Regionen vor (Müller *et al.* 1998; Tackmann *et al.* 1998). Im Gegensatz dazu sind bei Humankrankheiten meist komplette Datensätze der Bevölkerungsgröße und der Zahl der Krankheitsfälle verfügbar (Haining 1998; Lawson *et al.* 1999; Elliott *et al.* 2000).

Speziell in der medizinischen Epidemiologie sind unterschiedliche räumliche Filter, Glättungsmechanismen und parametrische Regressionsmethoden zur Lösung obiger Probleme vorgeschlagen worden (Pfeiffer & Morris 1994; Haining 1998; Lawson *et al.* 1999; Elliott *et al.* 2000). Einige Autoren verwenden nur die Raten und verzichten auf die komplette Information der Fallzahlen (Elliott *et al.* 1995; Smans & Estéve 1996).

## 3.2 Vergleich zweier räumlicher Modelle

Im Folgenden stellen wir ein volles Bayesianisches Modell zur räumlichen Analyse von Prävalenzdaten vor. Es basiert auf einer räumlichen Glättungspriori und wird mit Hilfe moderner MCMC-Methoden geschätzt. Damit werden die Probleme gelöst, welche durch spärliche Daten oder räumlicher Abhängigkeit auftreten. Wir vergleichen die Ergebnisse der Prävalenzschätzungen des vollen Bayesianischen Modells mit einem nicht-räumlichen Beta-Binomial-Modell am Beispiel von Surveillance-Daten der Pseudorabies (PRV)-Infektion von Wildschweinen (Erreger der Aujeszky'schen Krankheit).

### 3.2.1 Daten

Im Rahmen eines Wildtierkrankheiten-Monitoringprojekts in Ostdeutschland zwischen 1991 und 1994 wurde in Brandenburg eine Untersuchung über das Auftreten ausgewählter Viruskrankheiten von Wildschweinen durchgeführt. Das Untersuchungsgebiet, bestehend aus dem Bundesland Brandenburg im östlichen Teil Deutschlands, umfaßt ungefähr 29 530 km<sup>2</sup>. Das Bundesland Brandenburg ist in 1700 Verwaltungsbezirke (Gemeinden) mit einer durchschnittlichen Größe von 17.4 km<sup>2</sup> aufgeteilt. Gemeinden können Exklaven mit der gleichen Kennnummer haben, die nicht direkt räumlich mit der Hauptgemeinde verbunden sind. Die topographische Karte besteht daher aus einer Gesamtmenge von 1902 geographischen Gebieten. Die Großstädte Berlin und Potsdam wurden wegen des hohen Anteils urbanen Gebiets aus der Analyse entfernt.

Aus logistischen Gründen war es nicht möglich, alle in dieser Zeit geschossenen Tiere zu untersuchen. Daher wurden über das ganze Jahr verteilt, speziell aber in der Hauptjagdzeit, Serenproben einer Zufallsstichprobe von etwa 5% der geschossenen Wildschweine gesammelt. Die Seren wurden entweder von Mitgliedern lokaler Jagdvereine, von staatlichen Förstern oder Tierärzten zur obligatorischen Untersuchung auf *Trichinella spiralis* gesammelt. Danach wurden die Seren zum Labor geschickt und in einem vollantigenen *enzyme-linked immunosorbent assay* (ELISA) auf Antikörper gegen PRV untersucht. Datenbank, Untersuchungsverfahren und Resultate des kompletten PRV-Datensatzes werden in Müller *et al.* (1998) beschrieben.

Es liegen Zählzahlen der diagnostizierten positiven und negativen Fälle vor, die direkt in einem *Geographischen Informationssystem* (GIS) mit den Verwaltungsbezirken verbunden wurden. Bei Gemeinden mit Exklaven wurden die positiven und negativen Resultate proportional auf jede Subregion verteilt.

Wir verwenden die Daten der PRV-Infektion von Wildschweinen des Jahres 1993 als Beispiel für diese Studie. In 370 Gemeinden wurden 1364 Wildschweine überprüft (Abbildung 3.1 zeigt die Verteilung der Stichprobengröße). Im Diagnostetest waren 119 Tiere serologisch positiv; die Werte der Prävalenz in den Regionen liegen zwischen 0.0 und 1.0.

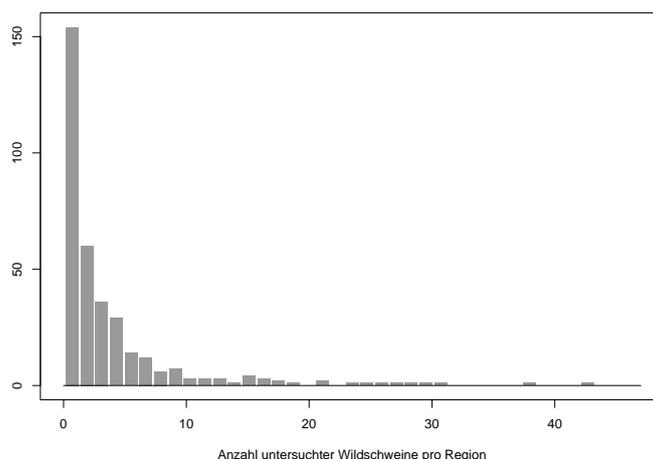


Abbildung 3.1: Verteilung der Stichprobengröße der auf PRV untersuchten Wildschweine pro Region.

### 3.2.2 Modelle

In einer Karte mit  $N$  geographischen Einheiten sei  $n_i$  die Stichprobengröße und  $y_i$  die Anzahl der positiven Resultate in der  $i$ -ten Region. Die beobachtete Prävalenz in jeder Region ist durch das Verhältnis  $p_i = y_i/n_i$  gegeben. Jedoch kann in vielen Fällen aufgrund des Fehlens von Daten keine Prävalenzschätzung durchgeführt werden ( $n_i = 0$ ). Dennoch kann davon ausgegangen werden, daß auch in diesen Regionen Wildschweine leben, da Regionen mit einem hohen Anteil für Wildschweine ungeeigneter Landschaft (z.B. Stadtgebiete) von vornherein aus der Analyse entfernt wurden. Wir verwenden räumliche Interpolation zur Schätzung der Prävalenz in den Gebieten mit fehlenden Daten.

Im Folgenden nehmen wir die Anzahl der Fälle in jeder Region als binomialverteilt

$$y_i \sim \text{Binomial}(n_i, \pi_i) \quad (3.1)$$

an, wobei  $\pi_i$  die unbekannte wahre Prävalenz bezeichnet. Das heißt, daß unsere Daten eine Zufallsstichprobe der Tiere in der jeweiligen Region bilden. Diese Annahme ist in einigen Anwendungen nicht unkritisch. Zum Beispiel könnte die Prävalenz in einer Region heterogen sein und die Stichproben aus einem lokalen Cluster kommen, entsprechend wäre die durchschnittliche Prävalenz in dieser geographischen Einheit über- oder unterrepräsentiert. Aufgrund der großen Anzahl an räumlichen Einheiten, der geringen Größe der

einzelnen Einheiten und der jeweils geringen Größe der  $n_i$ , können wir jedoch in diesem Fall erwarten, daß dieser Effekt minimal ist. In beiden Modellen fangen die Parameter unstrukturierte regionenspezifische Variation auf, das heißt die marginale Verteilung von  $\mathbf{y}$  kann sehr wohl auch möglicherweise existierende Überdispersion zeigen. Natürlich könnte man bei genauerer Information über regionale Heterogenität das Modell entsprechend anpassen und entsprechende Daten mit einbeziehen. Im vorliegenden Fall, ohne entsprechende Daten, ist uns aber keine alternative Modellformulierung bekannt, welche dieses Problem besser lösen könnte.

### Beta-Binomial-Modell

Das in der Veterinärepidemiologie häufig verwendete Beta-Binomial-Modell nutzt keinerlei Nachbarschaftsbeziehungen aus. Auf die unbekannte Prävalenzwahrscheinlichkeit  $\pi_i$  wird eine Beta-Priori gelegt. Eine Beta( $\alpha, \beta$ )-Verteilung hat Mittelwert  $\alpha/(\alpha + \beta)$  und Dichte

$$p(\pi_i) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \pi_i^{\alpha-1} (1 - \pi_i)^{\beta-1}. \quad (3.2)$$

Unter Anwendung des Bayes'schen Theorems (2.1), d.h. durch Multiplikation der Prioriverteilung (3.2) mit der Likelihood der Binomialverteilung

$$p(y_i | \pi_i) = \binom{n_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i}, \quad (3.3)$$

erhält man die Posterioriverteilung, welche aufgrund der Verwendung der konjugierten Prioriverteilung wieder eine Betaverteilung ist:

$$\pi_i | y_i \sim \text{Beta}(\alpha + y_i, \beta + n_i - y_i).$$

Speziell ist der Posteriorimittelwert der Prävalenz

$$\hat{\pi}_i = \frac{\alpha + y_i}{\alpha + \beta + n_i}.$$

Wir verwenden den Posteriorimittelwert als Punktschätzer der wahren Prävalenz. Auch die Posteriori-Wahrscheinlichkeit, daß die Prävalenz  $\pi_i$  in der  $i$ -ten Region die aggregierte Prävalenz  $\sum y_i / \sum n_i$  überschreitet, kann daraus berechnet werden. Damit ist eine graphische Darstellung der statistischen Kreditabilität des Punktschätzers möglich.

### BYM-Modell

Der nach Besag, York & Mollié (1991) oft BYM-Modell genannte Ansatz findet in der räumlichen (Human-)Epidemiologie häufig Anwendung. Es handelt sich hierbei um ein hierarchisches Bayesianisches Modell für binomiale Beobachtungen. Im Unterschied zum Beta-Binomial-Modell berücksichtigt dieser Ansatz die räumliche Struktur der Daten. Zusätzlich ist das Modell dazu in der Lage, für Regionen ohne Beobachtungen die Prävalenzoberfläche zu interpolieren.

Analog zu Besag *et al.* (1991) (allerdings mit einem logistischen Binomial-Modell statt eines log-linearen Poisson-Modells, siehe Abschnitt 2.1) zerlegen wir den Logit von  $\pi_i$  in einen Intercept  $\mu$  und zwei Heterogenitätsparameter, einen a priori räumlich strukturierten ( $u_i$ ) und einen für unstrukturierte Heterogenität ( $v_i$ ):

$$\log \left( \frac{\pi_i}{1 - \pi_i} \right) = \eta_i = \mu + u_i + v_i. \quad (3.4)$$

Für den räumlich strukturierten Parameter  $\mathbf{u} = (u_1, \dots, u_N)'$  verwenden wir ein Gauss-Markov-Zufallsfeld als Priori mit Nachbarschaftsmatrix  $\mathbf{M}$  und Präzisionsparameter  $\tau$ . Nach Abschnitt 2.2.1 lautet die gemeinsame Prioriverteilung

$$p(\mathbf{u}) \propto \tau^{rg(\mathbf{M})} \exp \left( -\frac{\tau}{2} \mathbf{u}' \mathbf{M} \mathbf{u} \right).$$

Die unstrukturierten Effekte  $v_i$  werden a priori als unabhängig normalverteilt mit Erwartungswert Null und Präzision  $\delta^{-1}$  angenommen. Dabei ist zu beachten, daß wegen der Priori-Unabhängigkeit der  $v_i$  Regionen ohne Beobachtungen ( $n_i = 0$ ) keine Information über die unstrukturierte Komponente liefern können. Sei  $I$  die Anzahl der räumlichen Einheiten mit Daten und nehmen wir an, die Einheiten seien in absteigender Reihenfolge nach  $n_i$  geordnet. Um potentielle Identifikationsprobleme zu vermeiden, setzen wir den unstrukturierten räumlichen Parameter für die  $N - I$  Regionen ohne Daten gleich Null:

$$v_i = 0 \quad \text{für } i = I + 1, \dots, N.$$

Sowohl für den Präzisionsparameter des strukturierten räumlichen Effekts  $\tau$  als auch für den des unstrukturierten Heterogenitätsparameters  $\delta$  verwenden wir eine üblicherwei-

se verwendete flache  $\text{Ga}(1,0.001)$ -Priori (Besag, Green, Higdon & Mengersen 1995). Wir nehmen also an, daß  $u_i$  und  $v_i$  a priori sehr glatt sind. Die Formulierung ist aber trotzdem flexibel genug, um Heterogenität in den Daten abzubilden. Dabei hat  $\tau$  die Rolle eines räumlichen Glättungsparameters, der die Variabilität der räumlich strukturierten Komponente  $u_i$  festlegt.

Rechentechnisch ist es günstiger, das Modell von  $v_i$  auf  $\eta_i$  zu reparametrisieren (Besag *et al.* 1995). Dies hat den Vorteil, daß die Full Conditional von  $\mathbf{u}$  multivariat normal ist. Wir können daher unter Verwendung der in Rue (2001) beschriebenen Algorithmen den Vektor  $\mathbf{u}$  blockweise ziehen. Um  $\eta_i$  zu ziehen, verwenden wir einen Metropolis-Schritt mit Gauss-Proposal, der entsprechend getunt wird, um Akzeptanzraten zwischen 35 und 45% zu erhalten. Die Full Conditional der Präzisionsparameter sind Gammaverteilungen und die Parameter können in Gibbs-Schritten gezogen werden. Weitere Einzelheiten über die technische Umsetzung in hierarchischen Modellen mit Markov-Zufallsfeldern sind in Knorr-Held & Rue (2002) und in Abschnitt 2.3 zu finden.

### Technische Details

Die Schätzungen der Parameter im Beta-Binomial-Modell erfolgte mit FORTRAN-Code nach Smith (1983). Der Algorithmus des BYM-Modells wurde in C++ implementiert. Für unsere Analyse lief der Algorithmus nach einer Burn-In-Phase von 10 000 weitere 100 000 Iterationen mit einem Ausdünnungsfaktor von 100. Die Konvergenz des Algorithmus wurde durch visuelle Inspektion der Samplepfade und Standard-Diagnose-Verfahren der CODA/BOA-Bibliothek in *S-Plus* (Mathsoft, Seattle, WA, USA) überprüft. Zur Erstellung der Karten und zur Gewinnung der Nachbarschaftsinformationen des geographischen Gebiets kam *ArcView 3.2a* (ESRI, Redlands, CA, USA) und dessen objektorientierte Programmiersprache *Avenue* zur Anwendung.

### 3.2.3 Anwendung

Im Folgenden vergleichen wir die zwei obigen Methoden empirisch am Beispiel der Daten für PRV-Infektionen in Brandenburg. Abbildung 3.2 zeigt die beobachtete Prävalenz  $y_i/n_i$  und Abbildung 3.3 den Posteriori-Mittelwert im Beta-Binomial-Modell. Die geschätzte Prävalenz ist im Vergleich zur Beobachteten deutlich geglättet, das heißt in Richtung der mittleren Prävalenz gesenkt. Die Schätzungen im Beta-Binomial-Modell liegen in der Tat immer zwischen der beobachteten Rate  $y_i/n_i$  und dem Priori-Mittelwert  $\alpha/(\alpha + \beta)$ , siehe Gelman, Carlin, Stern & Rubin (2000). Zu beachten ist, daß für Einheiten ohne Stichprobe keine Schätzung angegeben werden kann.

Die Ergebnisse der Schätzung mit dem BYM-Modell zeigt Abbildung 3.4. Die räumliche Struktur ist hier offensichtlicher; sie ist in weiten Teilen konstant mit erhöhten Werten im Osten Brandenburgs. Die räumliche Variabilität ist im vollen Bayes-Modell weitaus geringer als bei der einfachen Schätzung oder beim empirischen Bayes-Ansatz in den Abbildungen 3.2 und 3.3. Aufgrund der Einbeziehung der räumlichen Struktur und der räumlichen Einheiten ohne Daten erhält man eine starke Glättung der beobachteten Prävalenz. Die Abbildungen 3.5 und 3.6 zeigen für beide Modelle die Posteriori-Schätzung für die Wahrscheinlichkeit, daß die Prävalenz in der Gemeinde höher als die durchschnittliche Prävalenz ist. Die Karten zeigen erhöhte Infektions-Prävalenz im östlichen Teil Brandenburgs. Dies unterstützt die Hypothese, daß die Epidemie in Brandenburg nahe an der Grenze zu Polen ihren Anfang nahm.

Natürlicherweise stimmen die Posteriori-Wahrscheinlichkeiten (für die 370 Gemeinden mit Daten) für beide Modelle im wesentlichen überein. Allerdings hat das Beta-Binomial-Modell eine Residuenquadratsumme von 13.23 bei einer totalen Quadratsumme von 24.03, erklärt also 44.9% der Varianz. Im Vergleich dazu bringt es das volle Bayesianische Modell auf eine Residuenquadratsumme von 10.27, erklärt also 57.3% der Varianz. Der bessere Fit des hierarchischen Modells erklärt sich aber zum Teil durch die hohe Anzahl der Parameter der zusätzlichen räumlich strukturierten Komponente.

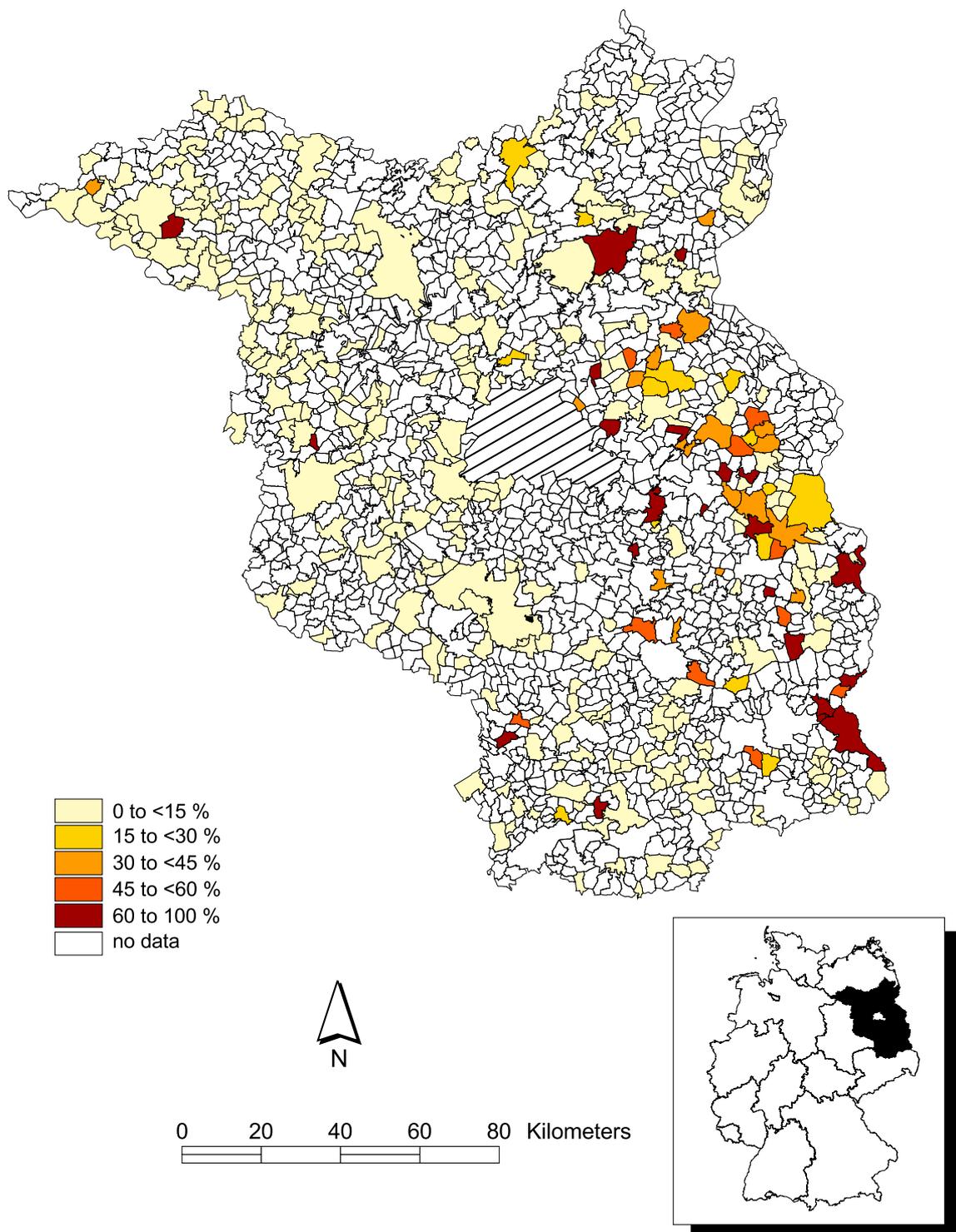


Abbildung 3.2: Beobachtete Prävalenz der PRV-Infektion von Wildschweinen je Gemeinde bzw. deren Exklaven basierend auf der serologischen Untersuchung im Jahr 1993. Die eingesetzte Karte zeigt das Untersuchungsgebiet (Brandenburg) innerhalb der Bundesrepublik Deutschland. Das Gebiet der Städte Berlin und Potsdam (schraffiert) wurde nicht in das Untersuchungsgebiet einbezogen.

### 3.2.4 Diskussion

Eine große Anzahl epidemiologischer Studien verwendet explorative räumliche Datenanalyse um geographische Verteilungen verschiedener Arten zu beschreiben (z.B. virologische oder serologische Prävalenz, Inzidenz, Proportionen von biologischen Markern). Krankheitsdaten von Wildtieren werden meist in administrativen Einheiten aggregiert erhoben, da allein der logistische Aufwand für die Erhebung der exakten Koordinaten enorm wäre.

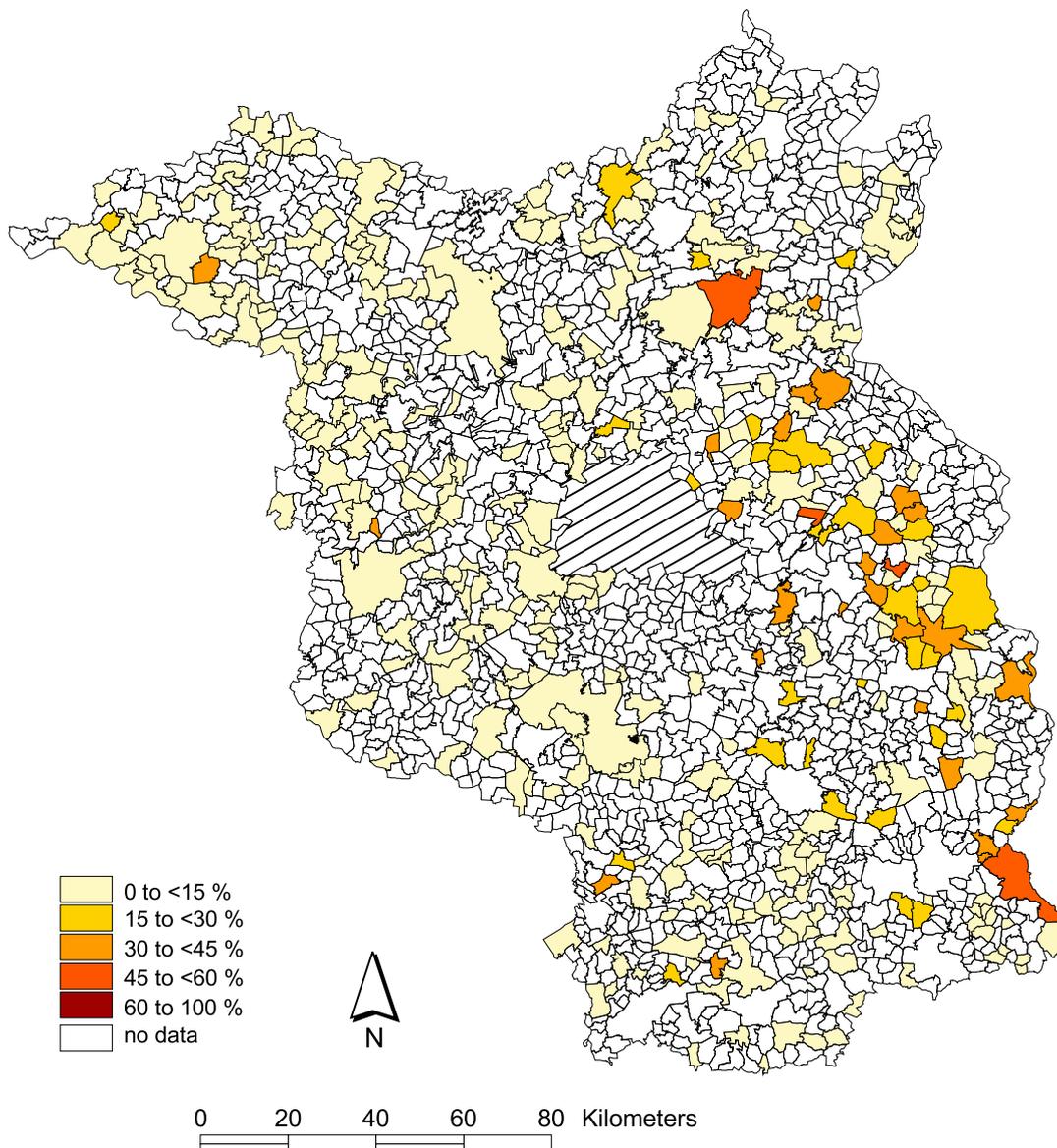


Abbildung 3.3: Prävalenz der PRV-Infektion geschätzt mit dem Beta-Binomial-Modell (Posterior-Median).

Die Erfassung der exakten geographischen Position, an der ein Tier erlegt wurde, ist speziell bei Studien notwendig, welche die Auswirkung von Umweltfaktoren auf eine Krankheit untersuchen (z.B. Biotop und Mikroklima, siehe Staubach *et al.* 2001). Daher werden Surveillance-Daten von Wildtier-Krankheiten oft unabhängig für jede räumliche Einheit abgebildet. Obwohl es Bemühungen zur Erhöhung der Stichprobenanzahl gibt, bleibt es meist bei geringen Stichprobenzahlen, die von der Individuendichte der Spezies, der Jagd-

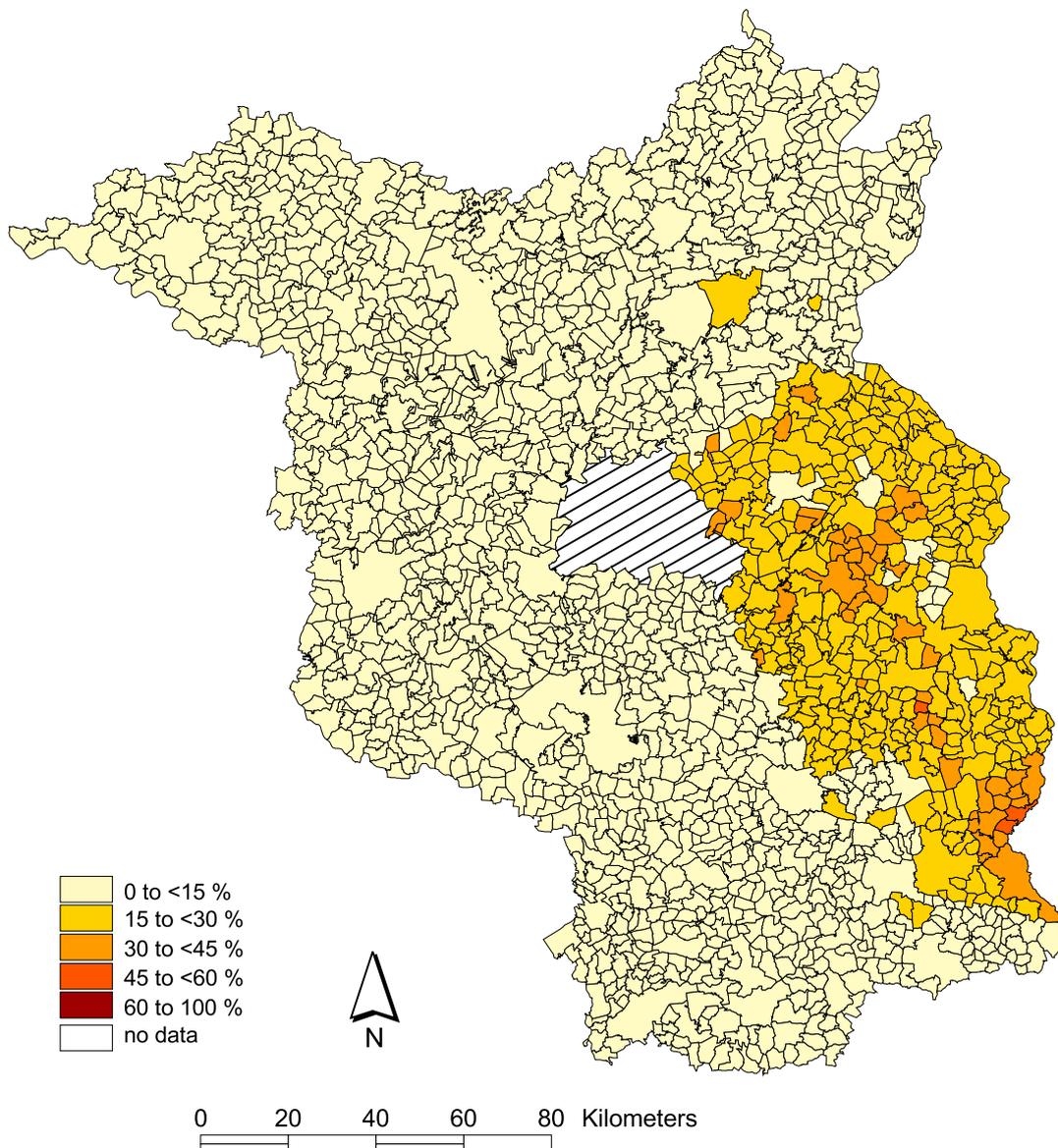


Abbildung 3.4: Prävalenz der PRV-Infektion geschätzt mit dem BYM-Modell (Posteriori-Median).

aktivität und der Größe der räumlichen Einheiten abhängt, vgl. Abbildung 3.1. Außerdem erlaubten die logistischen und Laborkapazitäten in unserer Studie nicht die komplette Untersuchung der erlegten Tiere im Bundesland Brandenburg.

Da es sich hier nicht um eine Vollerhebung, sondern nur um Stichproben handelt, hängt die Güte unserer statistischen Methode wesentlich von der Erhebungsmethode der Stichprobe ab, davon wie sich die PRV-Infektion ausbreitet und vom Verhalten und der Beweg-

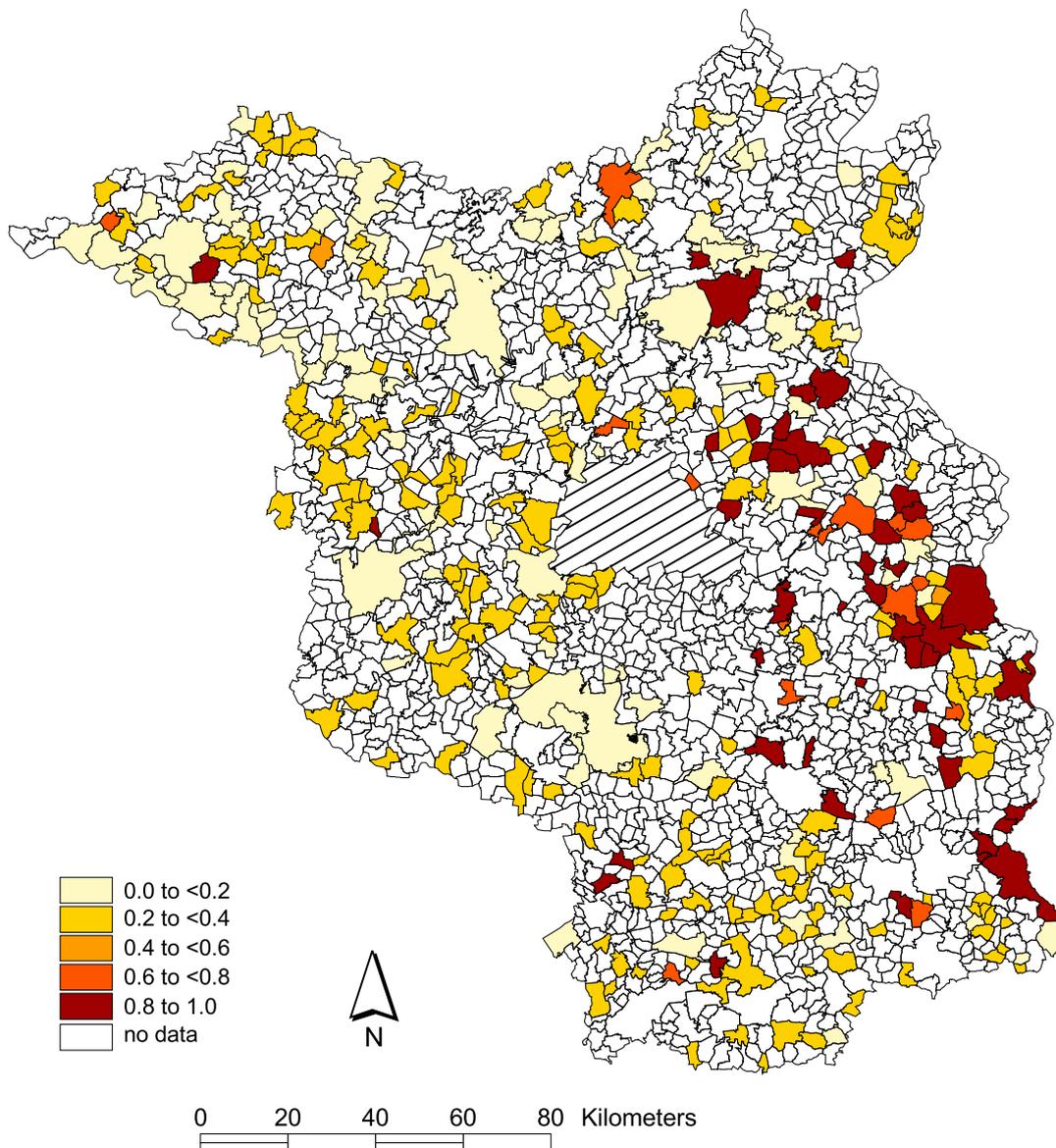


Abbildung 3.5: Posterioriwahrscheinlichkeit, daß die wahre Prävalenz über der mittleren Prävalenz der PRV-Infektion liegt, basierend auf dem Beta-Binomial-Modell.

lichkeit der Tiere in der Untersuchungsregion. Der erste wichtige Punkt ist die Position, an welcher der Jäger das Tier erlegt. Diese Positionen sind verständlicherweise nicht zufällig über die verschiedenen Landschaftstypen wie landwirtschaftliche Flächen, Weideflächen, Wald, Wasser und Stadtgebiet verteilt. Aber das mittlere Streifgebiet von Wildschweinen in der Untersuchungsregion ist kleiner oder gleich der mittleren Größe der Gemeinden (etwa  $5 - 15\text{km}^2$ , siehe Stubbe, Mehltitz, Peukert, Goretzki, Stubbe & Meynhardt 1989). Auch

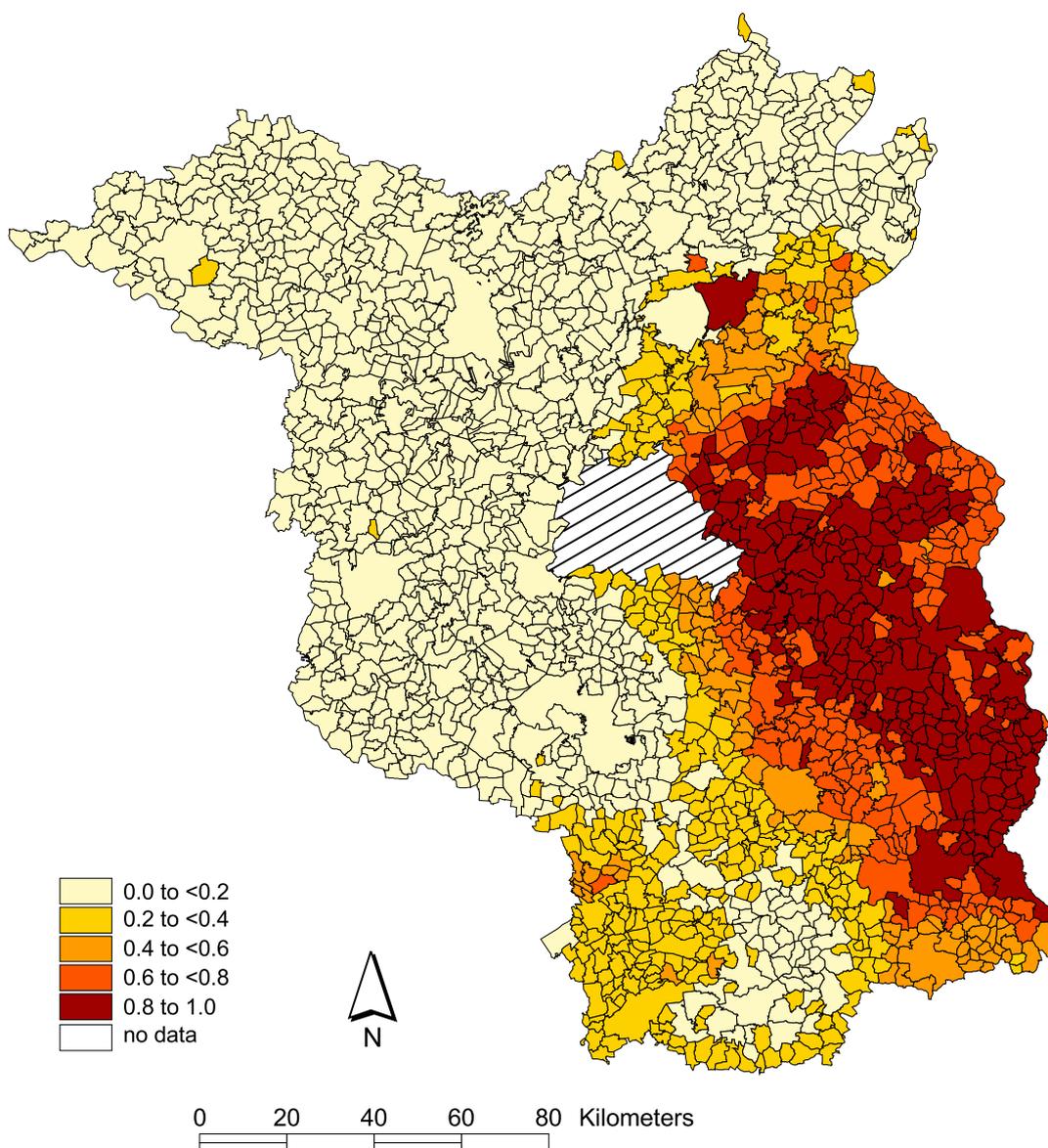


Abbildung 3.6: Posterioriwahrscheinlichkeit, daß die wahre Prävalenz über der mittleren Prävalenz der PRV-Infektion liegt, basierend auf dem BYM-Modell.

bewegt sich nur ein kleiner Anteil von Einzeltieren oder Wildschweinrotten über größere Distanzen (20-60 km, 3-5% der Tiere nach Stubbe *et al.* 1989). Zieht man schließlich die durchschnittliche tägliche Bewegungsweite von 8 km und die Dichte an Jägern in Betracht, kommt man zur Erkenntnis, daß die administrative Einheit, in der ein Wildschwein geschossen wurde, in etwa seinem Streifgebiet entspricht. Auch wenn die Sammlung der Seren natürlich von der Jagdintensität, der Individuendichte und dem Lebensraum abhängt, kann die Stichprobe als repräsentativ genug angesehen werden, um Schlüsse über die Verteilung der PRV-Infektion von Wildschweinen in der Untersuchungsregion zu ziehen. Aus der Sicht der Praxis muß hinzugefügt werden, daß die vorliegende Methode die einzige Möglichkeit ist, eine Datenstichprobe für diesen Zweck zu erheben. Weiterhin zeigt die homogene räumliche Verteilung der Stichprobe, daß die Untersuchungsregion einen optimalen Lebensraum für Wildschweine darstellt.

Die Aujeszky'sche Krankheit bei Wildschweinen ist eine Viruserkrankung, welche durch direkten Tier-zu-Tier-Kontakt übertragen wird (Müller *et al.* 1998). Wegen des ansteckenden Charakters der PRV-Infektion und der angegebenen Informationen über das Tierverhalten kann angenommen werden, daß die Verteilung innerhalb jeder räumlichen Einheit im wesentlichen homogen ist. Es ist daher anzunehmen, daß die Infektion zwischen den Rotten oder zwischen Einzeltieren an Orten übertragen wird, an denen sich Wildschweine gegenseitig treffen, an der Grenze ihrer Streifgebiets oder während der Paarungszeit. Die Wahrscheinlichkeit, daß sich die Krankheit schnell über große Distanzen ausbreitet, ist wegen der geringen Wahrscheinlichkeit einer Infektion und der geringen Wahrscheinlichkeit von großen Ortsveränderungen eines Wildschweins relativ gering. Damit ist die Grundannahme für den strukturierten räumlichen Effekt — daß die Prävalenz in benachbarten Einheiten voneinander abhängt — begründbar. Zusätzlich enthält unser Modell unstrukturierte Heterogenitätsparameter für Situationen, in denen das Markovzufallsfeld sich nicht eignet.

Experimentelle Untersuchungen haben gezeigt, daß nicht direkt übertragene PRV-Viren nicht zum Krankheitsausbruch bei Wildschweinen führen und daß die ersten Antikörper gegen PRV 3-6 Wochen nach der Infektion gebildet werden (Müller *et al.* 1998; Müller,

Teuffert, Zellmer & Conraths 2001). Eine Verzerrung durch fehlerhafte Zuordnung von infizierten und nicht infizierten Tieren ist daher zu vernachlässigen.

Ein großer Vorteil des vollen Bayesianischen Modells im Vergleich zum Beta-Binomial-Modell ist die Möglichkeit, die Prävalenz auch für räumliche Einheiten mit fehlenden Daten zu schätzen. Die Daten weiter zu aggregieren (z.B. auf Bezirks- oder Landesebene), ist meist nicht empfehlenswert, da dies die vorhandene räumliche Information verringert und sogar zu falschen Interpretationen über Krankheitsschwerpunkte oder infektionsfreie Gebiete führen kann (Tackmann *et al.* 1998). Auf dem untersten räumlichen Level (z.B. Gemeinden) wird man dagegen aus logistischen oder administrativen Gründen oft Regionen ohne Stichproben erhalten.

Zeitraubende Simulationsstudien sind die einzige Möglichkeit, die statistischen Eigenschaften der unterschiedlichen Modelle für räumliche binomiale Daten weiter im Detail zu erforschen (Lawson, Biggeri, Boehning, Lesaffre, Viel, Clark, Schlattmann & Divino 2000). Doch hat die Untersuchung verschiedener Daten mit verschiedenen statistischen Eigenschaften gezeigt, daß der volle Bayesianische Ansatz als erster Schritt einer deskriptiven und explorativen räumlichen Datenanalyse nützlich sein kann. Die produzierte Karte der geschätzten Prävalenz ist eine bessere Darstellung als die naive empirische Prävalenzschätzung, speziell wenn die Stichprobenlage spärlich ist — wie dies bei Datensätzen aus der Wildtier-Epidemiologie häufig der Fall ist. Für die abschließende Interpretation der Karte und der Hyperparameter ist es selbstverständlich notwendig, den Typ der Krankheit mit einzubeziehen.

### 3.3 Ein Ansatz mit Hilfsvariablen

Nachteil des BYM-Modells ist der hohe rechentechnische Aufwand. Durch moderne Verfahren lassen sich die Samples aber effizient ziehen. Während Besag *et al.* (1991) noch jeden Parameter ihres Modells einzeln ziehen, bekommen wir mit dem Blockupdate nach Rue (2001) ein sehr viel besseres Mixing. Es läßt sich weiter verbessern, wenn man einen Ansatz

mit Hilfsvariablen verwendet, den Holmes & Held (2003) vorgeschlagen haben. Auch dabei wird der räumliche Effekt blockweise mit den Algorithmen von Rue (2001) gezogen.

Dieser Ansatz hat den weiteren Vorteil, daß die Formulierung auf Basis der Individuen geschieht. Damit lassen sich ohne weiteres auch individuenspezifische Kovariablen mit in das Modell aufnehmen. Wir wollen in diesem Abschnitt den Ansatz von Holmes & Held (2003) mit der Formulierung in Abschnitt 3.2.2 vergleichen. Im darauf folgenden Abschnitt wird eine Anwendung auf einen weiteren Datensatz aus der Wildtier-Epidemiologie gezeigt, wobei das Modell um einen zeitlichen Trend und Kovariableneffekte erweitert wird.

### 3.3.1 Modellierung

Formuliert man das Binomialmodell (3.1) mit der Zerlegung (3.4) auf Basis der Individuen um, so lautet es

$$\begin{aligned}
 y_j &\sim \text{Bernoulli}(\exp(\eta_j)) & (3.5) \\
 \eta_j &= u_{i(j)} + v_{i(j)} \\
 \mathbf{u} &\sim \text{N}(0, (\tau \mathbf{M})^{-1}) \\
 v_i &\sim \text{N}(0, \delta) \text{ für alle } i = 1, \dots, N
 \end{aligned}$$

wobei  $y_j \in \{0, 1\}$ ,  $j = 1, \dots, n$  die binäre Responsevariable des Individuums  $j$  und  $i(j)$  den Index der Gemeinde des Individuums  $j$  bezeichnet. Der Vektor der Regressionskoeffizienten  $\mathbf{u}$  ist wie oben a priori ein Gauss-Markov-Zufallsfeld. Auf die Präzisionsparameter  $\tau$  und  $\delta$  legen wir jeweils Gamma-Prioris.

Das Modell läßt sich nach Holmes & Held (2003) wie folgt mit Hilfsvariablen formulieren:

$$\begin{aligned}
y_j &= \begin{cases} 1 & \text{für } z_j > 0 \\ 0 & \text{sonst} \end{cases} & (3.6) \\
z_j &= u_{i(j)} + v_{i(j)} + \epsilon_j \\
\epsilon_j &\sim N(0, \lambda_j) \\
\lambda_j &= (2\psi_j)^2 \\
\psi_j &\sim \text{KS} \\
\mathbf{u} &\sim N(0, (\tau\mathbf{M})^{-1}) \\
v_i &\sim N(0, \sigma_v^2) \text{ für alle } i = 1, \dots, N.
\end{aligned}$$

Dabei bezeichnet KS die Kolmogorov-Smirnov-Verteilung. Setzt man

$$\lambda_j = 1 \text{ für } j = 1, \dots, n$$

fest, so erhält man ein Probitmodell. Durch geeignete Wahl der Verteilung von  $\lambda$  sind auch andere Links möglich. Die marginale Likelihood  $L(\mathbf{u}, \mathbf{v} | \mathbf{y})$  ist in den beiden Modellformulierungen (3.5) und (3.6) gleich.

