Ludwig-Maximilians-Universität München

Institut für Statistik

Masterarbeit

Analyse der Wählerwanderung zwischen der Oberbürgermeisterwahl (2015) und Bundestagswahl (2013) in Mannheim anhand von Aggregat- und Individualdaten

Autor:
Petra Kopecki

Betreuer:
Prof. Dr. Helmut Küchenhoff
André Klima

17. August 2016

Danksagung

An dieser Stelle möchte ich mich bei all denjenigen bedanken, die mich bei der Erstellung meiner Masterarbeit unterstützt haben.

Prof. Dr. Helmut Küchenhoff danke ich für die Beratung und die Themenstellung dieser Arbeit. Bei Dr. Barbara Felderer bedanke ich mich für die Bereitstellung der Individualdaten. Dem Wahlbüro der Stadt Mannheim möchte ich für die Tabellen mit der Zuordnung der Straßen zu den Wahlbezirken bei den betrachteten Wahlen danken.

Weiterhin möchte ich mich bei André Klima für die umfangreiche Betreuung bedanken. Vielen Dank, dass Sie immer Zeit gefunden haben, alle meine Fragen zu beantworten und alle Unklarheiten zu beseitigen.

Ein herzliches Dankeschön geht an Daniela Sebald für das präzise Korrekturlesen. Insbesondere vielen Dank für die Zeit, die sie sich genommen hat, mir meine Fehler zu erklären.

Ganz besonders möchte ich mich bei meinem Mann bedanken, der mich moralisch und finanziell unterstützt hat und an mich geglaubt hat, auch wenn ich selbst gezweifelt habe.

Abstract

Zielsetzung dieser Arbeit ist die Schätzung der Wählerwanderung zwischen der Bundestagswahl im Jahr 2013 und der Oberbürgermeisterwahl im Jahr 2015 in Mannheim anhand von Aggregat- und Individualdaten. Hierfür wurden die amtlichen Ergebnisse in Form von Aggregatdaten und eine Nachwahlbefragung in Form der Individualdaten zur Verfügung gestellt. Es kommen zwei Modelle zur Anwendung, das *Multinomial-Dirichlet-Modell* von Rosen et al. (2001) und das *Multinomial-Log-Normal-Modell* von Greiner und Quinn (2009, 2010). Beide hierarchischen Modelle basieren auf der Bayesianischen Inferenz. Die Analyse erfolgt zum einen anhand der Aggregatdaten durch die ökologischen Versionen der Modelle und zum anderen anhand der Kombination von Individual- und Aggregatdaten durch die hybriden Versionen der Modelle. Das Multinomial-Dirichlet-Modell wurde von Schlesinger (2013) zum Hybridmodell ergänzt. In seiner Version ermöglicht er Vorwissen in das Modell zu integrieren. Alle Berechnungen und Grafiken in der Arbeit werden mit der Statistiksoftware `R` (R Core Team, 2015) erzeugt. Konkret dienen zwei Pakete für die Analyse. Das Erste, `eiwild` Paket, wurde von Schlesinger (2014) in `R` implementiert und das Zweite, `RxCEcolInf` Paket, wurde von Greiner et al. (2013) entwickelt. Die Güte der Schätzung lässt sich nicht überprüfen, da der wahre Zustand der Wählerwanderung zwischen zwei betrachteten Wahlen nicht bekannt ist. Infolgedessen werden die Modelle durch Konvergenzdiagnose und Vergleich der erzeugten Ketten bewertet. Anhand der betrachteten Daten wird die Konvergenz beim Multinomial-Log-Normal-Modell nicht erkannt. Das Multinomial-Dirichlet-Modell ist für die praktische Umsetzung besser geeignet und hat in allen Versionen des Modells die präziseren und zuverlässigeren Ergebnisse im Vergleich zum Multinomial-Log-Normal-Modell erzeugt. Die Individualdaten tragen in dieser Arbeit zur Stabilität der Ketten bei beiden Modellen bei. Einen Zuschuss leistet hierbei auch das integrierbare Vorwissen beim Multinomial-Dirichlet-Modell. Letztendlich werden die Ergebnisse anhand des Multinomial-Dirichlet-Hybridmodells mit dem Hyperpriori-Parameter $Gamma(\lambda_1 = 30, \lambda_2 = 1)$ für die Zellen der Loyalen ermittelt und interpretiert. Es zeigt sich eine sehr starke Wanderung der Wähler aller Parteien zur Kategorie *Nichtwähler*. Die Ausnahme sind die Wähler der *Grünen* und der *FDP*. Daraus lässt sich schließen, dass die Bedeutung der *Nichtwähler* bei der Wählerwanderungsanalyse weder von den Politikern noch von den Statistikern ignoriert werden darf.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