Mit einer Gauss-Markov-Zufallsfeld-Priori für  $\mathbf{u}$

$$p(\mathbf{u}) \propto \tau^{(N-1)/2} \exp\left(-\frac{\tau}{2} \mathbf{u}' \mathbf{M} \mathbf{u}\right) \quad (3.7)$$

wie in Abschnitt 3.2.2 lautet die Full Conditional von  $\mathbf{u}$

$$\begin{aligned}
\mathbf{u} | \mathbf{v}, \mathbf{z}, \boldsymbol{\lambda}, \mathbf{y} &\sim N(\mathbf{V}^{-1} \mathbf{b}, \mathbf{V}^{-1}) \\
\mathbf{b} &= (b_1, \dots, b_I) \\
b_i &= \sum_{j \in R_i} \frac{z_j - v_i}{\lambda_j} \\
\mathbf{V} &= \tau \mathbf{M} + \mathbf{W} \\
\mathbf{W} &= \text{diag} \left( \sum_{j \in R_1} \lambda_j^{-1}, \dots, \sum_{j \in R_I} \lambda_j^{-1} \right).
\end{aligned}$$

Dabei bezeichnet  $R_i$  die Menge der Individuen in Pixel  $i$ .

Entsprechend lautet die Full Conditional von  $\mathbf{v}$

$$\begin{aligned}\mathbf{v}|\mathbf{u}, \mathbf{z}, \boldsymbol{\lambda}, \mathbf{y} &\sim N(\tilde{\mathbf{V}}^{-1}\tilde{\mathbf{b}}, \tilde{\mathbf{V}}^{-1}) \\ \tilde{\mathbf{b}} &= (\tilde{b}_1, \dots, \tilde{b}_I) \\ \tilde{b}_i &= \sum_{j \in R_i} \frac{z_j - u_i}{\lambda_j} \\ \tilde{\mathbf{V}} &= \delta \mathbf{I} + \mathbf{W}\end{aligned}$$

Im Unterschied zum BYM-Modell in Abschnitt 3.2.2 kann hier  $v_i$  in jeder Region  $i \in \{1, \dots, I\}$  geschätzt werden. Liegen in einer Region keine Daten vor, so ist die Posteriori von  $v_i$  einfach gleich der Priori. Für die Ziehung der Hilfsvariablen  $z_j$  und  $\lambda_j$  siehe Holmes & Held (2003).

Für die Präzisionen  $\tau$  und  $\delta$  werden wie oben  $\text{Ga}(1, 0.001)$ -Verteilungen als Priori verwendet, die Full Conditional-Verteilungen sind dann wieder Gamma-Verteilungen mit Parametern  $a = 1 + \frac{N-1}{2}$  und  $b = 0.001 + \frac{1}{2}\mathbf{u}'\mathbf{M}\mathbf{u}$  bzw.  $\tilde{a} = 1 + \frac{N}{2}$  und  $\tilde{b} = 0.001 + \frac{1}{2}\sum_i^N v_i^2$ .

### 3.3.2 Vergleich

In Abschnitt 3.2.2 wird in Regionen ohne Beobachtungen kein unstrukturierter Effekt geschätzt. Die Schätzung der Präzision des unstrukturierten Effekts  $\delta$  ist mit 30.41 im Holmes-Held-Modell daher niedriger als in BYM-Modell mit 87.74. Damit wird im Ansatz mit Hilfsvariablen der strukturierte Effekt stärker geglättet. Setzt man im Holmes-Held-Modell ebenfalls in Regionen ohne Beobachtungen  $v_i = 0$ , so sind die Effektschätzungen und die Schätzungen der Glättungsparameter in beiden Modellen identisch (bis auf den Simulationsfehler). Die Korrelation zwischen den Punktschätzungen der Prävalenz (Median) beträgt im Durchschnitt bei 10 Läufen 0.9962. Abbildung 3.7 zeigt die Punktschätzungen der Prävalenz im Vergleich.

Die Laufdauer des Ansatzes mit Hilfsvariablen hängt nicht nur von der Anzahl der Gemeinden, sondern auch stark von der Anzahl der Beobachtungen ab. Für den vorliegenden Datensatz dauert eine Iteration des Holmes-Held-Modells auf einer SUN Sparc Ultra 5

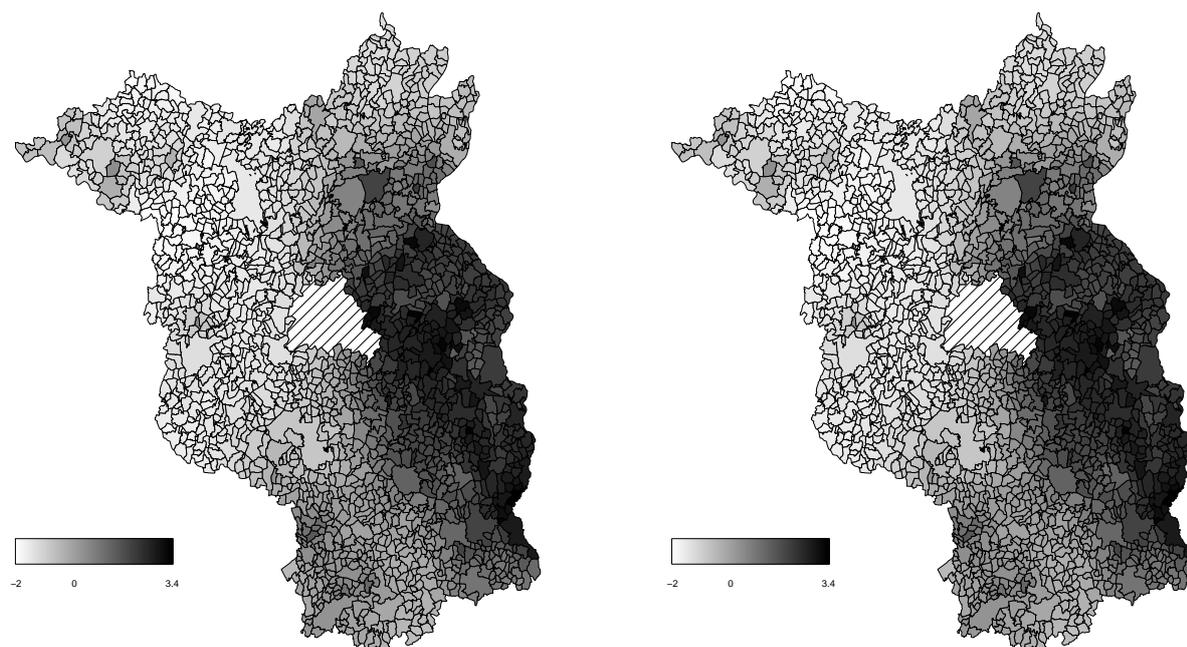


Abbildung 3.7: Räumliche Effekte des herkömmlichen und des Holmes-Held-Modells, jeweils Summe des strukturierten und unstrukturierten Effekts.

Workstation durchschnittlich etwa 0.255 Sekunden, was bei 100 000 Iterationen mit einem Ausdünnungsfaktor von 100 und einer Burn-In-Phase von 10 000 Iterationen eine Laufzeit von 7:47 Stunden ergibt.

Demgegenüber benötigt eine Iteration des in Abschnitt 3.2.2 vorgestellten Modells nur durchschnittlich 0.171 Sekunden und das Modell hat eine Laufzeit von 5:13 Stunden, also nur etwa zwei Drittel des Holmes-Held-Modells. Der Nachteil des größeren Zeitbedarfs wird jedoch durch den Vorteil des wesentlich besseren Mixings aufgehoben.

Als Maß der Effizienz des Samplers berechnen wir die euklidische Distanz zwischen den Iterationen für  $\mathbf{u}$

$$\text{Dist.} = \frac{1}{N-1} \sum_{i=1}^{N-1} \|\mathbf{u}^{(i)} - \mathbf{u}^{(i+1)}\|,$$

wobei  $\mathbf{u}^{(i)}$  der Wert des Parametervektors  $\mathbf{u}$  nach der  $i$ -ten Iteration ist. Im Mittel ergibt sich bei 10 Läufen beim herkömmlichen Modell 1.249, beim Holmes-Held-Modell 0.840. Dies entspricht einer relative Effizienzverbesserung von 48.6%. Holmes & Held (2003) zeigen ein weiteres Beispiel für die Effizienz ihres Modells.

Wenn auch bei diesem Datensatz die Effizienz den erhöhten Zeitbedarf des Holmes-Held-Modells aufwiegt, so muß doch kritisch angemerkt werden, daß dieser stark von der Anzahl der Beobachtungen abhängt. Der Zeitbedarf steigt hier linear mit der Anzahl der Beobachtungen an, während er im reinen Binomialmodell (bei gleichbleibender Anzahl an Regionen) aufgrund der Aggregation gleich bleibt.

### 3.3.3 Erweiterung um Kovariablen

Ein weiterer Vorteil des Holmes-Held-Modells ist die leichte Erweiterbarkeit des Modells. Wir zeigen im Folgenden die Anwendung des Modells auf einen Datensatz zur Echinokokkose bei Füchsen in Thüringen. Die Daten liegen auf Ebene der 1080 Gemeinden vor. Im Zeitraum vom 1990 bis 2003 wurden 17935 Füchse untersucht, davon wurden 3990 (22.2%) serologisch positiv auf *echinococcus multilocularis* getestet. Als zusätzliche Kovariablen lagen für jedes Tier Alter (binär kodiert, 0=juvenil, 1=adult) und Geschlecht (0=weiblich, 1=männlich) sowie das Jahr der Untersuchung vor.

#### Modellierung

Wir setzen für diesen Datensatz als Zerlegung der Hilfsvariablen  $z_j$  in Modell (3.6)

$$z_j = u_{i(j)} + v_{i(j)} + \phi_{t(j)} + \beta \mathbf{x}_j + \epsilon_j$$

an. Wie oben ist  $\mathbf{u}$  ein strukturierter räumlicher Effekt,  $\mathbf{v}$  ein unstrukturierter. Die Hilfsvariable  $\epsilon_j$  ist definiert wie in (3.6). Der Zeittrend ist mit  $\phi = (\phi_1, \dots, \phi_T)'$  bezeichnet, wobei  $t(j)$  das Jahr bezeichnet, zu dem das Tier  $j$  erlegt wurde. Für die  $p = 2$  Kovariablen  $x_j$  werden die Regressionsparameter  $\beta_1$  und  $\beta_2$  angesetzt. Die Regressionsparameter seien a priori normalverteilt mit Erwartungswert 0 und Varianz  $\delta^{-1} = 1000$ .

Als Priori für den strukturierten räumlichen Effekt verwenden wir wieder ein Gauss-Markov-Zufallsfeld (vgl. Abschnitt 2.2.1) mit Präzisionsparameter  $\tau \sim \text{Ga}(1, 10^{-5})$ , für den unstrukturierten unabhängige Normalverteilungen wie oben mit  $\delta \sim \text{Ga}(1, 10^{-3})$ . Als Priori

für den Zeittrend dient ein Random Walk (vgl. Abschnitt 2.2.2) mit Präzisionsparameter  $\kappa \sim \text{Ga}(1, 10^{-5})$ .

Wie oben ergeben sich für die Full Conditionals aller Haupteffekte und Hyperparameter Standardverteilungen (multivariate Normalverteilungen bzw. Gammaverteilungen) und die Parameter können in Gibbs-Schritten gezogen werden. Bei den Raum- und Zeit-Parametern lassen sich die Bandstrukturen der Präzisionsmatrizen im Rue (2001)-Algorithmus ausnutzen (vgl. Abschnitt 2.3.1). Hingegen ist die Präzisionsmatrix der Full Conditional der Regressionsparameter  $\beta$  voll besetzt (bei den hier vorliegenden zwei Kovariablen ist die Cholesky-Zerlegung aber auch nicht sehr rechenaufwendig). Die Ziehung der Hilfsvariablen  $z_j$  und  $\lambda_j$  erfolgt wie oben in Metropolis-Hastings-Schritten, verkompliziert sich aber durch die Erweiterung des Modells nicht. Dieser Ansatz läßt also eine große Flexibilität in der Modellierung zu.

Zur Schätzung der Parameter lief der Algorithmus über 50 000 Iterationen bei einer Ausdünnung von 50 und einem Burn-In von 1000.

## Ergebnisse

Tabelle 3.1 zeigt Median und 90%-Intervall der beiden Regressionsparameter und der drei Hyperparameter. Das Geschlecht hat keinen signifikanten Einfluß auf die Infektion. Ebenso

	5%-Quantil	Median	95%-Quantil
$\beta_{\text{Alter}}$	-1.579	0.024	1.629
$\beta_{\text{Geschl.}}$	-1.641	-0.025	1.617
$\tau$	0.198	0.240	0.288
$\delta$	4.93	10.69	28.46
$\kappa$	2.76	6.51	17.01

Tabelle 3.1: Median und 90%-Vertrauensintervall der Regressionsparameter und der Hyperparameter.

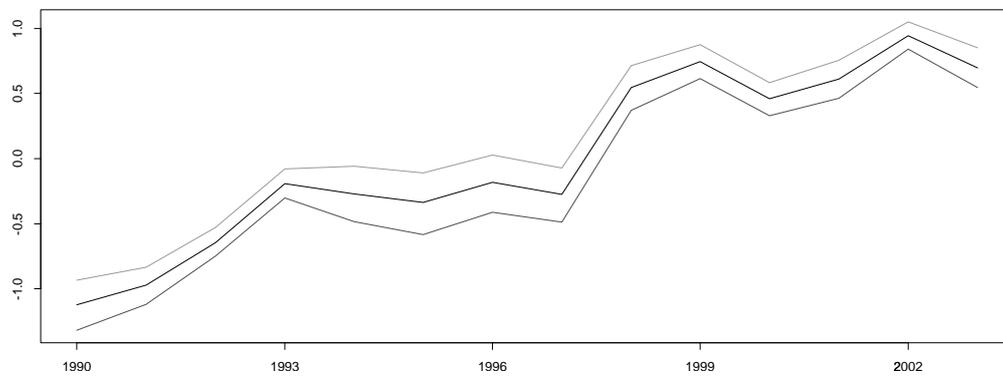


Abbildung 3.8: Median und 90%-Vertrauensintervall des zeitlichen Effekts.

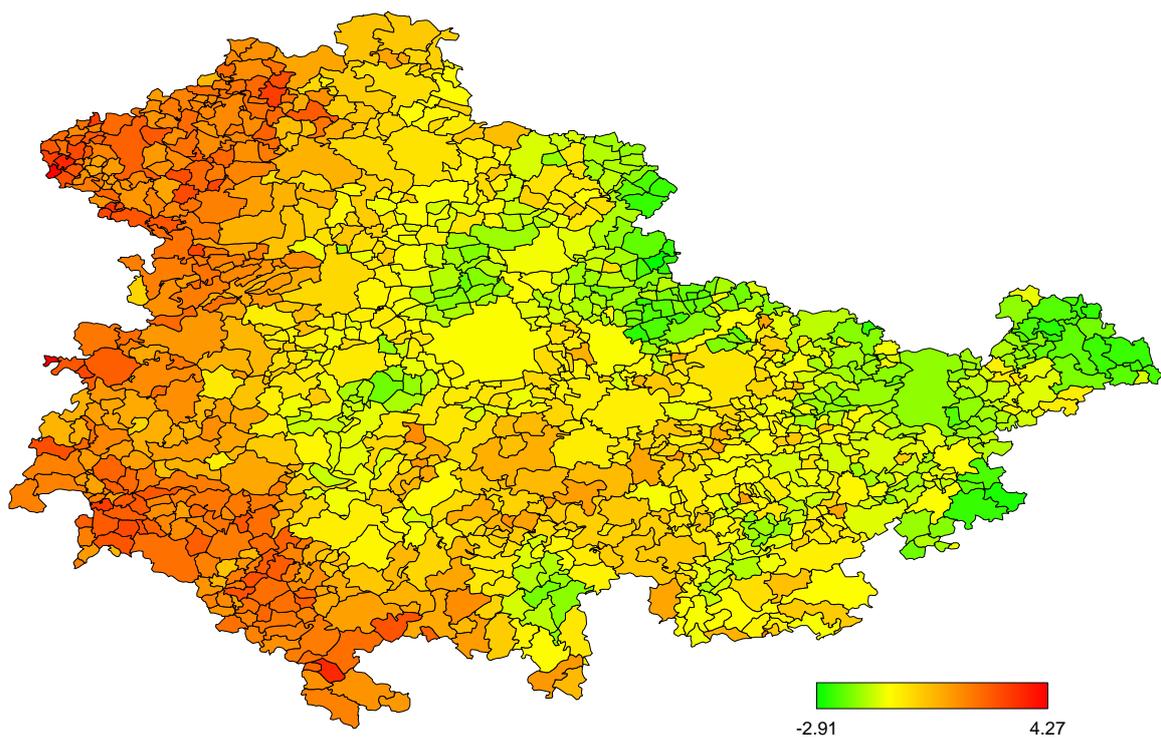


Abbildung 3.9: Median des räumlichen Effekts (Summe strukturierter und unstrukturierter Effekt).

hat das Alter keinen Einfluß, zumindest mit der gegebenen Einteilung in juvenile und adulte Tiere.

Die Schätzungen der Hyperparameter zeigen, daß ein starker strukturierter räumlicher Effekt und eine zeitliche Entwicklung in den Daten vorliegt. Dies entspricht den Erwartungen der Veterinäre. Abbildung 3.8 zeigt den zeitlichen Effekt, der im Beobachtungszeitraum einen deutlichen Anstieg aufweist.

Abbildung 3.9 zeigt den mittleren aufsummierten räumlichen Effekt. Erhöhtes Infektionsrisiko ist besonders im Westen Thüringens festzustellen. Hierbei handelt es sich hauptsächlich um das Gebiet des Thüringer Waldes. In den eher flachen Gebieten im Osten ist der Effekt dagegen negativ.

### 3.4 Fazit und Ausblick

Während in der Humanepidemiologie die Daten oft für alle Fälle eines Gebietes vorliegen, ist dies in der Wildtierepidemiologie nicht möglich. Die wahre Anzahl an Fällen und die Population sind in diesen Anwendungen nicht exakt bekannt. Eine räumliche Analyse mit Gauss-Markov-Zufallsfeldern ist aber möglich, wenn bei der Stichprobenziehung sorgfältig gearbeitet wird.

Die Auswertung mit GMRFs ist hier insbesondere nützlich, weil sie auch für spärliche Daten ein glatte, strukturierte Schätzung durch Ausnutzung des *borrowing strength* liefert. Die Möglichkeit, Posteriori-Wahrscheinlichkeiten zu berechnen und abzubilden ist wichtig, um die Kreditibilität der geschätzten Prävalenz darzustellen und die geographische Variation der Krankheit besser einzuschätzen. Mit den vorgestellten MCMC-Methoden lassen sich die Samples effizient gewinnen. So haben wir für den räumlichen Effekt  $\mathbf{u}$  einen Blockupdate verwendet.

Mit dem Ansatz von Holmes & Held (2003) ist die Einbeziehung von individuen- und regionenspezifischen Kovariablen möglich. Speziell für die Überprüfung von Impferfolgen

soll diese Methode in Zukunft verstärkt eingesetzt werden.

Die Anwendung voller Bayesianischer Modelle mit Gauss-Markov-Zufallsfeldern war in der Veterinärepidemiologie bisher praktisch unbekannt. Inzwischen wendet das Bundesforschungsinstitut für Tiergesundheit Wusterhausen die im Rahmen dieser Arbeit entstandenen Programme für eigene Analysen an. Neben der genannten Überprüfung von Impferfolgen ist hier auch die Modellierung von Raum-Zeit-Interaktionen geplant, die jedoch wegen der spärlichen Datenlage nicht unproblematisch ist.



# Kapitel 4

## Räumliche Effekte auf regulären Gittern in der Epidemiologie

Nach Cressie (1993) lassen sich drei Arten räumlicher Statistik unterscheiden (vgl. Abschnitt 1.1). In dieser Arbeit werden fast ausschliesslich Daten mit diskreter räumlicher Verteilung behandelt, Daten also, welche in einer abzählbaren Menge von räumlichen Einheiten vorliegen. Davon zu unterscheiden sind stetige Daten mit stetiger räumlicher Verteilung, auch *geostatistische* Daten genannt, sowie *räumliche Punktprozesse*. Letztere sind Zähldaten, bei denen bei jeder Beobachtung die Koordinaten stetig vorliegen.

Daten, welche aus räumlichen Punktprozessen stammen, werden für gewöhnlich mit mehrdimensionalen Glättern wie z.B. Kernschätzern oder Nächste-Nachbarn-Schätzern analysiert. Alternativ lassen sich die Daten auch künstlich in Sub-Regionen aggregieren und als diskret räumlich verteilt ansehen. Das folgende Kapitel zeigt dieses Vorgehen und vergleicht dies kurz mit der Analyse als räumlicher Punktprozeß. Wir verwenden hierbei ein reguläres Gitter zur Aggregation. Die sonst in dieser Arbeit analysierten Daten liegen alle auf irregulären Gittern, d.h. auf natürlichen Landkarten vor. Während bei letzteren die Nachbarschaft zweier Regionen als Vorhandensein einer gemeinsamen Grenze definiert wird, lassen sich auf regulären Gittern leicht auch andere Nachbarschaftsbeziehungen definieren.

## 4.1 Fallkontrollstudie Porto Alegre

In Porto Alegre, Bundesstaat Rio Grande do Sul in Brasilien, wurde eine Fall-Kontroll-Studie zur Kindersterblichkeit durchgeführt. Im Untersuchungsjahr 1998 wurden die Daten aller 183 Fälle von neonatalem (d.h. innerhalb der ersten 28 Tage nach der Geburt) Kindstod erhoben. Als Kontrolle diente eine Zufallsstichprobe von 600 Geburten im gleichen Jahr. Erhoben wurden dabei Alter und Schulbildung der Mutter, Dauer und Art (Einlings- oder Zwillingsgeburt) der Schwangerschaft, Art der Entbindung (natürlich, Kaiserschnitt, Zangengeburt), Geschlecht und Gewicht des Neugeborenen, sowie die Koordinaten des Wohnorts der Mutter auf 10 Meter genau. Abbildung 4.1 zeigt die räumliche Verteilung der Daten.

Shimakura *et al.* (2001) analysieren die räumliche Varianz der Daten (bezogen auf den Wohnort der Mutter) als räumlichen Punktprozeß nach Kelsall & Diggle (1998). Dabei erfolgt die Schätzung des stetig über die Untersuchungsregion variierenden Risikos mit Hilfe eines generalisierten additiven Modells (GAM). Sei die Wahrscheinlichkeit eines Kindstodes  $p(x, s)$  abhängig von den Kovariablen  $x$  und den räumlichen Koordinaten  $s$  des Wohnorts der Mutter. Dann lautet das generalisierte additive Modell

$$\log \left( \frac{p(s, x)}{1 - p(s, x)} \right) = \beta x + g(s). \quad (4.1)$$

Die Schätzung wird mittels eines Schaukelalgorithmus ausgeführt. Dazu wird abwechselnd einmal  $\hat{\beta}$  bei angenommenem festen räumlichen Effekt  $g(s)$  mit logistischer Regression und das andere Mal der räumliche Effekt  $\hat{g}(s)$  bei festem  $\beta$  mit zweidimensionalem Kernschätzer geschätzt. Dies wird bis zur Konvergenz der Parameter wiederholt.

Nachteil dieses Modells ist, daß die Konvergenz des Algorithmus nicht gesichert ist. Ein weiterer Nachteil ist die notwendige Wahl eines Glättungsparameters, also der Bandweite des Kernschätzers. Die Wahl des Parameters beeinflusst dabei stark die Schätzung des räumlichen Effekts. Als Ausweg verwenden Shimakura *et al.* (2001) einen Kreuzvalidierungs-Algorithmus zur Bestimmung einer optimalen Bandweite.

Alternativ dazu diskretisieren wir die räumlichen Daten. Im Bayesianischen hierarchi-

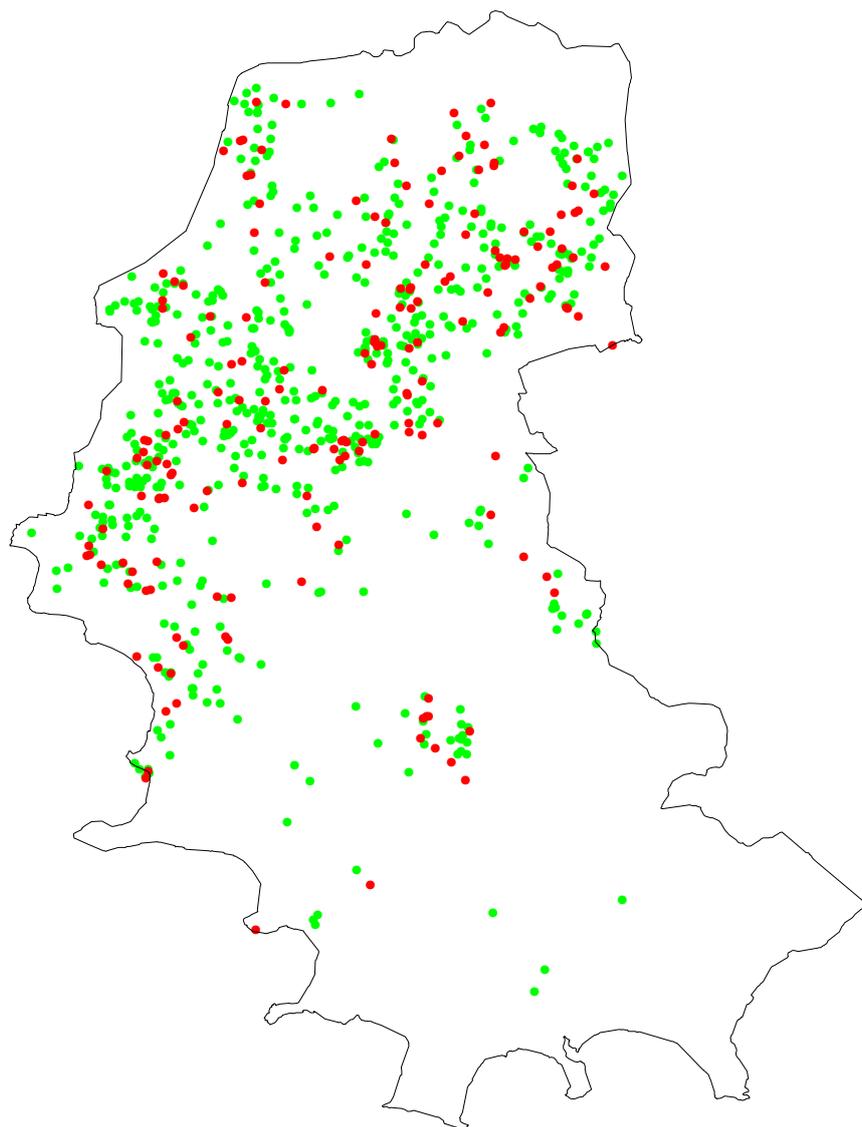


Abbildung 4.1: Wohnorte der Mütter in der Fall-Kontrollstudie über Kindersterblichkeit in der Stadt Porto Alegre. Rote Punkte sind Fälle, Grüne Punkte sind Kontrollfälle.

schen Modell lassen sich dann die Regressionsparameter und die räumliche Struktur simultan schätzen. Als Priori bieten sich Gauss-Markov-Zufallsfelder an, wobei man die Nachbarschaftsstruktur unterschiedlich definieren kann. Zusätzlich läßt sich auch der Glättungsparameter des GMRFs ohne weiteres mitschätzen.

Wir diskretisieren die Untersuchungsregion mit einem quadratischem Gitter bzw. vergrößern die Koordinaten des Wohnorts der Mutter. Wir fassen dazu jeweils 25 Koordi-

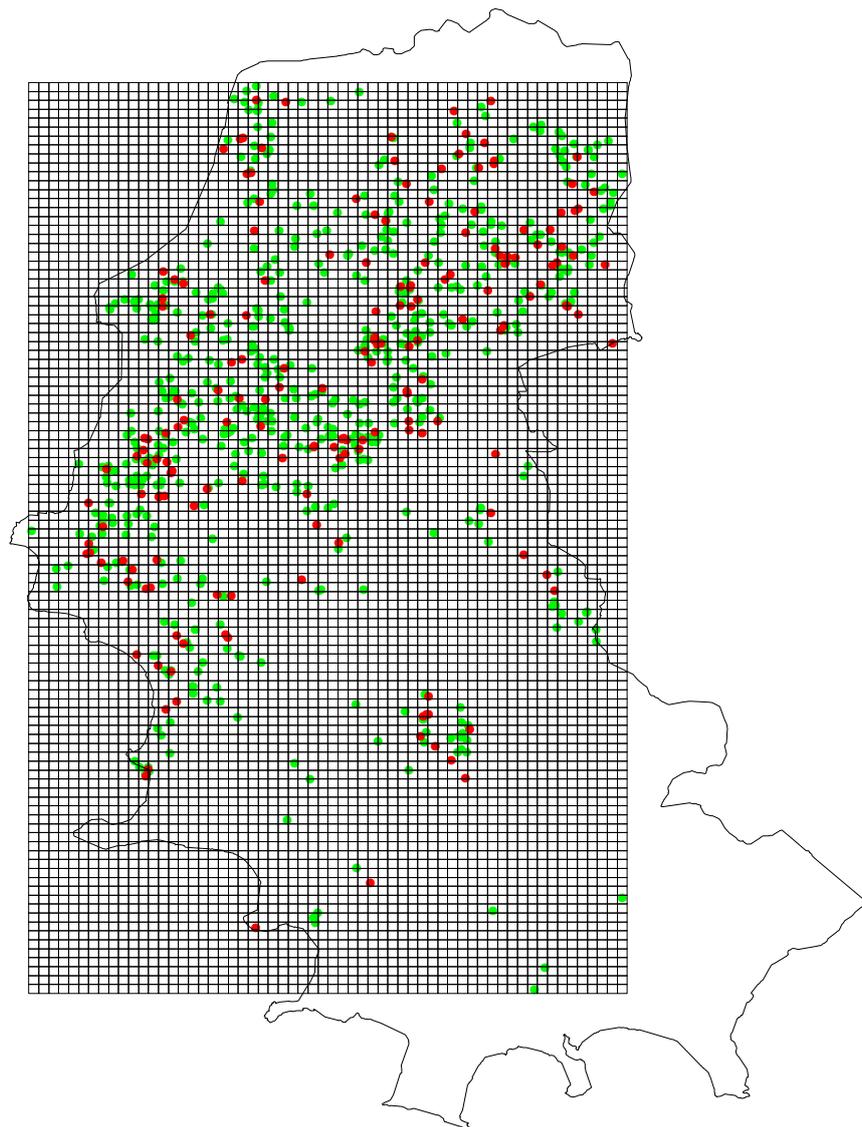


Abbildung 4.2: Daten der Fall-Kontrollstudie mit dem künstlichen regulären Gitter.

natenpunkte in jede Richtung zu einem Pixel zusammen, bekommen also ein Gitter mit Kantenlänge 250 Meter eines Pixels, siehe Abbildung 4.2. Die Untersuchungsregion umfaßt  $X = 60$  mal  $Y = 102$ , insgesamt also  $N = X \cdot Y = 6120$  Pixel. Sie hat eine Breite von 15 km und eine Höhe von 25.5 km. Im Folgenden seien die Pixel des regulären Gitters mit  $i = (x, y)$  bezeichnet, mit  $x \in \{1, \dots, X\}$  und  $y \in \{1, \dots, Y\}$ .

## 4.2 Modellformulierung

Wir setzen folgendes Bayesianisches logistisches Modell an

$$\begin{aligned}
 y_j &\sim \text{Bernoulli}(g^{-1}(\eta_j)) \\
 \eta_j &= \mathbf{x}_j \boldsymbol{\beta} + u_{i(j)} \\
 \boldsymbol{\beta} &\sim \text{N}(0, \delta \mathbf{I}) \\
 \mathbf{u} &= (u_1, \dots, u_N)' \\
 \mathbf{u} &\sim \text{N}(0, (\tau \mathbf{M})^{-1}).
 \end{aligned}$$

Dabei ist  $y_j \in \{0, 1\}$ ,  $j = 1, \dots, n = 783$  die binäre Responsevariable „neonataler Kindstod eingetreten“ für die 183 Fälle ( $y_j = 1$ ) und die 600 Kontrollfälle ( $y_j = 0$ ). Als Linkfunktion dient die logistische Funktion  $g(v) = \log(v/(1-v))$ . Die  $p = 7$  Kovariablen sind mit  $\mathbf{x}'_j = (x_{j1}, \dots, x_{j7})$  bezeichnet, der Index des Wohnorts der Mutter in obigem Gitter ist  $i(j)$ ,  $\eta_j$  ist der lineare Prädiktor,  $\boldsymbol{\beta}$  sei der Vektor der Regressionskoeffizienten. Der räumliche Effekt  $\mathbf{u}$  ist a priori ein Gauss-Markov-Zufallsfeld mit Präzisionsparameter  $\tau$  und Nachbarschaftsmatrix  $\mathbf{M}$ . Dabei verwenden wir eine allgemeinere Formulierung als in Abschnitt 2.2.1, siehe Abschnitt 4.3.

Holmes & Held (2003) zeigen, daß sich dieses Modell unter Verwendung von Hilfsvariablen umformulieren läßt. Durch die Umformulierung des Modells ist ein effizientes Ziehen mit üblichen MCMC-Methoden möglich. Ähnlich wie bei der Reparametrisierung von Besag *et al.* (1995) für Gauss-Markov-Zufallsfelder mit zufälligen Effekten sind die Full Conditionals des räumlichen Effekts dann multivariate Normalverteilungen und lassen sich mittels Gibbs-Sampler ziehen. Das umformulierte Modell lautet

$$\begin{aligned}
 y_j &= \begin{cases} 1 & \text{für } z_j > 0 \\ 0 & \text{sonst} \end{cases} \\
 z_j &= \mathbf{x}_j \boldsymbol{\beta} + u_{i(j)} + \epsilon_j
 \end{aligned}$$

$$\begin{aligned}
\epsilon_j &\sim N(0, \lambda_j) \\
\lambda_j &= (2\psi_j)^2 \\
\psi_j &\sim \text{KS} \\
\boldsymbol{\beta} &\sim \pi(\boldsymbol{\beta}) \\
\mathbf{u} &= (u_1, \dots, u_N)' \\
\mathbf{u} &\sim N(0, (\tau\mathbf{M})^{-1}).
\end{aligned}$$

Dabei bezeichnet KS die Kolmogorov-Smirnov-Verteilung. Die bedingten Verteilungen  $\pi(\boldsymbol{\beta}|\mathbf{y}, \mathbf{u})$  bzw.  $\pi(\mathbf{u}|\mathbf{y}, \boldsymbol{\beta})$  ändern sich durch die Umformulierung nicht.

Für die Regressionskoeffizienten  $\boldsymbol{\beta}$  verwenden wir a priori eine  $N(0, \delta\mathbf{I})$ -Verteilung, wobei  $\delta$  ein fest gewählter Varianzparameter ist. Der Parametervektor  $\boldsymbol{\beta}$  läßt sich gegeben allen anderen Parametern nach Holmes & Held (2003) wie folgt ziehen

$$\begin{aligned}
\boldsymbol{\beta}|\mathbf{u}, \mathbf{z}, \boldsymbol{\lambda}, \mathbf{y} &\sim N(\mathbf{V}^{-1}\mathbf{b}, \mathbf{V}^{-1}) \\
\mathbf{b} &= \mathbf{x}'\mathbf{W}(\mathbf{z} - \mathbf{u}^{(j)}) \\
\mathbf{u}^{(j)} &= (u_{i(1)}, \dots, u_{i(n)})' \\
\mathbf{V} &= (1/\delta)\mathbf{I} + \mathbf{x}'\mathbf{W}\mathbf{x} \\
\mathbf{W} &= \text{diag}(\lambda_1^{-1}, \dots, \lambda_n^{-1})
\end{aligned}$$

Für den räumlichen Effekt  $\mathbf{u}$  verwenden wir verschiedene GMRF-Prioris. Diese sind singuläre Normalverteilungen, deren Dichte sich nach Besag & Higdon (1999) schreiben läßt als

$$p(\mathbf{u}) \propto \tau^{rg(\mathbf{M})/2} |\mathbf{M}|^{1/2} \exp\left(-\frac{\tau}{2} \mathbf{u}'\mathbf{M}\mathbf{u}\right) \quad (4.2)$$

wobei  $\tau$  ein Präzisionsparameter und  $\mathbf{M}$  eine Nachbarschaftsmatrix ist. Dabei bezeichnet  $|\mathbf{M}|$  das Produkt der Eigenwerte ungleich Null von  $\mathbf{M}$ .

Der räumliche Effekt  $\mathbf{u}$  wird dann wie folgt gezogen:

$$\mathbf{u}|\boldsymbol{\beta}, \mathbf{z}, \boldsymbol{\lambda}, \mathbf{y} \sim N(\mathbf{V}^{-1}\mathbf{b}, \mathbf{V}^{-1}) \quad (4.3)$$

$$\mathbf{b} = (b_1, \dots, b_I) \quad (4.4)$$

$$b_i = \sum_{j \in R_i} ((z_j - \beta x_j) / \lambda_j)$$

$$\mathbf{V} = \tau \mathbf{M} + \mathbf{W}$$

$$\mathbf{W} = \text{diag} \left( \sum_{j \in R_1} \lambda_j^{-1}, \dots, \sum_{j \in R_I} \lambda_j^{-1} \right) \quad (4.5)$$

wobei  $R_i$  die Menge der  $n_i$  Beobachtungen im Gitterpunkt  $i$  bezeichnet. Die Ziehung aus der multivariaten Normalverteilung (4.3) erfolgt mit dem Algorithmus von Rue (2001), siehe dazu Abschnitt 2.3.1.

Für die Präzision  $\tau$  wird eine  $\text{Ga}(a, b)$ -Verteilung als Priori verwendet. Die Full Conditional ist dann ebenfalls eine Gammaverteilung mit Parametern  $a + \frac{\text{rg}(\mathbf{M})}{2}$  und  $b + \frac{1}{2} \mathbf{u}' \mathbf{M} \mathbf{u}$ . Aus den Full Conditionals der Hilfsvariablen  $z_j$  und  $\lambda_j$  kann leicht gezogen werden, für Details siehe Holmes & Held (2003).

### 4.3 GMRF-Prioris

Für die GMRF-Prioris haben Besag & Higdon (1999) drei verschiedene Vorschläge gemacht. Allgemein verwenden wir als Priori jedes Pixels  $u_i = u_{x,y}$  die bedingte Verteilung

$$u_i | \mathbf{u}_{-i}, \mathbf{w}, \tau \sim N \left( \sum_{j \in \partial_i} \frac{w_j}{w_\bullet} u_j, \frac{1}{\tau w_\bullet} \right) \quad (4.6)$$

mit Präzisionsparameter  $\tau$ , Gewichten  $\mathbf{w} = \{w_j | j \in \partial_i\}$  und  $w_\bullet = \sum_{j \in \partial_i} w_j$ . Dabei bezeichnet  $\partial_i$  die Menge der Pixel, welche dem Pixel  $i$  benachbart sind.

#### Modell 1

Die erste Möglichkeit ist, die Nachbarschaft nur bedingt auf die nächsten vier Nachbarn zu definieren, siehe Abbildung 4.3. Im einfachsten Fall wählt man  $w_j = 1$  für alle  $j \in \partial_i$

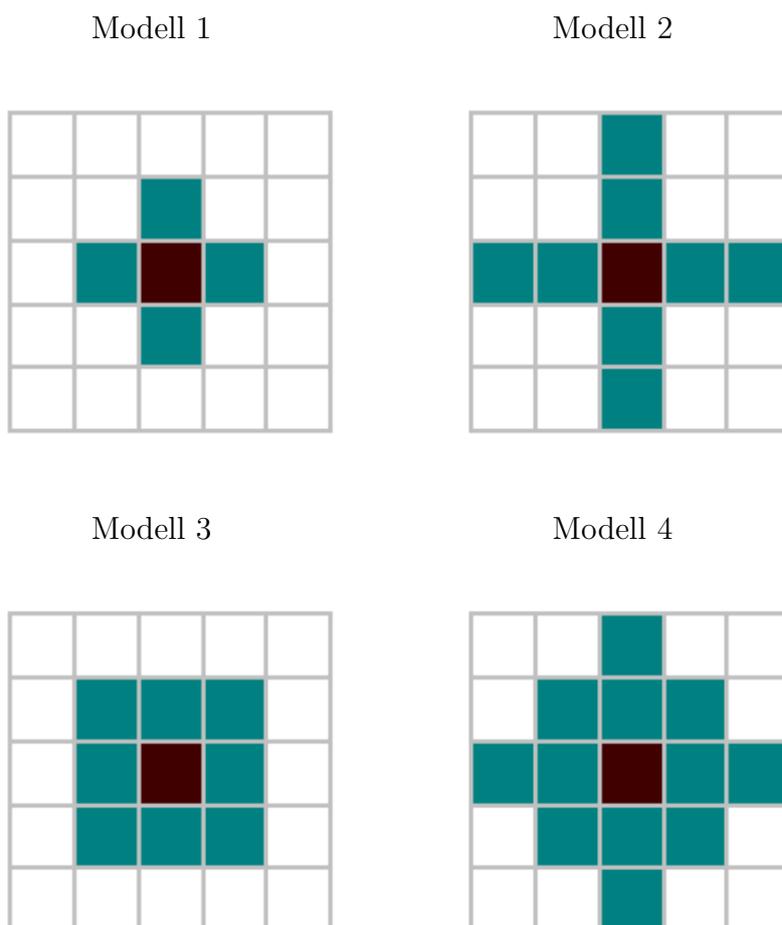


Abbildung 4.3: Definitionen der lokalen Nachbarschaften. Jeweils die grau hinterlegten Pixel werden als Nachbarn des mittleren schwarzen Pixels angesehen.

(Modell 1). Man erhält diese bedingte Randverteilung für jedes Element, wenn man in der gemeinsame Verteilung (4.2) von  $\mathbf{u}$  als Präzisionsmatrix die Kroneckermatrixsumme aus  $\mathbf{R}_Y^{(1)}$  und  $\mathbf{R}_X^{(1)}$ , also

$$\mathbf{M} = \mathbf{M}_1 = \mathbf{I}_X \otimes \mathbf{R}_Y^{(1)} + \mathbf{R}_X^{(1)} \otimes \mathbf{I}_Y$$

wählt. Dabei bezeichnet  $\mathbf{I}_n$  die Einheitsmatrix der Dimension  $n \times n$  und  $\mathbf{R}_n^{(1)}$  die Präzisionsmatrix der Dimension  $n \times n$  einer Random Walk-Priori erster Ordnung (RW 1, vgl. dazu Abschnitt 2.2.2). Nach den Rechenregeln für die Kroneckersumme gilt  $\text{rg}(\mathbf{M}_1) = XY - 1$ , vgl. z.B. Bellman (1970).

### Modell 1b

Modell 1 glättet gleichermaßen in X und Y-Richtung. Besag & Higdon (1999) schlagen vor, die Glättung in X-Richtung bzw. Y-Richtung mit einem Parameter  $\gamma \in (0, 2)$  zu gewichten, um Trends, welche nur in einer Richtung auftreten, zu erkennen. Der bedingte Erwartungswert der Priori im Pixel  $u_{x,y}$  lautet

$$E(u_{x,y} | \mathbf{u}_{-(x,y)}, \gamma) = \frac{1}{4} \left[ \gamma (u_{x-1,y} + u_{x+1,y}) + (2 - \gamma) (u_{x,y-1} + u_{x,y+1}) \right].$$

Die gemeinsame Prioriverteilung (4.2) läßt sich nach Besag & Higdon (1999) mit der Präzisionsmatrix

$$\mathbf{M} = \mathbf{M}_1(\gamma) = \gamma \mathbf{R}_X^{(1)} \otimes \mathbf{I}_Y + (2 - \gamma) \mathbf{I}_X \otimes \mathbf{R}_Y^{(1)}$$

definieren. Als Prioriverteilung für  $\gamma$  verwenden wir eine Gleichverteilung auf  $(0, 2)$ .

Die bedingte Verteilung von  $\gamma$  gegeben aller anderen Parameter lautet dann

$$\begin{aligned} p(\gamma | \cdot) &\propto I_{(0,2)}(\gamma) |\mathbf{M}_1(\gamma)|^{1/2} \exp \left( -\frac{\tau}{2} \mathbf{u}' \mathbf{M}_1(\gamma) \mathbf{u} \right) \\ &\propto I_{(0,2)}(\gamma) |\mathbf{M}_1(\gamma)|^{1/2} \\ &\quad \cdot \exp \left\{ -\frac{\tau}{2} \left( \gamma \sum_{x=2}^X \sum_{y=1}^Y (u_{x,y} - u_{x-1,y})^2 + (2 - \gamma) \sum_{x=1}^X \sum_{y=2}^Y (u_{x,y} - u_{x,y-1})^2 \right) \right\}. \end{aligned}$$

Dabei ist  $I_{(0,2)}$  die Indikatorfunktion auf  $(0, 2)$ . Während bei Modell 1 die Full Conditionals nicht von  $|\mathbf{M}_1|$  abhängen, geht hier  $|\mathbf{M}_1(\gamma)|$ , also das Produkt der Nicht-Null-Eigenwerte, in die bedingte Verteilung von  $\gamma$  ein. Die Eigenwerte  $\lambda_i$  von  $\mathbf{R}_n^{(1)}$  sind nach Künsch (1994)

$$\lambda_i = 2 \left[ 1 - \cos \left( \pi(i-1)/n \right) \right].$$

Die Eigenwerte der Kroneckersumme von  $\gamma \mathbf{R}_X^{(1)}$  und  $(2 - \gamma) \mathbf{R}_Y^{(1)}$  sind aber die Kroneckersummen der Vektoren der Eigenwerte von  $\gamma \mathbf{R}_X^{(1)}$  und  $(2 - \gamma) \mathbf{R}_Y^{(1)}$ , siehe Besag & Higdon (1999). Die Eigenwerte  $\lambda_{xy}$  von  $\mathbf{M}_1(\gamma)$  lauten also

$$\lambda_{xy} = 2 \left[ \gamma \left( 1 - \cos \left( \pi(x-1)/X \right) \right) + (2 - \gamma) \left( 1 - \cos \left( \pi(y-1)/Y \right) \right) \right] \quad (4.7)$$

mit  $x = 1, \dots, X$ ,  $y = 1, \dots, Y$  und  $|\mathbf{M}_1(\gamma)|$  läßt sich daraus explizit berechnen.

Wie sich zeigt, führt ein separater Update von  $\gamma$  zu sehr schlechtem Mixing. Wir ziehen daher  $\gamma$  und den räumlichen Effekt  $\mathbf{u}$  gemeinsam, siehe auch Abschnitt 2.3.3. Für das Gewicht  $\gamma$  verwenden wir ein Random Walk-Proposal für den Logit von  $\gamma$ . Sei also  $y^{(t-1)}$  der Zustand, in dem sich  $\gamma$  momentan befindet, dann ist das Proposal

$$\text{logit}(\gamma^*) \sim N\left(\text{logit}(\gamma^{(t-1)}), v\right).$$

Wir ziehen  $\mathbf{u}^*$  nun gegeben  $\gamma^*$  gemäß (4.3). Schließlich akzeptieren wir  $\gamma^*$  und  $\mathbf{u}^*$  mit Akzeptanzwahrscheinlichkeit  $\alpha$  oder verwerfen sie.

Die Akzeptanzwahrscheinlichkeit berechnet sich dabei wie folgt:

$$\alpha = \min \left\{ \frac{l(\mathbf{u}^*, \gamma^*)q(\gamma|\gamma^*)q(\mathbf{u}|\mathbf{u}^*, \gamma)}{l(\mathbf{u}, \gamma)q(\gamma^*|\gamma)q(\mathbf{u}^*|\mathbf{u}, \gamma^*)}, 1 \right\}$$

Die Terme  $q(\gamma|\gamma^*)$  und  $q(\gamma^*|\gamma)$  heben sich aufgrund der Verwendung der symmetrischen Normalverteilung als Proposal weg. Die Proposalverteilung von  $\mathbf{u}$  ist die Normalverteilung nach (4.3)

$$\begin{aligned} q(\mathbf{u}^*|\mathbf{u}, \gamma^*) &= (2\pi)^{-rg(\mathbf{V}(\gamma^*))} |\mathbf{V}(\gamma^*)|^{-1/2} \\ &\cdot \exp\left(-\frac{1}{2}(\mathbf{u}^* - \mathbf{V}(\gamma^*)^{-1}\mathbf{b})' \mathbf{V}(\gamma^*) (\mathbf{u}^* - \mathbf{V}(\gamma^*)^{-1}\mathbf{b})\right) \end{aligned}$$

mit  $\mathbf{V}(\gamma^*) = \tau\mathbf{M}(\gamma^*) + \mathbf{W}$  und  $\mathbf{b}$  bzw.  $\mathbf{W}$  nach (4.4) bzw. (4.5). Die gemeinsame Full Conditional von  $\mathbf{u}$  und  $\gamma$  ist

$$l(\mathbf{u}, \gamma) \propto I_{(0,2)}(\gamma) (2\pi\tau)^{-rg(\mathbf{M}(\gamma))} |\mathbf{M}(\gamma)|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{u} - \mathbf{V}(\gamma)^{-1}\mathbf{b})' \mathbf{V}(\gamma) (\mathbf{u} - \mathbf{V}(\gamma)^{-1}\mathbf{b})\right].$$

Damit bleibt für die Akzeptanzwahrscheinlichkeit

$$\alpha = \min \left\{ \frac{|\mathbf{V}(\gamma^*)|^{1/2} |\mathbf{M}(\gamma)|^{1/2}}{|\mathbf{V}(\gamma)|^{1/2} |\mathbf{M}(\gamma^*)|^{1/2}}, 1 \right\}$$

stehen. Während wir  $|\mathbf{M}_1(\gamma)|$  und  $|\mathbf{M}_1(\gamma^*)|$  als Produkt der Nicht-Null-Eigenwerte (4.7) berechnen können, ergeben sich  $\mathbf{V}(\gamma)$  und  $\mathbf{V}(\gamma^*)$  als Produkt der Elemente auf der Hauptdiagonale  $|\mathbf{V}|^{1/2} = \prod_i \mathbf{L}_{ii}$  der Cholesky-Zerlegung  $\mathbf{L}\mathbf{L}' = \mathbf{V}$  aus dem Rue-Algorithmus (siehe auch Abschnitt 2.3.2). Wir tunen den Parameter  $v$  der Proposalverteilung von  $\gamma$  so, daß der Metropolis-Hastings-Algorithmus eine Akzeptanzrate von 35% bis 40% hat.

## Modell 2

Modell 1 läßt sich als räumliches Analogon zum Random Walk erster Ordnung interpretieren. Ein räumliches Analogon zum Random Walk zweiter Ordnung (RW 2) bezieht die jeweils zwei nächsten Nachbarn in vertikaler und horizontaler Richtung mit ein, siehe Abbildung 4.3. Der bedingte Erwartungswert kann also formuliert werden als

$$\begin{aligned} E(u_{x,y} | \mathbf{u}_{-(x,y)}) &= \frac{1}{4} [(2u_{x-1,y} - u_{x-2,y}) + (2u_{x,y-1} - u_{x,y-2}) \\ &\quad + (2u_{x+1,y} - u_{x+2,y}) + (2u_{x,y+1} - u_{x,y+2})]. \end{aligned}$$

Analog zu Modell 1 läßt sich die Präzisionsmatrix der gemeinsamen Prioriverteilung (4.2) als Kroneckersumme

$$\mathbf{M} = \mathbf{M}_2 = \mathbf{R}_X^{(2)} \otimes \mathbf{I}_Y + \mathbf{I}_X \otimes \mathbf{R}_Y^{(2)}$$

schreiben, wobei  $\mathbf{R}_n^{(2)}$  die Präzisionsmatrix einer RW 2-Priori bezeichnet (siehe Abschnitt 2.2.2). Es gilt nach Bellman (1970)

$$rg(\mathbf{M}_2) = XY - 4.$$

## Modell 2b

Analog zu Modell 1 kann man auch bei Modell 2 unterschiedliche Gewichtungen für X- und Y-Richtung einführen. Die Präzisionsmatrix der gemeinsamen Prioriverteilung (4.2) ist definiert als

$$\mathbf{M} = \mathbf{M}_2(\gamma) = \gamma \mathbf{R}_X^{(2)} \otimes \mathbf{I}_Y + (2 - \gamma) \mathbf{I}_X \otimes \mathbf{R}_Y^{(2)}.$$

Auch in diesem Modell muß wieder  $|\mathbf{M}_2(\gamma)|$  gleich dem Produkt der  $(XY - 4)$  Nicht-Null-Eigenwerte berechnet werden. Die Eigenwerte von  $\mathbf{R}_X^{(2)}$  und  $\mathbf{R}_Y^{(2)}$  werden dazu vorher explizit berechnet. Die Eigenwerte von  $\mathbf{M}_2(\gamma)$  ergeben sich dann wie oben. Analog werden der Parametervektor  $\mathbf{u}$  und der Gewichtungparameter  $\gamma$  in einem gemeinsamen Block gezogen.

### Modell 3

Neben dem Modell mit den nächsten vier Nachbarn schlagen Besag & Higdon (1999) auch ein Modell mit den nächsten acht Nachbarn vor. Damit glättet man lokal quadratisch, während Modell 1 lokal konstant glättet. Besag & Higdon (1999) schlagen als bedingten Erwartungswert

$$\begin{aligned} E(u_{x,y} | \mathbf{u}_{-(x,y)}) &= \frac{1}{4} [2(u_{x,y+1} + u_{x-1,y} + u_{x+1,y} + u_{x,y-1}) \\ &\quad - (u_{x+1,y+1} + u_{x-1,y+1} + u_{x+1,y-1} + u_{x-1,y-1})] \end{aligned}$$

vor. In diesem Modell gehen also jeweils auch die diagonalen Nachbarn in die Priori jeden Pixels ein, vgl. Abbildung 4.3. Die Präzisionsmatrix  $\mathbf{M}$  der gemeinsamen Verteilung (4.2) läßt sich schreiben als Kroneckerprodukt

$$\mathbf{M} = \mathbf{M}_3 = \mathbf{R}_X^{(1)} \otimes \mathbf{R}_Y^{(1)}.$$

Nach den Rechenregeln für das Kroneckerprodukt (Bellman 1970) ist der Rang von  $\mathbf{M}_3$  gleich  $(X - 1)(Y - 1)$ .

Mit dieser Formulierung ist die gemeinsame Verteilung (4.2) invariant gegenüber der Addition von Konstanten zu einer beliebigen Spalte oder Reihe (Besag & Higdon 1999). Das ist unproblematisch, solange die Daten starke Zeilen bzw. Spaltenstruktur aufweisen. Bei unserem eher spärlichen Datensatz (783 Fälle auf 6120 Pixel) ist dies aber nicht der Fall. Technisch gesehen wirkt sich das so aus, daß die Präzisionsmatrix  $\mathbf{V}$  in (4.3) nicht vollen Rang hat, mithin also nicht invertierbar ist. Dieses Modell läßt sich daher nicht auf unsere Daten anwenden.

### Modell 4

Kombiniert man die Modelle 2 und 3, so erhält man ein Gauss-Markov-Zufallsfeld für die nächsten zwölf Nachbarn, vgl. Abbildung 4.3. Die Gewichte sind hier  $1/4$ ,  $1/8$  und  $-1/8$  für die ersten, zweiten und dritten Nachbarn, d.h. für die Nachbarn mit Abstand  $1$ ,  $\sqrt{2}$  und

2. Besag & Kooperberg (1995) weisen darauf hin, daß dieses Modell sich nicht für grobe Gitter geringer Größe eignet, sondern eher auf feineren Gittern nützlich ist.

Der bedingte Erwartungswert der Priori im Pixel  $(x, y)$  lautet also

$$\begin{aligned} E(u_{x,y} | \mathbf{u}_{-(x,y)}) &= \frac{1}{8} \left[ 2(u_{x,y+1} + u_{x-1,y} + u_{x+1,y} + u_{x,y-1}) \right. \\ &\quad + (u_{x+1,y+1} + u_{x-1,y+1} + u_{x+1,y-1} + u_{x-1,y-1}) \\ &\quad \left. - (u_{x,y-2} + u_{x,y+2} + u_{x+2,y} + u_{x-2,y}) \right]. \end{aligned}$$

Das entspricht dem bedingten Erwartungswert bei Modell 2 minus dem bedingten Erwartungswert bei Modell 3. Folglich ergibt sich auch für die Präzisionsmatrix

$$\mathbf{M} = \mathbf{M}_4 = \mathbf{M}_2 - \mathbf{M}_3 = \mathbf{R}_X^{(2)} \otimes \mathbf{I}_Y + \mathbf{I}_X \otimes \mathbf{R}_Y^{(2)} - \mathbf{R}_X^{(1)} \otimes \mathbf{R}_Y^{(1)}. \quad (4.8)$$

Anhang A.1 zeigt, daß sich wirklich die gewünschte Randverteilung von  $u_{x,y} | \mathbf{u}_{-(x,y)}$  bei dieser gemeinsamen Prioriverteilung ergibt. Der Rang von  $\mathbf{M}_4$  ist  $XY - 1$ .

## 4.4 Ergebnisse

Für die Schätzung der Parameter lief der MCMC-Algorithmus nach einem Burn-In von 10 000 Iterationen 1 000 000 mal, die Stichproben wurden mit dem Faktor 400 ausgedünnt. Auf den Hyperparameter  $\tau$  wird eine  $\text{Ga}(1, 0.1)$ -Priori gelegt. Als globalen Varianzparameter der Priori von  $\beta$  verwenden wir  $\delta = 1000$ .

Die sieben Kovariablen gehen linear in das Modell ein. Zwei der Kovariablen sind Eigenschaften der Mutter, Alter als quasistetiges Merkmal und Schulbildung. Letztere Variable ist dummykodiert (1 = keine abgeschlossene Schulausbildung, 0 = Schulausbildung abgeschlossen). Geschlecht (1 = weiblich, 0 = männlich) und Gewicht (stetig) beziehen sich auf das Neugeborene. Zwei Kovariablen beschreiben die Schwangerschaft, dabei ist die Dauer kodiert als 1 = Frühgeburt (Schwangerschaft kürzer als 37 Wochen), 0 = keine Frühgeburt, die Art der Schwangerschaft mit 1 = Zwillingsgeburt, 0 = Einlingsgeburt. Eine weitere Variable beschreibt die Art der Entbindung, diese ist kodiert mit 1 = vaginale, 0 = operative Entbindung.

Bei allen Modellen sind die Schätzungen für die Regressionskoeffizienten der Kovariablen in etwa gleich. Wie bei Shimakura *et al.* (2001) stellen sich nur das Gewicht des Neugeborenen und die Dauer der Schwangerschaft durchgehend als signifikante Merkmale heraus. Tabelle 4.1 zeigt die Punktschätzungen und die 90%-Vertrauensintervalle der Koeffizienten.

Die Schätzungen des räumlichen Effekts des Wohnorts der Mutter mit den Modellen 1 und 1b zeigt Abbildung 4.4. Wie bei Shimakura *et al.* (2001) finden sich Bereiche erhöhten Mortalitätsrisikos im Nord-Osten sowie im Westen Porto Alegres und deutlich vermindertes Risiko im Bereich dazwischen. Im Süden ist im Gegensatz zur GAM-Analyse von Shimakura *et al.* (2001) kein erhöhtes Risiko festzustellen. Die relativ geringe Anzahl von Daten in diesem Gebiet führt bei unserem Modell dazu, daß die Priori-Annahme der konstanten Oberfläche in diesem Bereich überwiegt. Bei der Glättung mit dem Kernschätzer führen die wenigen Fälle dagegen zu einer hohen Schätzung des Risikos auf Grundlage weniger Daten.

Auffällig ist die unterschiedliche Variabilität des räumlichen Effekts. Während bei Shimakura *et al.* (2001) der räumliche Effekt Werte von  $-2$  bis  $+1.5$  annimmt, schwankt der Wert bei Modell 1/1b zwischen  $-0.25$  und  $+0.25$  (bei Modell 2 im Stadtbereich zwischen  $-0.8$  und  $+0.8$ ). Der Präzisionsparameter  $\tau$  wird im Modell 1 auf 8.15 geschätzt, im Modell 1b auf 7.11.

Bei Modell 1b ergibt sich keine spezielle Gewichtung der Glättung nach der Achsenrichtung. Posteriori-Mittelwert des Gewichts  $\gamma$  ist 0.994 bei einer Standardabweichung von 0.104. Abbildung 4.5 zeigt Samplepfad und Posterioriverteilung von  $\gamma$ . Das Mixing ist hier aufgrund des gemeinsamen Updates von  $\gamma$  und  $\mathbf{u}$  sehr gut, während ein getrennter Update zu sehr schlechtem Mixing und stark verzerrtem Schätzer führt.

Modell 2 zeigt die gleichen Mortalitätsschwerpunkte wie oben. Die Schätzung ist aber glatter. Der Präzisionsparameter  $\tau$  hat den Posteriori-Median 47.19. Der Gewichtungsparemeter  $\gamma$  in Modell 2b hat einen Posteriori-Mittelwert von 0.994 bei einer Standardabweichung von 0.024 (siehe Abbildung 4.7), eine bevorzugte Glättungsrichtung wird also nicht

Kovariablen	Modell 1		Modell 1b	
	Schätzer	90%-Intervall	Schätzer	90%-Intervall
Alter der Mutter	-0.0062	(-0.0331, 0.0199)	-0.0068	(-0.0325, 0.0196)
Schulbildung Mutter	0.1062	(-0.2479, 0.4384)	0.1132	(-0.2475, 0.4708)
Dauer der Schwangerschaft	0.8942	( 0.3305, 1.3977)	0.8912	( 0.3314, 1.4580)
Art der Schwangerschaft	0.0229	(-0.7628, 0.7238)	0.0415	(-0.6380, 0.7477)
Art der Entbindung	-0.3154	(-0.7069, 0.0629)	-0.3163	(-0.6675, 0.0520)
Geschlecht	-0.3407	(-0.7038, 0.0106)	-0.3563	(-0.7340,-0.0021)
Gewicht des Neugeborenen	-0.0014	(-0.0017,-0.0012)	-0.0014	(-0.0017,-0.0011)
Kovariablen	Modell 2		Modell 2b	
	Schätzer	90%-Intervall	Schätzer	90%-Intervall
Alter der Mutter	-0.0038	(-0.0299, 0.0227)	-0.0058	(-0.0329, 0.0210)
Schulbildung Mutter	0.1172	(-0.2449, 0.4804)	0.1324	(-0.2415, 0.4917)
Dauer der Schwangerschaft	0.9167	( 0.3543, 1.4563)	0.8972	( 0.3403, 1.4619)
Art der Schwangerschaft	0.0294	(-0.6832, 0.7882)	0.0057	(-0.7432, 0.7766)
Art der Entbindung	-0.3286	(-0.7514, 0.0697)	-0.3155	(-0.7203, 0.0714)
Geschlecht	-0.3333	(-0.6956, 0.0336)	-0.3637	(-0.7266, 0.0245)
Gewicht des Neugeborenen	-0.0015	(-0.0018,-0.0012)	-0.0015	(-0.0018,-0.0012)
Kovariablen	Modell 4			
	Schätzer	90%-Intervall		
Alter der Mutter	-0.0082	(-0.0354, 0.0183)		
Schulbildung Mutter	0.0953	(-0.2410, 0.4399)		
Dauer der Schwangerschaft	0.8650	( 0.3573, 1.4109)		
Art der Schwangerschaft	0.0233	(-0.7026, 0.7478)		
Art der Entbindung	-0.3054	(-0.6761, 0.0631)		
Geschlecht	-0.3354	(-0.6722, 0.0183)		
Gewicht des Neugeborenen	-0.0014	(-0.0017,-0.0011)		

Tabelle 4.1: Geschätzte Regressionsparameter der Kovariablen mit 90%-Vertrauensintervallen in den verschiedenen Modellen.

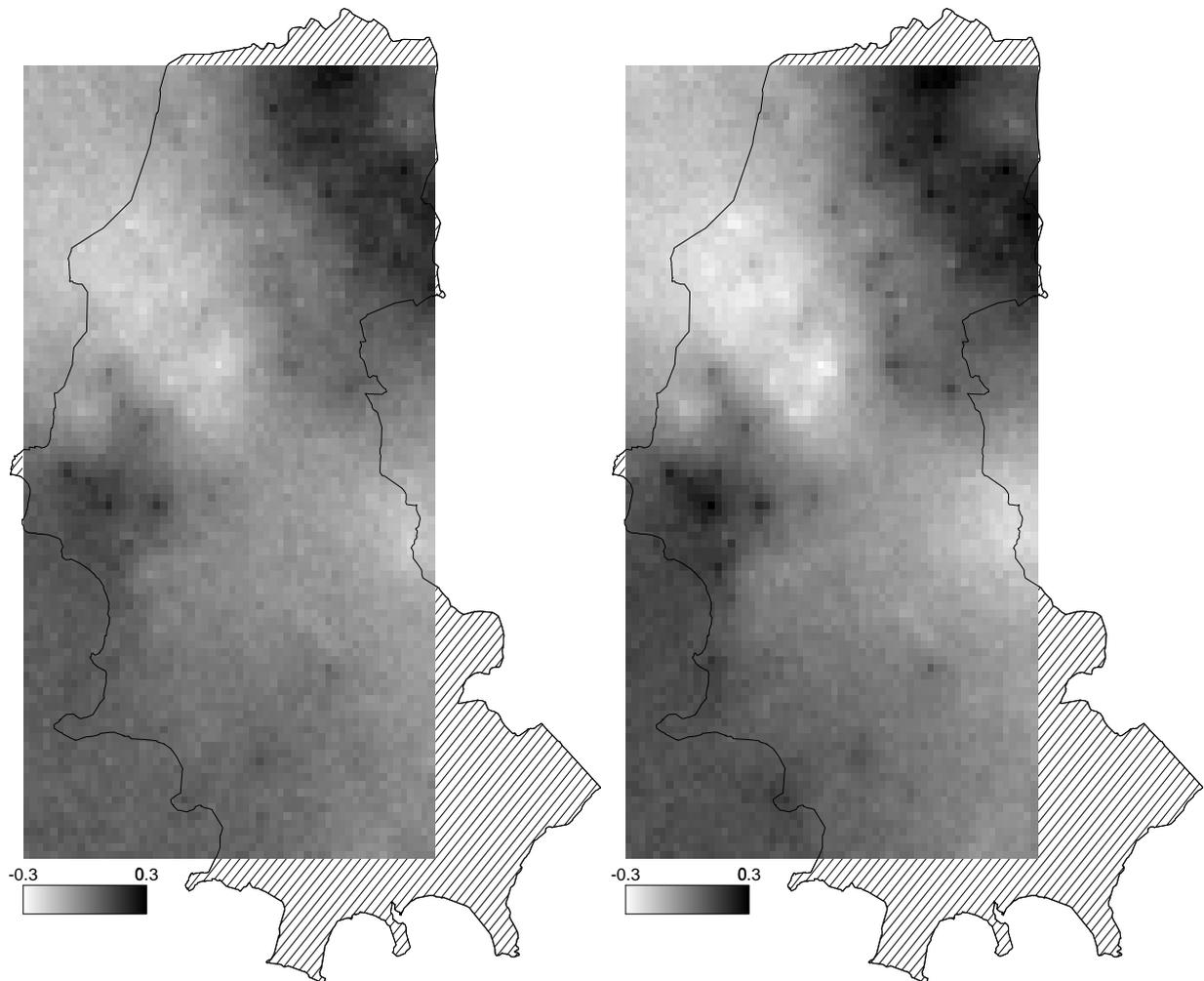


Abbildung 4.4: Posteriori-Median des räumlichen Effekts in den Modellen 1 und 1b.

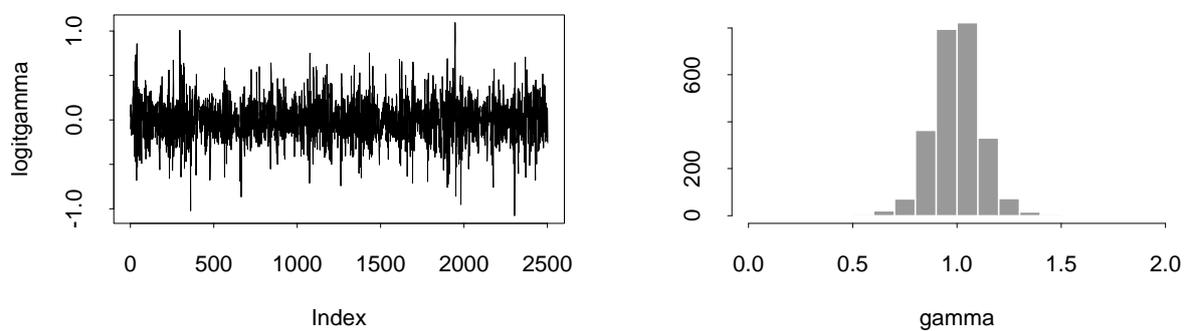


Abbildung 4.5: Samplefad und Histogramm der Posteriori von Gamma bei Modell 1b.

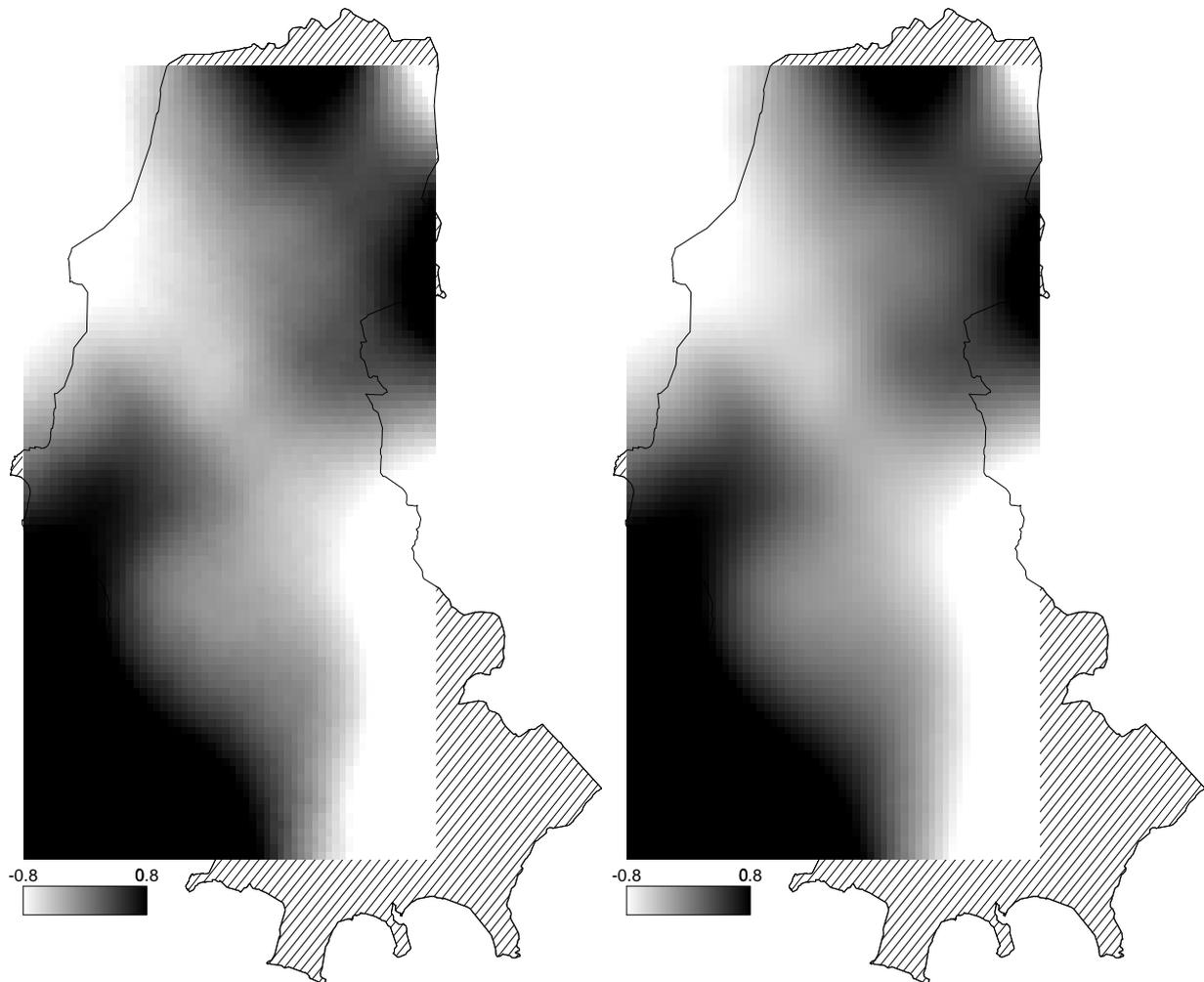


Abbildung 4.6: Posteriori-Median des räumlichen Effekts in den Modellen 2 und 2b.

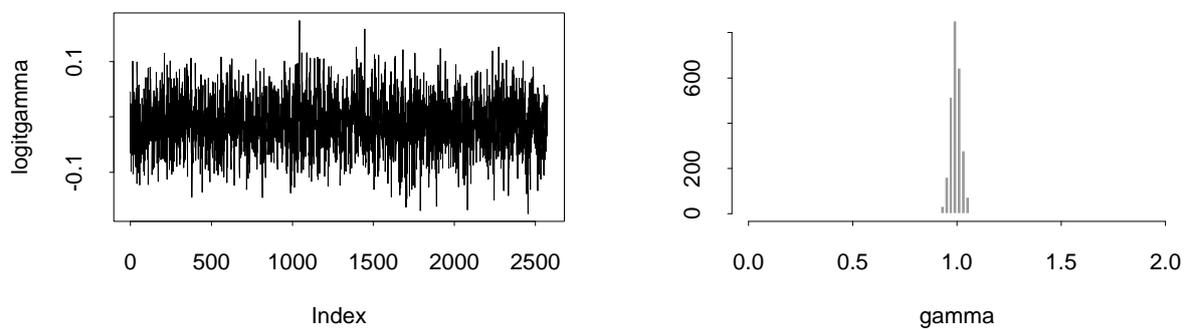


Abbildung 4.7: Samplepfad und Histogramm der Posteriori von Gamma bei Modell 2b.

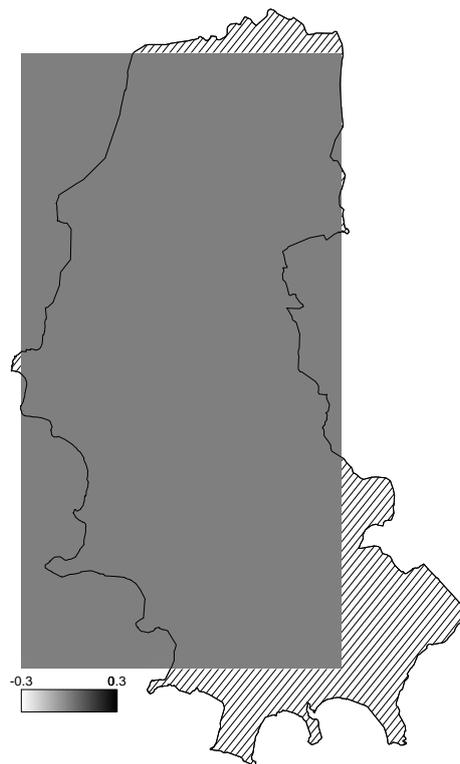


Abbildung 4.8: Posteriori-Median des räumlichen Effekts im Modell 4.

festgestellt. Dementsprechend unterscheiden sich die Schätzungen des räumlichen Effekts der Modelle 2 und 2b kaum, siehe Abbildung 4.6. Der räumliche Effekt bei Modell 2 gleicht am ehesten den Ergebnissen des Modells von Shimakura *et al.* (2001) mit Kreuzvalidierung.

Bei Modell 4 wird der räumliche Effekt praktisch konstant auf 0 geschätzt, siehe Abbildung 4.8. Die eher informative Priori dieses Modells führt zu einer starken Glättung. Bei der hier vorliegenden sehr geringen Anzahl an Daten werden diese durch das Modell komplett geglättet.

## 4.5 Diskussion

Daten, welche als räumliche Punktprozesse vorliegen, lassen sich nach Diskretisierung auch mit den üblichen Methoden für Daten mit diskreter räumlicher Verteilung analysieren. Wie

gezeigt ist dabei die simultane Schätzung sowohl des Glättungsparameter  $\tau$  als auch der Glättungsrichtung  $\gamma$  möglich, während dies bei der Modellierung als räumlicher Punktprozeß nicht möglich ist.

Im stetigen Fall hängt die Glättung vom vorher gewählten Glättungsparameter ab. Im diskreten Fall ist die Glättung stark von der Wahl der Nachbarschaftsstruktur abhängig, eine größere Anzahl von Nachbarn impliziert eine stärkere Glättung. Das Problem der Modellwahl wurde hier nicht genauer betrachtet. Möglich wäre einerseits die Auswahl eines Nachbarschaftsmodells durch ein geeignetes Gütekriterium wie z.B. das DIC (siehe Abschnitt 2.4.2). Eine andere Möglichkeit wäre die Mitschätzung der Nachbarschaftsstruktur, wobei man im Sinne des *posterior model averaging* zusätzlich über die Modelle glättet (Hoeting, Madigan, Raftery & Volinsky 1999).

Als Unterschied in den Ergebnissen der beiden Modellen zeigt sich, daß das diskrete Modell stärker in Richtung der Priori glättet. So wird z.B. bei wenigen Daten in einem Bereich die konstante Parameteroberfläche der Priori bevorzugt. Bei der Schätzung als räumlicher Punktprozeß wird dagegen die Punktschätzung auf Grundlage weniger Daten gemacht.



# Kapitel 5

## Alters-Perioden-Kohorten-Modelle

Eine häufig verwendete Analysemethode zur Beschreibung und Prognose zeitlicher Verläufe von Krankheitsraten sind Alters-Perioden-Kohorten-Modelle (*APC-Modelle*). Die eigentlich für eine Krankheit verantwortlichen Faktoren sind oft nicht bekannt, meist aber auch gar nicht direkt meßbar. Man analysiert die Krankheitsraten daher im Hinblick auf drei verschiedene Zeitskalen, Alter (Zeit zwischen Geburt und Infektion oder Tod), Periode (Infektions- oder Todeszeitpunkt) und Kohorte (Geburtszeitpunkt). Die Zeitskalen sind zusammenfassende Surrogatmaße, welche einen Hinweis auf die eigentlichen Gründe einer Krankheit liefern können (vgl. z.B. Armitage & Colton 1998, S. 82–99).

Das **Alter** wird in vielen Analysen als Einflußfaktor für das Risiko einer Erkrankung angenommen. Es steht für immer wieder auf eine Person einwirkende Faktoren. Der Effekt der **Periode** beruht auf Faktoren, welche auf alle unter Risiko stehenden Personen unabhängig vom Alter einwirken. Als Beispiel kann man die Umweltverschmutzung nennen, welche sich auf alle Personen relativ gleich auswirkt. Aber auch Fortschritte in der medizinischen Diagnostik und Technologie wirken sich als (positive) Periodeneffekte aus. Die **Kohorte** bezeichnet dagegen Effekte die eine *Generation* betreffen. Diese können einerseits durch äußere Umstände verursacht werden. So kann sich eine durch Krieg oder Seuchen verursachte Unterernährung bei Neugeborenen später als Kohorteneffekt zeigen. Aber auch sich ändernde Lebensgewohnheiten schlagen sich in den Kohorteneffekten nieder.

Als Beispiel sei hier die in den letzten Jahrzehnten steigende Zahl junger Raucherinnen genannt, welche sich als Kohorteneffekt vor allem bei der Lungenkrebsmortalität zeigen kann (Knorr-Held & Rainer 2001).

Im Rahmen dieser Arbeit entstand die Software **BAMP** zur **B**ayesianischen **A**lters-**P**erioden-**K**ohorten-**M**odellierung und **P**rognose. Alle in diesem Kapitel vorgestellten Modellansätze sind in BAMP implementiert. Die Analysen im letzten Abschnitt wurden mit BAMP durchgeführt. Die Software ist frei im Internet verfügbar. Weitere Details und das BAMP-Handbuch finden sich in Anhang B.

Dieses Kapitel enthält einen umfassenden Überblick über die Bayesianische Schätzung von APC-Modellen. Das folgende Kapitel 6 bespricht eine Erweiterung um räumliche Effekte und Raum-Zeit-Interaktionen.

## 5.1 Daten auf dem Lexis-Diagramm

Daten über Erkrankungen oder Sterbefälle liegen meist geschichtet nach Altersgruppen vor. Hat man solche Daten über mehrere Jahre (bezogen auf das Sterbe- oder Erkrankungsjahr), so kann man die Daten in einem Lexis-Diagramm (Lexis 1875) darstellen. Hierbei wird die Variable Alter auf der Ordinate und die Periode auf der Abszisse abgetragen. Beide Zeitskalen können dabei in verschieden lange Zeitintervalle aufgeteilt sein. Wie Abbildung 5.1 zeigt, können die Rohdaten in den entstehenden Rechtecken eingetragen werden (in der Abbildung sind Fallzahlen eines Beispieldatensatz eingetragen, der mit den APC-Effekten aus Abbildung 5.2 erzeugt wurde; pro Altersgruppe und Periode ist eine Bevölkerungszahl von 10 000 angenommen).

Zieht man von der linken unteren Ecke eines Datenrechtecks im 45 Grad-Winkel eine Linie zur Abszisse, so kommt man auf die Mitte der Kohorte des entsprechenden Datenpunkts. Zieht man dagegen von der linken oberen und der rechten unteren Ecke eines Datenrechtecks eine Linie nach unten, so erhält man Anfang und Ende der Kohorte, aus der die Personen, auf die sich die Daten beziehen, stammen. So gehören in diesem Beispiel

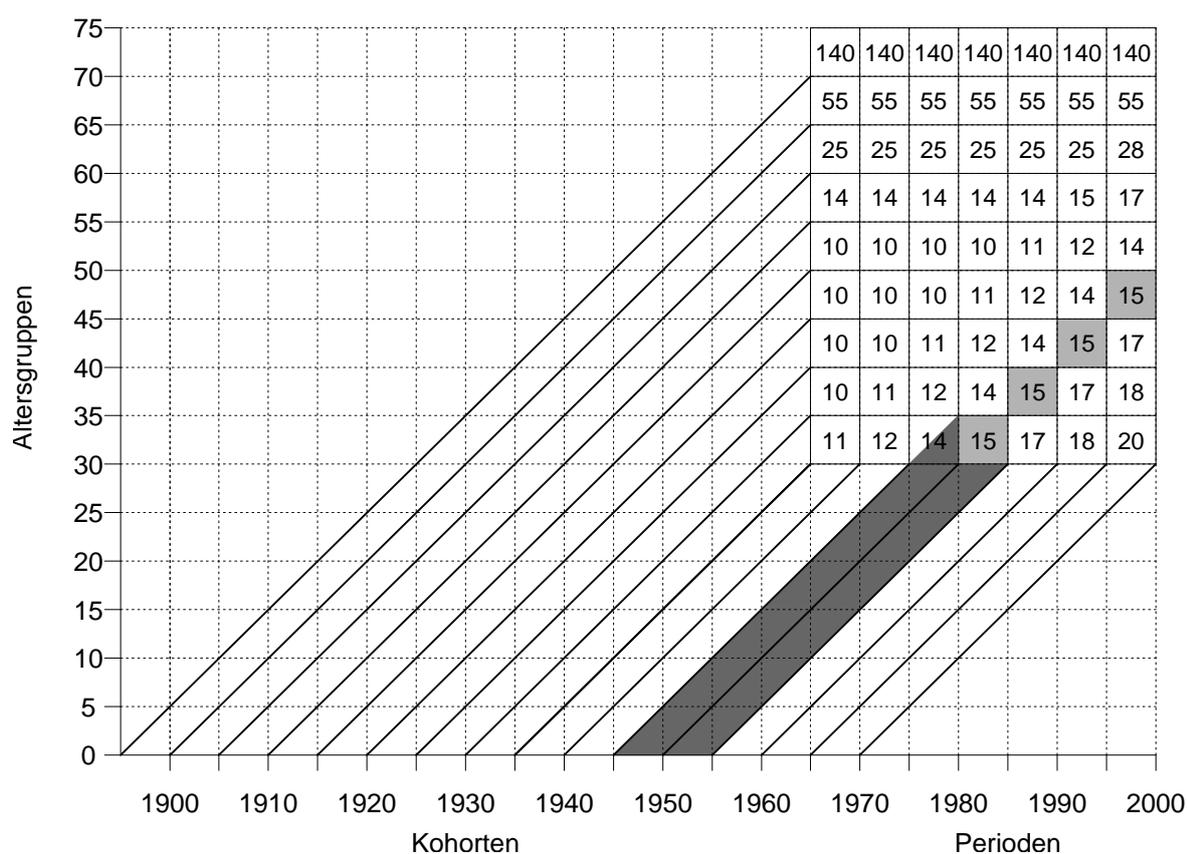


Abbildung 5.1: Lexis-Diagramm eines konstruierten Datensatzes (siehe Text). Die Zeitskala auf der Abszisse bezieht sich dabei sowohl auf Periode (vertikal abgelesen) als auch auf Kohorte (diagonal abgelesen).

die grau unterlegten Daten zur Kohorte 1945 bis 1955. Wie man sieht, überlappen sich die einzelnen Kohortengruppen.

Die Kohortenzugehörigkeit läßt sich eindeutig aus Altersgruppe und Periode bestimmen. Der Kohortenindex berechnet sich bei gleicher Skalierung von Altersgruppen und Perioden als

$$k = k(i, j) = (I - i) + j \quad (5.1)$$

wobei  $i$  der Index der Altersgruppe,  $j$  der Index der Periode und  $I$  die Anzahl der Perioden ist. Für unterschiedliche Skalierungen von Alter und Periode ist (5.1) entsprechend zu modifizieren (siehe dazu auch Holford 1983). So liegen zum Beispiel die Daten des in Abschnitt 5.3 analysierten COPD-Datensatzes jährlich vor, jedoch umfassen die Altersgruppen jeweils fünf Jahre. Die Kohortengruppen gehen damit über einen Zeitraum von

sechs Jahre, wobei sich aufeinanderfolgende Kohorten jeweils um fünf Jahre überlappen. Der Kohortenindex ergibt sich über die Formel

$$k = k(i, j) = 5 \cdot (I - i) + j. \quad (5.2)$$

Zur Übersichtlichkeit bezeichne der Buchstabe  $k$  im Folgenden immer den Kohortenindex in Abhängigkeit von  $i$  und  $j$ . Die Anzahl der Perioden sei  $J$ , die Anzahl der Kohorten  $K$ .

### 5.1.1 Das APC-Modell

Alters-Perioden-Kohortenmodelle werden in der Literatur oft als generalisiertes lineares Modell mit Likelihood-Methoden geschätzt, z.B. als log-lineares Poisson-Modell (für einen Überblick siehe Heuer 1994). Das heißt die Anzahl der Sterbe- oder Inzidenzfälle  $y_{ij}$  der Altersgruppe  $i$  in Periode  $j$  ist Poissonverteilt mit Rate  $n_{ij}\lambda_{ij}$ . Dabei bezeichnet  $n_{ij}$  die Population der Altersgruppe  $i$  zum Zeitpunkt  $j$  und  $\lambda_{ij}$  einen unbekanntem Parameter. Der Logarithmus dieses Parameters wird zerlegt in einen Intercept  $\mu$ , Alterseffekte  $\theta_i$ , Periodeneffekte  $\phi_j$  und Kohorteneffekte  $\psi_k$ ,

$$\log(\lambda_{ij}) = \eta_{ij} = \mu + \theta_i + \phi_j + \psi_k.$$

Um den Intercept  $\mu$  identifizierbar zu machen, werden Zentrierungsrestriktionen für die Haupteffekte in das Modell eingeführt:

$$\sum_i \theta_i = \sum_j \phi_j = \sum_k \psi_k = 0.$$

Eine Alternative zum log-linearen Poisson-Modell ist ein Binomialmodell mit Logit-Link. Beide Modelle entsprechen sich approximativ. Das Binomialmodell lautet

$$y_{ij} \sim B(n_{ij}, p_{ij}) \quad (5.3)$$

$$\eta_{ij} = \log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \mu + \theta_i + \phi_j + \psi_k. \quad (5.4)$$

### 5.1.2 Identifizierbarkeitsproblem

Analysen mit APC-Modellen haben alle mit einem Problem zu kämpfen: Die Effekte in diesem Modell sind nicht identifizierbar. Wie man aus (5.1) leicht sieht, lassen sich die Werte der Effekte für jedes beliebige  $c \in \mathbb{R}$  in der Art

$$\mu \rightarrow \mu - c \cdot I; \quad \theta_i \rightarrow \theta_i + c \cdot i; \quad \phi_j \rightarrow \phi - c \cdot j; \quad \psi_k \rightarrow \psi + c \cdot k \quad (5.5)$$

für alle  $i, j$  bzw.  $k$  transformieren, ohne daß sich der lineare Prädiktor  $\eta_{ij}$  dabei verändert (eine analoge Transformation läßt sich auch bei unterschiedlichen Skalierungen von Alter und Periode finden). Der lineare Prädiktor kann also bei der Schätzung identifiziert und auch interpretiert werden, nicht aber die einzelnen Haupteffekte.

Abbildung 5.2 veranschaulicht die Problematik der Nicht-Identifizierbarkeit. Die Abbildung zeigt zwei mögliche Effektschätzungen für den Beispieldatensatz aus Abbildung 5.1. Je nach Transformation ergibt sich nun entweder ein steigender oder ein fallender Periodeneffekt. Parallel dazu ergibt sich für die Kohorten ein fallender oder steigender Effekt. Der „Knick“ bei der achten Kohorte bleibt aber bei beiden Transformationen erhalten. Das Nicht-Identifizierbarkeitsproblem betrifft also nur den linearen Trend, Änderungen des Trends können identifiziert werden. Auch quadratische Trends wie im Beispiel beim Alterseffekt ab der fünften Altersgruppe können, bis auf die lineare Komponente, erkannt werden. Die Interpretation von geschätzten Effekten im APC-Modell muß also auf nicht-lineare Trends beschränkt bleiben.

In der Literatur werden verschiedene Ansätze präsentiert, wie mit dem Nicht-Identifizierbarkeitsproblem umgegangen werden kann. Heuer (1994, S. 40–54) bietet einen Überblick über Parametrisierungsansätze im Rahmen der Likelihoodinferenz. Robertson & Boyle (1998) vergleichen vier in der Epidemiologie populäre Ansätze. Berzuini, Clayton & Bernardinelli (1993) schlagen vor, nur die zweiten Differenzen der Effekte zu betrachten.

Nakamura (1986) weist darauf hin, daß beim empirischen Bayes-Ansatz eine auf ersten Differenzen basierende Priori die Zeiteffekte identifizierbar macht. Gemäß Knorr-Held & Rainer (2001) impliziert ein Modell mit Random Walks erster Ordnung eine stochastische

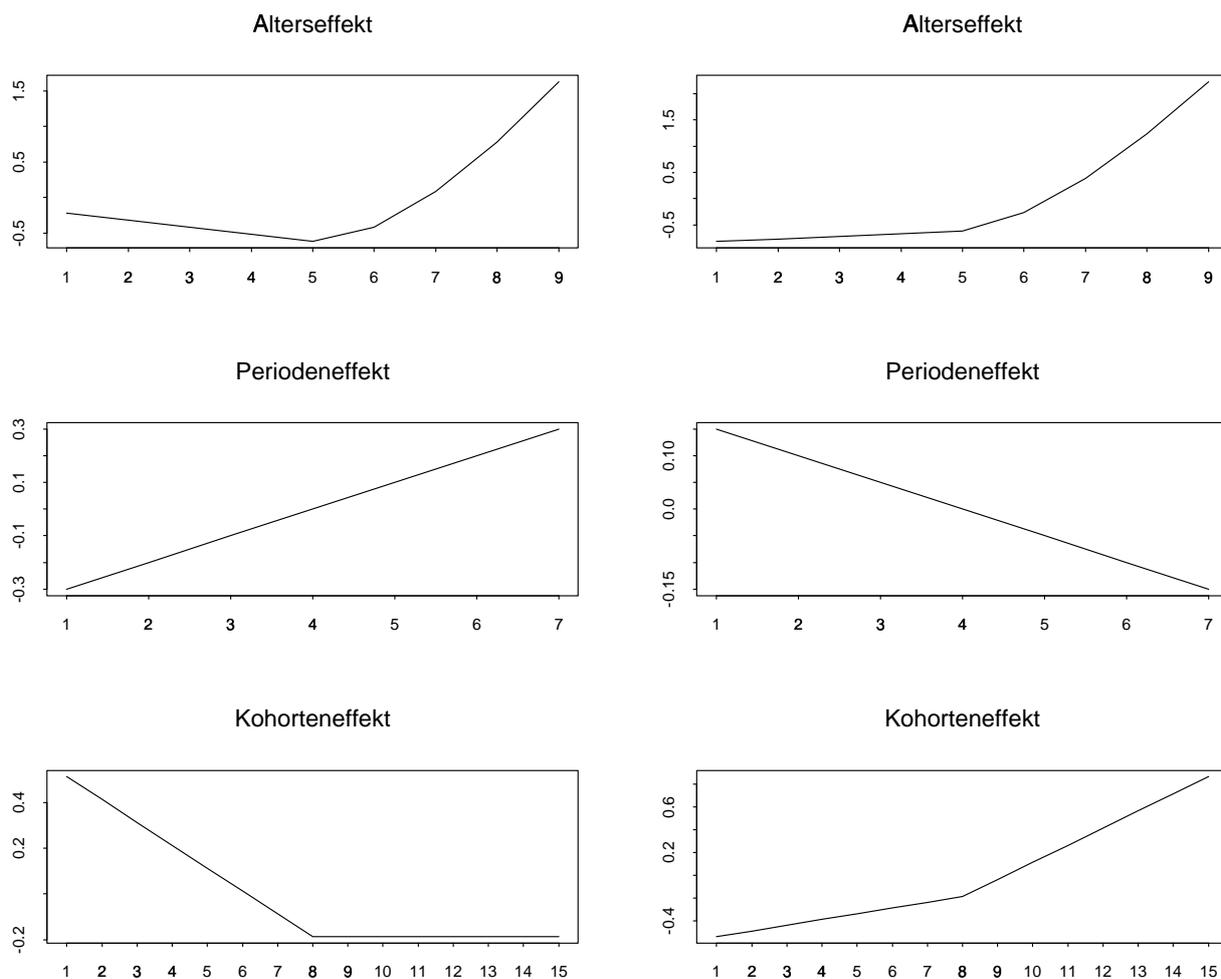


Abbildung 5.2: Zwei kongruente Effektschätzungen für die Beispieldaten aus Abb. 5.1; links ist  $\mu = -6.1$ , rechts  $\mu = -5.2$ .