1.1 Motivation

Einige Wähler bleiben ihrer Partei nach einer Legislaturperiode nicht loyal. Das Interesse der Politiker, Soziologen, Politologen und der Öffentlichkeit an der Neuorientierung und der Wegrichtung bei der Stimmenvergabe ist sehr groß. Bleibt ein Wähler, der beispielsweise die Partei P bei der Bundestagswahl 2009 gewählt hat, treu oder wählt er eine andere Partei? Zu welcher Partei wandert seine Stimme und wie hoch sind die Wahrscheinlichkeiten, dass der Wähler zu anderen Parteien wechselt? In der Statistik haben die Wissenschaftler unterschiedliche Methoden entwickelt, um solche Fragestellungen beantworten zu können.

Eine Methode stellt die Analyse der Individualdaten dar, die mithilfe einer Nachwahlbefragung (Eng. *Exit-Poll*) erhoben werden können. Die Bürgerinnen und Bürger werden nach der Wahl am Ausgang der Wahllokale gebeten, sich zu äußern, wie sie gewählt haben (Payne et al., 1986; Greiner und Quinn, 2012). Um die Wählerwanderung zu schätzen, können die Wähler auch über ihre letzte Wahl befragt werden. Die Schätzung durch diese Methode ist aus mehreren Gründen anfällig für Verzerrungen. Zum einen weigern sich viele Wähler an der Befragung teilzunehmen (Greiner und Quinn, 2012; Payne et al., 1986). Dies ist problematisch, da sich die Wähler, die ihre Teilnahme an der Studie verweigern, in ihrem Wahlverhalten in der Regel von den Wählern unterscheiden, die teilgenommen haben. Zum zweiten können sich einige der Befragten an die Vergabe ihrer Wahlstimme bei der letzten Wahl nicht mehr erinnern, wodurch fehlende oder falsche Angaben auftreten können. Himmelwelt et al. (1978) untersuchen das Problem und ermitteln, dass die falschen Angaben mit dem Zeitabstand zunehmen. Die Wähler der großen Parteien und die Wähler, die loyal geblieben sind, konnten sich dabei besser an ihre letzte Wahl erinnern. Obwohl ihre Studie das nicht bestätigt, berichten Himmelwelt et al. (1978), dass gemäß anderen Autoren die Angabe einer gesellschaftlich akzeptierten Antwort eine weitere Quelle für Fehler darstellen kann. Dies umfasst beispielsweise die Unterschätzung der Nichtwähler

oder die Überschätzung der Partei, die gewonnen hat. Payne et al. (1986) geben an, dass die Briefwähler bei einer Umfrage im Vorfeld der Wahlen nicht betrachtet werden können. Darüber hinaus ist die Durchführung einer Nachwahlbefragung am Ausgang der Wahllokale kosten- und arbeitsaufwändig (Greiner und Quinn, 2012). Aus erwähnten Gründen sind Individualdaten oft unzuverlässig oder nicht verfügbar.