Restriktion auf die Alters-, Perioden- und Kohortenparameter. Das Modell präferiert a priori unter allen möglichen Werten von  $c \in \mathbb{R}$  in (5.5) denjenigen, bei dem die quadratischen ersten Differenzen (gewichtet mit dem entsprechenden Präzisionsparameter) minimal werden. Das erlaubt, (nichtlineare) Trends in den Alters-, Perioden- und Kohortenparametern graphisch zu untersuchen. Ohne Restriktion ist es dagegen gar nicht möglich, die Haupteffekte explizit zu schätzen.

## 5.2 Bayesianische Alters-Perioden-Kohorten-Modelle

### 5.2.1 Random Walks und Prognose

Random Walk Prioris finden im APC-Modell für die Parameter  $\boldsymbol{\theta}$ ,  $\boldsymbol{\phi}$  und  $\boldsymbol{\psi}$  Anwendung. Grundlegende Idee der Random Walk Prioris ist die Annahme, daß der Effekt zum Zeitpunkt  $j$  nicht sehr stark vom Effekt zum Zeitpunkt  $j-1$  abweicht. Für genauere Diskussion von Random Walk Prioris siehe Abschnitt 2.2.2. Es gilt für Random Walks erster Ordnung (RW 1)

$$\phi_j | \phi_{j-1} \sim N(\phi_{j-1}, \kappa^{-1}), \quad (5.6)$$

wobei  $\kappa^{-1}$  ein Präzisionsparameter ist. Für Random Walks zweiter Ordnung (RW 2) gilt

$$\begin{aligned} p(\phi_1) &= p(\phi_2) \propto \text{const.} \\ \phi_j | \phi_{j-1}, \phi_{j-2} &\sim N(2\phi_{j-1} - \phi_{j-2}, \kappa^{-1}) \quad \text{für alle } j = 3, \dots, J. \end{aligned}$$

Während der RW 1 den Trend konstant von Zeitpunkt zu Zeitpunkt fortsetzt, geht man beim RW 2 a priori von einer linearen Entwicklung des Trends aus. Dementsprechend wählt man auch die Ordnung der Random Walk-Priori in Abhängigkeit davon, ob man a priori einen konstanten oder linearen Effekt erwartet.

Die gemeinsame Verteilung von  $\boldsymbol{\phi}$  lautet

$$p(\boldsymbol{\phi}) \propto \lambda^{rg(\mathbf{R})/2} \exp\left(-\frac{\lambda}{2} \boldsymbol{\phi}' \mathbf{R} \boldsymbol{\phi}\right),$$

wobei  $\mathbf{R} = \mathbf{R}^{(1)}$  für RW 1 und  $\mathbf{R} = \mathbf{R}^{(2)}$  für RW 2 mit  $\mathbf{R}^{(1)}$  und  $\mathbf{R}^{(2)}$  definiert wie in Abschnitt 2.2.2. Die Prioris der Alterseffekte  $\boldsymbol{\theta}$  und der Kohorteneffekte  $\boldsymbol{\psi}$  wird analog definiert.

Wie erwähnt, bringt ein Random Walk erster Ordnung durch die Priori auf die linearen Trends eine stochastische Restriktion auf die linearen Zeittrends in das Modell ein. Damit werden die Effekte trotz der oben beschriebenen Problematik identifizierbar. Die genannten Einschränkungen bei der Interpretation bleiben aber bestehen.

Die Präzision  $\kappa$  wird im Modell mitgeschätzt. Sie spielt dabei die Rolle eines Glättungsparameters. Je kleiner die Präzision, desto geringer die Glättung. Als Priori verwendet man meist eine  $\text{Ga}(a, b)$ -Verteilung. Standardwerte für die Hyperparameter sind  $a = 1$  und  $b = 0.001$  oder  $a = b = 0.001$ . Für den RW 2 empfiehlt es sich,  $b$  noch kleiner zu wählen ( $b = 0.00001$ ), um eine genügend glatte Schätzung zu erhalten.

Der Random Walk erster Ordnung geht von einem konstanten zeitlichen Trend aus. Dagegen setzt ein Random Walk zweiter Ordnung a priori lineare Trends weiter fort. Da in beiden Fällen die Verteilung des Effekts nur von den vorherigen Zeitpunkten abhängt, lassen sich beide Prioris gut dazu verwenden, Prognosen für zukünftige Zeitpunkte zu erstellen. Die zukünftigen Effekte ergeben sich dabei gemäß

$$\begin{aligned}\phi_{J+1} &\sim \text{N}(\phi_J, \kappa^{-1}) && \text{für RW 1,} \\ \phi_{J+1} &\sim \text{N}(2\phi_J - \phi_{J-1}, \kappa^{-1}) && \text{für RW 2.}\end{aligned}$$

Im APC-Modell lassen sich Perioden- und Kohorteneffekt auf diese Weise fortsetzen. Auch die Alterseffekte lassen sich so für jüngere und ältere Altersgruppen prognostizieren, dies ist für die Prognose zukünftiger Raten aber nicht notwendig.

Man erhält die Prognose einer zukünftigen Rate  $\eta_{ij}$  für  $j > J$ , indem man die prognostizierten Perioden- und Kohortentrends in (5.4) einsetzt. Die Prognose zukünftiger Fälle erhält man durch Anwendung der Binomialverteilung

$$y_{ij} \sim \text{B}\left(n_{ij}, \frac{\exp(\eta_{ij})}{1 + \exp(\eta_{ij})}\right), \quad (5.7)$$

dazu muß die Bevölkerungszahl  $n_{ij}$  bekannt sein (oder gegebenenfalls durch ein passendes Verfahren prognostiziert werden).

Mit einer Random Walk Priori erster Ordnung tritt allerdings ein Problem bezüglich der Nicht-Identifizierbarkeit auf. Sei im folgenden  $\tilde{\mu} = \mu - c \cdot I$ ,  $\tilde{\theta}_i = \theta_i + c \cdot i$ ,  $\tilde{\phi} = \phi - c \cdot j$  und  $\tilde{\psi}_k = \psi_k + c \cdot k$  für alle  $i, j$  bzw.  $k$  eine Transformation der APC-Effekte nach (5.5) und  $\phi_{J+1} = \phi_J + \epsilon$  mit  $\epsilon \sim \text{N}(0, \kappa^{-1})$  eine Prognose für den Periodeneffekt zum Zeitpunkt  $J + 1$ . Berechnet man nun den prognostizierten linearen Prädiktor  $\eta_{i, J+1}$  für die nächste

Periode für eine Altersgruppe  $i > 1$ , so gilt offensichtlich

$$\begin{aligned}
\eta_{i,J+1} &= \mu + \theta_i + \phi_{J+1} + \psi_k \\
&= \mu + \theta_i + \phi_J + \epsilon + \psi_{(I-i)+J+1} \\
&= \tilde{\mu} + c \cdot I + \tilde{\theta}_i - c \cdot i + \tilde{\phi}_J + c \cdot J + \epsilon + \tilde{\psi}_{(I-i)+J+1} - c \cdot (I - i + J + 1) \\
&= \tilde{\mu} + \tilde{\theta}_i + \tilde{\phi}_J + \epsilon + \tilde{\psi}_{(I-i)+J+1} - c \\
&= \tilde{\eta}_{i,J+1} - c.
\end{aligned}$$

Die Prognose des linearen Prädiktors ist also nicht von der Transformation unabhängig. Nur für die Altersgruppe  $i = 1$  geht bei der Prognose von  $\eta_{1,J+1}$  auch die Prognose des Kohorteneffekts  $\psi_{I-1+J+1} = \psi_{K+1}$  ein und der Transformationsfaktor  $c$  fällt heraus.

Das heißt also, daß die Prognose von der Transformation abhängt. Sie kann bei einer RW 1-Priori nur dann sinnvolle Ergebnisse liefern, wenn wirklich ein relativ konstanter Trend für den Periodeneffekt vorliegt. Beim Random Walk zweiter Ordnung tritt das Problem nicht auf. Hier gilt für die Prognose

$$\phi_{J+1} = 2\phi_J + \phi_{J-1} + \epsilon$$

und folglich

$$\begin{aligned}
\eta_{i,J+1} &= \mu + \theta_i + \phi_{J+1} + \psi_k \\
&= \mu + \theta_i + 2\phi_J - \phi_{J-1} + \epsilon + \psi_{(I-i)+J+1} \\
&= \tilde{\mu} + c \cdot I + \tilde{\theta}_i - c \cdot i + 2\tilde{\phi}_J + 2c \cdot J - \tilde{\phi}_{J-1} - c \cdot (J - 1) + \epsilon + \tilde{\psi}_{(I-i)+J+1} \\
&\quad - c \cdot (I - i + J + 1) \\
&= \tilde{\mu} + \tilde{\theta}_i + \tilde{\phi}_J + \epsilon + \tilde{\psi}_{(I-i)+J+1} \\
&= \tilde{\eta}_{i,J+1},
\end{aligned}$$

die Prognose des linearen Prädiktors und damit auch die Anzahl prognostizierter Fälle ist also unabhängig von der Transformation.

### 5.2.2 Inferenz

Im Folgenden sei  $\tau$  die Präzision des Random Walks für  $\boldsymbol{\theta}$ ,  $\kappa$  die für  $\boldsymbol{\phi}$  und  $\nu$  die für  $\boldsymbol{\psi}$ . Die Matrix  $\mathbf{R}$  steht alternativ für die Präzisionsmatrix des Random Walks erster ( $\mathbf{R} = \mathbf{R}^{(1)}$ ) oder zweiter Ordnung ( $\mathbf{R} = \mathbf{R}^{(2)}$ ). Die gemeinsame Posteriori aller Parameter des Modells lautet

$$p(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\psi}, \tau, \kappa, \nu | y_{ij}) \propto \left( \prod_{i,j} l(y_{ij} | \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\psi}) \right) p(\boldsymbol{\theta} | \tau) p(\tau) p(\boldsymbol{\phi} | \kappa) p(\kappa) p(\boldsymbol{\psi} | \nu) p(\nu).$$

Die Likelihood  $l(y_{ij} | \cdot)$  ergibt sich aus dem gewählten Modell. Im Folgenden verwenden wir nur das Binomialmodell

$$\begin{aligned} l(y_{ij} | \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\psi}) &\propto p_{ij}^{y_{ij}} (1 - p_{ij})^{n_{ij} - y_{ij}} \\ &= \left( \frac{\exp \eta_{ij}}{1 + \exp \eta_{ij}} \right)^{y_{ij}} \left( \frac{1}{1 + \exp \eta_{ij}} \right)^{n_{ij} - y_{ij}} \\ &= \exp \left( y_{ij} (\mu + \theta_i + \phi_j + \psi_k) \right) \left( 1 + \exp(\mu + \theta_i + \phi_j + \psi_k) \right)^{-n_{ij}}. \end{aligned}$$

Die Full Conditional von  $\boldsymbol{\theta}$  berechnet sich als

$$\begin{aligned} p(\boldsymbol{\theta} | \cdot) &\propto \prod_{i,j} l(y_{ij} | \boldsymbol{\theta}) p(\boldsymbol{\theta} | \tau) \\ &\propto \exp \sum_{i,j} \left( y_{ij} \theta_i + n_{ij} \log (1 + \exp(\mu + \theta_i + \phi_j + \psi_k)) \right) \cdot \exp \left( -\frac{\tau}{2} \boldsymbol{\theta}' \mathbf{R} \boldsymbol{\theta} \right) \\ &= \exp \left( -\frac{\tau}{2} \boldsymbol{\theta}' \mathbf{R} \boldsymbol{\theta} + \sum_{i,j} \theta_i y_{ij} - \sum_{i,j} n_{ij} \cdot f_j(\theta_i) \right) \end{aligned} \quad (5.8)$$

mit  $f_j(\theta_i) = \log (1 + \exp(\mu + \theta_i + \phi_j + \psi_k))$ . Hierbei handelt es sich offensichtlich nicht um eine Standardverteilung. Wir verwenden für die Ziehung des Parametervektors einen Metropolis-Hastings-Algorithmus. Als Proposal findet eine Normalverteilung Anwendung. Sei  $\theta_i^0$  der aktuelle Wert des Parameters. Dann approximieren wir  $f_j(\theta_i)$  mit einer Taylor-Entwicklung zweiten Grades um  $\theta_i^0$  (vergleiche Abschnitt 2.3.2):

$$\begin{aligned} f_j(\theta_i) &\approx f_j(\theta_i^0) + (\theta_i - \theta_i^0) f_j'(\theta_i^0) + \frac{1}{2} (\theta_i - \theta_i^0)^2 f_j''(\theta_i^0) \\ &= \theta_i \left( f_j'(\theta_i^0) - \theta_i^0 f_j''(\theta_i^0) \right) + \frac{1}{2} \theta_i^2 f_j''(\theta_i^0) + \text{const.} \end{aligned}$$

mit

$$f'_j(\theta_i^0) = \frac{\exp(\mu + \theta_i^0 + \phi_j + \psi_k)}{1 + \exp(\mu + \theta_i^0 + \phi_j + \psi_k)}$$

und

$$f''_j(\theta_i^0) = f'_j(\theta_i^0) - f_j'(\theta_i^0)^2.$$

Die Proposalverteilung ist also

$$\boldsymbol{\theta}|\boldsymbol{\theta}^0 \sim N(\mathbf{V}_\theta^{-1}\mathbf{m}_\theta, \mathbf{V}_\theta^{-1}) \quad (5.9)$$

mit

$$\mathbf{V}_\theta = \mathbf{R} + \text{diag}\left(\sum_j n_{1j} f''_j(\theta_1^0), \dots, \sum_j n_{Ij} f''_j(\theta_I^0)\right),$$

$\mathbf{m}_\theta = (m_1, \dots, m_I)'$  und

$$m_i = \sum_j \left( y_{ij} - n_{ij} f'_j(\theta_i^0) + n_{ij} \theta_i^0 f''_j(\theta_i^0) \right) \text{ für } i = 1, \dots, I.$$

Aus der multivariaten Normalverteilung (5.9) läßt sich mit dem Algorithmus von Rue (2001) ziehen (siehe Abschnitt 2.3.1). Die Präzisionsmatrix  $\mathbf{V}_\theta$  ist wie  $\mathbf{R}$  eine Bandmatrix mit Bandweite 1 für RW 1 und Bandweite 2 für RW 2. Die Cholesky-Zerlegung läßt sich daher schnell ausführen. Die Akzeptanzwahrscheinlichkeit im Metropolis-Hastings-Schritt lautet

$$\alpha = \frac{p(\boldsymbol{\theta}|\cdot)q(\boldsymbol{\theta}^0|\boldsymbol{\theta})}{p(\boldsymbol{\theta}^0|\cdot)q(\boldsymbol{\theta}|\boldsymbol{\theta}^0)}.$$

mit der Full Conditional (5.8) und der Dichte der Normalverteilung (5.9) für  $q(\boldsymbol{\theta}|\boldsymbol{\theta}^0)$ . Zur Berechnung von  $q(\boldsymbol{\theta}^0|\boldsymbol{\theta})$  und weitere technische Details vergleiche Abschnitt 2.3.2.

Als Proposal für den Periodeneffekt  $\boldsymbol{\phi}$  ergibt sich analog eine Normalverteilung mit Präzision

$$\mathbf{V}_\phi = \mathbf{R} + \text{diag}\left(\sum_i n_{i1} g''_i(\phi_1^0), \dots, \sum_i n_{iJ} g''_i(\phi_J^0)\right)$$

und Mittelwertsvektor  $\mathbf{V}_\phi^{-1}\mathbf{m}_\phi$  mit  $\mathbf{m}_\phi = (m_1, \dots, m_J)'$  und

$$m_j = \sum_i \left( y_{ij} - n_{ij} g'_i(\phi_j^0) + n_{ij} \phi_j^0 g''_i(\phi_j^0) \right) \text{ für } j = 1, \dots, J,$$

sowie

$$g'_i(\phi_j^0) = \frac{\exp(\mu + \theta_i + \phi_j^0 + \psi_k)}{1 + \exp(\mu + \theta_i + \phi_j^0 + \psi_k)}$$

und

$$g''_i(\phi_j^0) = g'_i(\phi_j^0) - g'_i(\phi_j^0)^2.$$

Beim Proposal für den Kohorteneffekt ist zu beachten, daß die Summation über alle Altersgruppen und Perioden der entsprechenden Kohorte erfolgen muß. Das heißt, als Proposal wird eine Normalverteilung mit Präzisionsmatrix

$$\mathbf{V}_\psi = \mathbf{R} + \text{diag}\left(\sum_{\{i,j:k(i,j)=1\}} n_{ij} h''_{ij}(\psi_1^0), \dots, \sum_{\{i,j:k(i,j)=K\}} n_{ij} h''_{ij}(\psi_K^0)\right)$$

und Mittelwertsvektor  $\mathbf{V}_\psi^{-1} \mathbf{m}_\psi$  mit  $\mathbf{m}_\psi = (m_1, \dots, m_K)'$  und

$$m_l = \sum_{\{i,j:k(i,j)=l\}} (y_{ij} - n_{ij} h'_{ij}(\psi_l^0) + n_{ij} \psi_l^0 h''_{ij}(\psi_l^0)) \quad \text{für } l = 1, \dots, K$$

und

$$h'_{ij}(\psi_k) = \frac{\exp(\mu + \theta_i + \phi_j + \psi_k^0)}{1 + \exp(\mu + \theta_i + \phi_j + \psi_k^0)}$$

sowie

$$h''_{ij}(\psi_k^0) = h'_{ij}(\psi_k^0) - h'_{ij}(\psi_k^0)^2$$

verwendet.

Die Full Conditional des Hyperparameters  $\tau$  ist

$$\begin{aligned} p(\tau|\cdot) &\propto p(\boldsymbol{\theta}|\tau)p(\tau) \\ &\propto \tau^{rg(\mathbf{R})/2} \exp\left(-\frac{\tau}{2} \boldsymbol{\theta}' \mathbf{R} \boldsymbol{\theta}\right) \cdot \tau^{a-1} \exp(-b\tau) \\ &\propto \tau^{rg(\mathbf{R})/2+a-1} \exp\left(-\tau\left(b + \frac{1}{2} \boldsymbol{\theta}' \mathbf{R} \boldsymbol{\theta}\right)\right). \end{aligned}$$

Dies entspricht der Dichte einer Gammaverteilung mit den Parametern  $a + rg(\mathbf{R})/2$  und  $b + \frac{1}{2} \boldsymbol{\theta}' \mathbf{R} \boldsymbol{\theta}$ . Analog ergeben sich auch für  $\kappa$  und  $\nu$  Gammaverteilungen als Full Conditionals und die Hyperparameter lassen sich mittels Gibbs-Sampling ziehen.

Für den Intercept  $\mu$  verwendet man eine flache Priori

$$p(\mu) \propto \text{const.}$$

Der Parameter läßt sich dann wie oben in einem Metropolis-Hastings-Schritt mit approximiertem Normalverteilungsproposal schätzen. Alternativ dazu datiert man z. B.  $\mu + \boldsymbol{\theta}$  statt  $\boldsymbol{\theta}$  auf, indem man auf die Zentrierung des Alterseffekts verzichtet.

### 5.2.3 Modell mit zusätzlicher Heterogenität

Knorr-Held & Rainer (2001) verwenden statt (5.4) ein Modell mit einem zusätzlichem Parameter  $z_{ij}$  für Heterogenität, welche nicht von den Alters-, Perioden- und Kohorteneffekten erklärt werden kann,

$$\eta_{ij} = \mu + \theta_i + \phi_j + \psi_k + z_{ij}. \quad (5.10)$$

Als Priori von  $z_{ij}$  wird  $z_{ij} \sim N(0, \delta^{-1})$  verwendet, mit einer  $\text{Ga}(a_\delta, b_\delta)$ -Priori für die Präzision  $\delta$ . Bei der Prognose zukünftiger Raten muß neben  $\phi_j$  und  $\psi_{k(i,j)}$  zusätzlich eine Prognose des Parameters  $z_{ij}$  geschätzt und in (5.10) eingesetzt werden.

Das Modell hat einen entscheidenden technischen Vorteil. Reparametrisiert man das Modell wie Besag *et al.* (1995) von  $z_{ij}$  nach  $\eta_{ij}$  für alle  $i$  und  $j$ , so sind die Full Conditionals der Haupteffekte Normalverteilungen und können in Gibbs-Schritten gezogen werden (vgl. Knorr-Held & Rainer 2001, Appendix — statt dem dort beschriebenen *forward filtering-backward sampling*-Algorithmus verwenden wir aber den Algorithmus nach Rue 2001).

Die Full Conditionals der Haupteffekte sind nach Rainer (1998)

$$\begin{aligned} \boldsymbol{\theta} | \cdot &\sim N(\mathbf{w}_\theta, \mathbf{U}_\theta) \\ \mathbf{U}_\theta &= \tau \mathbf{R} + \delta \mathbf{J} \mathbf{I} \\ \mathbf{w}_\theta &= \delta \mathbf{U}_\theta^{-1} \left( \sum_j \xi_{1j}^\theta, \dots, \sum_j \xi_{Ij}^\theta \right)' \\ \xi_{ij}^\theta &= \eta_{ij} - \phi_j - \psi_k \text{ für alle } i, j \end{aligned}$$

$$\begin{aligned}
\phi| \cdot &\sim N(\mathbf{w}_\phi, \mathbf{U}_\phi) \\
\mathbf{U}_\phi &= \kappa \mathbf{R} + \delta \mathbf{\Pi} \\
\mathbf{w}_\phi &= \delta \mathbf{U}_\phi^{-1} \left( \sum_i \xi_{i1}^\phi, \dots, \sum_i \xi_{iJ}^\phi \right)' \\
\xi_{ij}^\phi &= \eta_{ij} - \theta_i - \psi_k \text{ für alle } i, j \\
\\
\psi| \cdot &\sim N(\mathbf{w}_\psi, \mathbf{U}_\psi) \\
\mathbf{U}_\psi &= \nu \mathbf{R} + \delta \text{diag}(n_1, \dots, n_K) \\
n_k &= \sum_{\{i,j:k(ij)=l\}} 1 \quad \text{für } l = 1, \dots, K \\
\mathbf{w}_\psi &= \delta \mathbf{U}_\psi^{-1} \left( \sum_{k(ij)=1} \xi_{ij}^\psi, \dots, \sum_{k(ij)=K} \xi_{ij}^\psi \right)' \\
\xi_{ij}^\psi &= \eta_{ij} - \theta_i - \phi_j \text{ für alle } i, j
\end{aligned}$$

Die Parameter  $\eta_{ij}$  werden unabhängig in Metropolis-Hastings-Schritten gezogen. Die Full Conditional lautet nach Rainer (1998)

$$p(\eta_{ij}|\cdot) \propto \frac{\exp(\eta_{ij} y_{ij} - \frac{\delta}{2}(\eta_{ij} - \mu - \theta_i - \phi_j - \psi_k)^2)}{(1 + \exp(\eta_{ij}))^{n_{ij}}}. \quad (5.11)$$

Einem Ansatz von Knorr-Held & Raßer (2000, Appendix) folgend, verwenden wir ein Log-Gamma-Proposal für das Aufdatieren von  $\eta_{ij}$ , d.h. wir verwenden den Logarithmus einer  $\text{Ga}(y_{ij} + \nu_{ij}^2/\sigma_{ij}^2, n_{ij} + \nu_{ij}/\sigma_{ij}^2)$ -verteilten Zufallsvariable. Dabei sind

$$\nu_{ij} = \exp(\mu + \theta_i + \phi_j + \psi_k + 0.5 \delta^{-1})$$

und

$$\sigma_{ij}^2 = \nu_{ij}^2 (\exp(\delta^{-1}) - 1)^2 (\exp(\delta^{-1}) - 1)^{-1}$$

Erwartungswert und Varianz der bedingten (lognormalen) Prioriverteilung von  $\exp(\eta_{ij})|\delta$ . Im Regelfall erhält man damit Akzeptanzwahrscheinlichkeiten von etwa 99%.

Die Akzeptanzwahrscheinlichkeit sinkt aber bei sehr geringen Fallzahlen stark ab. In diesem Fall verwenden wir den Logarithmus einer  $\text{Ga}(\nu_{ij}^2/\sigma_{ij}^2, \nu_{ij}/\sigma_{ij}^2)$ -verteilten Zufallsvariable.

riable als Proposal und erreichen wieder Akzeptanzwahrscheinlichkeiten um die 99% (die Auswahl des Proposals erfolgt in einem Tuning-Schritt zu Beginn des Algorithmus).

### 5.2.4 Unstrukturierte Zeiteffekte

In der Praxis tritt gelegentlich das Problem auf, daß die APC-Effekte zu wenig glatt geschätzt werden, weil zusätzliche Heterogenität in den Effekten oder auch nur einzelne Ausreißer auftreten. Damit wird auch die Schätzung der Varianz des Effekts größer und die Prognoseintervalle breiter.

Eine Verbesserung erhält man, indem man einen Parameter für zusätzliche Heterogenität einer Zeitskala einführt. Auch beim Modell mit globalen Heterogenitätsparameter  $z_{ij}$  (Abschnitt 5.2.3) kann ein zusätzlicher Heterogenitätsparameter für eine Zeitskala sinnvoll sein, um die Varianz von  $z_{ij}$  möglichst gering zu halten.

Wir zeigen dies am Beispiel zusätzlicher Heterogenität beim Periodeneffekt. Statt der Random Walk Priori verwenden wir nun

$$\phi_j \sim N(\phi_j^*, \lambda_1) \text{ für } j = 1, \dots, J$$

und eine Random Walk Priori mit Präzisionsparameter  $\lambda_2$  für  $\phi^* = (\phi_1^*, \dots, \phi_J^*)'$ . Sei im Folgenden  $\tilde{\phi} = (\tilde{\phi}_1, \dots, \tilde{\phi}_{(2J)})'$  mit

$$\begin{aligned} \bar{\phi}_{(2j)-1} &= \phi_j \\ \bar{\phi}_{(2j)} &= \phi_j^* \end{aligned}$$

für  $j = 1, \dots, J$ . Wir können dann die gemeinsame Priori von  $\phi$  und  $\phi^*$  als Priori von  $\tilde{\phi}$  angeben:

$$p(\tilde{\phi}) \propto \exp \left[ -\frac{1}{2} \tilde{\phi}' (\lambda_1 \mathbf{I} \otimes \mathbf{A} + \lambda_2 \mathbf{R} \otimes \mathbf{B}) \tilde{\phi} \right]$$

mit

$$\mathbf{A} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \text{ und } \mathbf{B} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

Die Präzisionsmatrix

$$\mathbf{P} = \lambda_1 \mathbf{A} \otimes \mathbf{I} + \lambda_2 \mathbf{B} \otimes \mathbf{R}$$

hat dabei Bandweite 2 beim Random Walk erster Ordnung und 4 beim Random Walk zweiter Ordnung.

Damit können die Parameter  $\phi$  und  $\phi^*$  auch gemeinsam in einem Block gezogen werden. Wie oben approximieren wir die Full Conditional mittels Taylor-Entwicklung durch eine Normalverteilung. Diese hat mit den Bezeichnungen von oben Präzision

$$\tilde{\mathbf{V}}_\phi = \mathbf{P} + \text{diag} \left( \sum_i n_{i1} g_i''(\phi_1^0), 0, \dots, \sum_i n_{iJ} g_i''(\phi_J^0), 0 \right)$$

und Mittelwertsvektor  $\tilde{\mathbf{V}}_\phi^{-1}(\mathbf{m}_\phi \otimes (1, 0))$ .

Im Modell mit Heterogenitätsparameter  $z_{ij}$  ist die Full Conditional wieder eine Normalverteilung mit Präzisionsmatrix

$$\tilde{\mathbf{U}}_\phi = \mathbf{P} + \delta J \left[ \mathbf{I} \otimes \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \right]$$

und Mittelwertsvektor

$$\tilde{\mathbf{w}}_\phi = \tilde{\mathbf{U}}_\phi^{-1} \delta \left( \sum_i \xi_{i1}^\phi, 0, \dots, \sum_i \xi_{iJ}^\phi, 0 \right),$$

wir können also Gibbs-Schritte zur Aufdatierung verwenden. Die Full Conditional der anderen Parameter ändern sich durch die Erweiterung des Periodeneffekts nicht.

Diese Modifizierung der Priori ist analog für den Kohorteneffekt und für den Alterseffekt möglich, bei letzteren in der Praxis aber kaum nötig. Auch ein Modell mit zusätzlicher Heterogenität für mehrere Zeitskalen ist denkbar. Dabei entfällt allerdings der Vorteil der Identifizierbarkeit bei Verwendung von RW 1-Prioris, da die oben genannte stochastische Restriktion nicht mehr vorliegt.

### 5.3 Anwendung auf COPD-Daten

Die vorgestellten Modelle werden im Folgenden zur Analyse eines Datensatzes über *chronic obstructive pulmonary disease* (COPD) in England und Wales verwendet. COPD bezeichnet eine Gruppe von chronischen, entzündlich-destruktiv verlaufenden Lungenerkrankungen und gehört zu den häufigsten Erkrankungen weltweit. Verursacht wird die Krankheit

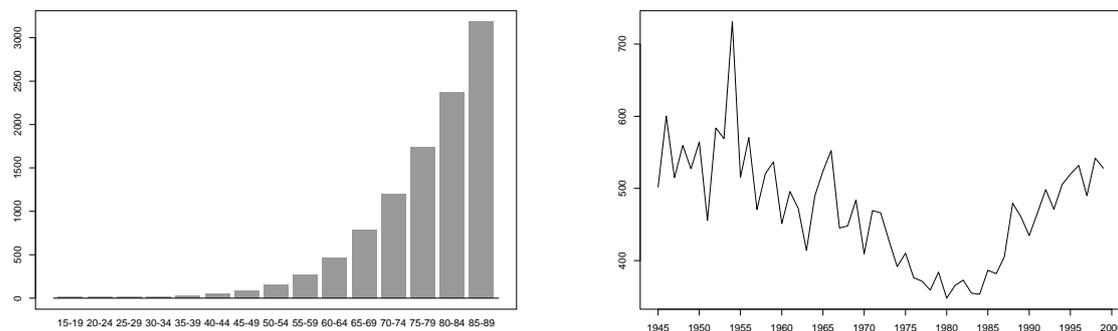


Abbildung 5.3: Altersverteilung und zeitlicher Verlauf der Mortalitätsraten von COPD unter Frauen in England und Wales. Abgebildet sind Fälle pro 1 000 000 Einwohner.

durch das Zusammenwirken von genetischer Veranlagung und der Belastung durch Luftverschmutzung und Nikotingebrauch (Pauwels, Buist, Calverley, Jenkins & Hurd 2001).

Für die Analyse lagen die Mortalitätsdaten für COPD und Asthma für 15 Fünf-Jahres-Altersgruppen (beginnend mit der Gruppe der 15–19-Jährigen) der Jahre 1940 bis 1999 für Frauen in England und Wales vor. Die Daten wurden bezüglich der Änderungen des ICD-Codes adjustiert (Hansell, Schmid & Knorr-Held 2003). Die deskriptive Analyse der Daten zeigt einen Anstieg der Mortalitätsraten seit etwa 1985, während der Trend vorher fallend war, siehe Abbildung 5.3. Es zeigt sich eine deutliche Abhängigkeit der Mortalität vom Alter. Im gesamten Beobachtungszeitraum wurden 502 977 Sterbefälle gezählt, was einer durchschnittlichen Rate von 470 Sterbefällen pro 1 000 000 Einwohner entspricht. Wie bereits erwähnt sind bei diesem Datensatz Alter und Periode nicht gleich skaliert sind, der Kohortenindex berechnet sich wie in (5.2).

### 5.3.1 Schätzungen

Bei der Analyse mit den Alters-Perioden-Kohorten-Modelle arbeiten wir im Folgenden jeweils mit  $\text{Ga}(1, 10^{-5})$ -Prioris für die Präzisionen der Random Walks und  $\text{Ga}(1, 0.05)$ -Prioris für die Präzisionen der zufälligen Effekte. Zur Schätzung der Parameter lief der Algorith-

mus jeweils 300 000 Iterationen bei einem Burn-In von 50 000 und einer Ausdünnung von 250. Damit verbleiben 1000 Iterationen als Sample. Bei den Modellen mit zusätzlicher Heterogenität ist noch ein Tuningschritt notwendig, um eine der beiden Proposalverteilungen zu wählen (siehe Seite 84); der Tuningschritt wird auf Grund praktischer Erfahrung nach einer Laufzeit von 800 Iterationen durchgeführt. Bei den Modellen ohne  $z_{ij}$  ist kein Tuning notwendig, hier müssen jedoch die Startwerte speziell der Präzisionen sorgfältig gewählt werden (dies erledigt BAMP automatisch in einem *trial and error*-Verfahren).

Der Alterseffekt im einfachen APC-Modell mit RW 1 ist, wie zu erwarten, stark ausgeprägt (Abbildung 5.4). Der Periodeneffekt hat einen Knick Anfang der Achtziger Jahre, was zum Teil die steigende Mortalität seit dieser Zeit erklären kann. Hierbei sei nochmals darauf hingewiesen, daß wegen des Nicht-Identifizierbarkeitsproblems nicht entschieden werden kann, ob der Periodentrend seit dieser Zeit ansteigt oder nur nicht mehr fällt. Auffallend ist die starke Schwankung des Effekts, was auf zusätzliche Heterogenität schließen läßt. Der Kohorteneffekt hat drei Knickpunkte etwa bei den Geburtsjahrgängen um 1915, um 1935 und um 1950. Da es sich bei den Geburtsjahrgängen zwischen den beiden Weltkriegen um die altersmäßig relevanten Gruppen für die Sterbefälle der Jahre 1980–1999 handelt, hat offensichtlich auch der Kohorteneffekt Einfluß auf die seit 1980 steigenden Mortalitätsraten.

Das Modell mit zusätzlichem Heterogenitätsparameter (APC+z) zeigt praktisch die gleichen Effekte (Abbildung 5.5). Perioden- und Kohorteneffekt sind etwas glatter als im einfachen APC-Modell, da die Variation in den Daten zum Teil durch den zusätzlichen Heterogenitätsparameter erklärt wird. Durch die Verwendung des Gibbs-Samplers ist die Laufzeit erheblich kürzer (ca. 18 Minuten auf einem Pentium 4, 2.66 GHz, gegenüber 80 Minuten beim Modell ohne Heterogenität). Die Heterogenitätsparameter  $z_{ij}$  sind in 10.8% der Altersgruppen-Perioden-Kombinationen signifikant von Null verschieden (95%-Kredibilitätsniveau), Abbildung 5.6 zeigt die Kombinationen mit signifikanten Abweichungen. Die Anpassung ist bei diesem Modell natürlich besser (vgl. Tabelle 5.1), da viel mehr Parameter geschätzt werden. Auch die effektive Zahl der Parameter nach Spiegelhalter *et al.* (2002) ist deutlich höher, das Modell ist dennoch DIC-besser (für Devianz, effektive

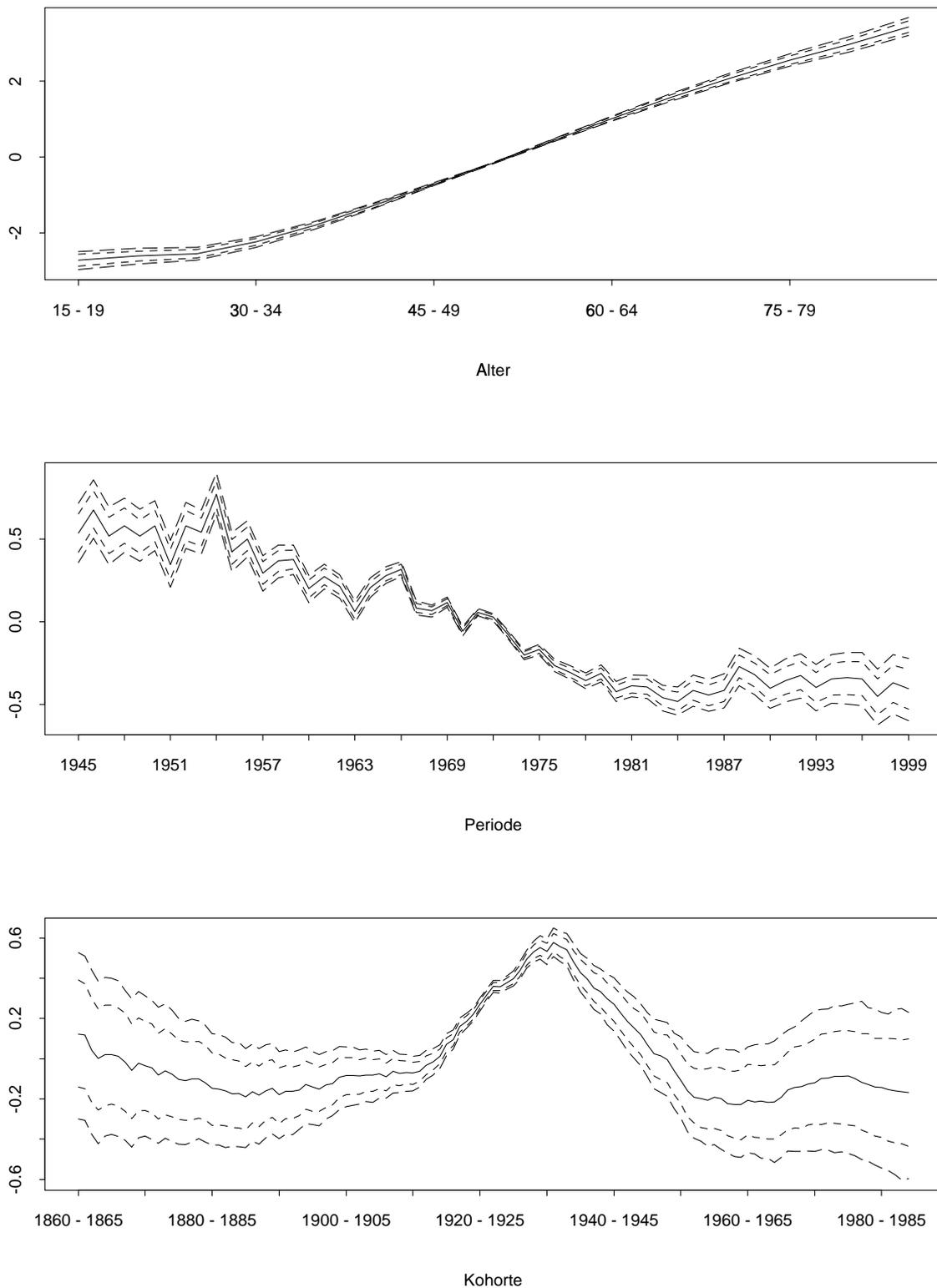


Abbildung 5.4: Haupteffekte im einfachen APC-Modell. Abgebildet sind jeweils Median, 80%- und 95%-Kredibilitätsintervalle.

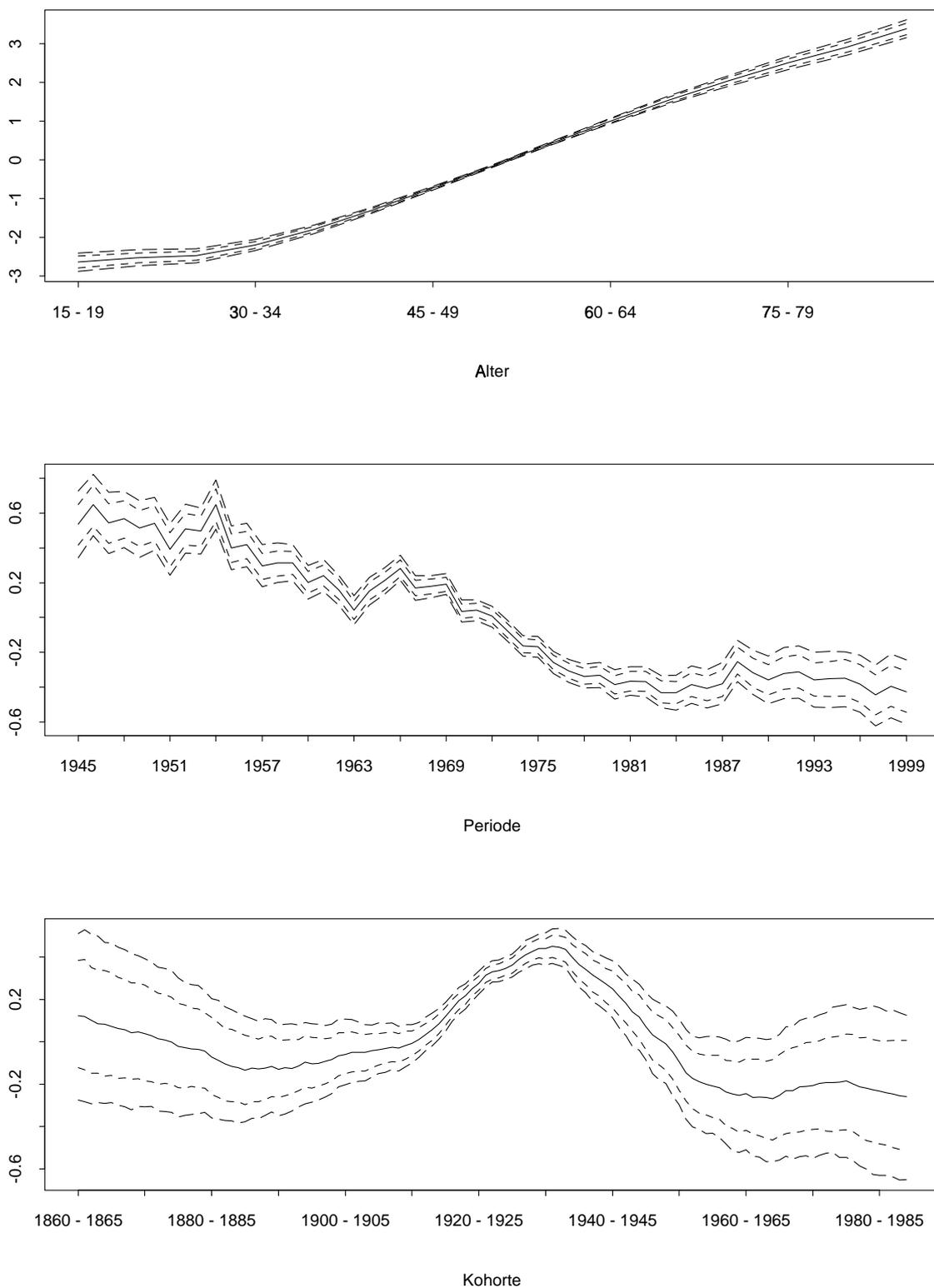


Abbildung 5.5: Haupteffekte im APC-Modell mit Heterogenitätsparameter. Abgebildet sind jeweils Median, 80%- und 95%-Kredibilitätsintervalle.

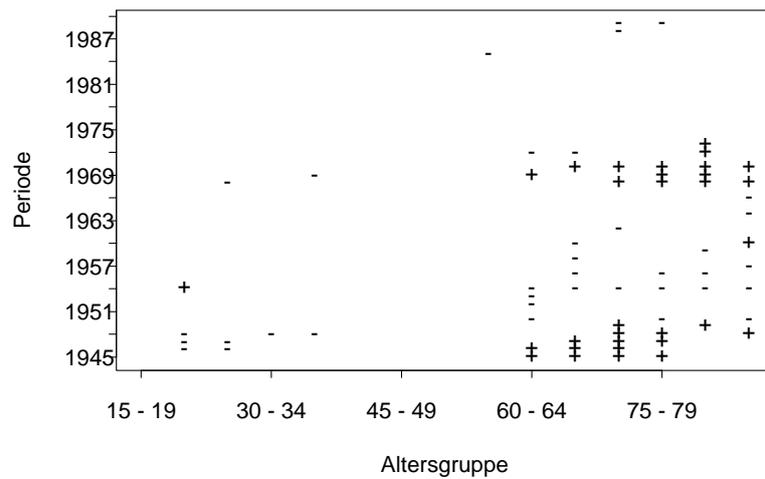


Abbildung 5.6: Abweichung des Heterogenitätsparameters von Null nach oben (+) oder unten (-) bei einem Kreditätsniveau von 95%.