Andererseits stehen die offiziellen amtlichen Wahlergebnisse kostenlos jedem zur Verfügung. Diese Daten sind vollständig, jedoch nach den Wahlgebiet oder Wahlbezirk gruppiert, sprich aggregiert (Ambühl, 2003, S. 8). Die „Beziehungen zwischen Variablen auf der Aggregatebene können, müssen aber nicht ähnliche Beziehungen auf der Individualebene widerspiegeln“ (Gschwend, 2006, S. 227). Deswegen besteht bei der Schlussfolgerung von Aggregatdaten auf das individuelle Verhalten die Gefahr, den sogenannten *ökologischen* [1] *Fehlschluss* (Pappi, 1977 in: Gschwend, 2006) zu treffen. Zum Beispiel stellt man sich vor, dass in einem Stadtbezirk eine positive Korrelation zwischen den Asylbewerbern und den Anstieg der Angriffe mit gefährlichen Körperverletzungen beobachtet wurde. Daraus könnte man schließen, dass die Asylbewerber für solche Angriffe verantwortlich sind. Hypothetisch wäre es jedoch möglich, dass Asylbewerber in diesen Stadtbezirken öfter von Rechtsextremen angegriffen wurden. Vor dem ökologischen Korrelationsproblem warnte Robinson schon im Jahr 1950. Von diesem Zeitpunkt an sind viele Methodiker auf der Suche nach einem fehlerfreien Verfahren der ökologischen Inferenz. Um die Vorteile der Aggregat- und Individualdaten ausnutzen zu können, wurden letztendlich die neuen Hybridmodelle entwickelt, die zur Bestimmung der Wählerwanderung die beiden Datenquellen kombinieren.

Für die Analyse in dieser Arbeit werden zwei hierarchische Modelle, die auf Bayesianischer Inferenz basieren, in ihrer ökologischen und hybriden Version mithilfe der Statistiksoftware `R` (R Core Team, 2015) angewendet. Das ökologische Multinomial-Dirichlet-Modell von Rosen et al. (2001) wurde von Schlesinger (2013) zum Hybridmodell ergänzt. Die beiden Versionen werden hier in seinem Paket `eiwild` (Schlesinger, 2014) berechnet. Das ökologische und hybride Multinomial-Log-Normal-Modell von Greiner und Quinn (2009, 2010) wurde von Autoren im `RxCEcolInf` Paket (Greiner et al., 2013) implementiert.

[1] Als *ökologisch* werden die Daten bezeichnet, wenn die Sub-Gruppen von Individuen bezüglich der geographischen bzw. ökologischen Einheiten oder Regionen (Stadtbezirk, Stadt, Land, usw.) aufgebaut werden (Robinson, 1950; Cho und Manski, 2009).

1.2 Struktur der Arbeit

Da die Schätzungen der betrachteten hierarchischen Modelle auf *Markov-Chain-Monte-Carlo-Verfahren* basieren, wird im Abschnitt 2.1 des Kapitels 2 die Basis der Bayesianischen Inferenz und die Funktionsweise der Markov-Chain-Monte-Carlo-Verfahren vorgestellt. Im Abschnitt 2.2 werden ferner die möglichen Vorgehensweisen bei der Konvergenzdiagnose der Markov-Ketten erläutert. Dazu werden die Begriffe *Burn-In*, *Thinning* und *Sample* erklärt, welche für die Interpretation der Konvergenz und der Ergebnisse von Relevanz sind.

Eine theoretische Einführung und die Darstellung der Grundprinzipien der ökologischen Inferenz erfolgen im Kapitel 3. Im Abschnitt 3.1 werden zuerst die grundlegenden Modelle der ökologischen Inferenz aufgezeigt. Nach einer kurzen Beschreibung deren Vormodelle erfolgt im Abschnitt 3.2 die Darstellung der interessierenden ökologischen, hierarchischen Modelle, des *Multinomial-Dirichlet-Modells* von Rosen et al. (2001) und des *Multinomial-Log-Normal-Modells* von Greiner und Quinn (2009, 2010). Schließlich befasst sich der Abschnitt 3.3 mit deren Erweiterung auf die *Hybridmodelle*, welche durch die Individualdaten ergänzt werden.

Im Abschnitt 4.1 des Kapitels 4 wird die Datengrundlage vorgestellt und beschrieben. Der Unterabschnitt 4.1.1 erläutert den Inhalt und Ursprung der Aggregatdaten, die amtlichen Ergebnisse der Bundestagswahl (2013) und der Oberbürgermeisterwahl (2015). Folglich beschreibt der Unterabschnitt 4.1.2 die Ergebnisse und die Problematik der Individualdaten, die durch eine Nachwahlbefragung in Mannheim (Juni, 2015) erhoben worden sind. Im Abschnitt 4.2 wird die theoretische Begründung und die Beschreibung der Datenaufbereitung dargelegt.

Im Kapitel 5 werden die wichtigen Funktionen der verwendeten Pakete beschrieben. Unterdessen wird im Unterabschnitt 5.1.4 erläutert, wie beim Multinomial-Dirichlet-Modell die Hyperpriori-Parameter die Priori-Verteilung beeinflussen und wie das Vorwissen (Schlesinger, 2013), falls vorhanden, für die Verbesserung der Schätzung verwendet werden kann.

Im Kapitel 6 erfolgt die Konvergenzdiagnose der erzeugten Ketten und Vergleich der Ketten und Modelle für das Multinomial-Dirichlet-Modell im Abschnitt 6.1 und für das Multinomial-Log-Normal-Modell im Abschnitt 6.2. Letztendlich werden im Abschnitt 6.3 die Ergebnisse des gewählten Multinomial-Dirichlet-Hybridmodells mit Hyperpriori-Parameter $Gamma(\lambda_1 = 30, \lambda_2 = 1)$ dargestellt und beschrieben.

2 Einführung in die Bayes-Inferenz und MCMC Verfahren

2.1 Bayes-Inferenz

Der grundlegende Unterschied zwischen Bayesianischer und frequentistischer Inferenz stammt aus der unterschiedlichen Betrachtung des unbekannten Parameters θ. In der Bayesianischen Inferenz wird θ als zufällige Variable betrachtet, hingegen ist θ in der frequentistischen Inferenz eine feste Größe (Held und Bové, 2014, S. 167). Eine andere bedeutende Eigenschaft besteht in der Quantifizierung der Unsicherheit in der Inferenz durch die Wahrscheinlichkeitsmodelle, wodurch die Anpassung komplexer Modelle mit vielen Parametern möglich ist (Gelman et al., 2014, S. 3 f.). Die Information über den unbekannten Parameter θ lässt sich *a priori* und *a posteriori* als Dichte einer Wahrscheinlichkeitsverteilung darstellen. Vor der Beobachtung der Daten wird eine Priori-Verteilung $p(\theta)$ definiert, während die Posteriori-Verteilung $f(\theta|x)$ das vollständige Wissen über den unbekannten Parameter enthält, welches durch die Datenbeobachtung verfügbar wird (Held und Bové, 2014, S. 167).

Die Berechnung der Posteriori-Verteilungsfunktion beruht auf dem Theorem von Thomas Bayes. Gegeben seien Ereignisse A und B, wobei $P(B) > 0$. Die Wahrscheinlichkeit, dass ein Ereignis A eintritt, wenn wir wissen, dass Ereignis B bereits eingetreten ist, ist nach dem Bayes Theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{2.1}$$

Damit lässt sich die Posteriori-Verteilung als

$$f(\theta|x) = \frac{f(x|\theta)p(\theta)}{f(x)} \tag{2.2}$$

bestimmen, wobei die sogenannte *marginale Likelihood* $f(x)$ für stetige θ gleich $\int f(x|\theta)p(\theta)d\theta$ ist. Die Funktion $f(x|\theta)$ stellt die Likelihood $L(\theta)$ dar. Hierbei ist die marginale Likelihood $f(x)$ unabhängig von θ, respektive eine Konstante, und kann

demzufolge weggelassen werden. Die Posteriori-Verteilung reduziert sich dadurch letztendlich auf

$$f(\theta|x) \propto f(x|\theta)p(\theta). \tag{2.3}$$

(Held und Bové, 2014, S. 170; Gelman et al., 2014, S. 7)