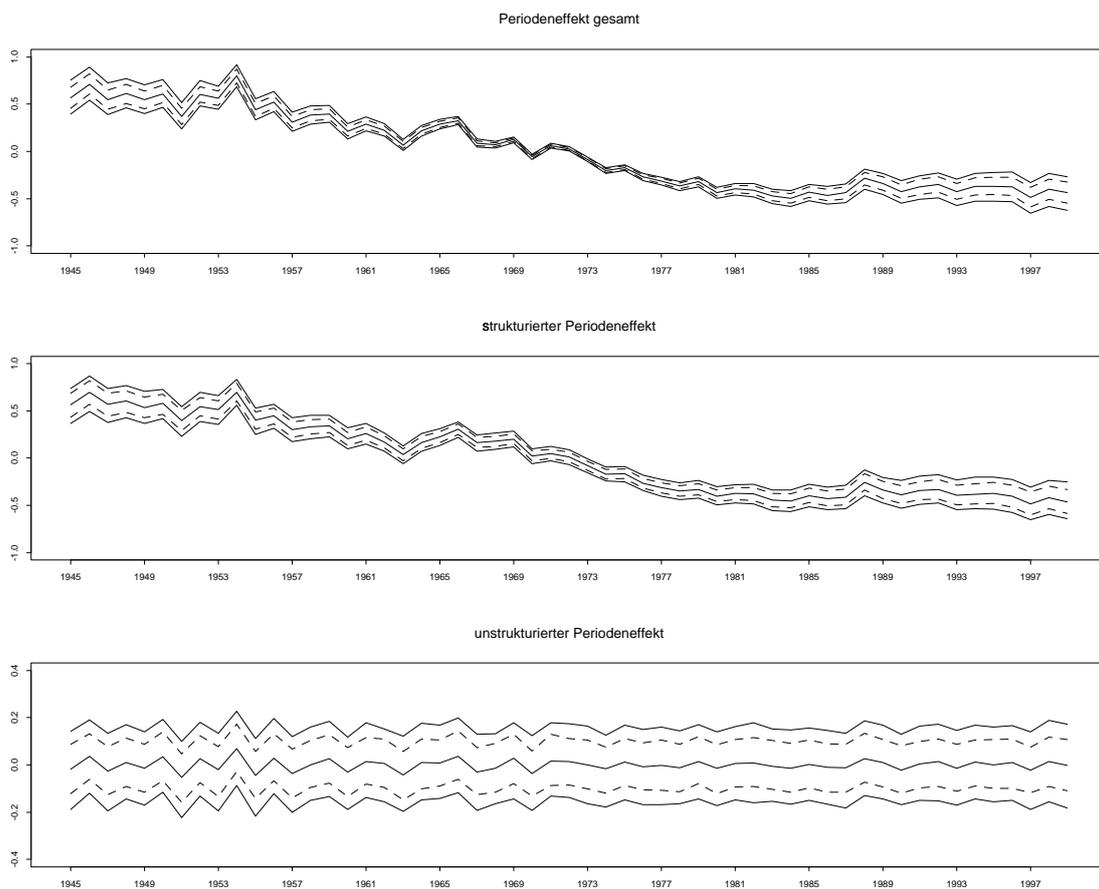


Abbildung 5.7: Aufteilung des Periodeneffekts in strukturierten und unstrukturierten Teil im Modell AP\*C. Abgebildet sind jeweils Median, 80%- und 95%-Kreditätsintervalle.

Modell	Prioris	Devianz	$p_D$	DIC
APC	RW 1	2650.5	132.6	2783.0
APC	RW 2	2684.9	96.2	2781.1
APC+z	RW 1	1091.9	538.1	1629.9
APC+z	RW 2	1072.0	540.0	1612.1
AP*C	RW 1	2649.8	135.6	2785.5
AP*C	RW 2	2675.9	102.0	2777.9
AP*C+z	RW 1	1104.4	532.6	1637.0
AP*C+z	RW 2	1114.7	525.4	1640.1

Tabelle 5.1: Median der Posteriori-Devianz, effektive Anzahl der Parameter und DIC verschiedener Modelle.

Anzahl der Parameter und DIC siehe Abschnitt 2.4.2).

Wie man in Abbildung 5.4 sieht, hat der Periodeneffekt große Heterogenität. Es bietet sich an, ein Modell mit zusätzlichem unstrukturierten Periodeneffekt zu modellieren (AP\*C-Modell). Abbildung 5.7 zeigt die Aufteilung des gesamten Periodeneffekt  $\phi$  in den strukturierten Teil  $\phi^*$  und den unstrukturierten, zufälligen Teil  $\phi - \phi^*$  im AP\*C-Modell. Der strukturierte Teil ist dabei geringfügig glatter als im APC-Modell. Gleiches gilt für das Modell mit zusätzlicher (globaler) Heterogenität (AP\*C+z). Die Schätzung der Haupteffekte ändert sich praktisch nicht, daher ändert sich auch die Datenanpassung kaum (siehe Tabelle 5.1).

Die APC-Effekte sind für die Modelle mit RW 2-Priori aufgrund des Nichtidentifizierbarkeitsproblems nicht darstellbar. Die Effekte werden zwar im MCMC-Algorithmus mitgeschätzt, jedoch tritt bei jeder Iteration eine Transformation gemäß (5.5) auf. Ein Posteriori-Schätzer der APC-Effekte kann damit nicht angegeben werden. Die Güte der Anpassung entspricht im wesentlichen der bei den RW 1-Prioris.

### 5.3.2 Prognose

Zum weiteren Vergleich erstellen wir mit jedem der Modelle eine rückwirkende Prognose. Wir passen also jedes Modell mit den Daten von 1945-1989 an, prognostizieren die Mortalitätsfälle der Jahre 1990-1999 gemäß (5.7) und vergleichen diese mit den Beobachteten. Tabelle 5.2 zeigt den Anteil der  $15 \cdot 10 = 150$  Beobachtungen aller Altersgruppen in den zehn Prognosejahren, welche im 50%-, 80%- und 95%-Prognoseintervall liegen. Zusätzlich berechnen wir die Wurzel des mittleren quadratischen Abstand zwischen beobachteten und prognostizierten Fallzahlen

$$\chi = \sqrt{\frac{1}{150} \sum_{i=1}^{15} \sum_{j=46}^{55} (y_{ij} - \hat{y}_{ij})^2}.$$

Die Ergebnisse zeigen deutlich, daß die Modelle mit RW 1-Prioris sehr viel bessere Punktschätzer liefern als diejenigen mit RW 2-Prioris. Die wahre Anzahl der Fälle wird bei letzteren immer unterschätzt (Abbildung 5.8 als Beispiel). Die Überdeckungswahrscheinlichkeit ist beim RW 2 immer größer als das Niveau, die Prognoseintervalle also zu breit.

Bei den Modellen mit zusätzlicher Periodenheterogenität sind die Punktschätzer zwar

Modell	Prioris	50%	80%	95%	$\chi$
APC	RW 1	43.3	62.0	74.0	154.3
APC	RW 2	95.3	98.0	99.3	2169.8
APC+z	RW 1	48.7	63.3	76.0	285.8
APC+z	RW 2	93.3	97.3	99.3	2684.8
AP*C	RW 1	63.3	88.7	98.7	267.2
AP*C	RW 2	86.7	97.3	98.0	1367.6
AP*C+z	RW 1	63.3	88.7	98.7	65.4
AP*C+z	RW 2	61.3	89.3	96.0	520.8

Tabelle 5.2: Überdeckungshäufigkeit der Prognoseintervalle und mittlere Abweichung der Punktprognosen; bezogen auf alle prognostizierten Altersgruppen-Perioden-Kombinationen.

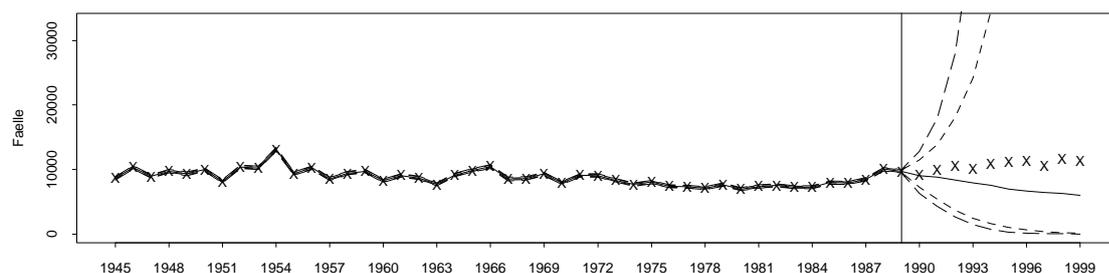


Abbildung 5.8: Schätzung und beobachtete Fälle mit 80%- und 95%-Kreditabilitätsintervallen im APC-Modell mit RW 2-Priori.

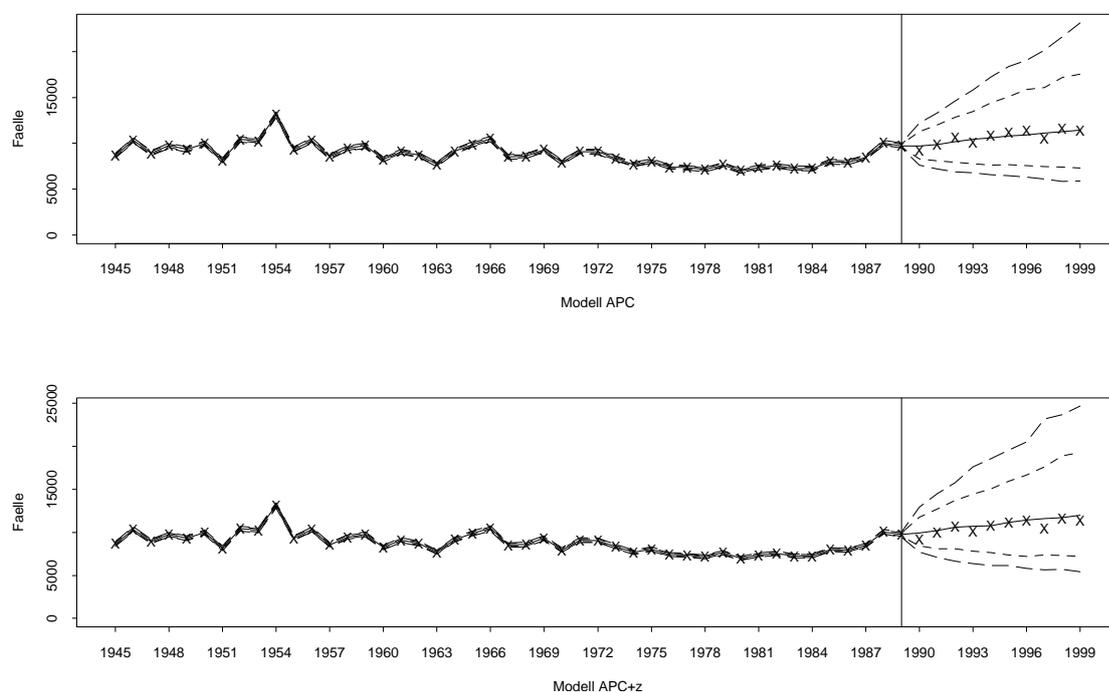


Abbildung 5.9: Schätzung und beobachtete Fälle mit 80%- und 95%-Kreditabilitätsintervallen im APC- und APC+z-Modell mit RW 1-Priori.

akzeptabel, die Prognoseintervalle sind aber auch hier deutlich zu breit. Am besten schneiden hier die einfachen Modelle mit RW 1-Prioris ab. Dabei unterscheiden sich das Modell mit und ohne Heterogenität kaum. Abbildung 5.9 zeigt die Prognosen beider Modelle. Die konstante Fortsetzung des Trends scheint hier eine sinnvolle Annahme zu sein (vgl. Abschnitt 5.2.1).

## 5.4 Zusammenfassung

Bayesianische Alters-Perioden-Kohortenmodelle sind als Werkzeug zur Analyse von Mortalitäts- und Inzidenzraten nützlich. Die Verwendung von RW 1-Prioris macht die APC-Effekte identifizierbar. Obwohl lineare Trends nicht interpretierbar sind, so kann man doch Trendänderungen auf den Zeitskalen interpretieren. APC-Modelle können damit ein erster Schritt zur genaueren Untersuchung der Inzidenzursachen sein.

Prognosen sind im Bayesianischen APC-Modell durch konsequente Fortsetzung der autoregressiven Prioris zugänglich. Die Wahl des zugrundeliegenden Modells muß dabei gut gewählt werden. Modelle mit RW 1-Prioris können einen starken Bias haben, falls die Annahme konstanter Zeittrends nicht zutrifft, Modelle mit RW 2-Prioris neigen zu größeren Kreditabilitätsintervallen (vgl. dazu auch Knorr-Held & Rainer (2001) oder Kapitel 6). Beim vorliegenden Datensatz liefert ein Modell mit Random Walk erster Ordnung die beste Prognosegüte.

Knorr-Held & Rainer (2001) versuchen durch Einbeziehung eines zeitvariierenden Kovariableneffekts ihr Bayesianisches APC-Modell mit zusätzlicher Heterogenität zu verbessern. Das Modell lautet dann

$$\eta_{ij} = \mu + \theta_i + \beta_j x_{j-L} + \psi_k + z_{ij} \quad (5.12)$$

mit einer Random Walk Priori für  $\beta$ . Die Kovariable  $\mathbf{x}$  geht dabei zeitverzögert mit einem Lag  $L$  ein (zum Beispiel wirkt sich das Rauchverhalten auf die Entstehung einer Lungenerkrankung meist erst nach etwa 20-25 Jahren aus). Die Inferenz für die Parameter ergibt sich aus dem in diesem Kapitel beschriebenen Vorgehen. Unklar ist bei diesem Modell jedoch, inwieweit die Kovariable zu standardisieren ist. Hinzu kommt, daß bei Werten der Kovariable nahe Null die Kreditabilitätsintervalle sehr groß werden (Abbildung 5.10 zeigt die Schätzung von  $\beta$  für den COPD-Datensatz mit der Anzahl des (geschätzten) konsumierten Tabaks von Frauen als Kovariable. Hierbei liegen die Werte der Kovariablen in den ersten Jahren des Beobachtungszeitraums nahe bei Null); bei Kovariablenwerten gleich Null ist die Schätzung sogar nicht möglich. Die Schätzung des Modells (5.12) ist in BAMP möglich,

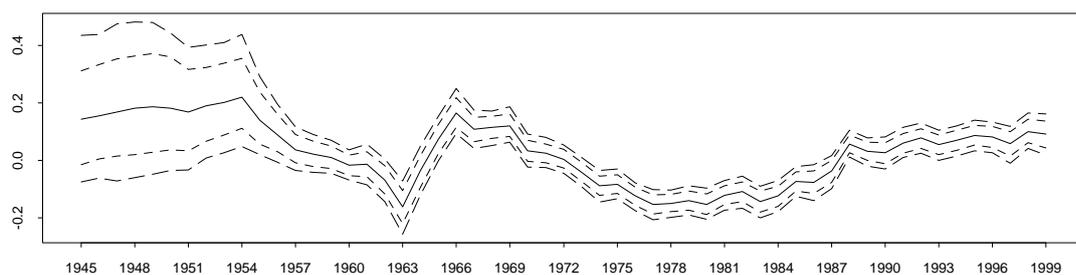


Abbildung 5.10: Schätzung des zeitvariierenden Kovariableneffekts im Modell (5.12) mit einem Datensatz zum Tabakkonsum.

das Modell wird in dieser Arbeit wegen der noch offenen Fragen aber nicht weiter behandelt. Der Ansatz ist es aber sicher wert, weiterverfolgt zu werden, weil die Modellanpassung meist besser wird (so erreicht man mit obigem Kovariablendatzensatz eine Devianz von 1007.0 und einen DIC von 1586.2 beim Modell APC+z).

Die im Rahmen dieser Arbeit entstandene Software **BAMP** ist ein umfassendes Tool zur Durchführung Bayesianischer Alters-Perioden-Kohorten-Analysen. Alle in diesem Kapitel beschriebenen Modelle sind in BAMP verfügbar. Durch die Anwendung moderner, effizienter MCMC-Methoden ist die Parameterschätzung mit akzeptablem Zeitaufwand möglich.

# Kapitel 6

## Extrapolation von Raum-Zeit-Trends bei Krebsregisterdaten

Sterbezahlen von Krebspatienten werden in Deutschland im Krebsregister erfaßt (Becker & Wahrendorf 1997). Sie liegen jährlich auf Kreis- oder Bezirksebene vor. Im Folgenden entwickeln wir ein Modell für einen Datensatz von Magenkrebssterbezahlen in westdeutschen Regierungsbezirken. Wir erweitern dazu das in Kapitel 5 vorgestellte Alters-Perioden-Kohorten-Modell um eine zusätzliche räumliche Komponente. Weiterhin untersuchen wir, inwieweit Raum-Zeit-Interaktionen in diesem Datensatz auftreten. Wir werden entscheiden, ob hierfür Modelle mit Perioden-Raum- oder Kohorten-Raum-Interaktionen besser geeignet sind, um Prognosen für zukünftige Sterbezahlen zu erstellen. Die Haupteffekte des Modells werden mit einer Gauss-Markov-Zufallsfeld-Priori versehen. Inferenz wird mit den bereits besprochenen effizienten MCMC-Blockalgorithmen betrieben (siehe Abschnitt 2.3). Dieses Kapitel basiert auf einem Artikel, der in englischer Sprache in *Biometrics* erschien (Schmid & Held 2004).

## 6.1 Einführung

Magenkrebs gehört immer noch zu den häufigsten Krebsarten in Deutschland, die Magenkrebsmortalität hat aber in den letzten drei Jahrzehnten signifikant abgenommen, wie Becker & Wahrendorf (1997) feststellen. Hauptursache für diesen zeitlichen Trend dürften die sich verändernden Ernährungsgewohnheiten sein. Die Mortalitätsraten sind besonders in Bayern sehr hoch, wenn auch mit stark fallendem Trend. Boeing, Frentzel-Beyme & Berger (1991) nennen verschiedene Risikofaktoren, welche die erhöhte Mortalität erklären können, z.B. die Konservierung von Fleischprodukten durch Räuchern oder die große Anzahl privater, vom öffentlichen Netz unabhängiger Wasserquellen. Viele dieser Faktoren trafen auf Bayern zu, wurden mit der Zeit aber immer weniger. Daher ist es von speziellem Interesse zu untersuchen, ob sich Hinweise auf Raum-Zeit-Interaktionen in den Magenkrebsmortalitätsdaten zeigen.

### 6.1.1 Datensatz zur Magenkrebsmortalität

Im Folgenden analysieren wir Daten zur Magenkrebsmortalität bei Männern in West-Deutschland zwischen 1976 und 1990. Die Daten sind, wie bei Krebsregisterdaten häufig der Fall, nach Alter, Todesjahr und Regierungsbezirken stratifiziert.

Bezeichne  $y_{ijl}$  bzw.  $n_{ijl}$  die Anzahl von Todesfällen bzw. Personen unter Risiko in Altersgruppe  $i = 1, \dots, I$ , Periode  $j = 1, \dots, J$  und Region  $l = 1, \dots, L$ . In unserer Anwendung ist das Alter kategorisiert in die Altersgruppen *bis* 20, 20 – 24, 25 – 29, ..., 70 – 74 und *über* 74, zusammen also in  $I = 13$  Altersgruppen. Die Daten liegen für  $J = 15$  Jahre 1976 bis 1990 in  $L = 30$  Regierungsbezirken in West-Deutschland (alte Bundesländer) vor.

Ist das genaue Alter einer Person zum Todeszeitpunkt bekannt, kann man den Geburtszeitpunkt der Person berechnen. Für aggregierte Daten muß aber die Geburtskohorte genauer definiert werden. Bei unserem Datensatz kommt das zusätzliche Problem hinzu, daß das Alter (in 5-Jahres-Gruppen) und die Perioden (auf Jahresbasis) auf unterschiedlichem Skalenniveau gemessen werden. Wir berechnen also den Kohortenindex als

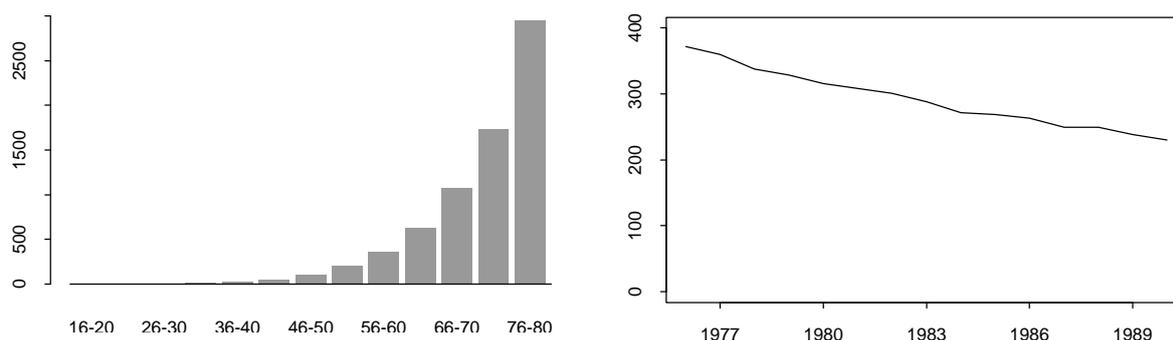


Abbildung 6.1: Altersverteilung und zeitliche Entwicklung der Mortalitätsraten von Magenkrebs in Westdeutschland. Aufgetragen sind jeweils Fälle pro 1 000 000 Einwohner.

$k = 5 \cdot (I - i) + j$  (vgl. Abschnitt 5.1). Knorr-Held & Rainer (2001) weisen darauf hin, daß mit dieser Konstruktion jede der  $K = 75$  Kohortengruppen sechs Jahre umfaßt und zwei aufeinanderfolgende Kohorten sich um fünf Jahre überlappen.

Im Beobachtungszeitraum starben 125 086 Männer in West-Deutschland an Magenkrebs. Das entspricht einer durchschnittlichen Mortalitätsrate von 291.5 Toten pro 1 000 000 Einwohnern.

Eine erste deskriptive Analyse (Abbildung 6.1) zeigt eine höhere Mortalität mit steigendem Alter, was für viele Krebsarten charakteristisch ist. Im Beobachtungszeitraum 1976-1990 fiel die Mortalität weitgehend linear.

Die deskriptive räumliche Analyse in Abbildung 6.2 zeigt eine erhöhte Mortalität im Südosten und geringere Mortalität im Westen. Die durchschnittliche Mortalität im Regierungsbezirk mit den höchsten Mortalitätsraten (Niederbayern) ist dabei 1.75-fach höher als im Regierungsbezirk mit den niedrigsten Mortalitätsraten (Düsseldorf). Vergleicht man aber die zeitliche Entwicklung der Mortalitätsraten in diesen beiden Bezirken (Abbildung 6.3), so gleichen sich diese bis 1990 fast an. Die Berücksichtigung von Raum-Zeit-Interaktionen im Modell scheint für diesen Datensatz also unbedingt notwendig.

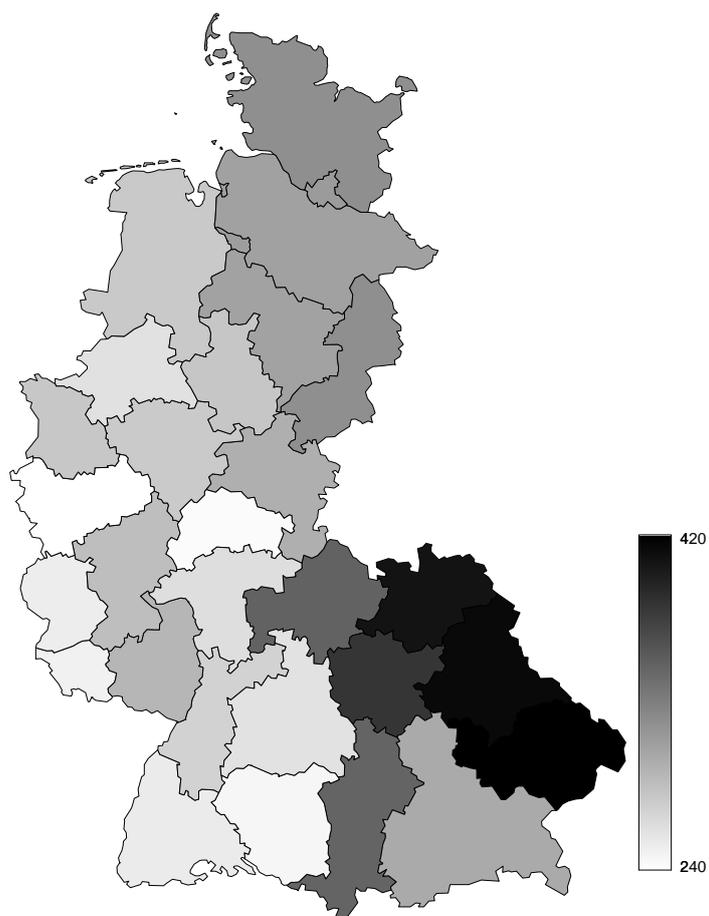


Abbildung 6.2: Verteilung der Mortalitätsraten von Magenkrebs auf die westdeutschen Regierungsbezirke. Abgebildet sind Fälle pro 1 000 000 Einwohner.

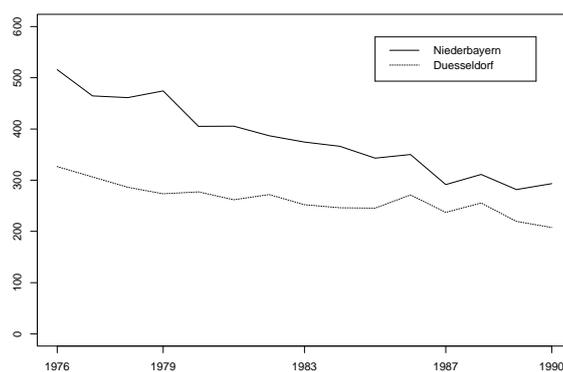


Abbildung 6.3: Vergleich der zeitlichen Entwicklung der Mortalitätsraten in den Regierungsbezirken Niederbayern und Düsseldorf. Abgebildet sind jeweils Fälle pro 1 000 000 Einwohner.

### 6.1.2 Literaturüberblick

Wir führen eine modellbasierte Analyse mit direkter Einbeziehung von Kohorteneffekten durch. Dieses Bayesianische Modell basiert auf Gauss-Markov-Zufallsfeldern. Ein Spezialfall davon sind Random Walks für zeitliche Effekte (Knorr-Held 2000; Fahrmeir & Lang 2001). Mit zeitlichen und räumlichen Effekten sowie spezifischen Raum-Zeit-Interaktionen können Trends in den Daten extrapoliert werden, um Vorhersagen von zukünftigen Mortalitätsraten in den genannten Regionen zu erhalten.

Viele der früheren Arbeiten im Bereich modellbasierter statistischer Analyse von Sterbezahlen auf dem Lexis-Diagramm (vgl. Abschnitt 5.1) wurden mit Likelihood-Inferenzkonzepten durchgeführt (Holford 1983; Clayton & Schifflers 1987a; Clayton & Schifflers 1987b). Der allgemeinste Ansatz, der in diesen Artikeln vorgestellt wird, ist das Alters-Perioden-Kohorten-Modell (siehe auch Kapitel 5). Später wurden Bayesianische Ansätze mit hierarchischen Modellen vorgeschlagen, welche den Vorteil haben, daß sie zusätzliche Parameter für unstrukturierte Variation mit einbeziehen können, um Überdispersion aufzufangen. Außerdem können sie Alters-, Perioden- und Kohortenparameter erfassen, was die Schätzeigenschaften verbessert und die Möglichkeit zur Vorhersage von zukünftigen Raten zuläßt (Berzuini *et al.* 1993; Besag *et al.* 1995; Knorr-Held & Rainer 2001; Abschnitt 5.2.1).

Im Bayesianischen Ansatz werden rein räumlich vorliegende Sterbezahlen üblicherweise mit Markov-Zufallsfeld-Prioris analysiert, siehe Besag *et al.* (1991) und Clayton & Bernardinelli (1992). Diese Modelle lösen Probleme, welche durch die Heterogenität in den Bevölkerungszahlen auftreten und liefern eine Glättung der Rohdaten, falls räumliche Strukturen vorliegen. Alternativ kann auch eine zusätzliche unstrukturierte räumliche Variation in das Modell integriert werden.

Zur modellbasierten Raum-Zeit-Analyse von Sterbezahlen wurden bisher nur ansatzweise Vorschläge gemacht. Bernardinelli, Clayton, Pascutto, Montomoli, Ghislandi & Songini (1995) schlagen ein Modell vor, welches bereits Raum-Zeit-Interaktionen mit linearen Zeit-trends beinhaltet. Sie arbeiten mit altersstandardisierten Raten, was ihnen erlaubt, Alters- und Kohorteneffekte in ihrer Analyse zu ignorieren. Allerdings wird in diesen zweistufigen

Analysen die Unsicherheit der entsprechenden Schätzungen vernachlässigt. Ein weiteres praktisches Problem ist die Wahl der Referenzbevölkerung für die Standardisierung. Assunção, Reis & Oliveira (2001) haben eine Erweiterung des Modells von Bernardinelli *et al.* (1995) mit quadratischen statt linearen Zeittrends vorgeschlagen.

Knorr-Held (2000) stellt vier verschiedene Typen von Interaktionen von Raum-Zeit-Trends vor, welche nicht nur die nichtparametrische Schätzung der räumlichen Struktur, sondern auch der Zeittrends erlauben. Das Modell beinhaltet jedoch keinerlei Alters- oder Kohorteneffekte. Einen anderen Weg schlagen Knorr-Held & Besag (1998) ein, indem sie (zeitveränderliche) Alterseffekte modellieren, welche prinzipiell die Schätzung beliebiger Alters-Perioden-Interaktionen erlauben, eingeschlossen Kohorteneffekte. Das Modell enthält aber keinerlei Raum-Zeit-Interaktionen.

Wir verwenden die von Knorr-Held (2000) vorgeschlagenen Interaktionstypen, um in unseren Daten zusätzliche Raum-Zeit-Interaktionen zu den Alters-, Perioden-, Kohorten- und räumlichen Effekten zu modellieren. Desweiteren werden wir entscheiden, ob für diese Daten Modelle mit Raum-Perioden- oder Raum-Kohorten-Interaktionen besser Datenanpassung und Prognosen liefern. Zum Vergleich der verschiedenen Modelle verwenden wir das von Spiegelhalter *et al.* (2002) vorgeschlagene *deviance information criterion*, um Modellanpassung und Komplexität zu messen.

Kürzlich wurde der Ansatz von Knorr-Held (2000) von Lagazio, Dreassi & Biggeri (2001) zur Analyse von Lungenkrebsraten in der Toskana verwendet. Sie betonen die Wichtigkeit, Interaktionen mit Kohortenparametern zu modellieren. Damit erhielten sie eine deutlich bessere Datenanpassung als bei Interaktionen von Periode und Raum. Hier wurde jedoch wiederum kein volles Modell verwendet, die Daten wurden vielmehr im Voraus altersstandardisiert.

Eine volle modellbasierte Analyse wurde auch von Lagazio, Biggeri & Dreassi (2003) präsentiert, ebenfalls zur Analyse obiger Lungenkrebsraten in der Toskana. Sie verwenden jedoch einen *single-site* MCMC-Algorithmus, was hier nachteilig sein kann. Erstens verhindern Blockalgorithmen, daß der Markov-Kern in den Enden der Posterioriverteilung

hängen bleibt, was sogar für sehr lange Läufe zu verzerrten Schätzungen führen kann, vgl. Knorr-Held (2003). Zweitens können in *single-site*-Algorithmen Zentrierungsrestriktionen nur durch ad hoc-Zentrierung berücksichtigt werden (d.h. jeder Parameter wird aus der Full Conditional ohne Berücksichtigung der Restriktion gezogen und dann erst werden die Parameter so zentriert, daß sie Mittelwert Null haben).

Im Gegensatz dazu verwenden wir die von Rue (2001) vorgeschlagenen Blockupdate-Algorithmen für Markov-Zufallfelder (vgl. Abschnitt 2.3.1). Dies verbessert das Mixing des Markov-Kerns deutlich. Zusätzlich haben diese Algorithmen den Vorteil, daß Identifizierbarkeitsbedingungen von latenten Parametern explizit in der Priori berücksichtigt werden können. Wie wir zeigen werden, kann das Modell unter Ausnutzung des von Rue (2001) vorgeschlagenen Algorithmus für restringierte Markov-Zufallfelder dazu verwendet werden, zukünftige Magenkrebssterberaten zu prognostizieren.

## 6.2 Modell

Alters-Perioden-Kohorten-Modelle sind sehr nützlich, um zeitliche Trends z.B. bei Krebsinzidenz im Hinblick auf drei Zeitvariablen zu verstehen: Das individuelle Alter, das Geburtsdatum (Kohorte) und der Todeszeitpunkt (Periode). Siehe dazu Kapitel 5.

### 6.2.1 Das Binomialmodell

Wir schlagen ein Modell für binomiale Beobachtungen  $y_{ijl}$  bei gegebenem  $n_{ijl}$  mit unbekannter Krankheitswahrscheinlichkeit  $\pi_{ijl}$  vor. Wir zerlegen die logarithmierten Chancen

$$\eta_{ijl} = \log \left( \frac{\pi_{ijl}}{1 - \pi_{ijl}} \right)$$

dieser Wahrscheinlichkeiten additiv in

- (a) Haupteffekte für Alter, Periode, Kohorte und Raum,
- (b) Interaktionen zwischen Periode und Raum oder Kohorte und Raum und

(c) Parameter, welche zusätzliche unstrukturierte Heterogenität in jeder Zelle  $(i, j, l)$  beschreiben.

Genauer gehen wir im vollständigen Modell von folgender Formulierung aus

$$\eta_{ijl} = \mu + \theta_i + \varphi_j + \psi_k + \xi_l + \left\{ \begin{array}{c} \delta_{jl} \\ \delta_{kl} \end{array} \right\} + z_{ijl} \quad (6.1)$$

mit einem Intercept  $\mu$ , dem Alterseffekt  $\theta_i$ , dem Periodeneffekt  $\varphi_j$ , dem Kohorteneffekt  $\psi_k$  und dem räumlichen Effekt  $\xi_l$ . Die Parameter für die Raum-Zeit-Interaktion werden mit  $\delta_{jl}$  oder  $\delta_{kl}$  bezeichnet (entweder Perioden oder Kohorten mit Raum, die Klammern in (6.1) bedeuten dabei, daß nur eine der beiden Optionen in das Modell eingeht) und  $z_{ijl}$  sind die zusätzlichen Parameter für unstrukturierte Heterogenität. Diese Parameter fangen nach Adjustierung auf die Haupteffekte und Interaktionen zusätzliche Überdispersion ab.

Um die Identifizierbarkeit der Parameter zu gewährleisten, nehmen wir eine Zentrierungsrestriktion für alle Haupteffekte mit in das Modell auf:

$$\sum_i \theta_i = \sum_j \phi_j = \sum_k \psi_k = \sum_l \xi_l = 0.$$

Allerdings sind Alters-, Perioden- und Kohorteneffekt trotzdem nicht separat identifizierbar, da jede Transformation der Art

$$\theta_i \rightarrow \theta_i + c \cdot i, \quad \phi_j \rightarrow \phi_j - c \cdot j, \quad \psi_k \rightarrow \psi_k + c \cdot k, \quad \mu \rightarrow \mu - c \cdot I \quad (6.2)$$

mit beliebigem  $c \in \mathbb{R}$  den Prädiktor  $\eta_{ijl}$  unverändert läßt (vgl. Abschnitt 5.1.2).

### 6.2.2 Prioriannahmen

Für den Intercept  $\mu$  verwenden wir eine flache Priori, d.h.

$$p(\mu) \propto \text{const.}$$

Alters-, Perioden- und Kohorteneffekt werden mit Random Walk-Prioris modelliert, wie in vielen APC-Modellen üblich (Berzuini *et al.* 1993; Besag *et al.* 1995; Knorr-Held &

Rainer 2001; Bray 2002). Random Walk Prioris werden in Abschnitt 2.2.2 allgemein sowie in Abschnitt 5.2.1 für die APC-Effekte ausführlich besprochen. Im Folgenden seien die Präzisionsmatrizen der Prioris der Alters-, Perioden- bzw. Kohorteneffekte  $\lambda\mathbf{K}_\theta$ ,  $\tau\mathbf{K}_\phi$  bzw.  $\nu\mathbf{K}_\psi$

Den räumlichen Effekt modellieren wir mit einer Markov-Zufallsfeld-Priori

$$p(\boldsymbol{\xi}|\omega) \propto \omega^{rg(\mathbf{K}_\xi)} \exp\left(-\frac{\omega}{2}\boldsymbol{\xi}'\mathbf{K}_\xi\boldsymbol{\xi}\right)$$

wobei  $\mathbf{K}_\xi$  durch die Nachbarschaftsstruktur der Bezirke vorgegeben ist, siehe Abschnitt 2.2.1.

Die Prioris für die Interaktion  $\boldsymbol{\delta}$  werden nach einem erstmals von Clayton (1996) vorgestellten Prinzip formuliert. Die Idee dabei ist, das Kroneckerprodukt der Präzisionsmatrizen der Haupteffekte als Präzisionsmatrix für den Interaktionseffekt zu verwenden. Wir wenden hier nur Interaktionsprioris an, welche zeitliche Abhängigkeit implizieren. In der Terminologie von Knorr-Held (2000) werden diese als *Typ II*- und *Typ IV*-Interaktionen bezeichnet. Wir unterscheiden weiterhin zwischen den Interaktionen von Periode und Raum sowie von Kohorte und Raum.

*Typ II*-Interaktionen können als unabhängige regionenspezifische Random Walks angesehen werden. Sei  $\boldsymbol{\delta}_l = (\delta_{1l}, \dots, \delta_{Jl})'$  der Vektor für Raum-Perioden-Interaktionen oder  $\boldsymbol{\delta}_l = (\delta_{1l}, \dots, \delta_{Kl})'$  der Vektor für Raum-Kohorten-Interaktionen. Dann kann die Priori für  $\boldsymbol{\delta} = (\boldsymbol{\delta}_1, \dots, \boldsymbol{\delta}_L)$  wie folgt formuliert werden:

$$p(\boldsymbol{\delta}|\rho) \propto \exp\left(-\frac{\rho}{2}\sum_{l=1}^L \boldsymbol{\delta}'_l \left\{ \begin{array}{c} \mathbf{K}_\phi \\ \mathbf{K}_\psi \end{array} \right\} \boldsymbol{\delta}_l\right).$$

Wie oben zeigen die Klammern an, daß nur eine Option in die Formulierung eingeht. Die Matrizen  $\mathbf{K}_\phi$  und  $\mathbf{K}_\psi$  sind dieselben wie oben und können für Random Walks erster oder zweiter Ordnung spezifiziert werden.

Die *Typ II*-Interaktion impliziert, daß die Parameter  $\boldsymbol{\delta}_l$ ,  $l = 1, \dots, L$  zwischen den Bezirken unabhängig sind. Bei den *Typ IV*-Interaktionen wird dagegen davon ausgegangen, daß zeitliche Trends in benachbarten Bezirken ähnlich sind.

In Knorr-Held (2000) wird das *Typ IV*-Interaktionsmodell für  $\boldsymbol{\delta}$  nur für den Random Walk erster Ordnung angegeben. Die Formulierung kann aber einfach für den Random Walk zweiter Ordnung angepaßt werden, indem man die ersten Differenzen durch die zweiten Differenzen ersetzt. Damit wird auch die Präzisionsmatrix  $\mathbf{K}_\phi$  entsprechend angepaßt.

Für Raum-Perioden-Interaktionen mit  $\boldsymbol{\delta} = (\delta_{11}, \dots, \delta_{1L}, \delta_{21}, \dots, \delta_{2L}, \dots, \delta_{J1}, \dots, \delta_{JL})'$  lautet die Priori dann

$$\begin{aligned} p(\boldsymbol{\delta}|\rho) &\propto \rho^{(L-1)(T-2)/2} \\ &\cdot \exp\left(-\frac{\rho}{2} \sum_{l \sim m} \sum_{j=3}^J ((\delta_{jl} - 2\delta_{j-1,l} + \delta_{j-2,l}) - (\delta_{jm} - 2\delta_{j-1,m} + \delta_{j-2,m}))^2\right) \\ &\propto \rho^{(L-1)(T-2)/2} \exp\left(-\frac{\rho}{2} (\boldsymbol{\delta}'(\mathbf{K}_\phi \otimes \mathbf{K}_\xi)\boldsymbol{\delta})\right). \end{aligned} \quad (6.3)$$

Diese Priori ist also ein Gauss-Markov-Zufallsfeld, bei dem nicht nur die räumlichen Nachbarn, sondern auch die ersten und zweiten zeitlichen Nachbarn sowie auch die zeitlichen Nachbarn der räumlichen Nachbarn in die bedingte Verteilung von  $\delta_{jl}$  eingehen.

Die Dichte (6.3) ist invariant gegenüber der Addition von beliebigen Konstanten zu jedem Zeitpunkt  $j$  und jedem Bezirk  $m$ . Daher müssen wir  $J + L - 1$  zusätzliche Zentrierungsrestriktionen mit einbauen, um  $\boldsymbol{\delta}$  identifizierbar zu machen. Zum Beispiel verwenden wir

$$\sum_{l=1}^L \delta_{jl} = 0 \quad \text{für } 1 \leq j \leq J \quad \text{und} \quad \sum_{j=1}^J \delta_{jl} = 0 \quad \text{für } 1 \leq l \leq L - 1.$$

Wie leicht zu sehen ist, impliziert dies  $\sum_{j=1}^J \delta_{jL} = 0$ , d.h. alle Zeilen- und Spaltensummen sind gleich Null, der Interaktionsterm ist also bezüglich jeder Periode und bezüglich jedes Bezirks um Null zentriert. Das selbe Modell kann für die Raum-Kohorten-Interaktionen verwendet werden.

Für die unstrukturierten Parameter  $z_{ijl}$  verwenden wir eine *white noise* Priori

$$z_{ijl} \sim N(0, \epsilon^{-1}).$$

Für die Hyperparameter  $\lambda, \kappa, \nu, \omega, \rho$  und  $\epsilon$  gehen wir von unabhängigen Gammaprioris  $\text{Ga}(a, b)$  mit fest gewählten  $a = 1$  und  $b = 0.05$  aus.

### 6.2.3 MCMC Simulation

Analog zu Besag *et al.* (1995) ziehen wir nicht  $z_{ijl}$ , sondern reparametrisieren das Modell und betrachten den linearen Prädiktor  $\eta_{ijl}$  statt  $z_{ijl}$  als den zu schätzenden Parameter. Dies hat den Vorteil, daß die Full Conditional aller oben genannten Effekte Normalverteilungen sind und wir damit Gibbs-Schritte verwenden können. Der lineare Prädiktor selbst kann in unabhängigen Metropolis-Hastings-Schritten gezogen werden (vgl. Abschnitt 5.2.3).

Die Full Conditionals der Haupteffekte haben alle dieselbe Form, zum Beispiel lauten sie für den Alterseffekt  $\boldsymbol{\theta}$

$$\begin{aligned} p(\boldsymbol{\theta}|\cdot) &\propto \exp\left(-\frac{\lambda}{2}(\boldsymbol{\theta}'\mathbf{K}_\theta\boldsymbol{\theta})\right) \cdot \exp\left(-\frac{\epsilon}{2}(\boldsymbol{\theta}'\mathbf{I}\boldsymbol{\theta}) + \epsilon(\boldsymbol{\beta}'\boldsymbol{\theta})\right) \\ &\propto \exp\left(-\frac{1}{2}(\boldsymbol{\theta}'\mathbf{A}_\theta\boldsymbol{\theta}) + (\boldsymbol{\beta}'\boldsymbol{\theta})\right), \end{aligned} \quad (6.4)$$

wobei  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_I)^T$  ein  $I \times 1$ -Vektor ist mit

$$\beta_i = \sum_{j,l} \left( \eta_{ijl} - \mu - \phi_j - \psi_k - \xi_l - \begin{bmatrix} \delta_{jl} \\ \delta_{jk} \end{bmatrix} \right)$$

und  $\mathbf{A}_\theta = \lambda\mathbf{K}_\theta + \epsilon\mathbf{I}$ . Die Full Conditional  $p(\boldsymbol{\theta}|\cdot)$  ist also die Dichte einer multivariaten Normalverteilung mit Erwartungswertvektor  $\mathbf{A}_\theta^{-1}\boldsymbol{\beta}$  und Präzisionsmatrix  $\mathbf{A}_\theta$ . Wie man leicht sieht, ist  $\mathbf{A}_\theta$  eine Bandmatrix mit der gleichen Bandweite wie die Priori-Präzisionsmatrix  $\mathbf{K}_\theta$ . Analog sind die Full Conditionals der Perioden-, Kohorten- und Raumeffekte ebenfalls multivariat normal, wobei die Bandweite der Präzisionsmatrix gleich der Bandweite der Präzisionsmatrix der korrespondierenden Priori ist.

Für den räumlichen Effekt verwenden wir einen in Rue (2001) beschriebenen Trick und ordnen die Indizes der Bezirke so an, daß  $\mathbf{K}_\xi$  und folglich auch die Präzisionsmatrix der Full Conditional eine Bandmatrix mit minimaler Bandweite  $m$  ist (siehe Rue 2001 für weitere Details). In unserer Anwendung ist  $m = 10$ . Unter Verwendung der Bandstruktur der Präzisionsmatrix können wir effizient aus (6.4) ziehen, wobei die von Rue (2001) beschriebenen Algorithmen Anwendung finden. Die Idee dabei ist, die Präzisionsmatrix per Cholesky-Zerlegung zu zerlegen, was für Bandmatrizen sehr schnell möglich ist. Wie bereits

erwähnt, müssen wir Zentrierungsrestriktionen berücksichtigen. Das birgt jedoch keine weiteren Probleme, auch in diesem Fall können wir die auf effizienter Cholesky-Zerlegung der Präzisionsmatrix basierenden Algorithmen verwenden, siehe dazu Rue (2001).

Die Bandweite der Präzisionsmatrizen der Full Conditionals ist im *Typ II*-Interaktionsmodell, abhängig von der verwendeten Random Walk Priori, entweder 1 oder 2. Bei *Typ IV*-Interaktionen ist die Bandweite in der Regel größer. Auch hier verwenden wir die umgeordneten Indizes der Bezirke, die Bandweite von  $\mathbf{K}_\phi \otimes \mathbf{K}_\xi$  ist  $L + m$  beim Random Walk erster Ordnung und  $2 \cdot L + m$  beim Random Walk zweiter Ordnung. Im Allgemeinen kann die Bandweite durch erneutes Umordnen der Indizes nicht weiter verringert werden.

Die Full Conditionals der Hyperparameter sind jeweils Dichten von Gammaverteilungen und die Hyperparameter können in Gibbs-Schritten gezogen werden (vgl. Abschnitt 5.2.2). Schließlich ist die Full Conditional des linearen Prädiktors

$$p(\eta_{ijl}|\cdot) \propto \frac{\exp(\eta_{ijl}y_{ijl})}{(1 + \exp(\eta_{ijl}))^{n_{ijl}}} \exp \left( -\frac{\epsilon}{2} \left( \eta_{ijl} - \mu - \theta_i - \phi_j - \psi_k - \xi_l - \begin{bmatrix} \delta_{jl} \\ \delta_{kl} \end{bmatrix} \right)^2 \right),$$

keine Standardverteilung, vgl. (5.11). Analog zu Abschnitt 5.2.2 verwenden wir ein Log-Gamma-Proposal im Metropolis-Hastings-Schritt. Mit diesem Ansatz erhalten wir Akzeptanzraten von 90%-99%.