Als geeigneter Punktschätzer kann der Posteriori-Mittelwert

$$E(\theta|x) = \int \theta f(\theta|x) d\theta, \tag{2.4}$$

sowie der Posteriori-Median und der Posteriori-Mode berechnet werden. Zur Berechnung des Erwartungswertes und anderer Kennzahlen einer Posteriori-Verteilung ist die Integration einer Funktion erforderlich, die in einigen Fällen analytisch nicht lösbar ist. In solchen Fällen können verschiedene numerische Verfahren zur Berechnung dienen. Falls die Dimension des unbekannten Parametervektors niedrig ist, bietet *Monte-Carlo-Integration* eine Alternative. Ansonsten können *Markov-Chain-Monte-Carlo-Verfahren* angewendet werden. (Held und Bové, 2014, S. 171, 247 f., 258)

2.1.1 Monte-Carlo-Integration

Unter der Annahme, dass es möglich wäre, die unabhängigen Zufallszahlen $\theta^{(1)}, \ldots, \theta^{(M)}$ aus der Posteriori-Verteilung $f(\theta|x)$ zu ziehen, lässt sich der Posteriori-Erwartungswert aus der Gleichung 2.4 folgendermaßen approximieren:

$$\hat{E}(\theta|x) = \frac{1}{M} \sum_{m=1}^{M} \theta^{(m)} \tag{2.5}$$

Dank des Gesetzes der großen Zahlen konvergiert die Schätzung zum wahren Wert für $M \to \infty$. Mit anderen Worten, für sehr viele Ziehungen sollte der approximierte Erwartungswert konsistent sein. (Held und Bové, 2014, S. 258)

2.1.2 Markov-Chain-Monte-Carlo- (MCMC-) Verfahren

Wenn es hingegen nicht möglich ist, die unabhängigen Zufallszahlen $\theta^{(1)}, \ldots, \theta^{(M)}$ aus der Posteriori-Verteilung $f(\theta|x)$ zu ziehen, kann die Simulation der *Markov-Kette* zur Anwendung kommen. Eine Reihe von Zufallsvariablen $\theta^{(1)}, \ldots, \theta^{(m)}, \ldots$ heißt Markov-Kette, wenn für jedes m die bedingte Verteilung $f(\theta^{(m)}|\theta^{(1)}, \ldots, \theta^{(m-1)})$ nur vom vorherigen Wert $\theta^{(m-1)}$ abhängt. Bei der Simulation wird θ^* aus einer *Vorschlagsdichte* $f^*(\theta|\theta^{(m-1)})$ gezogen, mit einer Wahrscheinlichkeit α akzeptiert und als neuer Zustand

$\theta^{(m)}$ eingesetzt. Bei einer Ablehnung von θ^* wird der vorherige Zustand der Kette $\theta^{(m-1)}$ erneut verwendet, sprich $\theta^{(m)} = \theta^{(m-1)}$. Das Ziel ist, durch Iterationen eine stationäre Posteriori-Verteilung zu erreichen. Das heißt, eine Markov-Kette zu erzeugen, die gegen die Posteriori-Verteilung $f(\theta|x)$ konvergiert. Dann kann der Posteriori-Erwartungswert mithilfe der gezogenen Werte $\theta^{(m)}$, wie in der Gleichung 2.5, bestimmt werden. (Gelman et al., 2014, S. 275; Held und Bové, 2014, S. 269 f.)

Beim *Metropolis-Hastings-Algorithmus* lässt sich die *Akzeptanzwahrscheinlichkeit* durch

$$\alpha = \min \left\{ 1, \underbrace{\frac{f(\theta^*|x)}{f(\theta^{(m-1)}|x)}}_{\text{Posteriori-Ratio}} \times \underbrace{\frac{f^*(\theta^{(m-1)}|\theta^*)}{f^*(\theta^*|\theta^{(m-1)})}}_{\text{Vorschlags-Ratio}} \right\} \tag{2.6}$$

bestimmen. *Metropolis-Algorithmus* und *Gibbs-Sampler* gelten als die Sonderfälle der Metropolis-Hastings-Methode. Beim Metropolis-Algorithmus ist $f^*(\theta^{(m-1)}|\theta^*) = f^*(\theta^*|\theta^{(m-1)})$. Das heißt, die Vorschlags-Ratio besitzt den Wert eins und dementsprechend reduziert sich die Akzeptanzwahrscheinlichkeit auf

$$\alpha = \min \left\{ 1, \underbrace{\frac{f(\theta^*|x)}{f(\theta^{(m-1)}|x)}}_{\text{Posteriori-Ratio}} \right\}. \tag{2.7}$$

Bei dem Gibbs-Sampler ist die Vorschlagsdichte gleich der Posteriori-Dichte, respektive $f(\theta^*|x) = f(\theta^{(m-1)}|x)$, und demzufolge gilt $\alpha = 1$. Dies ist der Fall, wenn Ziehungen aus vollständig bedingten Dichten $f(\theta_j|x, \theta_{-j})$ möglich sind, da $f(\theta_j|x, \theta_{-j}) \propto f(\theta|x)$ gilt. Eine vollständig bedingte Dichte entspricht der Dichte eines Subvektors θ_j bedingt auf alle anderen Subvektoren von θ, außer θ_j. Bei den komplexen Modellen kann es vorkommen, dass einige vollständig bedingte Dichten bekannten Verteilungen zugeordnet werden können und die anderen nicht. In dem Fall ist es möglich, den Metropolis-Algorithmus und den Gibbs-Sampler zu kombinieren, was oftmals *Metropolis-within-Gibbs-Sampler* genannt wird. (Gelman et al., 2014, S. 276 ff.; Held und Bové, 2014, S. 270)

2.1.3 Priori-Verteilung

Ein essenzieller Schritt der Bayesianischen Inferenz ist die Bestimmung der Priori-Verteilung $p(\theta)$. Das Vorwissen über den Parameter θ ist dabei selten ausreichend um

eine Verteilung zu definieren, die den unbekannten Parameter präzise beschreibt. Die mangelhaften Informationen müssen deswegen oftmals mit subjektiven Auswertungen ergänzt werden. Diese Unsicherheit motiviert die Kritiker des Bayesianischen Ansatzes, denn schließlich beeinflusst die Wahl der Priori-Verteilung die Posteriori-Inferenz. (Robert, 2007, S. 105 f.)

Wenn das Priori-Wissen über θ unzulänglich wird, lässt sich dennoch der Einfluss von der Priori-Verteilung auf die Posteriori-Inferenz durch verschiedenen Methoden kontrollieren oder unterdrücken (mehr zum Thema in Held und Bové, 2014, S. 179-191). Die Methode, die hier von Relevanz ist, entspricht der Wahl einer *nichtinformativen* Priori-Verteilung (Held und Bové, 2014, S. 183). Im Unterabschnitt 5.1.4 (Seite 50) und 5.2.4 (Seite 56) des Kapitels 5 wird im Kontext der betrachteten ökologischen und hybriden Modelle die nichtinformative Priori-Verteilung sowie das Einsetzen des Vorwissens in die Analyse durch die *informative* Priori-Verteilung weiter diskutiert.

Die hierarchischen Modelle nutzen das Prinzip der Bayesianischen Inferenz und setzen eine zusätzliche Priori-, die sogenannte *Hyperpriori-Verteilung* ein, um den unbekannten Parameter θ genauer zu bestimmen. Dabei sind mehrere bedingte Niveaus der Verteilung möglich, indem das jeweilige Niveau die unzureichenden Informationen des vorherigen Niveaus ergänzt. Der Vorteil dieses Ansatzes ist die Verbesserung der Robustheit der erzeugten Schätzer. Allerdings kann die Interpretation der Parameter und deren Beziehungen über mehrere Niveaus abstrakt und schwierig nachvollziehbar werden. Die Komplexität überträgt sich ferner auf die Berechnung der Schätzer, die lediglich mithilfe von numerischen Verfahren umsetzbar ist. (Robert, 2007, S. 113, 458, 468)