#### 6.2.4 Vorhersage von Mortalitätsraten

Ein interessanter Aspekt unseres Ansatzes ist, daß es durch die Verwendung von Random Walk-Prioris möglich ist, zukünftige Sterblichkeitsraten unter Berücksichtigung von Raum-Zeit-Interaktionen zu prognostizieren, indem wir wie in Abschnitt 5.2.1 Perioden- und Kohorteneffekte zur Prädiktion zukünftiger Inzidenzraten fortsetzen. Die prognostizierte logarithmierte Chance der Sterblichkeit zur Periode  $J + 1$  lautet dann

$$\eta_{i,J+1,l} = \mu + \theta_i + \varphi_{J+1} + \psi_k + \xi_l + \left\{ \begin{array}{c} \delta_{J+1,l} \\ \delta_{kl} \end{array} \right\} + z_{i,J+1,l}. \quad (6.5)$$

Der Überdispersionsparameter  $z_{i,J+1,l}$  muß dabei mit prognostiziert werden. Er ergibt sich aus

$$z_{i,J+1,l} \sim N(0, \delta^{-1}).$$

In Modellen mit *Typ II*-Interaktionen kann  $\delta_{J+1,l}$  analog gezogen werden mit

$$\delta_{J+1,l} \sim N(\delta_{Jl}, \omega^{-1})$$

in jeder Periode  $l = 1, \dots, L$ . Ebenso erhält man  $\delta_{K+1,l}$  im Modell mit Kohorten-Regionen-Interaktionen analog. *Typ IV*-Interaktionen können nur gezogen werden, wenn man eine Restriktion der Art  $\sum_{l=1}^L \delta_{jl} = 0$  für jede Periode  $j = J + 1, J + 2, \dots$  berücksichtigt. Für diese anspruchsvolle Aufgabe schlagen wir einen sequentiellen Algorithmus vor. Die Priori der Typ IV-Perioden-Regionen-Interaktionen im RW 1-Modell lautet

$$p(\boldsymbol{\delta}|\rho) \propto \exp\left(-\frac{\rho}{2} \sum_{l \sim m} \sum_{j=2}^j \left((\delta_{jl} - \delta_{j-1,l}) - (\delta_{jm} - \delta_{j-1,m})\right)^2\right).$$

Ein logischer Ansatz für die Prädiktion zukünftiger Werte von  $\boldsymbol{\delta}$  ist, sie sequentiell in der Zeit fortzusetzen. Wir starten mit der Ziehung von  $\boldsymbol{\delta}_{J+1}|\boldsymbol{\delta}_J, \rho$ , wobei  $\boldsymbol{\delta}_{J+1} = (\delta_{J+1,1}, \dots, \delta_{J+1,L})$ :

$$\begin{aligned} p(\boldsymbol{\delta}_{J+1}|\boldsymbol{\delta}_J, \rho) &\propto \exp\left(-\frac{\rho}{2} \sum_{l \sim m} (\delta_{J+1,l} - \delta_{J+1,m} - \delta_{Jl} + \delta_{Jm})^2\right) \\ &\propto \exp\left(-\frac{1}{2} \boldsymbol{\delta}'_{J+1} (\rho \mathbf{K}_\xi) \boldsymbol{\delta}_{J+1} + \rho \boldsymbol{\delta}'_{J+1} \mathbf{b}\right) \end{aligned} \quad (6.6)$$

mit  $\mathbf{b} = b_1, \dots, b_L$  und  $b_l = \sum_{m:m \sim l} (\delta_{Jl} - \delta_{Jm})$ . Dies ist die Dichte eines Gauss-Markov-Zufallsfelds mit singulärer Präzisionsmatrix vom Rang  $L - 1$ . Daher können wir ohne weitere Restriktionen nicht aus dieser Dichte ziehen. Wie in Abschnitt 6.2.2 verwenden wir eine Zentrierungsrestriktion  $\mathbf{1}' \boldsymbol{\delta}_{J+1} = 0$ , unter welcher die Dichte (6.6) proportional ist zu

$$\exp\left(-\frac{1}{2} \boldsymbol{\delta}'_{J+1} (\rho \mathbf{K}_\xi + \mathbf{1}\mathbf{1}') \boldsymbol{\delta}_{J+1} + \rho \boldsymbol{\delta}'_{J+1} \mathbf{b}\right). \quad (6.7)$$

Dies ist leicht mit einem Lemma von Box & Tiao (1973, S. 419) nachzuweisen. Wir ziehen  $\boldsymbol{\delta}_{J+1}$  aus der unrestringierten Verteilung mit Dichte (6.7), welche eine multivariate

Normalverteilung mit Präzisionsmatrix  $\mathbf{Q} = (\rho\mathbf{K}_\xi + \mathbf{1}\mathbf{1}')$  und Erwartungswertvektor  $\rho\mathbf{Q}^{-1}\mathbf{b}$  ist. Die restringierte Stichprobe  $\tilde{\boldsymbol{\delta}}_{J+1}$  erhält man, indem man

$$\tilde{\boldsymbol{\delta}}_{J+1} = \boldsymbol{\delta}_{J+1} - \mathbf{Q}^{-1}\mathbf{1}(\mathbf{1}'\mathbf{Q}^{-1}\mathbf{1})^{-1}(\mathbf{1}'\boldsymbol{\delta}_{J+1})$$

berechnet, siehe Abschnitt 2.3.1. Für die Interaktionen mit Random Walk zweiter Ordnung kann die Dichte der Verteilung von  $\boldsymbol{\delta}_{J+1}$  aus (6.3) geschlossen werden. Analoge Berechnungen führen zu einer restringierten Stichprobe wie oben, jedoch mit

$$b_l = \sum_{m:m \sim l} 2(\delta_{Jl} - \delta_{Jm}) - (\delta_{J-1,l} - \delta_{J-1,m}).$$

Nach dem wir  $\boldsymbol{\delta}_{J+1}$  gezogen haben, können wir denselben Algorithmus verwenden, um  $\boldsymbol{\delta}_{J+2}$ ,  $\boldsymbol{\delta}_{J+3}$  und so weiter zu ziehen. Die Prognose der Kohorten-Regionen-Interaktionen wird analog vollzogen.

## 6.3 Anwendung auf Magenkrebsdaten

### 6.3.1 Modellwahl

Um die Güte der Anpassung des Modells an die Daten zu messen, berechnen wir den Median der Posteriori-Devianz  $\bar{D}$  und das *deviance information criterion* DIC. Das DIC ist definiert als

$$DIC = \bar{D} + p_D,$$

wobei  $p_D$  die effektive Anzahl der Parameter bezeichnet, siehe Abschnitt 2.4.2.

Tabelle 6.1 zeigt Devianz, effektive Anzahl der Parameter und DIC aller beschriebenen Modelle. Die Ergebnisse sind für RW 1- und RW 2-Prioris im wesentlichen identisch. Die Devianz ist am niedrigsten für die Modelle mit *Typ II*-Interaktionen und am höchsten für die Modelle ohne Interaktionen. Wie in Abschnitt 2.4.2 ausgeführt, führt eine hohe Anzahl von Parametern im Allgemeinen zu besserer Anpassung. Die Anpassung der Modelle mit *Typ IV*-Interaktionen ist weniger gut als bei *Typ II*-Interaktionen, da die ersten eine stärkere Abhängigkeit unter den Interaktionsparametern annehmen.

	Interaktion	Devianz	$p_D$	DIC
RW 1	keine	5623.1	39.9	5663.0
	Typ II Periode	5374.0	196.4	5570.4
	Typ II Kohorte	5235.5	1573.5	6809.0
	Typ IV Periode	5424.9	269.5	5694.4
	Typ IV Kohorte	5289.5	1483.0	6772.5
RW 2	keine	5656.8	19.4	5676.2
	Typ II Periode	5406.2	180.0	5586.2
	Typ II Kohorte	5309.7	1202.7	6512.4
	Typ IV Periode	5461.8	266.5	5728.3
	Typ IV Kohorte	5358.1	1312.4	6670.5

Tabelle 6.1: Median der Posteriori-Devianz, effektive Anzahl der Parameter und DIC aller Modelle.

Interessanterweise ist aber die effektive Zahl der Parameter in den Modellen mit *Typ IV*-Interaktionen höher als bei den Modellen mit *Typ II*-Interaktionen. Das kommt daher, daß in letzteren in manchen Regionen der Interaktionsparameter für fast alle Perioden auf Null geschätzt wird, während für erstere aufgrund des zugrundeliegenden Markovzufallsfelds mehr Parameter ungleich Null geschätzt werden.

Die Modelle mit Kohorten-Raum-Interaktionen haben ein deutlich höhere DIC-Werte als jene mit Perioden-Raum-Interaktionen. Daraus können wir schließen, daß die Modelle mit Perioden-Raum-Interaktionen für diese Daten besser geeignet sind.

### 6.3.2 Ergebnisse

Zuerst zeigen wir die geschätzten Haupteffekte des bezüglich des DIC-Kriteriums besten Modells, jenes mit *Typ II*-Perioden-Regionen-Interaktionen. Da wie erwähnt die Alters-, Perioden- und Kohorteneffekte für das RW 2-Modell nicht einzeln identifizierbar sind, zeigen wir nur die Schätzer für das RW 1-Modell.

Abbildung 6.5 zeigt jeweils Median und 90%-Kredibilitätsintervall der Alters-, Perioden- und Kohorteneffekte. Dabei sollte man sich in Erinnerung rufen, daß diese durch lineare Transformationen der Art (6.2) verändert werden können. Deswegen können nur nichtlineare Trends interpretiert werden. Während der Periodeneffekt einigermaßen linear erscheint, hat der Kohorteneffekt einen auffälligen Knick etwa bei der Geburtskohorte 1940. Nach 1940 ist der Effekt weitgehend linear. Allerdings basieren die Schätzungen der letzten Geburtskohorten auf sehr geringen Fallzahlen und sollten daher nicht überinterpretiert werden. Abbildung 6.4 zeigt den räumlichen Effekt. Die höchsten Werte sind, wie erwartet, im

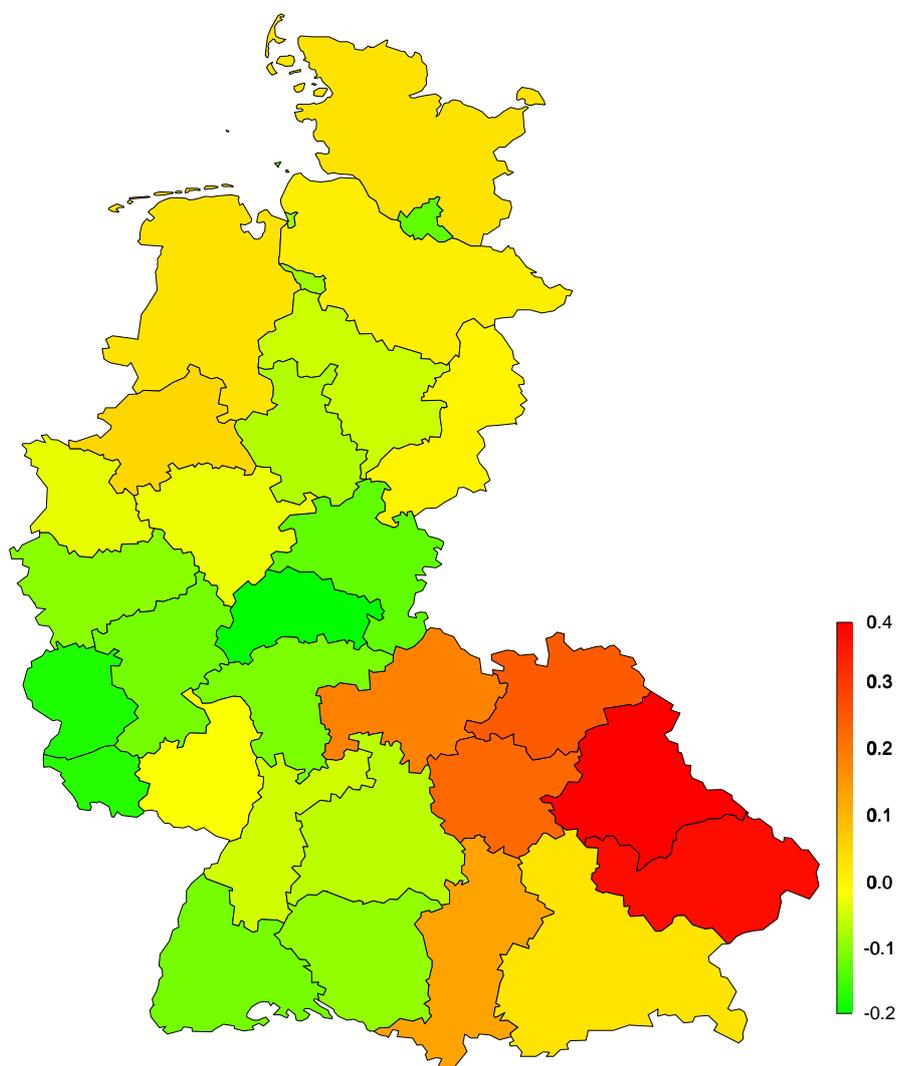


Abbildung 6.4: Posteriori-Median des räumlichen Effekts.

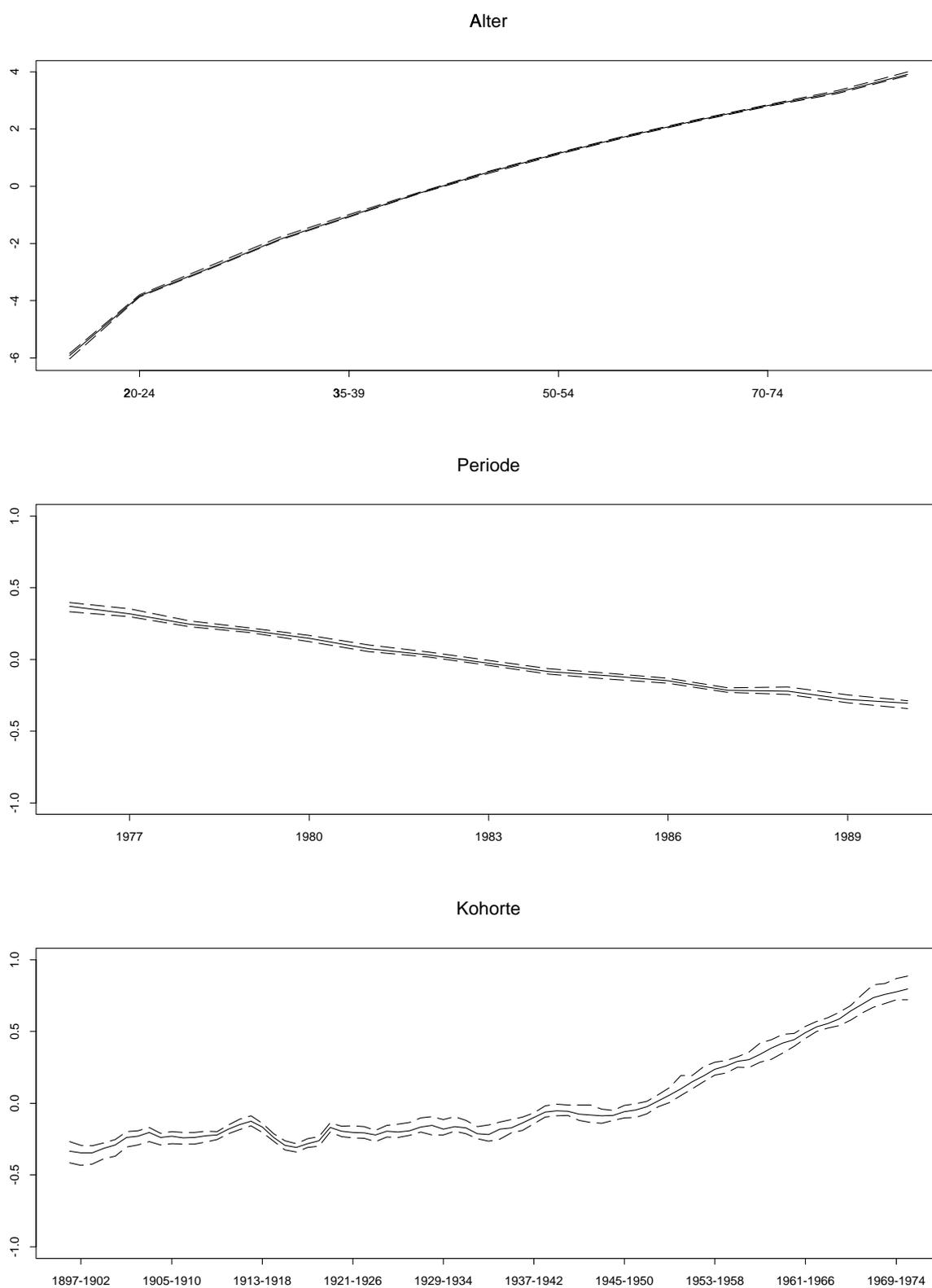


Abbildung 6.5: Posteriori-Mediane mit 90% Kredibilitätsintervallen der Alters-, Perioden- und Kohorteneffekte.

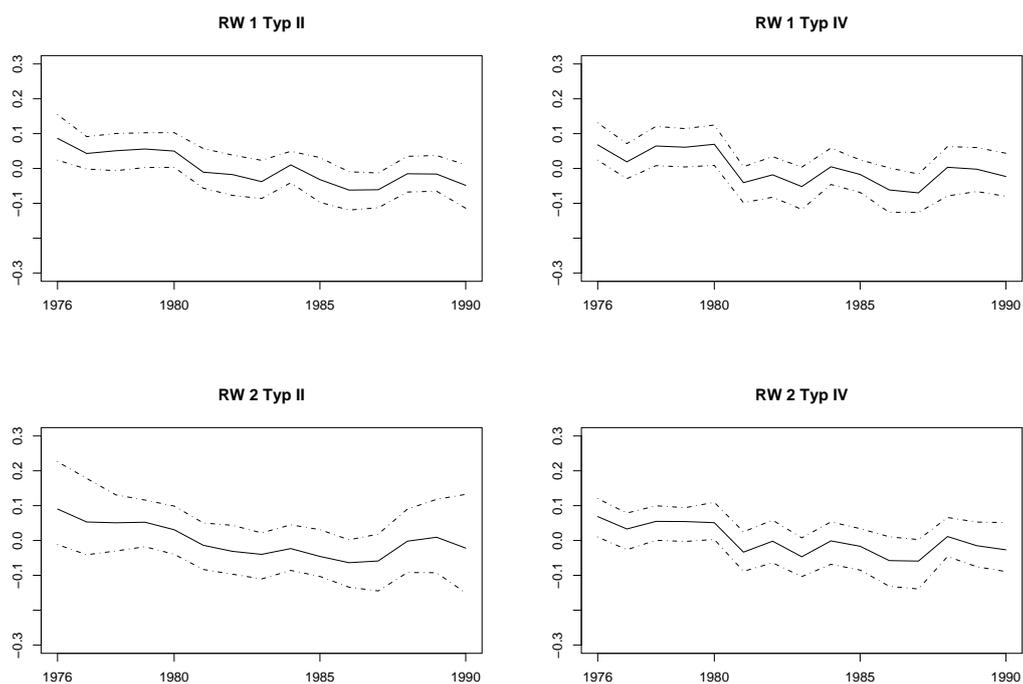


Abbildung 6.6: Posteriori-Median mit 90% Kreditabilitätsintervallen der Raum-Perioden-Interaktionen in Niederbayern.

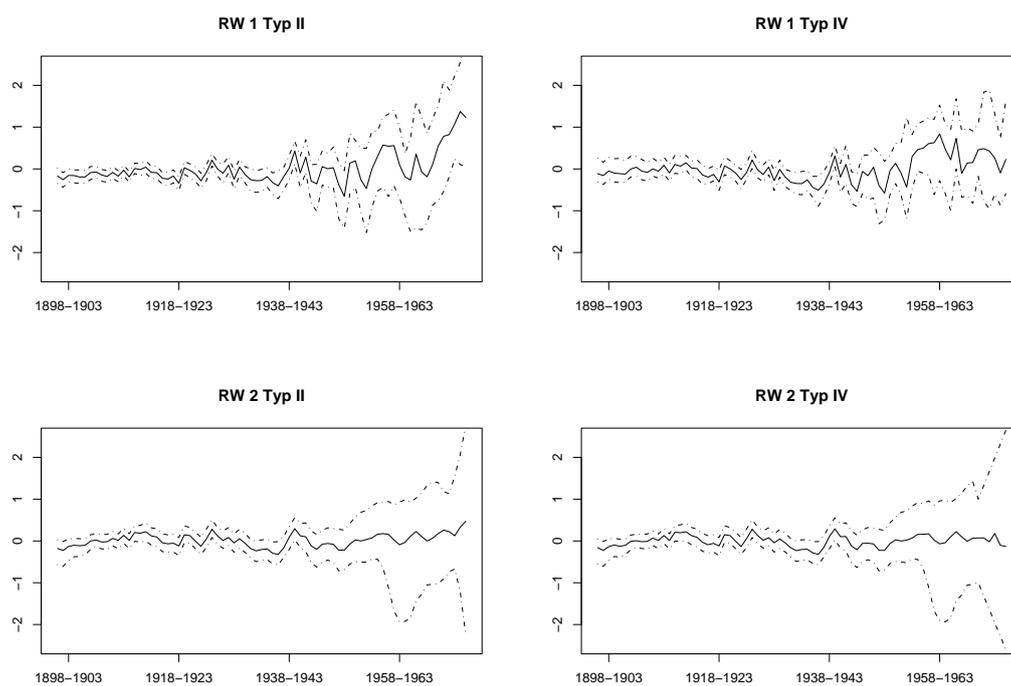


Abbildung 6.7: Posteriori-Median mit 90% Kreditabilitätsintervallen der Raum-Kohorten-Interaktionen in Niederbayern.

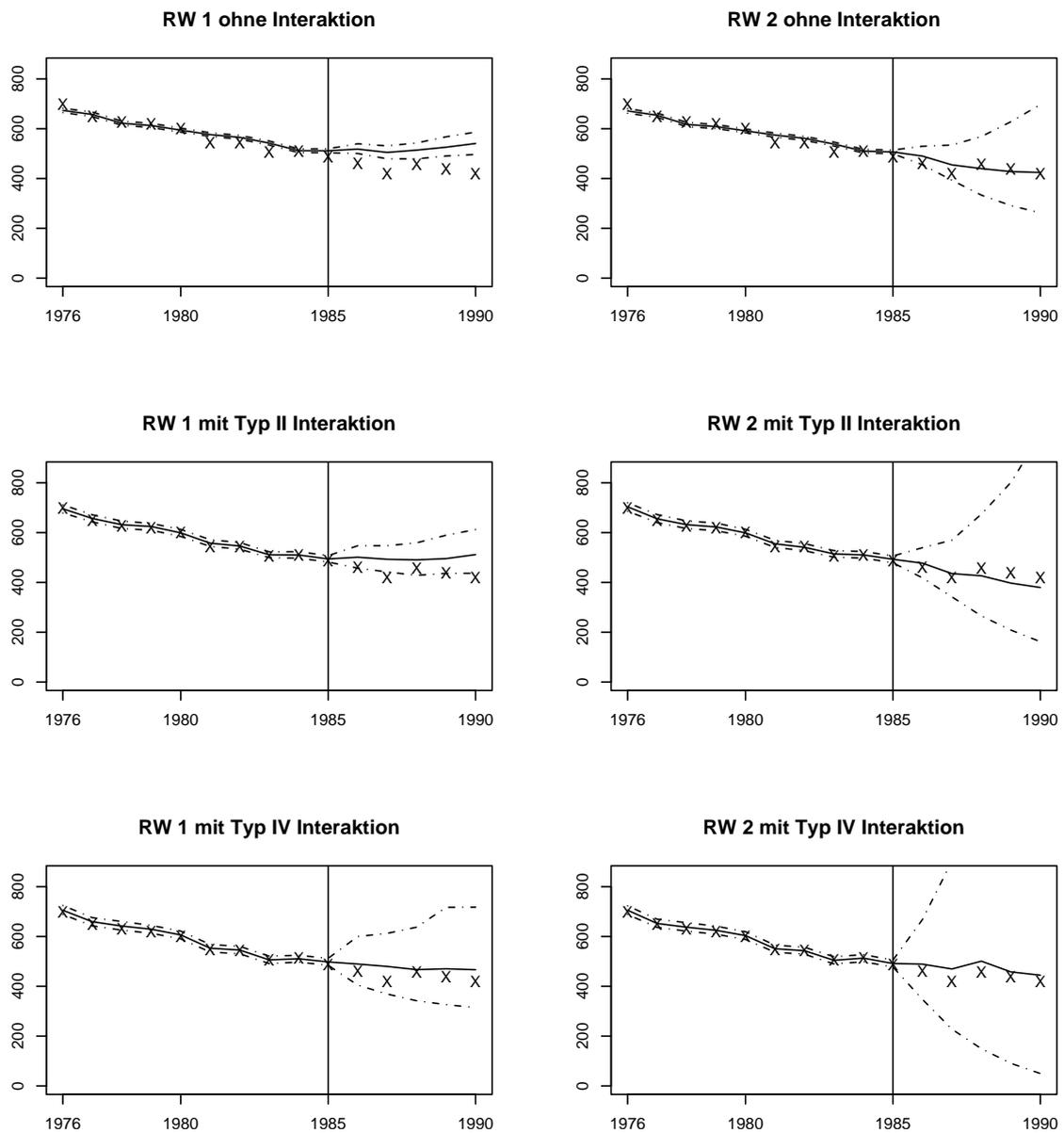


Abbildung 6.8: Beobachtete und prognostizierte Anzahl von Fällen mit punktweisen 90%-Kredibilitätsintervallen für Niederbayern, basierend auf dem RW 1-Modell (links) und dem RW 2-Modell (rechts) mit Raum-Perioden-Interaktionen. Die Kreuze kennzeichnen die Anzahl der beobachteten Fälle.

Südosten Deutschlands in den Regierungsbezirken Oberpfalz und Niederbayern zu sehen.

Abbildung 6.6 zeigt die Perioden-Raum-Interaktionen für den Bezirk Niederbayern. In allen Modellen hat diese Interaktion einen fallenden Trend. Ein ähnlicher Effekt kann in den benachbarten Bezirken beobachtet werden, was darauf hinweist, daß - nach Adjustierung auf den globalen Periodeneffekt - die Sterblichkeitsrate im Südosten Deutschlands sinkt.

Das auf den Raum-Kohorten-Interaktionen basierende Modell (Abbildung 6.7) zeigt weniger stabile Schätzungen als das mit Raum-Perioden-Interaktionen. Der globale Varianzparameter des Interaktionsterms wird sehr hoch geschätzt wird (Posteriori-Median  $\rho = 0.717$  im Modell mit Typ II-Raum-Kohorten-Interaktion und RW 1-Prioris, gegenüber  $\rho = 96.79$  im selben Modell mit Raum-Perioden-Interaktionen), da die Effekte in den letzten Geburtskohorten, welche auf sehr wenigen Fällen beruhen, hohe Variabilität zeigen. Wie zu erwarten, ist der Effekt im RW 2-Modell etwas glatter als im RW 1-Modell.

Um die prädiktive Qualität der verschiedenen Modelle genauer zu untersuchen, haben wir das Modell an die Daten der ersten zehn Jahre angepaßt und Prognosen für die letzten fünf Jahre erstellt. Da der Kohorten-Raum-Effekt für diese Daten offensichtlich unpassend ist, beschränken wir uns im Folgenden auf die Modelle mit Perioden-Raum-Interaktionen. Abbildung 6.8 zeigt den Median der prognostizierten Anzahl von Fällen mit 90%-Kredibilitätsintervallen (Linien) und die tatsächlich beobachteten Fälle (Kreuze) für Niederbayern. Dabei wurden die Anzahlen über alle Altersgruppen im Bezirk summiert.

Kredibilitätsniveau	keine Interaktion		Typ II		Typ IV	
	RW 1	RW 2	RW 1	RW 2	RW 1	RW 2
Alle Angaben in %						
50	11	41	22	50	42	62
80	20	55	39	62	56	68
95	33	62	50	66	64	71

Tabelle 6.2: Anteil der beobachteten Fälle, welche innerhalb des Prognosekredibilitätsintervalls liegen.

Die RW 1-Modelle haben einen zum Teil erheblichen Bias bei der Prognose (siehe dazu die Ausführungen in Abschnitt 5.2.1), dafür aber engere Kreditabilitätsintervalle. Die Modelle mit RW 2-Priori haben dagegen weitere Intervalle, sind dafür in den Punktprognosen besser. Die Modelle mit Interaktionen haben größere Intervalle, wobei jene der Modelle mit Typ II-Interaktionen etwas enger sind.

Zur genaueren Analyse zeigt Tabelle 6.2 den Anteil der beobachteten Fallzahlen, welche innerhalb des Kreditabilitätsintervalls zu einem bestimmten Niveau liegen. Die Tabelle basiert auf dem vorliegenden Alters-Perioden-Gitter, d.h. die Prognosen wurden in jedem der  $13 \cdot 5 \cdot 30 = 1750$  Kombinationen von Altersgruppen, Perioden und Bezirken mit den beobachteten Werten verglichen.

Zuerst fällt auf, daß die Modelle ohne Interaktionen eine sehr schlechte Überdeckung aufweisen. Die Modelle mit Interaktionen kommen nahe an das wahre Kreditabilitätsniveau heran, speziell jene mit RW 2-Prioris. Fast alle Modelle schöpfen aber das Prognoseniveau nicht aus, mit Ausnahme der *Typ IV*-Interaktionen mit RW 2-Priori, wo 62% der Daten im 50%-Kreditabilitätsintervall liegen. Dies weist darauf hin, daß dieses Modell die Daten überschätzt. Damit hat das RW 2-Modell mit *Typ II*-Interaktionen die besten Vorhersagequalitäten: Es kommt am nächsten an die wahren Kreditabilitätsniveaus heran, ohne die Daten zu überschätzen.

## 6.4 Diskussion und Ausblick

Die Formulierung eines Modells für Raum-Zeit-Interaktionen unter Berücksichtigung der Adjustierung für Alters-, Perioden- und Kohorteneffekte ist eine anspruchsvolle Aufgabe. In diesem Kapitel haben wir verschiedene Modelle vorgeschlagen, die Interaktionseffekte zwischen Periode und Raum oder Kohorte und Raum beinhalten. In unserer Anwendung konnte das Modell mit Raum-Perioden-Interaktionen die Schätzungen gegenüber dem einfachen Modell verbessern, selbst unter Berücksichtigung der größeren Komplexität des Modells. Eine weitere Bestätigung für die Überlegenheit dieses Modells liefert die Unter-

suchung der empirischen Überdeckungswahrscheinlichkeit der Prognosen.

Die Modelle haben eine große nominelle Anzahl an a priori abhängigen Parametern. Daher erweist es sich als vorteilhaft, Block-Algorithmen zu verwenden, um ein zu schlechtes Mixing der Samples zu verhindern. Außerdem erlaubt die Anwendung der Blockalgorithmen die Berücksichtigung von Identifizierbarkeitsrestriktionen, in diesem Fall also von Zentrierungsrestriktionen. Die komplexeste Formulierung erlaubt die Abhängigkeit des Interaktionsparameters über Raum und Zeit, unter Berücksichtigung der Zentrierungsrestriktionen in jeder Region und zu jedem Zeitpunkt. Eine weitere Verbesserung des Mixings könnte möglicherweise durch gemeinsames Aufdatieren von Parametern und Hyperparametern erreicht werden (vgl. Knorr-Held & Rue 2002 und Abschnitt 2.3.3).

# Kapitel 7

## Ein stochastisches Modell für Surveillance-Daten

Die in dieser Arbeit vorgestellten Raum-Zeit-Interaktionen lassen sich auch bei Daten von infektiösen Krankheiten verwenden. Die Raum-Zeit-Interaktion kann dort als Ausbreitung der Infektion angesehen werden.

Im Folgenden wollen wir aber einen anderen Ansatz für infektiöse Krankheiten verfolgen. Hofmann *et al.* (2004) modellieren die Anzahl von Infektionen über die Zeit als Verzweigungsprozeß mit Immigration. Dabei diskriminieren sie die beobachteten Fälle in epidemische und endemische Infektionen. Ziel des Modells ist es, den Ausbruch von Epidemien frühzeitig zu erkennen und ein Online-Warnsystem hierfür zu entwickeln. Wir erweitern das dort vorgestellte Modell für Raum-Zeit-Daten um die räumlich-zeitliche Modellierung der endemischen Komponente und vergleichen dies mit dem ursprünglichen Ansatz.

## 7.1 Ein Verzweigungsprozeß als Modell für infektiöse Krankheiten

### 7.1.1 Daten

Seit Inkrafttreten des Infektionsschutzgesetzes werden die Daten aller meldepflichtigen Infektionskrankheiten in Deutschland am Robert-Koch-Institut (RKI) Berlin gesammelt. Das RKI stellte uns die Zahlen der *Campylobacter*-Infektionen im Zeitraum Januar 2001 bis Juni 2003 ( $n = 129$  Wochen) in den deutschen Bundesländern ( $I = 16$ ) zur Verfügung.

Im gesamten Zeitraum wurden 97 132 Fälle registriert, das entspricht 753 Fällen pro Woche. Abbildung 7.1 zeigt die Anzahl der Infektionen über die Zeit. Auffallend ist der starke saisonale Verlauf, die Fallzahlen steigen jeweils in den Sommermonaten stark an, um im Winter wieder abzufallen. Jeweils zu den Weihnachtsfeiertagen ist ein Zeitraum von zwei Wochen zu beobachten, in dem eher weniger Fälle vorliegen. Hierbei handelt es sich offensichtlich um ein Meldeproblem. Abbildung 7.2 zeigt die standardisierten Inzidenzraten (SIR) in den Bundesländern. Die SIR berechnen sich als Quotient der Anzahl der beobachteten und der erwarteten Fälle. Es zeigen sich erhöhte Raten in den Stadtstaaten Hamburg und Berlin sowie in den nordwestlichen Bundesländern.

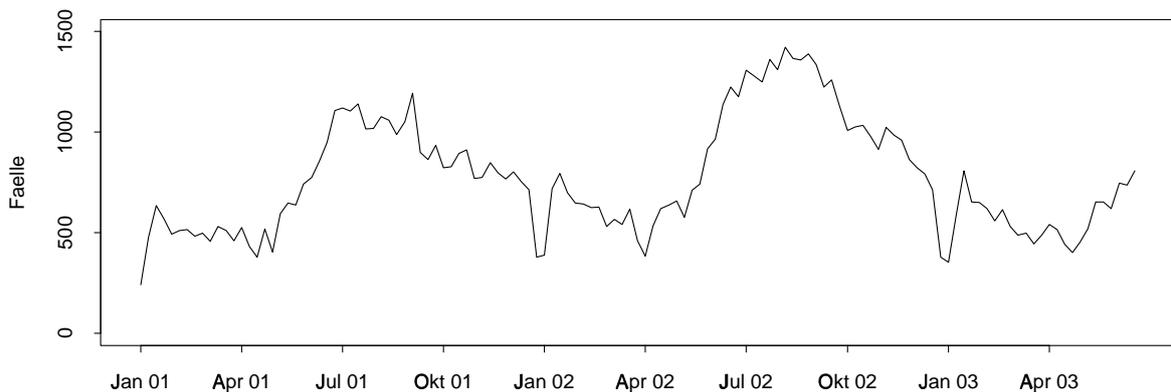


Abbildung 7.1: Zeitlicher Verlauf der Zahl der Infektionen an *Campylobacter* in Deutschland.

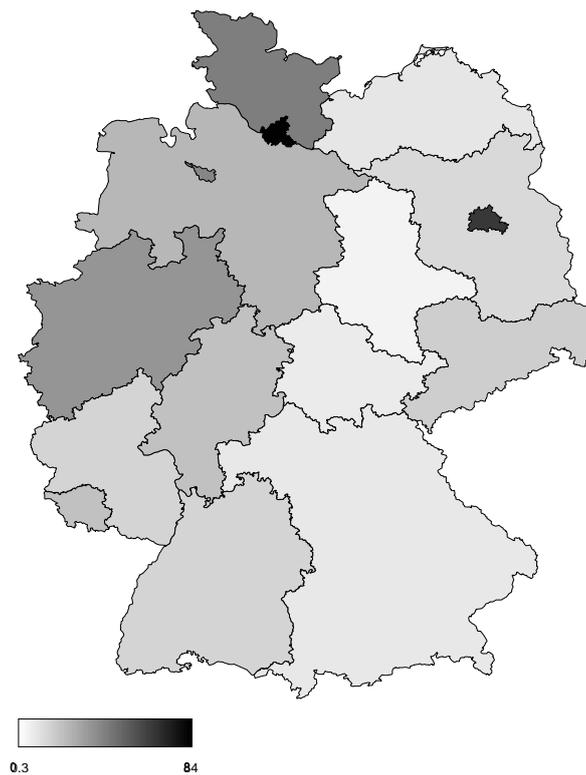


Abbildung 7.2: Standardisierte Inzidenzraten der *Campylobacter*-Infektionen in den deutschen Bundesländern auf der Log-Skala.

### 7.1.2 Modell

In unserem Modell gehen wir davon aus, daß die gesamte Anzahl von Inzidenzfällen  $Z_{it}$  in der Region  $i$  zum Zeitpunkt  $t$  sich aus einer Anzahl endemischer Fälle  $X_{it}$  und einer Anzahl epidemischer Fälle  $Y_{it}$  zusammensetzt. Die endemischen Fälle seien unabhängig Poissonverteilt mit einem von der Region und der Zeit abhängigen Parameter

$$X_{it} \sim \text{Po}(\nu_{it}).$$

Die epidemischen Fälle  $Y_{it}$  entstehen durch Ansteckung. Die Anzahl hängt also von der Anzahl der Infizierten zum vorherigen Zeitpunkt ab. Wir setzen hierfür an

$$Y_{it} \sim \text{Po}(\lambda(Y_{it-1} + X_{it-1})) \text{ für } t \geq 2.$$

Für den Startzeitpunkt des Beobachtungszeitraums  $t = 1$  muß  $Y_{i1}$  speziell modelliert werden:

$$Y_{i1} \sim \text{Po} \left( \frac{\omega_i^{(s)} \lambda \bar{\nu}_i}{1 - \lambda} \right),$$

wobei  $\bar{\nu}_i = \frac{1}{n} \sum_t \nu_{it}$  bezeichnet. Auf  $\omega_i^{(s)}$  wird eine Gamma-Priori

$$\omega_i^{(s)} \sim \text{Ga}(\bar{\nu}_i(1 + \lambda), \bar{\nu}_i(1 + \lambda))$$

gelegt. Hofmann *et al.* (2004) zeigen, daß für festes  $\nu_{it} = \nu$  die epidemische Komponente  $\mathbf{Y}$  im stationären Erwartungswert startet. Für variierendes  $\nu$  hat  $\mathbf{Y}$  keinen stationären Erwartungswert, wir verwenden daher eine Näherung zum stationären Erwartungswert bei festem  $\nu$ .

Das Modell erhebt nicht den Anspruch, den Verlauf der Ausbreitung einer Infektion genau zu modellieren. Vielmehr baut es auf dem stochastischen Modell des Verzweigungsprozesses mit Immigration auf, um eine möglichst gute Prognose zukünftiger Fälle zu erhalten und damit Ausbrüche frühzeitig zu erkennen.

### 7.1.3 Überdispersion

Typischerweise entsprechen die Zeitintervalle der Beobachtung nicht der Ansteckungszeit der Infektion. Zusätzlich ist die Zeit zwischen zwei Ansteckungen zufällig verteilt. Dadurch kommt zusätzliche Varianz in die Daten, die durch das Modell nicht erklärt wird. Es ist daher angebracht, zusätzliche Überdispersionsparameter  $\omega_{it}$  in das Modell aufzunehmen.

$$\begin{aligned} Y_{it} | \omega_{it} &\sim \begin{cases} \text{Po}(\omega_{it} \frac{\omega_i^{(s)} \lambda \nu_{it}}{1 - \lambda}) & \text{für } t = 1, \\ \text{Po}(\omega_{it} \lambda (X_{i,t-1} + Y_{i,t-1})) & \text{für } t \geq 2, \end{cases} \\ \omega_{it} &\sim \text{Ga}(\psi, \psi), \\ \psi &\sim \text{Ga}(1, 0.1). \end{aligned}$$

Man kann zeigen, daß die Randverteilung von  $Y_{it}$  eine Negativ-Binomial-Verteilung ist (DeGroot 1970, S. 119),

$$Y_{it} \sim \begin{cases} \text{NegBin} \left( \omega_i^{(s)} \frac{\lambda \bar{\nu}_i}{1-\lambda}, \psi \right) & \text{für } t = 1, \\ \text{NegBin} \left( \omega_i^{(s)} \lambda (X_{i,t-1} + Y_{i,t-1}), \psi \right) & \text{für } t = 2, \dots, n, \end{cases}$$

wobei  $\text{NegBin}(\mu, \psi)$  die Negativ-Binomial-Verteilung mit Erwartungswert  $\mu$  und Dispersionsparameter  $\psi$  ist.

## 7.2 Modell mit festem $\nu$

### 7.2.1 Priori-Annahmen

Für den Parameter der endemischen Fälle  $X_{it}$  gehen Hofmann *et al.* (2004) davon aus, daß dieser sich über die Zeit nicht ändert. In jeder Region ist der Parameter gleich dem Anteil  $\xi_i$  der Bevölkerungszahl im Land  $i$  an der Gesamtbevölkerung mal einem globalen, festen Parameter  $\nu$ . Dieser ist aufgrund der angenommenen Unabhängigkeit der Regionen der Poissonparameter der Verteilung von  $\sum_i X_{it}$ .

$$\begin{aligned} \nu_{it} &= \nu \xi_i & \text{für alle } i = 1, \dots, I, t = 1, \dots, n. \\ \nu &\sim \text{Ga}(\alpha_\nu, \beta_\nu) \end{aligned}$$

Auf den Parameter  $\lambda$  der Poissonverteilung der epidemischen Fälle  $Y_{it}$  legen Hofmann *et al.* (2004) eine Beta-Priori

$$\lambda \sim \text{Beta}(\alpha_\lambda, \beta_\lambda).$$

### 7.2.2 Inferenz

Die gemeinsame Verteilung des Modells mit Überdispersion lautet

$$p(\mathbf{Z}, \boldsymbol{\theta}) = \prod_{i=1}^I \left( \prod_{t=1}^n (p(Z_{it}|X_{it}, Y_{it})p(X_{it}, \nu_{it})) \prod_{t=2}^n (p(Y_{it}|\lambda, Y_{i,t-1}, X_{i,t-1}, \omega_{it})) \right) \\ \cdot \prod_{i=1}^I (p(Y_{i1}|\lambda, \nu_{i1}, \omega_{i1}, \omega_i^{(s)})) \cdot p(\boldsymbol{\nu})p(\lambda)p(\boldsymbol{\omega}^{(s)})p(\boldsymbol{\omega}|\boldsymbol{\psi})p(\boldsymbol{\psi})$$

Für das Modell ohne Überdispersion erhält man die Verteilung, indem man  $p(\omega_{it} = 1) = 1$  für alle  $i, t$  setzt.

Als Full Conditional des Parameters  $\nu$  ergibt sich damit

$$p(\nu|\cdot) \propto p(\nu) \prod (p(X_{it}|\nu, \xi_i)) \prod (p(Y_{i1}|\lambda, \nu, \xi_i, \omega_{i1}, \omega_i^{(s)})) \\ \propto \nu^{\alpha\nu-1} \exp(-\beta\nu) \prod_{i,t} ((\nu\xi_i)^{X_{it}} \exp(-\nu\xi_i)) \prod_i \left( (\nu\xi_i)^{Y_{i1}} \exp\left(-\frac{\lambda\nu\xi_i\omega_{i1}\omega_i^{(s)}}{1-\lambda}\right) \right) \\ \propto \nu^{\alpha\nu+\sum_{i,t}(X_{it})+\sum_i(Y_{i1})-1} \exp\left(-\left(\beta\nu+n+\frac{\lambda}{1-\lambda}\sum_{i,t}\xi_i\omega_{i1}\omega_i^{(s)}\right)\nu\right).$$

Hierbei handelt es sich um die Dichte einer Gammaverteilung mit den Parametern

$$\alpha\nu + \sum_{i,t} X_{it} + \sum_i Y_{i1} \quad \text{und} \quad \beta\nu + n + \sum_i \frac{\lambda\xi_i\omega_{i1}\omega_i^{(s)}}{1-\lambda}.$$

Der Parameter  $\nu$  kann also in einem Gibbs-Schritt gezogen werden.

Die Verteilung des Parameter  $\lambda$  gegeben aller anderen Parameter ist keine Standardverteilung. Zur Vereinfachung der Berechnung der Akzeptanzwahrscheinlichkeit ziehen wir statt  $\lambda$  den Logit des Parameters  $\lambda' = \text{logit}(\lambda)$ . Die Full Conditional von  $\lambda'$  lautet

$$p(\lambda'|\cdot) \propto p(\lambda') \prod_{i,t} p(Y_{it}|\lambda, X_{it}, Y_{it}, \omega_{it}) \prod_i p(Y_{i1}|\lambda, \nu_{i1}, \omega_{i1}, \omega_i^{(s)}) \\ = p(\lambda) \frac{d}{d\lambda'} \text{logit}^{-1}(\lambda') \prod_{i,t} p(Y_{it}|\lambda, X_{it}, Y_{it}, \omega_{it}) \prod_i p(Y_{i1}|\lambda, \nu_{i1}, \omega_{i1}, \omega_i^{(s)}).$$

Um aus  $\lambda'$  zu ziehen, verwenden wir ein Random Walk-Proposal

$$\lambda' \sim N(\lambda'_0, \sigma_{\lambda'}^2),$$

wobei  $\lambda'_0$  den momentanen Zustand von  $\lambda'$  bezeichnet. Die Varianz  $\sigma_{\lambda'}^2$  wird so gewählt, daß Akzeptanzwahrscheinlichkeiten von 35% – 40% erreicht werden.

Die (latenten) Parameter  $X_{it}$  und  $Y_{it}$  werden wegen der Beziehung  $Z_{it} = X_{it} + Y_{it}$  gemeinsam gezogen. Es gilt dabei

$$p(X_{it}, Y_{it} | \cdot) = p(Y_{it} | X_{it}, \cdot) p(X_{it} | \cdot). \quad (7.1)$$

Für die endemische Komponente gilt

$$X_{it} | \cdot \sim \begin{cases} \text{B} \left( Z_{i1}, \frac{\nu_{i1}}{\nu_{i1} + \frac{\lambda \nu_{i1} \omega_{i1}^{(s)}}{1-\lambda}} \right) & \text{für } t = 1, \\ \text{B} \left( Z_{it}, \frac{\nu_{it}}{\nu_{it} + \lambda \omega_{it} (Y_{i,t-1} + X_{i,t-1})} \right) & \text{für } t \geq 2. \end{cases}$$

Die epidemische Komponente ergibt sich aus

$$Y_{it} = Z_{it} - X_{it}.$$

Die Full Conditional der Überdispersionsparameter  $\omega_{it}$  und der Parameter  $\omega_i^{(s)}$  sind die Dichten von Gammaverteilungen, siehe Hofmann *et al.* (2004)

$$\begin{aligned} \omega_{it} | \cdot &\sim \begin{cases} \text{Ga} \left( \psi + Y_{i1}, \psi + \lambda \left( \frac{\bar{\nu}_i \omega_i^{(s)}}{1-\lambda} \right) \right) & \text{für } t = 1, \\ \text{Ga} \left( \psi + Y_{it}, \psi + \lambda (X_{i,t-1} + Y_{i,t-1}) \right) & \text{für } t \geq 2, \end{cases} \\ \omega_i^{(s)} | \cdot &\sim \text{Ga} \left( \nu(1+\lambda) + Y_{i1}, \bar{\nu}_i \left( (1+\lambda) + \frac{\omega_{i1} \lambda}{1-\lambda} \right) \right). \end{aligned}$$

### 7.3 Modell mit variierendem $\nu$

In Erweiterung des Modells gehen wir von unterschiedlichen Parametern in den Regionen und zu jedem Zeitpunkt aus. Wir zerlegen dann den Logarithmus des Parameters  $\nu_{it}$  in eine räumliche, eine zeitliche und eine Saisonkomponente. Wir setzen also für  $\nu_{it}$  an:

$$\log(\nu_{it}) = \alpha_i + \beta_t + \zeta(t) \quad (7.2)$$

Die restlichen Parameter werden wie im Modell mit festem  $\nu$  spezifiziert. Die Inferenz erfolgt für diese Parameter wie in Abschnitt 7.2.2.

### 7.3.1 Raum-Zeit-Modellierung für $\nu$

Die räumlichen Effekte  $\alpha_i$  seien a priori unabhängig normalverteilt. Um die unterschiedliche Größe der Regionen und damit die unterschiedliche Anzahl von Personen unter Risiko im Modell zu berücksichtigen, verwenden wir als Priori-Erwartungswert für den Regioneneffekt den logarithmierten Anteil der Bevölkerung in der Region  $i$  an der Gesamtbevölkerung  $\log(\xi_i)$ ,

$$\begin{aligned}\alpha_i &\sim \text{N}(\log(\xi_i), \sigma_\alpha^2) \\ \sigma_\alpha^2 &\sim \text{IG}(1, 0.001)\end{aligned}$$

Dabei entspricht die Verwendung des logarithmierten Bevölkerungsanteils als Erwartungswert der multiplikativen Berücksichtigung des Bevölkerungsanteils beim Modell mit festem  $\nu$ .

Für den Zeittrend  $\beta$  gehen wir von einem linearen Trend aus und verwenden daher einen Random Walk zweiter Ordnung als Priori

$$\begin{aligned}p(\beta_1) = p(\beta_2) &\propto \text{const.} \\ \beta_t &\sim \text{N}(2\beta_{t-1} - \beta_{t-2}, \sigma_\beta^2) \quad \text{für alle } t \geq 3 \\ \sigma_\beta^2 &\sim \text{IG}(1, 10^{-5})\end{aligned}$$

Die Prioriverteilung des Parametervektors  $\beta$  läßt sich schreiben als

$$p(\beta) = (2\pi\sigma_\beta^2)^{(n-2)/2} \exp\left(-\frac{1}{2\sigma_\beta^2} \beta' \mathbf{R} \beta\right)$$

wobei  $\mathbf{R} = \mathbf{R}^{(2)}$  die Präzisionsmatrix eines Random Walks zweiter Ordnung ist, siehe Abschnitt 2.2.2.

Den saisonalen Trend  $\zeta(t)$  modellieren wir parametrisch als Überlagerung von  $L$  Sinus-Schwingungen:

$$\zeta(t) = \gamma_0 + \sum_{l=1}^L \left( A_l \sin(\rho j t + \phi_l) \right). \quad (7.3)$$

Dabei bezeichnet  $A_l$  die Amplitude der jeweiligen Schwingung und  $\phi_l$  die Phasenverschiebung. Die Grundfrequenz ist  $\rho$ , für wöchentlich vorliegende Daten z.B.  $\rho = \frac{2\pi}{52}$ . Neben

der Grundschwingung über ein Jahr gehen also  $l - 1$  Oberschwingungen mit ein. Nach Bronstein & Semendjajew (1991), S. 183 läßt sich (7.3) auch schreiben als

$$\zeta(t) = \gamma_0 + \sum_{j=1}^J \left( \gamma_{2j-1} \sin(\rho jt) + \gamma_{2j} \cos(\rho jt) \right)$$

mit  $J = 2L$ . Im Folgenden sei

$$\begin{aligned} s_j(t) &= \sin\left(\frac{\rho t j}{2}\right) && \text{für } j \text{ gerade,} \\ s_j(t) &= \cos\left(\frac{\rho t(j+1)}{2}\right) && \text{für } j \text{ ungerade} \\ s_0(t) &= 1. \end{aligned}$$

Damit läßt sich (7.3) auf ein einfaches Regressionsproblem zurückführen

$$\zeta(t) = \sum_{j=0}^J (\gamma_j s_j(t)). \quad (7.4)$$

Als Priori für  $\gamma$  wählen wir  $\gamma \sim N(0, \sigma_\gamma^2 \mathbf{I})$  mit  $\sigma_\gamma^2 = 10$ . Die Priori wird hier relativ informativ gewählt, um sicherzustellen, daß nicht benötigte Frequenzen herausfallen.

### 7.3.2 Inferenz

Die vom Parameter  $\nu$  abhängigen Teile der Loglikelihood lauten

$$\begin{aligned} l(\nu_{it}) &\propto \sum_i \sum_t \left[ X_{it} \log(\nu_{it}) - \nu_{it} \right] + \sum_i \left[ Y_{i1} \log(\nu_{i1}) - \nu_{i1} \frac{\omega_i^{(s)} \omega_{i1} \lambda}{1 - \lambda} \right] \\ &= \sum_i \sum_t \left[ X_{it} \left( \alpha_i + \beta_t + \zeta(t) \right) - \exp \left( \alpha_i + \beta_t + \zeta(t) \right) \right] \\ &\quad + \sum_i \left[ Y_{i1} \left( \alpha_i + \beta_1 + \zeta(1) \right) - \exp \left( \alpha_i + \beta_1 + \zeta(1) \right) \frac{\omega_i^{(s)} \omega_{i1} \lambda}{1 - \lambda} \right]. \end{aligned}$$

Die gemeinsame Verteilung der Parameter  $\alpha$ ,  $\beta$  und  $\zeta$  gegeben  $\mathbf{X}$  und  $\mathbf{Y}$  lautet daher

$$\begin{aligned}
p(\alpha, \beta, \zeta | \mathbf{X}, \mathbf{Y}) \propto & \exp \left\{ \sum_i \sum_t \left[ X_{it} \left( \alpha_i + \beta_t + \zeta(t) \right) - \exp \left( \alpha_i + \beta_t + \zeta(t) \right) \right] \right. \\
& + \sum_i \left[ Y_{i1} \left( \alpha_i + \beta_1 + \zeta(1) \right) - \omega_i^{(s)} \omega_{i1} \frac{\lambda p}{1 - \lambda} \exp \left( \alpha_i + \beta_1 + \zeta(1) \right) \right] \\
& - \frac{I}{2} \log(\sigma_\alpha^2) - \frac{1}{2\sigma_\alpha^2} \sum_{i=1}^I \left( \alpha_i - \log(\xi_i) \right)^2 - \frac{n-2}{2} \log(\sigma_\beta^2) \\
& \left. - \frac{1}{2\sigma_\beta^2} \beta' \mathbf{R} \beta - \frac{J+1}{2} \log(\sigma_\gamma^2) - \frac{1}{2\sigma_\gamma^2} \sum_{j=0}^J \gamma_j^2 \right\}.
\end{aligned}$$

### Räumlicher Effekt

Damit ergibt sich als Full Conditional von  $\alpha_i$

$$\begin{aligned}
p(\alpha_i | \cdot, s) \propto & \exp \left[ -\frac{1}{2\sigma_\alpha^2} (\alpha_i - \log(\xi_i))^2 + \alpha_i \sum_t X_{it} \right. \\
& \left. - \sum_t \exp(\alpha_i + \beta_t + \zeta(t)) + \alpha_i Y_{i1} - \omega_i^{(s)} \omega_{i1} \frac{\lambda p}{1 - \lambda} \exp(\alpha_i + \beta_1 + \zeta(1)) \right] \\
\propto & \exp \left[ -\frac{1}{2\sigma_\alpha^2} \alpha_i^2 + \alpha_i \left( \sum_t X_{it} + Y_{i1} + \frac{\log(\xi_i)}{\sigma_\alpha^2} \right) \right. \\
& \left. - \sum_t \exp(\alpha_i + \beta_t + \zeta(t)) - \omega_i^{(s)} \omega_{i1} \frac{\lambda p}{1 - \lambda} \exp(\alpha_i + \beta_1 + \zeta(1)) \right] \quad (7.5)
\end{aligned}$$

Dies ist keine Standardverteilung. Wir approximieren die Full Conditional daher durch eine Normalverteilung, indem wir eine Taylorentwicklung zweiten Grades des Terms (7.5) durchführen (vgl. dazu Abschnitt 2.3.2). Da für die Ableitung von

$$f(\alpha_i) = - \sum_t \exp(\alpha_i + \beta_t + \zeta(t)) - \omega_i^{(s)} \omega_{i1} \frac{\lambda p}{1 - \lambda} \exp(\alpha_i + \beta_1 + \zeta(1))$$

gilt

$$f'(\alpha_i) = - \sum_t \exp(\alpha_i + \beta_t + \zeta(t)) - \omega_i^{(s)} \omega_{i1} \frac{\lambda p}{1 - \lambda} \exp(\alpha_i + \beta_1 + \zeta(1)) = f(\alpha_i)$$

lautet die Taylorentwicklung zweiten Grades von (7.5) um  $\alpha_i^0$

$$\begin{aligned}
f(\alpha_i) & \approx f(\alpha_i^0) + (\alpha_i - \alpha_i^0) f'(\alpha_i^0) + \frac{1}{2} (\alpha_i - \alpha_i^0)^2 f''(\alpha_i^0) \\
& = \alpha_i (1 - \alpha_i^0) f(\alpha_i^0) + \frac{1}{2} \alpha_i^2 f(\alpha_i^0) + \Gamma,
\end{aligned}$$

wobei  $\Gamma$  eine nicht von  $\alpha$  abhängige Konstante ist. Mithin können wir die Full Conditional von  $\alpha_i$  approximieren durch

$$\begin{aligned} q(\alpha_i|\alpha_i^0) &\propto \exp \left[ -\frac{1}{2\sigma_\alpha^2}\alpha_i^2 + \alpha_i \left( \sum_t X_{it} + Y_{i1} + \frac{\log(\xi_i)}{\sigma_\alpha^2} \right) \right. \\ &\quad \left. + \frac{1}{2}\alpha_i^2 f(\alpha_i) + \alpha_i(1 - \alpha_i^0)f(\alpha_i^0) \right] \\ &= \exp \left[ -\frac{1}{2}\alpha_i^2 \left( \frac{1}{\sigma_\alpha^2} - f(\alpha_i^0) \right) + \alpha_i \left( \sum_t X_{it} + Y_{i1} + \frac{\log(\xi_i)}{\sigma_\alpha^2} + (1 - \alpha_i^0)f(\alpha_i^0) \right) \right] \end{aligned}$$

Gegeben einem Zustand  $\alpha_i^0$  lautet unser Proposal

$$\begin{aligned} \alpha_i|\alpha_i^0 &\sim N(m, s^2) \tag{7.6} \\ \text{mit } s^2 &= 1/\left(\frac{1}{\sigma_\alpha^2} - f(\alpha_i^0)\right) \\ m &= \left(\sum_t X_{it} + Y_{i1} + (1 - \alpha_i^0)f(\alpha_i^0)\right) \cdot s^2 \end{aligned}$$

Um nun  $\alpha$  zu ziehen, verwenden wir einen Metropolis-Hastings-Algorithmus, wobei wir (7.6) als Vorschlagsverteilung verwenden. Dabei setzen wir  $\alpha_i^0 = \alpha_i^{(n-1)}$ , den momentanen Zustand von  $\alpha_i$ . Mit den Abkürzungen  $C = \sum_t X_{it} + Y_{i1}$ ,  $(s^0)^2 := 1/\left(\frac{1}{\sigma_\alpha^2} - f(\alpha_i)\right)$  und  $m^0 := \left(\sum_t X_{it} + Y_{i1} + (1 - \alpha_i)f(\alpha_i)\right) \cdot (s^0)^2$ , berechnet sich die Akzeptanzwahrscheinlichkeit  $a$  des Metropolis-Hastings-Algorithmus hierbei als

$$a = \min \left( \frac{\exp \left( -\frac{1}{2\sigma_\alpha^2}\alpha_i^2 + \alpha_i C + f(\alpha_i) \right) \frac{1}{\sqrt{2\pi(s^0)^2}} \exp -\frac{1}{2} \left( \frac{\alpha_i^0 - m^0}{s^0} \right)^2}{\exp \left( -\frac{1}{2\sigma_\alpha^2}(\alpha_i^0)^2 + \alpha_i^0 C + f(\alpha_i^0) \right) \frac{1}{\sqrt{2\pi s^2}} \exp -\frac{1}{2} \left( \frac{\alpha_i - m}{s} \right)^2}, 1 \right)$$

### Zeitlicher Trend

Während die räumlichen Effekte, welche a priori als unabhängig angenommen werden, einzeln gezogen werden, ziehen wir die Zeiteffekte blockweise. Die Full Conditional von  $\beta = (\beta_1, \dots, \beta_T)'$  lautet

$$\begin{aligned} p(\beta|\cdot) &\propto \exp \left[ -\frac{1}{2\sigma_\beta^2}\beta' \mathbf{R} \beta + \sum_t \left( \beta_t \sum_i X_{it} \right) + \beta_1 \sum_i Y_{i1} \right. \\ &\quad \left. - \sum_t \sum_i \exp(\alpha_i + \beta_t + \zeta(t)) - \omega_i^{(s)} \omega_{i1} \frac{\lambda p}{1 - \lambda} \sum_i \exp(\alpha_i + \beta_1 + \zeta(1)) \right] \end{aligned} \tag{7.7}$$

Mit

$$h(\boldsymbol{\beta}) = - \sum_t \sum_i \exp(\alpha_i + \beta_t + \zeta(t)) - \omega_i^{(s)} \omega_{i1} \frac{\lambda p}{1-\lambda} \sum_i \exp(\alpha_i + \beta_1 + \zeta(1))$$

gilt

$$\begin{aligned} \frac{\partial}{\partial \beta_t} h(\boldsymbol{\beta}) &= - \sum_i \exp(\alpha_i + \beta_t + \zeta(t)) \quad \text{für } t \geq 2 \\ \frac{\partial}{\partial \beta_1} h(\boldsymbol{\beta}) &= - \sum_i \exp(\alpha_i + \beta_1 + \zeta(1)) - \omega_i^{(s)} \omega_{i1} \frac{\lambda p}{1-\lambda} \sum_i \exp(\alpha_i + \beta_1 + \zeta(1)) \\ \frac{\partial^2}{\partial \beta_t \partial \beta_s} h(\boldsymbol{\beta}) &= 0 \quad \text{für } t \neq s \\ \frac{\partial^2}{\partial \beta_t^2} h(\boldsymbol{\beta}) &= - \sum_i \exp(\alpha_i + \beta_t + \zeta(t)) \quad \text{für } t \geq 2 \\ \frac{\partial^2}{\partial \beta_1^2} h(\boldsymbol{\beta}) &= - \sum_i \exp(\alpha_i + \beta_1 + \zeta(1)) - \omega_i^{(s)} \omega_{i1} \frac{\lambda p}{1-\lambda} \sum_i \exp(\alpha_i + \beta_1 + \zeta(1)) \end{aligned}$$

Bezeichne

$$k(\beta_t) = \begin{cases} - \sum_i \exp(\alpha_i + \beta_1 + \zeta(1)) - \omega_i^{(s)} \omega_{i1} \frac{\lambda p}{1-\lambda} \sum_i \exp(\alpha_i + \beta_1 + \zeta(1)) & \text{für } t = 1 \\ - \sum_i \exp(\alpha_i + \beta_t + \zeta(t)) & \text{für } t \geq 2 \end{cases}$$

Dann lautet die Taylorentwicklung von (7.7)

$$\begin{aligned} h(\boldsymbol{\beta}) &\approx h(\boldsymbol{\beta}^0) + \sum_t (\beta_t - \beta_t^0) \frac{\partial}{\partial \beta_t} h(\boldsymbol{\beta}) \\ &\quad + \frac{1}{2} \sum_s \sum_t (\beta_t - \beta_t^0) (\beta_s - \beta_s^0) \frac{\partial^2}{\partial \beta_t \partial \beta_s} h(\boldsymbol{\beta}) \\ &= \sum_t (\beta_t - \beta_t^0) k(\beta_t^0) + \frac{1}{2} \sum_t (\beta_t - \beta_t^0)^2 k(\beta_t^0) + \Gamma_1 \\ &= \sum_t \beta_t (1 - \beta_t^0) k(\beta_t^0) + \frac{1}{2} \sum_t \beta_t^2 k(\beta_t^0) + \Gamma_2, \end{aligned}$$

wobei  $\Gamma_1$  und  $\Gamma_2$  nicht von  $\boldsymbol{\beta}$  abhängige Konstanten sind. Damit können wir  $p(\boldsymbol{\beta}|\cdot)$  approximieren durch

$$\begin{aligned} q(\boldsymbol{\beta}|\boldsymbol{\beta}^0) &\propto \exp \left[ - \frac{1}{2} \left( \boldsymbol{\beta}' \mathbf{R} \boldsymbol{\beta} - \text{diag} \left( (\beta_1^0)^2 k(\beta_1^0), \dots, (\beta_n^0)^2 k(\beta_n^0) \right) \right) \right. \\ &\quad \left. + \sum_t \beta_t \left( \sum_i X_{it} + (1 - \beta_t^0) k(\beta_t^0) + \beta_1^0 \sum_i Y_{i1} \right) \right] \end{aligned} \quad (7.8)$$

Das heißt also,  $\boldsymbol{\beta}$  ist approximativ multivariat normalverteilt mit Präzisionsmatrix

$$\mathbf{Q} = \mathbf{R} - \text{diag}(k(\beta_1^0), \dots, k(\beta_n^0))$$

und Mittelwertsvektor  $\mathbf{m} = \mathbf{Q}^{-1}\mathbf{b}$  mit  $\mathbf{b} = (b_1, \dots, b_n)$  und

$$b_t = \sum_i X_{it} + (1 - \beta_t^0)k(\beta_t^0) + \beta_1^0 \sum_i Y_{i1}.$$

Der Parametervektor  $\boldsymbol{\beta}$  läßt sich damit mit dem Algorithmus von Rue (2001) in einem Metropolis-Hastings-Schritt ziehen. Man verwendet dazu (7.8) als Vorschlagsdichte, siehe Abschnitte 2.3.1 und 2.3.2.

Zur Berechnung der Akzeptanzwahrscheinlichkeit wird die Normalisierungskonstante der Proposalverteilung (7.8) benötigt. Die logarithmierte Normalisierungskonstante läßt sich nach Rue (2001) als  $-\frac{n}{2} \log 2\pi + \sum_t \log(L_{tt})$  schreiben, wobei  $\mathbf{L}$  die Cholesky-Zerlegung  $\mathbf{P} = \mathbf{L}\mathbf{L}'$  von  $\mathbf{P}$  ist. Sei nun  $\boldsymbol{\beta}^0$  der momentane Zustand des Parameters und bezeichne

$$\mathbf{Q}^0 = \mathbf{R} - \text{diag}(k(\beta_1), \dots, k(\beta_n)), \quad \mathbf{Q}^0 = \mathbf{L}^0(\mathbf{L}^0)',$$

$$\mathbf{b}^0 = (b_1^0, \dots, b_n^0), \quad b_t^0 = \sum_i X_{it} + (1 - \beta_t)k(\beta_t) + \beta_1 \sum_i Y_{i1},$$

$\boldsymbol{\mu}^0$  die Lösung des Gleichungssystems

$$\mathbf{L}^0 \mathbf{v}^0 = \mathbf{b}^0, (\mathbf{L}^0)' \boldsymbol{\mu}^0 = \mathbf{v}^0$$

und

$$\mathbf{c} = \left( \sum_i (X_{i1} + Y_{i1}), \sum_i (X_{i2}), \dots, \sum_i (X_{in}) \right)'$$

Dann gilt für den Logarithmus der Akzeptanzwahrscheinlichkeit  $a$

$$\begin{aligned} \log(a) &= -\frac{1}{2\sigma_{\boldsymbol{\beta}}^2} \boldsymbol{\beta}' \mathbf{R} \boldsymbol{\beta} + \boldsymbol{\beta} \mathbf{c} + h(\boldsymbol{\beta}) - \sum_t \log(L_{tt}) + \frac{1}{2} \mathbf{z}' \mathbf{z} \\ &\quad + \frac{1}{2\sigma_{\boldsymbol{\beta}}^2} \boldsymbol{\beta}^0' \mathbf{R} \boldsymbol{\beta}^0 - \boldsymbol{\beta}^0 \mathbf{c} + h(\boldsymbol{\beta}^0) + \sum_t \log(L_{tt}^0) - \frac{1}{2} (\boldsymbol{\beta} - \boldsymbol{\mu}^0)' \mathbf{Q}^0 (\boldsymbol{\beta} - \boldsymbol{\mu}^0) \end{aligned}$$

### Saisonkomponente

Die Full Conditional der Saisonkomponente lautet

$$\begin{aligned}
p(\zeta|\cdot) &= p(\gamma|\cdot) \\
&\propto \exp \left[ \sum_i \sum_t \left[ X_{it} \left( \alpha_i + \beta_t + \sum_{j=0}^J (\gamma_j s_j(t)) \right) - \exp \left( \alpha_i + \beta_t + \sum_{j=0}^J (\gamma_j s_j(t)) \right) \right] \right. \\
&\quad \left. + \sum_i \left[ Y_{i1} \left( \alpha_i + \beta_1 + \sum_{j=1}^J (\gamma_j s_j(t)) \right) - \omega_i^{(s)} \omega_{i1} \frac{\lambda p}{1-\lambda} \exp \left( \alpha_i + \beta_1 + \sum_{j=0}^J (\gamma_j s_j(1)) \right) \right] \right. \\
&\quad \left. - \frac{J+1}{2} \log(\sigma^2)_\gamma - \frac{1}{2\sigma_\gamma^2} \sum_{j=0}^J \gamma_j^2 \right]. \\
&\propto \exp \left[ -\frac{1}{2\sigma_\gamma^2} \gamma' \mathbf{I} \gamma + \sum_{j=0}^J \gamma_j \left( \sum_{i,t} (s_j(t) X_{it}) + \sum_i s_j(1) Y_{i1} \right) \right. \\
&\quad \left. - \sum_t \sum_i \exp \left( \alpha_i + \beta_t + \sum_j (\gamma_j s_j(t)) \right) \right. \\
&\quad \left. + \omega_i^{(s)} \omega_{i1} \frac{\lambda p}{1-\lambda} \sum_t \exp \left( \alpha_i + \beta_1 + \sum_j (\gamma_j s_j(1)) \right) \right]
\end{aligned}$$

Wie oben approximieren wir  $p(\gamma)$ , indem wir eine Taylorentwicklung zweiten Grades der beiden letzten Terme berechnen. Sei also

$$g(\gamma) = - \sum_t \sum_i \exp \left( \alpha_i + \beta_t + \sum_j (\gamma_j s_j(t)) \right) + \omega_i^{(s)} \omega_{i1} \frac{\lambda p}{1-\lambda} \sum_t \exp \left( \alpha_i + \beta_1 + \sum_j (\gamma_j s_j(1)) \right)$$

Die ersten partiellen Ableitungen lauten

$$\begin{aligned}
\frac{\partial}{\partial \gamma_k} g(\gamma) &= - \sum_{it} s_k(t) \exp \left( \alpha_i + \beta_t + \sum_j (\gamma_j s_j(t)) \right) \\
&\quad - \sum_i s_k(1) \omega_i^{(s)} \omega_{i1} \frac{\lambda p}{1-\lambda} \exp \left( \alpha_i + \beta_1 + \sum_j (\gamma_j s_j(1)) \right),
\end{aligned}$$

die zweiten partiellen Ableitungen sind

$$\begin{aligned}
\frac{\partial^2}{\partial \gamma_k \partial \gamma_l} g(\gamma) &= - \sum_{it} s_k(t) s_l(t) \exp \left( \alpha_i + \beta_t + \sum_j (\gamma_j s_j(t)) \right) \\
&\quad - \sum_i s_k(1) s_l(1) \omega_i^{(s)} \omega_{i1} \frac{\lambda p}{1-\lambda} \exp \left( \alpha_i + \beta_1 + \sum_j (\gamma_j s_j(1)) \right)
\end{aligned}$$

Damit läßt sich  $g(\boldsymbol{\gamma})$  um den Punkt  $\boldsymbol{\gamma}^0$  entwickeln:

$$\begin{aligned} g(\boldsymbol{\gamma}) &\approx g(\boldsymbol{\gamma}^0) + \sum_j (\gamma_j - \gamma_j^0) \frac{\partial}{\partial \gamma_j} g(\boldsymbol{\gamma}^0) + \frac{1}{2} \sum_j \sum_k (\gamma_j - \gamma_j^0) (\gamma_k - \gamma_k^0) \frac{\partial^2}{\partial \gamma_j \partial \gamma_k} g(\boldsymbol{\gamma}^0) \\ &= \sum_j \gamma_j \left( \frac{\partial}{\partial \gamma_j} g(\boldsymbol{\gamma}^0) + \sum_k \gamma_k \frac{\partial^2}{\partial \gamma_j \partial \gamma_k} g(\boldsymbol{\gamma}^0) \right) + \frac{1}{2} \sum_j \sum_k \gamma_j \gamma_k \frac{\partial^2}{\partial \gamma_j \partial \gamma_k} g(\boldsymbol{\gamma}^0) + C, \end{aligned}$$

mit  $C$  eine nicht von  $\boldsymbol{\gamma}$  abhängige Konstante. Als Proposal für den Vektor  $\boldsymbol{\gamma}$  verwenden wir also eine multivariate Normalverteilung mit Präzisionsmatrix

$$\mathbf{P} = \frac{1}{\sigma_\gamma^2} \mathbf{I} - \mathbf{H}_g(\boldsymbol{\gamma}^0),$$

wobei  $\mathbf{H}_g(\boldsymbol{\gamma}^0)$  die Hessematrix der Funktion  $g$  an der Stelle  $\boldsymbol{\gamma}^0$  bezeichnet, und Mittelwertsvektor  $\mathbf{P}^{-1} \mathbf{c}$  mit  $\mathbf{c} = (c_1, \dots, c_J)'$ ,

$$c_j = \sum_{i,t} (s_j(t) X_{it}) + \sum_i (s_j(1) Y_{i1}) + \frac{\partial}{\partial \gamma_j} g(\boldsymbol{\gamma}^0) - \sum_k \gamma_k^0 \frac{\partial^2}{\partial \gamma_j \partial \gamma_k} g(\boldsymbol{\gamma}^0) \text{ für alle } j = 1, \dots, J.$$

Die Saisonkomponente wird wieder mit dem Algorithmus von Rue (2001) gezogen. Dabei ist zu beachten, daß die Präzisionsmatrix  $\mathbf{P}$  voll besetzt ist, der Vorteil der Bandstruktur also nicht ausgenutzt werden kann. Der Rechenbedarf für die Ziehung steigt deswegen quadratisch mit der Anzahl der Frequenzen an. Die Akzeptanzwahrscheinlichkeit berechnet sich analog zur Akzeptanzwahrscheinlichkeit beim zeitlichen Trend.

Diese Vorgehensweise entspricht in diesem Fall, also wenn die Präzisionsmatrix voll besetzt ist, dem von Gamerman (1997) vorgestellten *iterative weighted least squares* Proposal.

### Hyperparameter

Die Full Conditionals der Hyperparameter  $\sigma_\alpha^2$  und  $\sigma_\beta^2$  sind Invers-Gammaverteilungen

$$\begin{aligned} \sigma_\alpha^2 | \cdot &\sim IG \left( 1 + \frac{I}{2}, 0.001 + \sum_i (\alpha_i - \log(\xi_i))^2 \right) \\ \sigma_\beta^2 | \cdot &\sim IG \left( 1 + \frac{n-2}{2}, 10^{-5} + \boldsymbol{\beta}' \mathbf{R} \boldsymbol{\beta} \right) \end{aligned}$$

und die Hyperparameter können in Gibbs-Schritten gezogen werden.

## 7.4 Ergebnisse

Im Folgenden vergleichen wir die Ergebnisse für vier verschiedene Modelle, mit festem und variierendem  $\nu$  und jeweils ohne und mit Überdispersion. Tabelle 7.1 zeigt die Mittelwerte der Posterioriverteilungen der Parameter.

Im Modell mit festem  $\nu$  ohne Überdispersion wird der Parameter der epidemischen Fälle  $\lambda$  auf 0.948 geschätzt. Das heißt, das Modell wird epidemisch dominiert. Die endemische Komponente ist dagegen sehr gering, obwohl *Campylobacter*-Infektionen eher als endemisch dominiert gelten. Die Anpassung an die Daten ist dementsprechend eher schlecht.

Beim Modell mit Überdispersion ist die Devianz sehr viel geringer, die Datenanpassung also entsprechend höher. Auch hier wird das Modell aber epidemisch dominiert. Abbildung 7.3 zeigt den Mittelwert der Überdispersion über die Zeit. Man sieht, daß die saisonale Entwicklung zum Teil durch die Überdispersion erklärt wird. Die „Weihnachtslöcher“ finden sich ebenfalls als Ausschläge in der Überdispersion wieder. In den Bundesländer ist der Mittelwert der Überdispersion jeweils nahe bei Null.

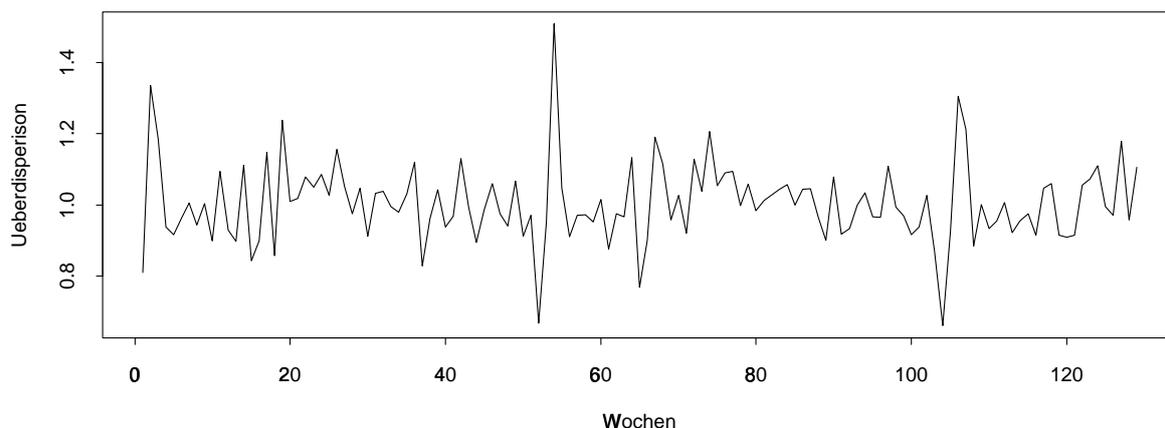


Abbildung 7.3: Mittlere Überdispersion über die Zeit im Modell mit festem  $\nu$ .

Überdisp.	$\nu$ fest		$\nu$ variierend	
	ohne	mit	ohne	mit
$\lambda$	0.948	0.956	0.424	0.594
$\nu$	36.8	32.2	749.9	533.4
$\psi$	-	11.83	-	7.51
Devianz	8228.7	2439.7	5287.4	2399.9
$p_D$	12.9	1418.5	3.6	928.4
DIC	8215.7	3858.2	5283.8	3328.3

Tabelle 7.1: Vergleich der geschätzten Parameter (Posteriori-Mittelwert) bei den Modellen mit festem und variierendem  $\nu$  sowie ohne und mit Überdispersion.

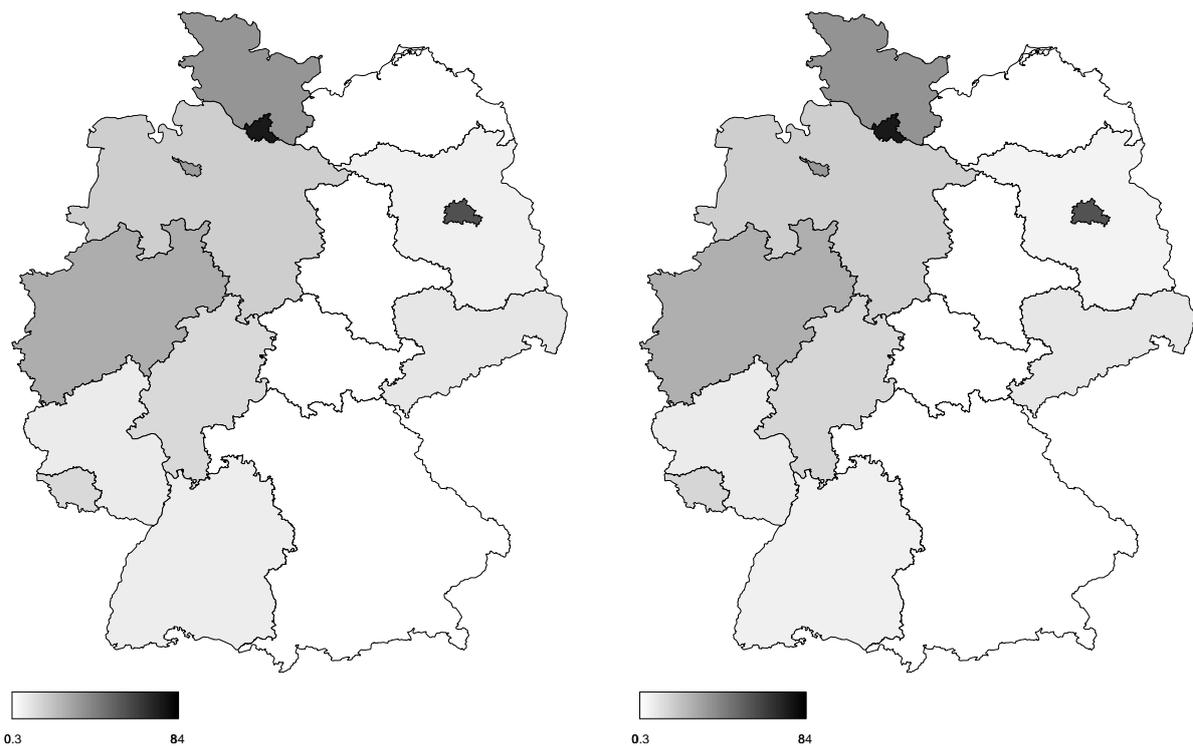


Abbildung 7.4: Räumlicher Effekt des variierenden  $\nu$  nach Berücksichtigung der Population bei den Modellen ohne (links) und mit (rechts) Überdispersion.

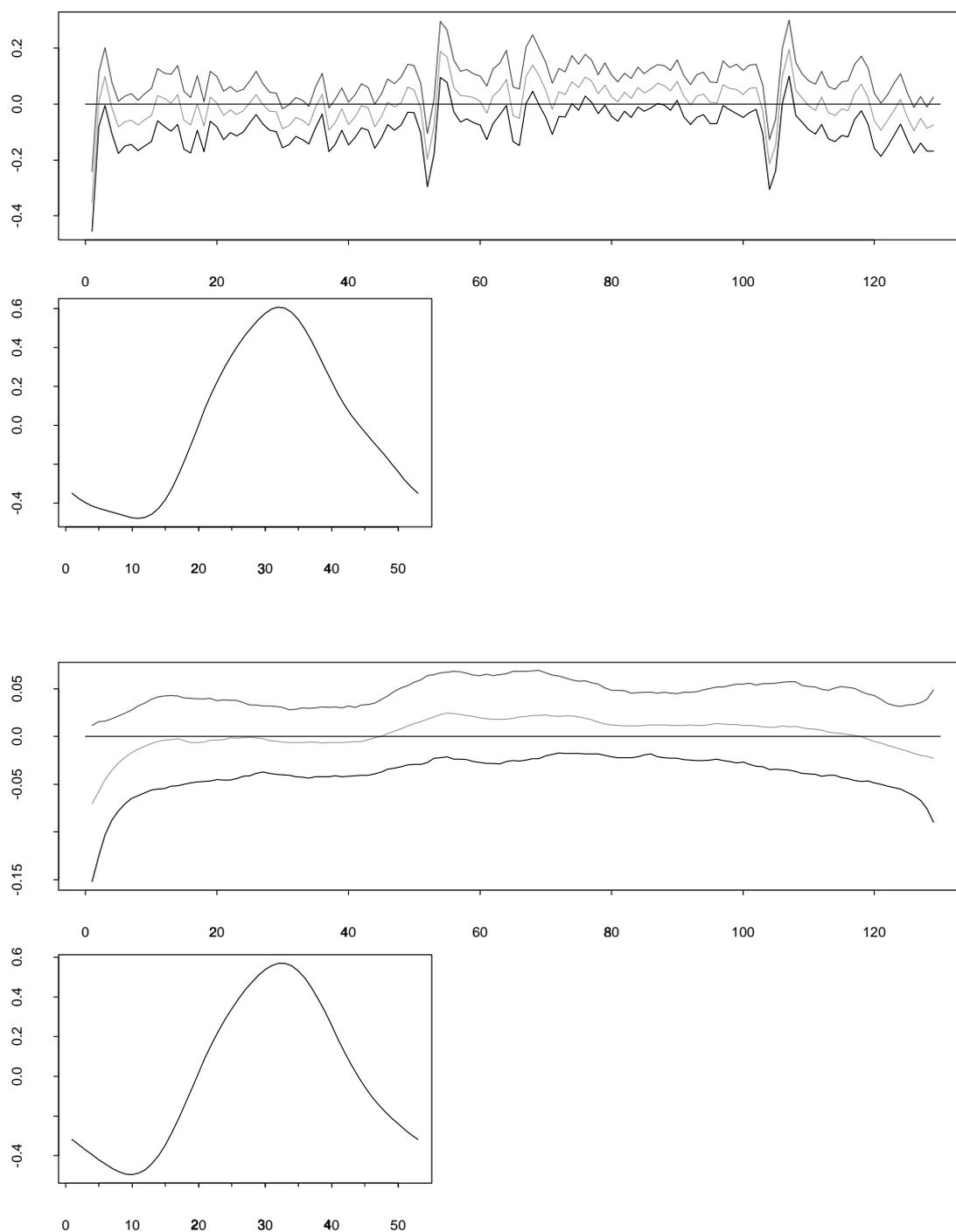


Abbildung 7.5: Zeittrend und Saisonkomponente des variierenden  $\nu$  beim Modell ohne (oben) und mit (unten) Überdispersion. Für den Zeittrend sind Median und 90%-Intervall abgebildet. Die Saisonkomponente wurde gemäß (7.4) aus den Medianen der Regressionsparameter  $\gamma$  berechnet.

Die saisonale Modellierung ist bei diesen Daten also angebracht. Läßt man  $\nu$  über Raum und Zeit variieren, so wird das Modell eher endemisch dominiert. Der Parameter  $\lambda$  wird auf 0.424 geschätzt, der Parameter  $\exp(\gamma_0)$ , der dem Intercept von  $\nu$  entspricht, dagegen auf 749.9. Abbildung 7.5 zeigt Zeittrend und Saisonkomponente im Modell ohne Überdispersion. Der Zeittrend weist hohe Varianz auf und spiegelt insbesondere auch die „Weihnachtslöcher“ wieder. Im zweiten Jahr läßt sich ein leicht erhöhtes Infektionsrisiko feststellen. Wichtiger ist aber die Saisonkomponente. Mit Beginn des Sommers steigt das Risiko schnell an, um im Herbst wieder abzusinken. Abbildung 7.4 zeigt den räumlichen Effekt des Modells nach Berücksichtigung der Population (also  $\exp(\hat{\alpha}_i)/\xi_i$ ). Hohe Raten sind in den Stadtstaaten Hamburg, Berlin und Bremen sowie in Schleswig-Holstein und Nordrhein-Westfalen festzustellen, während im restlichen Bundesgebiet das Infektionsrisiko fast identisch ist. Die Datenanpassung ist in diesem Modell zwar besser als im Modell mit festem  $\nu$ , aber immer noch nicht sehr gut.

Das Modell mit variierendem  $\nu$  und Überdispersion zeigt einen ähnlichen räumlichen Effekt wie das Modell ohne Überdispersion (Abb. 7.4). Der Zeittrend zeigt in diesem Modell wieder eine erhöhte Rate im zweiten Jahr, da das 95%-Intervall jedoch sogar punktweise die Null einschließt, könnte man den Trend hier vernachlässigen (Abb. 7.5). Die Saisonkomponente zeigt im Winter einen etwas anderen Verlauf als das Modell ohne Überdispersion. Dies kann damit erklärt werden, daß die „Weihnachtslöcher“ hier in der Überdispersion aufgefangen werden. Abbildung 7.6 zeigt den Mittelwert der Überdispersion über die Zeit. Eine saisonale Entwicklung kann hier nicht mehr festgestellt werden. Der Mittelwert der Überdispersion in den Ländern ist jeweils etwa bei Null.

Die epidemische Komponente ist in diesem Modell etwas wichtiger als im Modell ohne Überdispersion ( $\lambda = 0.594$ ), der Mittelwert von  $\nu$  dementsprechend geringer ( $\exp(\gamma_0) = 533.4$ ). Der Parameter  $\psi$  ist im Modell mit variierendem  $\nu$  etwas niedriger als im Modell mit festem  $\nu$ , das heißt die Varianz der Überdispersion ist im Modell mit variierendem  $\nu$  überraschenderweise etwas höher. Die Devianz ist in diesem Modell am geringsten, jedoch nur geringfügig niedriger als im Modell mit festem  $\nu$ . Jedoch ist die effektive Zahl der Parameter sehr viel geringer, das Modell stellt sich also als bestes Modell im Sinne des

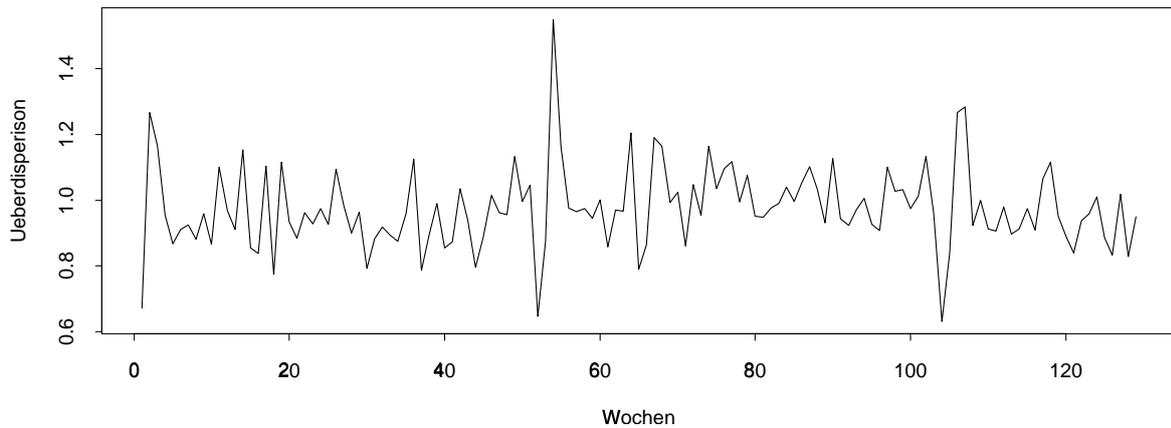


Abbildung 7.6: Mittlere Überdispersion über die Zeit im Modell mit variierendem  $\nu$ .

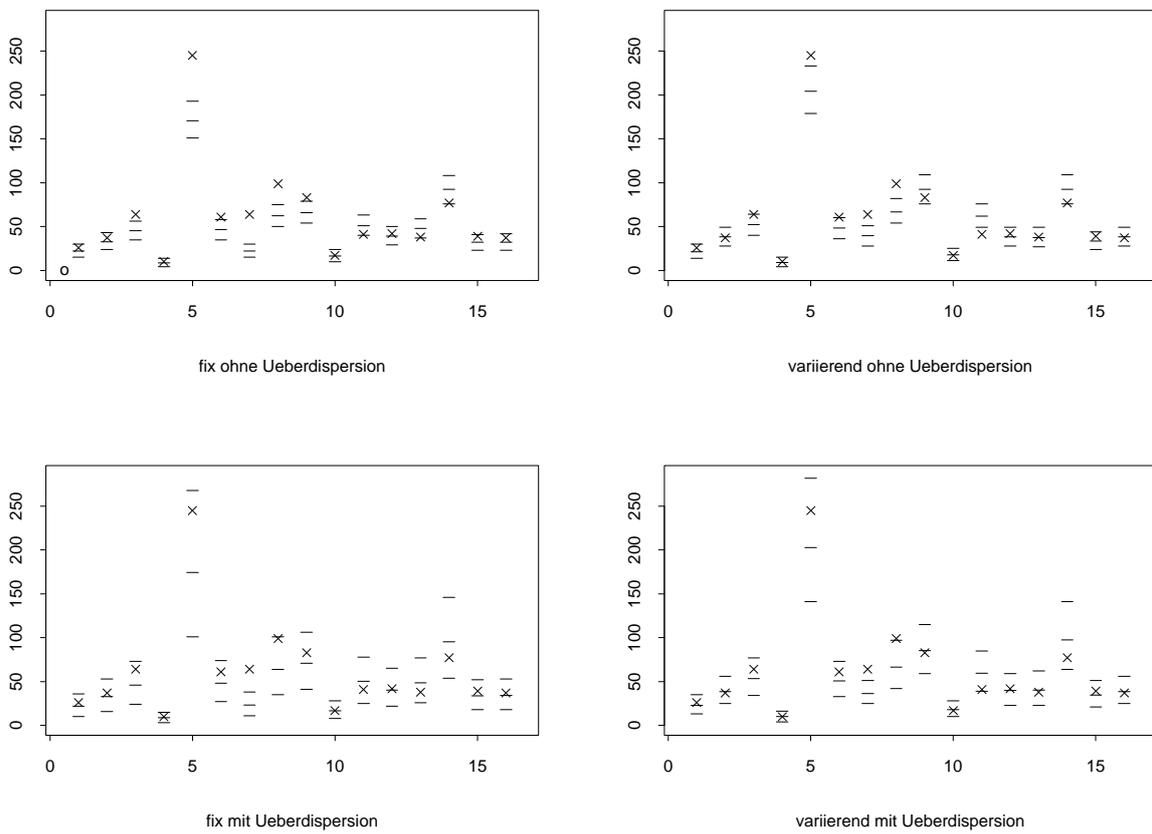


Abbildung 7.7: Vergleich der prädiktiven Eigenschaften der Modelle. Abgebildet sind für jedes Bundesland jeweils Median und 90%-Intervalle der prognostizierten Fallzahl der 130. Woche (Striche), sowie beobachtete Fallzahlen (Kreuze). Die Indizes der Bundesländer entsprechen denen auf S. 12.