2.2 Konvergenzdiagnose

Bei MCMC Verfahren ist in erster Linie wichtig, genug Iterationen durchzuführen, um die Konvergenz, das heißt eine stationäre Posteriori-Verteilung, zu erreichen. Die Theorie liefert jedoch kein Antwort auf die Frage, wie viele Iterationen notwendig sind. Stattdessen beschreiben einige Autoren, wie man die Konvergenz erkennen und überprüfen kann und welche Probleme dabei zu beachten und zu beheben sind (mehr zum Thema in: Cowles und Carlin, 1996; Gelman und Shirley, 2011; Gelman et al., 2014, S. 281-286; Geyer, 2011, S. 17-21). Generell sind zwei gegenläufige Richtungen zu erkennen. Zum einen, ob die Konvergenz auf der Basis einer längeren Kette (Geyer,

2011) oder zum anderen, anhand mehrerer kleinerer Ketten (Gelman et al., 2014; Gelman und Shirley, 2011) festgelegt sein soll. Gemäß Cowles und Carlin (1996, S. 903) steckt die Lösung in einem Kompromiss zwischen den beiden Ansätzen.

Unabhängig davon, welches Diagnoseverfahren verwendet wird, ist Vorsicht bei den Schlussfolgerungen geboten, denn *„Diagnostics can only reliably be used to determine a lack of convergence and not detect convergence per se.“*, wie Brooks et al. (2003, in: Gelman und Shirley, 2011, S. 165) betonen. Da sich die Meinungen und die Vorgehensweisen unterscheiden, können Entscheidungen teilweise von der subjektiven Auswertung der Wissenschaftler, von der Präferenz zu einigen Verfahren und Autoren oder sogar von den technischen Eigenschaften der verfügbaren Computerausstattung abhängen. Im weiteren Verlauf werden drei relevante Begriffe, *Burn-In*, *Thinning* und *Sample* erklärt.

2.2.1 *Burn-In*

Gelman et al. (2014, S. 282) warnen, dass die Startwerte die gewünschte Verteilung der simulierten Werte beeinflussen, weshalb die Iterationen am Anfang der Kette ignoriert werden sollten. Sie empfehlen, mit einer kleinen Anzahl von Iterationen anzufangen, die erste Hälfte der Kette zu verwerfen und das Vorgehen so lang zu wiederholen, bis die Konvergenz erreicht wird. Alternativ können die vorherigen Iterationen an der Stelle abgeschnitten werden, wo die stationäre Verteilung beginnt (Gelman et al., 2014, S. 282; Held und Bové, 2014, S. 272). Mit einem *Trace Plot* lassen sich die gezogenen Simulationen gegen die Iterationen grafisch darstellen und damit kann untersucht werden, ob die Konvergenz nach der *Burn-In-Phase* visuell erreicht wird (Held und Bové, 2014, S. 272).

Geyer (2011, S. 20 f.) steht dem *Burn-In* Konzept kritisch gegenüber. Obwohl er es als ungefährlich bezeichnet, ist dies seiner Meinung nach eine unnötige Methode zur Bestimmung eines guten Startwertes. Als Alternative schlägt er vor, die nächste Kette an dem Punkt anzufangen, wo die letzte Kette beendet wurde oder wo der Modus der stationären Verteilung liegt. Der Autor argumentiert, dass die Verzerrung unwesentlich bleibt, sofern die Kette lang genug ist.

2.2.2 *Thinning*

Unabhängig davon, ob die Kette konvergiert oder nicht, sind die Ziehungen aus MCMC Verfahren nicht unabhängig, wodurch die Genauigkeit und die Effizienz der Schätzer reduziert werden (Gelman et al., 2014, S. 282; Link und Eaton, 2012, S. 112). *Thinning* ist eine übliche Methode, bei der jede k-te Ziehung berücksichtigt wird und der Rest verworfen wird, um eine Verringerung der Autokorrelation zu erzielen (Link und Eaton, 2012, S. 112). *Thinning* ist bei Modellen mit vielen Parametern ein praktisches Verfahren, wenn die Speicherkapazität des Computers begrenzt ist. Deswegen schlagen Gelman et al. (2014, S. 283) vor, k so zu wählen, dass letztendlich 1 000 Iterationen gespeichert werden.