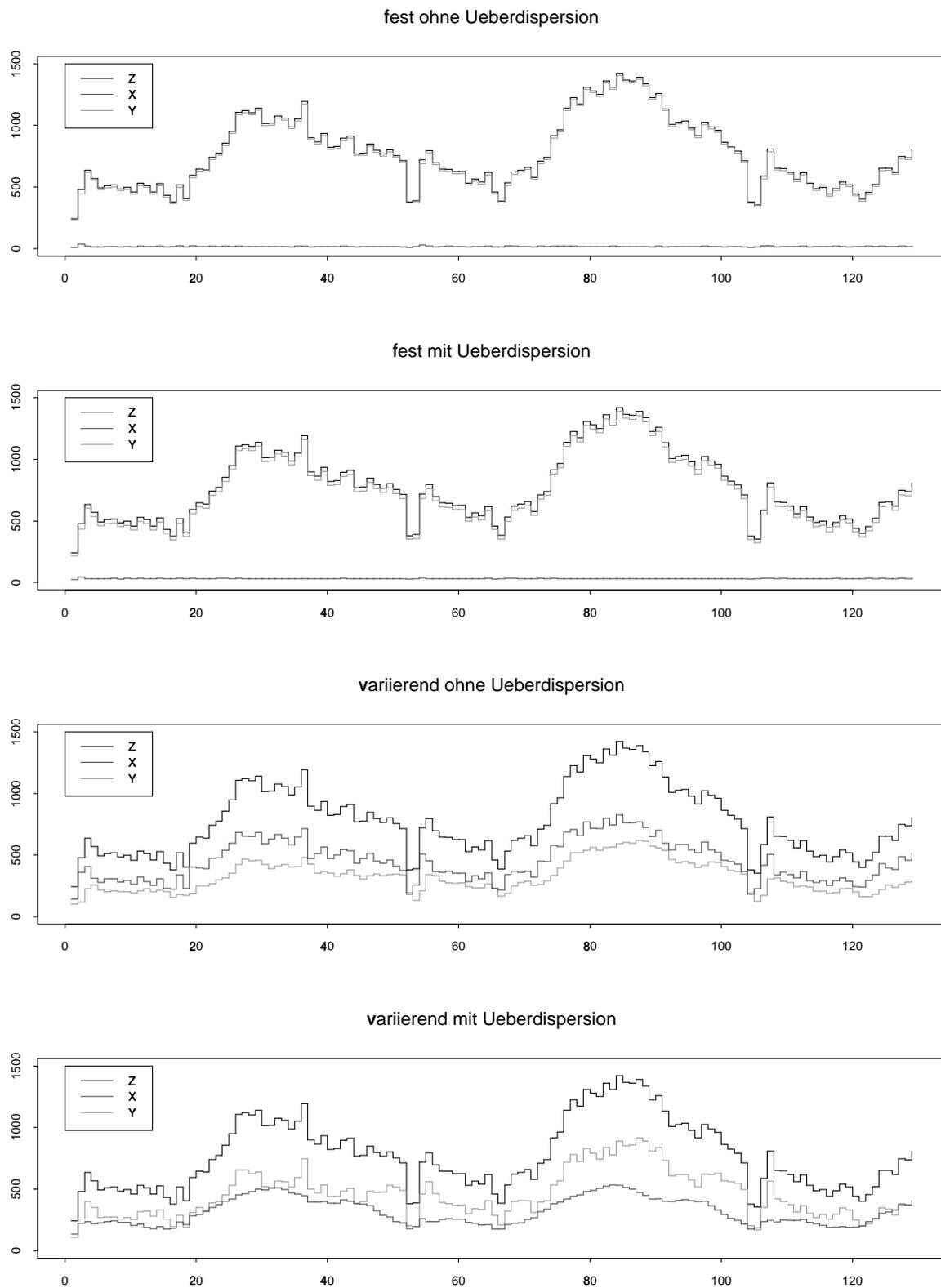


Abbildung 7.8: Aufteilung der gesamten beobachteten Fälle ( $Z$ ) in endemische ( $X$ ) und epidemische ( $Y$ ) Fälle in den vier Modellen.

DIC heraus.

Abbildung 7.8 zeigt die Aufteilung der gesamten Fälle in endemische und epidemische Fälle in der vier Modellen. In den Modellen mit festem  $\nu$  wird die endemische Komponente praktisch vernachlässigt und alles über die epidemische Komponenten erklärt. In den Modellen mit variierendem  $\nu$  wird dagegen sehr viel über die endemische Komponente erklärt.

Zur Überprüfung der prädiktiven Eigenschaften der Modelle wurden die Fallzahlen in der 130. Woche des Beobachtungszeitraums prognostiziert und mit den beobachteten Fallzahlen verglichen. Die Modelle sollen in der Praxis als Alarmsystem eingesetzt werden. Überschreitet die beobachtete Fallzahl ein vorgegebenes Prognoseintervall, ist von einem unvorhergesehenen Ausbruch auszugehen. Abbildung 7.7 zeigt die Prognosen der vier Modelle im Vergleich. Das Modell mit variierendem  $\nu$  und mit Überdispersion ist hier das Beste. Die Modelle mit festem  $\nu$  sind im Allgemeinen weiter von den wahren Werten entfernt. Die Modelle mit Überdispersion haben größere Prognoseintervalle als die ohne Überdispersion; damit wird hier ein zu früher Alarm verhindert.

## 7.5 Diskussion und Ausblick

Ziel des Modells ist die Bereitstellung eines Systems zur frühzeitigen Alarmierung bei epidemischen Ausbrüchen. Die Modellierung mit variierendem  $\nu$  und Überdispersion für die endemische Komponente scheint für endemisch dominierte Infektionen wie *Campylobacter* ein guter Ansatz zu sein. Die parametrische Modellierung der Saisonkomponente sowie die Schätzung räumlicher Effekte ist hierbei besonders wichtig.

Will man die Daten auf kleineren räumlichen Einheiten, z.B. auf Kreisebene analysieren, so empfiehlt es sich die räumlichen Effekte nicht mehr als unabhängig anzunehmen. Da die Daten in den Kreisen sehr spärlich sind, kann die Modellierung mit Hilfe von Nachbarschaftsstrukturen, z.B. mit einem Gauss-Markov-Zufallsfeld, unter Ausnutzung des *borrowing strength*-Prinzips zu stabileren Schätzungen führen, vergleiche Abschnitt 2.2.1. Das

Modell läßt sich leicht auch um Kovariablen erweitern. Eventuell kann damit das Alarmsystem weiter verbessert werden.

Sicher lassen sich für die Surveillancedaten anderer Infektionskrankheiten bessere Modelle finden, sei es durch Raum-Zeit-Modellierung der epidemischen Komponente oder durch die Berücksichtigung von Nachbarschaften zwischen Regionen bereits bei der Unterscheidung in endemische und epidemische Komponente.

Ein möglicher anderer Ansatz kann sein, Nachbarschaftsbeziehungen bereits in den zugrundeliegenden Verzweigungsprozeß mit einzubeziehen. Die epidemische Komponente könnte z.B. wie folgt modelliert werden:

$$Y_{it} \sim \text{Po}\left(\lambda_1(X_{it} + Y_{it}) + \lambda_2\left(\sum_{\{j:j\sim i\}} w_j(X_{jt} + Y_{jt})\right)\right) \text{ für } t \geq 2,$$

wobei  $w_j$  Gewichte mit  $\sum_j w_j = 1$  sind. Das heißt, auch die Fallzahlen der vorhergehenden Zeitpunkten in benachbarten Regionen führen in gewissem Grad zu neuen (epidemischen) Fällen in der entsprechenden Region  $i$ .

Mit den vorgestellten Methoden ist es möglich, ein Alarmsystem zu entwickeln. Dieses soll dem Robert-Koch-Institut dabei helfen, Ausbrüche epidemischer Krankheiten frühzeitig zu erkennen und Gegenmaßnahmen zu ergreifen.



# Kapitel 8

## Schlußbemerkungen

Erst die rasante Entwicklung der Rechengeschwindigkeit moderner Computer hat es ermöglicht, komplexe hierarchische Bayesianische Modelle zu berechnen. Inferenz ist in solchen Modellen nur über aufwendige MCMC-Methoden möglich. Neben der größeren Rechengeschwindigkeit sorgen auch in den letzten Jahren entwickelte effiziente Algorithmen wie der von Rue (2001) beschriebene Blockupdate oder auch die von Holmes & Held (2003) vorgeschlagene Formulierung mit Hilfsvariablen dafür, daß immer komplexere Modelle schätzbar sind.

Die in dieser Arbeit vorgestellten Modelle decken eine große Bandbreite epidemiologischer Datensätze ab. Die räumlichen und zeitlichen Trends werden meist mit Hilfe autoregressiver Prioris geglättet. Diese Prioris legen keine Vorinformation auf den Mittelwert der Effekte, sondern nur auf die Differenzen von Effekten benachbarter Regionen oder Zeitpunkte. Auf gleiche Weise lassen sich auch Raum-Zeit-Interaktionen modellieren.

Der nach Besag *et al.* (1991) als BYM-Modell bezeichnete Ansatz ist der Ausgangspunkt der räumlichen Analyse epidemiologischer Modelle. Das Modell wurde hier auf Daten aus der Veterinärepidemiologie angewandt. In Zusammenarbeit mit den Anwendern wurde das Modell erweitert. Eine weitere Erweiterung um Raum-Zeit-Interaktionen wird angestrebt, um die Ausbreitung von Krankheiten besser verstehen zu können.

Stetige räumliche Daten lassen sich nach Diskretisierung auch mit diskreten Gauss-Markov-Zufallsfeldern modellieren. Die Glättungsparameter lassen sich dabei zwar im Algorithmus mitschätzen, aber die Wahl des Nachbarschaftsmodells beeinflusst ebenfalls die Glättung.

Ein wichtiges Analyseinstrument in der Epidemiologie sind Alters-Perioden-Kohorten (APC)-Modelle. Die Modellierung mit Random Walks erster Ordnung bringt eine stochastische Restriktion in das Modell ein und löst damit zum Teil das Nicht-Identifizierbarkeitsproblem. Mit Random Walks lassen sich einfach und logisch Prognosen zukünftiger Raten erstellen. Die Erweiterung auf räumliche strukturierte Daten ist möglich. Mit Interaktionen läßt sich bei den vorgestellten Modellen die Prognosefähigkeit verbessern. Sicherlich gibt es kein allgemein bestes Modell, sondern für jeden Datensatz muß die Analyse mit allen Modellen verglichen werden. Die im Rahmen dieser Arbeit entstandene Software BAMP wird inzwischen von Anwendern für Bayesianische Alters-Perioden-Kohorten-Modellierung und Prognose verwendet.

Für Surveillancedaten von Infektionskrankheiten wurde ein stochastisches Modell vorgestellt, welches die beobachteten Fälle in endemische und epidemische Fälle zu unterscheiden sucht. Bei endemisch dominierten Krankheiten ist die Raum-Zeit-Modellierung der endemischen Komponente sinnvoll. Mit der prädiktiven Verteilung der geschätzten zukünftigen Fälle läßt sich ein Alarmsystem errichten.

Die vorgestellten Anwendungen zeigen, wie mit Bayesianischer räumlich-zeitlicher Statistik epidemiologische Problemstellungen behandelt werden können. Die Modelle geben Hinweise auf die Ursachen der untersuchten Krankheiten, sie zeigen Trendänderungen auf, arbeiten lokale Schwerpunkte heraus und zeigen mittels Interaktionen lokale Trends auf. Diese Modelle helfen also dem Anwender, sich einen Überblick über die Verbreitung einer Krankheit zu gewinnen. Sie ermöglichen es, sich auf zukünftig zu erwartende Krankheitsfälle einzustellen und mithin frühzeitig Gegenmaßnahmen zu veranlassen.

# Anhang A

## Beweis

### A.1 Gemeinsame Verteilung bei zwölf nächsten Nachbarn

Auf einem regulären Gitter soll ein Gauss-Markov-Zufallsfeld erzeugt werden. Für jeden Pixel gehen dabei die zwölf nächsten Nachbarn in die lokale Priori mit ein, wobei die Gewichte  $1/4$ ,  $1/8$  und  $-1/8$  für die ersten, zweiten bzw. dritten Nachbarn (also jene mit Abstand  $1$ ,  $\sqrt{2}$  bzw.  $2$ ) vergeben werden. Die gemeinsame Verteilung

$$p(\mathbf{u}) \propto \exp \left[ -\frac{\tau}{2} \left( \sum_{j_1=1}^{J_1} \sum_{j_2=3}^{J_2} (u_{j_1, j_2} - 2u_{j_1, j_2-1} + u_{j_1, j_2-2})^2 \right. \right. \\ \left. \left. + \sum_{j_1=3}^{J_1} \sum_{j_2=1}^{J_2} (u_{j_1, j_2} - 2u_{j_1-1, j_2} + u_{j_1-2, j_2})^2 \right. \right. \\ \left. \left. - \sum_{j_1=2}^{J_1} \sum_{j_2=2}^{J_2} (u_{j_1, j_2} - u_{j_1, j_2-1} - u_{j_1-1, j_2} + u_{j_1-1, j_2-1})^2 \right) \right]$$

hat dann die gewünschte gemeinsame Verteilung.

Beweis:

$$\begin{aligned}
p(u_{xy}|\cdot) &\propto \exp \left[ -\frac{\tau}{2} \left( \sum_{j_2=3}^{J_2} (u_{x,j_2} - 2u_{x,j_2-1} + u_{x,j_2-2})^2 \right. \right. \\
&\quad + \sum_{j_1=3}^{J_1} (u_{j_1,y} - 2u_{j_1-1,y} + u_{j_1-2,y})^2 \\
&\quad - \sum_{j_2=2}^{J_2} (u_{x,j_2} - u_{x,j_2-1} - u_{x-1,j_2} + u_{x-1,j_2-1})^2 \\
&\quad \left. \left. - \sum_{j_2=2}^{J_2} (u_{x+1,j_2} - u_{x+1,j_2-1} - u_{x,j_2} + u_{x,j_2-1})^2 \right) \right] \\
&\propto \exp \left[ -\frac{\tau}{2} \left( (u_{x,y} - 2u_{x,y-1} + u_{x,y-2})^2 + (u_{x,y+1} - 2u_{x,y} + u_{x,y-1})^2 \right. \right. \\
&\quad + (u_{x,y+2} - 2u_{x,y+1} + u_{x,y})^2 + (u_{x,y} - 2u_{x-1,y} + u_{x-2,y})^2 \\
&\quad + (u_{x+1,y} - 2u_{x,y} + u_{x-1,y})^2 + (u_{x+2,y} - 2u_{x+1,y} + u_{x,y})^2 \\
&\quad - (u_{x,y} - u_{x,y-1} - u_{x-1,y} + u_{x-1,y-1})^2 \\
&\quad - (u_{x,y+1} - u_{x,y} - u_{x-1,y+1} + u_{x-1,y})^2 \\
&\quad - (u_{x+1,y} - u_{x+1,y-1} - u_{x,y} + u_{x,y-1})^2 \\
&\quad \left. \left. - (u_{x+1,y+1} - u_{x+1,y} - u_{x,y+1} + u_{x,y})^2 \right) \right] \\
&\propto \exp \left[ -\frac{\tau}{2} \left( u_{x,y}^2 - 2u_{x,y}u_{x,y-1} + u_{x,y}u_{x,y-2} - 2u_{x,y}u_{x,y+1} + 4u_{x,y}^2 - 2u_{x,y}u_{x,y-1} \right. \right. \\
&\quad + u_{x,y}u_{x,y+2} - 2u_{x,y}u_{x,y+1} + u_{x,y}^2 + u_{x,y}^2 - 2u_{x,y}u_{x-1,y} + u_{x,y}u_{x-2,y} \\
&\quad - 2u_{x,y}u_{x+1,y} + 4u_{x,y}^2 - 2u_{x,y}u_{x-1,y} + u_{x,y}u_{x+2,y} - 2u_{x,y}u_{x+1,y} + u_{x,y}^2 \\
&\quad - u_{x,y}^2 + u_{x,y}u_{x,y-1} + u_{x,y}u_{x-1,y} - u_{x,y}u_{x-1,y-1} \\
&\quad + u_{x,y}u_{x,y+1} - u_{x,y}^2 - u_{x,y}u_{x-1,y+1} + u_{x,y}u_{x-1,y} \\
&\quad + u_{x,y}u_{x+1,y} - u_{x,y}u_{x+1,y-1} - u_{x,y}^2 + u_{x,y}u_{x,y-1} \\
&\quad \left. \left. - u_{x,y}u_{x+1,y+1} + u_{x,y}u_{x+1,y} + u_{x,y}u_{x,y+1} - u_{x,y}^2 \right) \right] \\
&\propto \exp \left[ -\frac{\tau}{2} \left( 8u_{x,y}^2 - 2u_{x,y}u_{x-1,y} - 2u_{x,y}u_{x,y-1} - 2u_{x,y}u_{x+1,y} - 2u_{x,y}u_{x,y+1} \right. \right. \\
&\quad + u_{x,y}u_{x-2,y} + u_{x,y}u_{x+2,y} + u_{x,y}u_{x,y-2} + u_{x,y}u_{x,y+2} \\
&\quad \left. \left. - u_{x,y}u_{x-1,y-1} - u_{x,y}u_{x+1,y-1} - u_{x,y}u_{x-1,y+1} - u_{x,y}u_{x+1,y+1} \right) \right]
\end{aligned}$$

# Anhang B

## BAMP

Im Rahmen dieser Arbeit entstand die Software BAMP zur **B**ayesianischen **A**lters-**P**erioden-Kohorten-**M**odellierung und **P**rognose. Das Programm eignet sich zur Erstellung von Bayesianischen APC-Modellen, wie in Kapitel 5 beschrieben. Alle dort beschriebenen Modellvarianten sind verfügbar. Die Schätzung der Parameter erfolgt mit effizienten MCMC-Verfahren (Rue 2001).

Die Software ist kostenfrei im Internet unter

`http://www.stat.uni-muenchen.de/sfb386/software/bamp.html`

verfügbar. Auf den folgenden Seiten findet sich das BAMP-Handbuch.

# BAMP

## Bayesianische Alters–Perioden–Kohorten- Modellierung und Prognose

Volker Schmid      Leonhard Held

Institut für Statistik  
Ludwig-Maximilians-Universität München  
Ludwigstr. 33, D-80539 München

Ver. 1.3.β.1 - 16.7.2004

---

### Lizenzvereinbarung

Die Autoren billigen jedem Individuum und jeder nicht kommerziellen Organisation das Recht zu, diese Software zu verwenden und eine beliebige Anzahl von Kopien zu erstellen. Kommerzielle Nutzung bedarf einer Lizenzierung durch die Autoren. Es ist nicht erlaubt, die Software zu dekompileieren oder zu modifizieren. Der Copyright-Hinweis darf nicht entfernt werden. Die Software darf nicht ohne Zustimmung der Autoren verkauft werden. Die Autoren sind nicht verantwortlich für Schäden, welche durch die Verwendung dieses Programms entstehen. Ebenso geben sie keinerlei Garantie, auch nicht bei ordnungsgemäßem Gebrauch. Die Software wird vertrieben wie sie ist und Sie, der Anwender, sind für alle bei der Nutzung entstehenden Risiken verantwortlich.

---

## B.1 Allgemeines

BAMP ist ein Softwarepaket zur Analyse von Inzidenz- oder Mortalitätsdaten auf dem Lexis-Diagramm mittels Bayesianischer Alters-Perioden-Kohorten-Modelle (APC-Modelle). Das Modell wird in Berzuini *et al.* (1993), Besag *et al.* (1995) und Knorr-Held & Rainer (2001) beschrieben. Für jeden Pixel auf dem Lexis-Diagramm (d.h. für die Kombination einer Altersgruppe und Periode) müssen Anzahl der Personen unter Risiko (z.B. Bevölkerungszahl) und die Anzahl der Krankheitsfälle (typischerweise Krebsinzidenz- oder -mortalitätsfälle) zur Verfügung stehen. Es wird ein hierarchisches Modell verwendet mit einem Binomial-Modell auf der ersten Stufe

$$y_{jt} \sim B(n_{jt}, \pi_{jt})$$
$$\log\left(\frac{\pi_{jt}}{1 - \pi_{jt}}\right) = \xi_{jt} = \mu + \theta_j + \phi_t + \psi_k,$$

wobei  $y_{jt}$  die Zahl der Fälle,  $n_{jt}$  die Zahl der Personen unter Risiko der Altersgruppe  $j = 1, \dots, J$  und Periode  $t = 1, \dots, T$  bezeichnet.

Random Walk Prioris erster oder zweiter Ordnung (RW 1 oder RW 2) sind als Glättungspriori für die Alters-, Perioden- und Kohortenparameter möglich. BAMP erlaubt das Modell zu reduzieren und zum Beispiel ein Alters-Perioden- oder Alters-Kohorten-Modell zu analysieren. Zusätzliche Heterogenität kann durch zufällige Effekte für jeden Pixel auf dem Lexis-Diagramm oder für jeden Haupteffekt mit aufgenommen werden. Zu beachten ist, daß im APC-Modell ein Nicht-Identifizierbarkeitsproblem besteht (Clayton & Schifflers 1987b). Dadurch entstehen Probleme in der Visualisierung und Interpretation der Alters-, Perioden- und Kohorteneffekte.

BAMP kann Folgendes leisten:

- BAMP berechnet Prognosen bei gegebenen zukünftigen Bevölkerungszahlen,
- BAMP berechnet retrospektive Prognosen, um das Modell zu überprüfen,

- Daten können unterschiedlich skaliert sein, z.B. können die Perioden in Jahres-Intervallen und die Altersgruppen in Fünf-Jahres-Intervallen vorliegen

Graphische Routinen um

- die Haupteffekte,
- einen Vergleich zwischen beobachteten und prognostizierten Fällen,
- prognostizierte Raten und
- die „Signifikanz“ zusätzlicher Heterogenität

abzubilden, sind für S-Plus und R erhältlich.

BAMP ermöglicht ebenfalls, Kovariablen in das Modell aufzunehmen, um eine bessere Datenanpassung und bessere Prognose zu ermöglichen. Diese Funktion ist noch in Entwicklung und wird in diesem Handbuch nur der Vollständigkeit halber erwähnt, aber nicht genauer beschrieben.

## B.2 Erste Schritte

Nach dem Herunterladen und Entpacken von BAMP muß die Datei `bamp.ini` geändert werden oder eine eigene `*.ini`-Datei erzeugt werden (siehe Abschnitt B.3.1). Das Programm wird mit

```
bamp ini-file
```

oder, falls die Datei `bamp.ini` verwendet wird, mit

```
bamp
```

auf der Kommandozeile gestartet.

Bei graphischen Oberflächen kann BAMP auch durch Doppelklick auf das BAMP-Icon gestartet werden. BAMP verwendet dann die Datei `bamp.ini`. In beiden Fällen muß die Daten `cygwin1.dll` im gleichen Verzeichnis liegen wie `bamp.exe` (gilt nur für die Windows-Version).

## B.3 Eingabedateien

### B.3.1 \*.ini-Datei

Das Modell und die Eingabedateien wurden durch Setzen von Variablen in einer \*.ini-Datei spezifiziert, standardmäßig in der Datei `bamp.ini`. Folgende Variablen können gesetzt werden:

cases file	Pfad und Name der Datei mit den Fallzahlen	†
population file	Pfad und Name der Datei mit der Zahl der Personen unter Risiko	†
dataorder	Anordnung der Daten in den Eingabedateien (siehe Abschnitt B.3.3)	0
output folder	Pfad zum Ordner, in dem die Ergebnisse ausgegeben werden	†
temp folder	Pfad zum Ordner für temporäre Dateien und Samples	†
ouput file	Dateiname für allgemeine Programmausgabe (0: im Programmfenster)	0
number of agegroups	Zahl der Altersgruppen	†
number of periods	Zahl der Perioden	†
periods per agegroup	Skalierung der Perioden relativ zu den Altersgruppen	†
number of predictions	Zahl der zu prognostizierenden Jahre	0
predictions	0: keine Prognose 1: Prognose der Raten 2: Prognose mit Vergleich beobachteter Daten 3: Prognose der Fälle bei gegebener Bevölkerungszahl	0
number of iterations	Zahl der Iterationen (einschließlich Burn-In)	†
burn in	Zahl der Iterationen für den Burn-In (einschließlich Tuning)	†

step	Ausdünnung	†
tuning	Zahl der Iterationen bis zum Tuning	†
samples	1 falls die Samples gespeichert werden sollen, 0 sonst	0
lock	0: Bestehende Samples werden überschrieben 1: Prüfung, ob Samples in „temp folder“ existieren	1
deviance	0: Devianz wird nicht berechnet 1: Devianz wird berechnet 2: Devianz wird auch für prognostizierte Fälle berechnet (falls predictions: 2)	0
age block	Priori des Alterseffekts, siehe Abschnitt B.3.2	0
age hyperpar. a	Parameter $a$ und $b$ der Priori der Präzision des Alterseffekts	‡
age hyperpar. b		‡
age hyperpar.2 a	Parameter $a$ und $b$ der Priori der Präzision des unstrukturierten Alterseffekts	‡
age hyperpar.2 b		‡
period block	Priori des Periodeneffekts, siehe Abschnitt B.3.2	0
period hyperpar. a	Parameter $a$ und $b$ der Priori der Präzision des Periodeneffekts	‡
period hyperpar. b		‡
period hyperpar.2 a	Parameter $a$ und $b$ der Priori der Präzision des unstrukturierten Periodeneffekts	‡
period hyperpar.2 b		‡
cohort block	Priori des Kohorteneffekts, siehe Abschnitt B.3.2	0
cohort hyperpar. a	Parameter $a$ und $b$ der Priori der Präzision des Kohorteneffekts	‡
cohort hyperpar. b		‡
cohort hyperpar.2 a	Parameter $a$ und $b$ der Priori der Präzision des unstrukturierten Kohorteneffekts	‡
cohort hyperpar.2 b		‡
z mode	0: ohne, 1: mit pixelweiser Heterogenität	1
z hyperpar. a	Hyperparameter $a$ und $b$ der Priori der Präzision der der pixelweisen Heterogenität	‡
z hyperpar. b		‡
quantile 1	Auszugebende Quantile (-1: keine Ausgabe)	-1
quantile 2		-1
quantile 3		-1
quantile 4		-1
quantile 5		-1
period covariate	Pfad und Name der Kovariablendatei für die Periode	‡
period start	Erster Datenpunkt, der als Kovariable verwendet werden soll	‡
cohort covariate	Pfad und Name der Kovariablendatei für die Kohorte	‡
cohort start	Erster Datenpunkt, der als Kovariable verwendet werden soll	‡

Die letzte Spalte zeigt jeweils die Standardbelegung. Mit † gekennzeichnete Zeilen müssen angegeben werden, mit ‡ gekennzeichnete Zeilen müssen unter Umständen angegeben werden (z.B. müssen die Parameter der Priori im Regelfall angegeben werden, es sein denn der entsprechende Effekt wird aus dem Modell entfernt).

Jede Zeile der \*.ini-Datei besteht aus der Angabe des Namens der zu setzenden Variable, gefolgt von einem Doppelpunkt und dem Wert der Variable. Die Reihenfolge ist beliebig. Beispiel für den Beginn einer \*.ini-Datei:

```
cancer data      :daten/lungcam.txt
population data  :daten/popm.txt
output folder    :daten/output/
temp folder      :./localtmp/
dataorder        :0
number of agegroups :9
...
```

### B.3.2 Mögliche Modellierungen der APC-Effekte

Für die Alters-, Perioden- und Kohorteneffekte sind folgende Angaben möglich:

- 1 RW 1-Priori
- 2 RW 2-Priori
- 0 Effekt wird aus dem Modell entfernt
- 3 RW 1-Priori und zusätzliche Heterogenität
- 4 RW 2-Priori und zusätzliche Heterogenität
- 8 RW 1-Priori mit Einbeziehung von Kovariableninformation
- 9 RW 2-Priori mit Einbeziehung von Kovariableninformation

### B.3.3 Daten

Zur Ausführung des Programms werden zwei Dateien benötigt, welche die zu untersuchenden Dateien enthalten: Einerseits die Fallzahlen (`cases file`), andererseits die Zahl der Personen unter Risiko (`population file`). Beide Dateien müssen in der Form einer Matrix mit  $T$  Zeilen und  $J$  Spalten vorliegen. Setzt man `dataorder:1`, so liegen die Dateien transponiert vor (also  $T$  Spalten und  $J$  Zeilen).

Für Prognosen bei gegebenen Daten für  $S$  Zeitpunkte (`predictions:2`) müssen beide Dateien  $T + S$  Zeilen haben. Sind nur die Bevölkerungszahlen gegeben (`predictions:3`), so hat auch nur die Datei mit den Bevölkerungszahlen  $T + S$  Zeilen (bzw. Spalten für `dataorder:1`). In allen Fällen werden überzählige Zeilen ignoriert.

## B.4 Ausgabedateien

Der MCMC-Algorithmus erzeugt Zufallsziehungen (Samples) der Posterioriverteilung der unbekannt Parameter. Die Samples werden in `temp folder` geschrieben. Nach Ende des MCMC-Algorithmus berechnet das Programm (punktweise) Quantile der Posterioriverteilung. Diese werden im Ordner `output folder` abgelegt. Die Variablen `quantile 1` bis `quantile 5` spezifizieren, welche Quantile berechnet werden.

Die Ausgabe besteht aus folgenden Dateien:

- `theta.txt`, `phi.txt`, `psi.txt`: Quantile der Alters, Perioden- bzw. Kohorteneffekte, jede Zeile ein Quantil, Ausgabe nur bei RW 1-Prioris
- `hyper.txt`: Quantile der Präzisionsparameter, in jeder Zeile Angabe des Parameters und die entsprechenden Quantile
- `theta2.txt`, `phi2.txt`, `psi2.txt`: Quantile der zweiten Differenzen der Haupteffekte, jede Zeile ein Quantil

- `z.txt`: Quantile der pixelweisen Heterogenität; jeder Block ist ein Quantil mit  $T$  Zeilen und  $J$  Spalten
- `my.txt`: Quantile des Intercepts  $\mu$
- `devianz.txt` (falls `deviance > 0`): Quantile der Devianz des Modells
- `devianz_prog.txt` (falls `prognosis:2` und `deviance:2`): Quantile der prädiktiven Devianz (pro Jahr eine Zeile)
- `dic.txt`: Berechnung des DIC
- `pr.txt`: Quantile der geschätzten Wahrscheinlichkeit ( $T$  Zeilen,  $J$  Spalten)
- `y_pred.txt`, `ysum_pred.txt` (falls `prediction:2` oder `3`, siehe Abschnitt B.5): Quantile prognostizierter Fälle pro Altersgruppe und summiert
- `beta_phi.txt`, `beta_psi.txt`: Perioden- und Kohorteneffekt, falls Kovariablen verwendet werden

## B.5 Devianz und Prognose

Um die Güte der Datenanpassung zu messen, gibt BAMP standardmäßig die Posteriori-Devianz aus. Die Devianz ist definiert als

$$D = 2 \cdot \sum_{j,t} [l(\hat{y}_{jt}) - l(y_{jt})],$$

wobei  $l(\hat{y}_{jt})$  die Log-Likelihood des Modells bezeichnet und  $l(y_{jt})$  die maximal erreichbare Log-Likelihood. Je kleiner die Devianz, um so besser die Datenanpassung. Für Modelle mit Heterogenität wird die Datenanpassung immer gut sein, es empfiehlt sich daher ein Vergleich über den DIC (Spiegelhalter *et al.* 2002) oder die prädiktive Devianz.

Um Prognosen zukünftiger Raten zu erhalten, setzt man `predictions:1` und `number of`

`predictions` auf die Zahl der Jahre, die man prognostizieren will. Hat man Bevölkerungszahlen für diesen Zeitraum, setzt man `predictions:3`, BAMP prognostiziert dann auch die Fälle. Um das Modell zu testen, kann man auch Prognosen mit vorhandenen Daten vergleichen. Die Eingabedateien müssen diese dann enthalten (siehe B.3.3). Setzt man `deviance:2`, berechnet BAMP die prädiktive Devianz für jedes prognostizierte Jahr.

## B.6 S-Plus- / R- Skript

Auf der BAMP-Homepage sind Skripte für die Programmpakete S-Plus und R zur Auswertung der Ausgabedateien zu finden. Damit können

- die Haupteffekte (nur für RW 1!)
- ein Vergleich zwischen beobachteten und prognostizierten Fällen
- prognostizierte Raten
- die „Signifikanz“ zusätzlicher Heterogenität

graphisch abgebildet werden. Um die Skripte zu verwenden, müssen die ersten Zeilen entsprechend modifiziert werden:

- `startingage`: Beginn der ersten Altersgruppe
- `startingperiod`: Beginn des Beobachtungszeitraums
- `yearsperperiod`: Zahl der Jahre pro Periode
- `inifile`: Pfad/Name der `*.ini`-Datei

Die Skripte können in S-Plus / R mit

```
source("Pfad/bamp.S")
```

gestartet werden. Die Ausgabe erfolgt in der Datei `bamp.ps` im Ausgabeverzeichnis.

# Literaturverzeichnis

- Armitage, P. & Colton, T. (Hrsg.) (1998). *Encyclopedia of Biostatistics*. Vol. 1. John Wiley & Sons, Chichester.
- Assunção, R. M., Reis, I. A. & Oliveira, C. (2001). Diffusion and prediction of Leishmaniasis in a large metropolitan area in Brazil with a Bayesian space-time model. *Statistics in Medicine* **20**: 2319–2335.
- Becker, N. & Wahrendorf, J. (1997). *Atlas of Cancer Mortality in the Federal Republic of Germany 1981–1990*. Berlin: Springer.
- Bellman, R. (1970). *Introduction to Matrix Analysis*. 2. Auflage. McGraw-Hill, New York.
- Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C., Ghislandi, M. & Songini, M. (1995). Bayesian analysis of space-time variation in disease risk. *Statistics in Medicine* **14**: 2433–2443.
- Berzuini, C., Clayton, D. & Bernardinelli, L. (1993). Bayesian inference on the Lexis diagram. *Bulletin of the International Statistical Institute* **50**: 149–164.
- Besag, J. E. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society Series B* **36**: 192–236.
- Besag, J. E., Green, P. J., Higdon, D. M. & Mengersen, K. L. (1995). Bayesian computation and stochastic systems (with discussion). *Statistical Science* **10**: 3–66.
- Besag, J. E. & Higdon, D. (1999). Bayesian analysis of agricultural field experiments. *Journal of the Royal Statistical Society Series B* **61**: 691–746.

- Besag, J. E. & Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika* **82**: 733–746.
- Besag, J. E., York, J. & Mollié, A. (1991). Bayesian image restoration with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics* **43**: 1–59.
- Boeing, H., Frentzel-Beyme, R. & Berger, M. (1991). Case-control study on stomach cancer in Germany. *International Journal of Cancer* **47**: 858–864.
- Box, G. E. P. & Tiao, G. C. (1973). *Bayesian inference in statistical analysis*. John Wiley & Sons, Reading.
- Bray, I. (2002). Application of Markov chain Monte Carlo methods to projecting cancer incidence and mortality. *Journal of the Royal Statistical Society Series C (Applied Statistics)* **50**: 151–164.
- Bronstein, I. & Semendjajew, K. (Hrsg.) (1991). *Taschenbuch der Mathematik*. 25. Auflage. Verlag Nauka, Moskau.
- Clayton, D. G. (1996). Generalized linear mixed models. in W. R. Gilks, S. Richardson & D. J. Spiegelhalter (Hrsg.), *Markov chain Monte Carlo in Practice*. Chapman & Hall, London. S. 275–301.
- Clayton, D. G. & Bernardinelli, L. (1992). Bayesian methods for mapping disease risks. in P. Elliot, J. Cuzick, D. English & R. Stern (Hrsg.), *Geographical and Environmental Epidemiology: Methods for Small Area Studies*. Oxford University Press, Oxford. S. 205–220.
- Clayton, D. & Schifflers, E. (1987a). Models for temporal variation in cancer rates. i: age-period and age-cohort models. *Statistics in Medicine* **6**: 449–467.
- Clayton, D. & Schifflers, E. (1987b). Models for temporal variation in cancer rates. ii: age-period-cohorts models. *Statistics in Medicine* **6**: 468–481.

- Cressie, N. (1993). *Statistics for Spatial Data*. Wiley, New York.
- DeGroot, M. (1970). *Optimal statistical decision*. 16. Auflage. McGraw-Hill.
- Elliott, P., Martuzzi, M. & Shaddick, G. (1995). Spatial statistical methods in environmental epidemiology: a critique. *Statistical Methods in Medical Research* **4**: 137–159.
- Elliott, P., Wakefield, J., Best, N. & Briggs, D. (2000). *Spatial epidemiology: Methods and Applications*. Oxford University Press, New York.
- Fahrmeir, L. & Lang, S. (2001). Bayesian inference for generalized additive mixed models based on Markov random field priors. *Journal of the Royal Statistical Society Series C (Applied Statistics)* **50**: 201–220.
- Gamerman, D. (1997). Sampling from the posterior distribution in generalized linear mixed models. *Statistics and Computing* **7**: 57–68.
- Gelfand, A. & Vounatsou, P. (2003). Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics* **4**: 11–25.
- Gelman, A., Carlin, J., Stern, H. & Rubin, D. (2000). *Bayesian Data Analysis*. Chapman & Hall/CRC Press, London/Boca Raton.
- Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. (Hrsg.) (1996). *Markov chain Monte Carlo in Practice*. Chapman & Hall, London.
- Haining, R. (1998). Spatial statistics and the analysis of health data. in A. Gatrell & M. Löytönen (Hrsg.), *GIS and Health*. Taylor & Francis, London. S. 29–47.
- Hammersley, J. & Clifford, P. (1971). Markov fields on finite graphs and lattices. *Unveröffentlicht*.
- Hansell, A., Schmid, V. & Knorr-Held, L. (2003). Projecting mortality from obstructive lung disease using a Bayesian age-period-cohort model. *Unveröffentlicht*.

- Heuer, C. (1994). *Analyse von Zeittrends in Mortalitäts- und Inzidenzraten mit Alters-Perioden-Kohortenmodellen und Regressionssplines*. Dissertation. Universität Dortmund.
- Hodges, J. S., Carlin, B. P. & Fan, Q. (2003). On the Precision of the Conditionally autoregressive Prior in Spatial Models. *Biometrics* **59**: 317–322.
- Hoeting, J. A., Madigan, D., Raftery, A. E. & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science* **14**(4): 382–401.
- Hofmann, M., Hoehle, M. & Held, L. (2004). A stochastic model for multivariate surveillance of infectious diseases. Sonderforschungsbereich 386, *Discussion Paper* 394.
- Holford, T. (1983). The estimation of age, period and cohort effects for vital rates. *Biometrics* **39**: 311–324.
- Holmes, C. C. & Held, L. (2003). On the simulation of Bayesian binary and polychotomous regression models using auxiliary variables. *Technical report*. Imperial College, London.
- Kelsall, J. E. & Diggle, P. J. (1998). Spatial variation in risk of disease: a nonparametric binary regression approach. *Journal of the Royal Statistical Society Series C (Applied Statistics)* **47**: 559–573.
- Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine* **19**: 2555–2567.
- Knorr-Held, L. (2003). Some remarks on Gaussian Markov random field models for disease mapping. in P. J. Green, N. L. Hjort & S. Richardson (Hrsg.), *Highly Structured Stochastic Systems*. Oxford University Press, Oxford. S. 260–264.
- Knorr-Held, L. & Besag, J. E. (1998). Modelling risk from a disease in time and space. *Statistics in Medicine* **17**: 2045–2060.
- Knorr-Held, L. & Rainer, E. (2001). Projections of lung cancer mortality in West Germany: a case study in Bayesian prediction. *Biostatistics* **2**: 109–129.

- Knorr-Held, L. & Raßer, G. (2000). Bayesian detection of clusters and discontinuities in disease maps. *Biometrics* **56**: 13–21.
- Knorr-Held, L. & Rue, H. (2002). On block updating in Markov random field models for disease mapping. *Scandinavian Journal of Statistics* **29**: 597–614.
- Krige, D. (1951). A statistical approach to some basic mine valuation problems on the witwatersrand. *Journal of Chemical, Metallurgical and Mining Society of South Africa* **52**(6): 119–139.
- Künsch, H. (1994). Robust priors for smoothing and image restoration. *Annals of the Institute of Statistical Mathematics* **46**: 1–19.
- Lagazio, C., Biggeri, A. & Dreassi, E. (2003). Age-period-cohort models and disease mapping. *Environmetrics* **14**: 475–490.
- Lagazio, C., Dreassi, E. & Biggeri, A. (2001). A hierarchical Bayesian model for space-time variation of disease risk. *Statistical Modelling* **1**: 17–29.
- Lawson, A., Biggeri, A., Boehning, D., Lesaffre, E., Viel, J.-F. & Bertollini, R. (1999). *Disease Mapping and Risk Assessment for Public Health*. Wiley, Chichester.
- Lawson, A., Biggeri, A., Boehning, D., Lesaffre, E., Viel, J.-F., Clark, A., Schlattmann, P. & Divino, F. (2000). Disease mapping models: an empirical evaluation. *Statistics in Medicine* **19**: 2217–2241.
- Lewis, J. (1982). Algorithm 582: The Gibbs-Poole-Stockmeyer and Gibbs-King algorithms for reordering sparse matrices. *ACM Transactions on Mathematical Software* **8**(2): 190–194.
- Lexis, W. (1875). *Einleitung in die Theorie der Bevölkerungsstatistik*. Trübner: Straßburg.
- Müller, T., Teuffert, J., Zellmer, R. & Conraths, F. (2001). Experimental infection of European wild boars and domestic pigs with pseudorabies viruses with differing virulence. *American Journal of Veterinary Research* **62**: 252–258.

- Müller, T., Teuffert, J., Ziedler, K., Possardt, C., Kramer, M., Staubach, C. & Conraths, F. (1998). Pseudorabies virus infection of the European wild boar from eastern Germany. *Journal of Wildlife Disease* **34**: 251–258.
- Nakamura, T. (1986). Bayesian Cohort Models for General Cohort Table Analysis. *Annals of the Institute of Mathematical Statistics* **38**: 353–370.
- Pauwels, R., Buist, A., Calverley, P., Jenkins, C. & Hurd, S. (2001). Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Pulmonary Disease. *American Journal of Respiratory and Critical Care Medicine* **163**(5): 1256–1276. NHLBI/WHO Global Initiative for Chronic Obstructive Lung Disease (GOLD) Workshop Summary.
- Pfeiffer, D. & Morris, R. (1994). Spatial analysis techniques in veterinary epidemiology. *The Kenya Vet* **18**: 483–485.
- Rainer, E. (1998). *Prognose der Lungenkrebsmortalität in Westdeutschland mit Bayesianischen Alters-Perioden-Kohorten-Modellen*. Diplomarbeit. Ludwig-Maximilians-Universität München.
- Raßer, G. (2003). *Clustering Partition Models for Discrete Structures with Application in Geographical Epidemiology*. Dissertation. Ludwig-Maximilians-Universität München.
- Robertson, C. & Boyle, P. (1998). Age-period-cohort analysis of chronic disease rates. I: modelling approach. *Statistics in Medicine* **17**(12): 1305–1323.
- Rue, H. (2001). Fast sampling of Gaussian Markov random fields. *Journal of the Royal Statistical Society Series B* **63**: 325–338.
- Rüger, B. (1999). *Test- und Schätztheorie, Band I: Grundlagen*. Oldenbourg, München.
- Schmid, V. & Held, L. (2004). Bayesian extrapolation of space-time trends for cancer registry data. *Biometrics* **60**: 1034–1042.

- Shimakura, S. E., Carvalho, M. S., Aerts, D. R. & Flores, R. (2001). Distribuição especial do risco: modelagem da mortalidade infantil em Porto Alegre, Rio Grande do Sul, Brasil. *Cadernos de Saúde Pública/Reports in Public Health* **17**(5): 1251–1261.
- Smans, M. & Estéve, J. (1996). Practical approaches to disease mapping. in P. Elliot, J. Cuzick, D. English & R. Stern (Hrsg.), *Geographical and Environmental Epidemiology*. Oxford University Press, Oxford. S. 141–150.
- Smith, D. (1983). Maximum likelihood estimation of the parameters of the beta binomial distribution - algorithms as 189. *Journal of the Royal Statistical Society Series C (Applied Statistics)* **32**: 196–204.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society Series B* **64**: 583–639.
- Staubach, C., Schmid, V., Knorr-Held, L. & Ziller, M. (2002). A Bayesian model for spatial wildlife disease prevalence data. *Preventive Veterinary Medicine* **56**: 75–87.
- Staubach, C., Tackmann, K., Thulke, H.-H., Hugh-Jones, M. & Conraths, F. (2001). Geographical information system-aided analysis of factors potentially influencing the spatial distribution of *echinococcus multilocularis* infections of foxes. *American Journal of Tropical Medicine and Hygiene* **65**: 943–948.
- Stubbe, C., Mehlitz, S., Peukert, R., Goretzki, J., Stubbe, W. & Meynhardt, H. (1989). Lebensraumnutzung und Populationsumsatz des Schwarzwilds in der DDR – Ergebnisse der Wildmarkierung. *Beiträge zur Jagd- und Wildforschung* **16**: 212–231.
- Tackmann, K., Löschner, U., Mix, H., Staubach, C., Thulke, H.-H. & Conraths, F. (1998). Spatial distribution patterns of *echinococcus multilocularis* (Leuckert 1863) (cestoda: Cyclophyllidea: Taniidea) among red foxes in an endemic focus in Brandenburg (Germany). *Epidemiology and Infection* **120**: 101–109.



# Lebenslauf

Volker Johann Schmid

03.08.1973	Geboren in Roding/Oberpfalz
1980 - 1984	Grundschule Roding
1984 - 1993	Josef-von-Fraunhofer Gymnasium Cham
07.07.1993	Abiturprüfung
Juli 1993 - Juni 1994	Grundwehrdienst in Feldkirchen b. Straubing und Roding
Okt. 1994 - Juni 2000	Studium der Statistik an der LMU München
28.11.1996	Vordiplomsprüfung Statistik
05.06.2000	Diplomprüfung Statistik
Juli 2000 - Dez. 2004	vollbeschäftigter wissenschaftlicher Mitarbeiter bei Prof. Dr. L. Fahrmeir bzw. Prof. Dr. L. Held und im Sonderforschungsbereich 386 "Statistische Analyse diskreter Strukturen"