Dennoch kritisieren Link und Eaton (2012, S. 114 f.) in ihrem Artikel *„On Thinning of Chains in MCMC“* die verbreitete Anwendung dieses Vorgehens. Sie argumentieren, dass die Approximation der Schätzer anhand von ganzen Ketten im Vergleich zu verdünnten Ketten genauer wird. Denn durch das Verdünnen gehen letztendlich viele Daten verloren. Trotzdem treten sie dem Vorgehen nicht ausschließlich kritisch gegenüber und bestätigen, dass in einigen Fällen, wie zum Beispiel bei der oben genannten begrenzten Speicherkapazität, das *Thinning* nützlich sein kann.

2.2.3 *Sample*

Nach dem Verwerfen der ersten Iterationen und der Anwendung des *Thinnings* werden die gespeicherten Werte als eine Stichprobe (Eng. *Sample*) betrachtet, die für die Berechnung des Posteriori-Erwartungswertes mittels Gleichung 2.5 verwendet wird. Der Stichprobenumfang ist somit geringer als die Anzahl der durchgeführten Iterationen. Hingegen wird der Stichprobenumfang gleich der Anzahl der Iterationen sein, falls die Schätzer ohne *Thinning* und *Burn-In* approximiert werden.

3 Ökologische Inferenz: Grundlagen und Entwicklung einiger Modelle

Die erste bekannte Verwendung der ökologischen Inferenz stammt aus dem Jahr 1919 von Wiliam Ogburn und Inez Goltra (Gow, 1985; Bulmer 1984; in: King, 1997, S. 3). Robinson (1950) kritisiert diese und andere Studien, die sich auf die ökologischen Korrelationen verlassen. Seine Warnung galt dem Unterschied zwischen der *individuellen Korrelation*, deren Variablen deskriptive Eigenschaften von Individuen darstellen und der *ökologischen Korrelation*, deren Variablen deskriptive Eigenschaften, wie Prozente oder Mittelwerte, von Gruppen abbilden (Robinson, 1950, S. 351). Gemäß dem Autor darf ein Wissenschaftler aus der ökologischen Korrelation nicht auf die individuelle Korrelation schließen, denn *„...there are a large number of individual correlations which might correspond to any given ecological correlation“* (Robinson, 1950, S. 354).

Wie können sich trotzdem Informationen über individuelles Verhalten aus Aggregatdaten gewinnen lassen? Die Suche nach der Antwort resultiert in einer Menge statistischer Verfahren, die unterschiedliche Wege zur interessierenden Schätzung anbieten. In diesem Kapitel werden im Abschnitt 3.1 die grundlegenden Modelle, die *Ökologische Regression* von Goodman (1953) und die *Methode der Ränder* von Duncan und Davis (1953), beschrieben. Im Abschnitt 3.2 befindet sich eine Darstellung des *EI* Modells von King (1997), ein Basismodell für die weitere Entwicklung der hierarchischen Modelle. Danach werden die zwei interessierenden Modelle, das *Multinomial-Dirichlet-Modell* von Rosen et al. (2001) und das *Multinomial-Log-Normal-Modell* von Greiner und Quinn (2009), erläutert. Die Erweiterung der ökologischen, hierarchischen Modelle zu *Hybridmodellen*, die mit Hilfe der Individualdaten eine Verbesserung der Schätzung erzielen können, wird anschließend im Abschnitt 3.3 dargelegt.

3.1 Grundlegende Modelle

3.1.1 Goodman: Ökologische Regression

Die erste Antwort auf das ökologische Korrelationsproblem von Robinson (1950) kam im Jahr 1953 von Goodman. In seinem Artikel *„Ecological Regressions and Behavior of Individuals“* geht er davon aus, dass die Feststellung von Robinson im Allgemeinen gilt, dennoch sollte ein Regressionsmodell möglich sein, wenn bestimmte Bedingungen erfüllt sind (Goodman, 1953, S. 663). Im Nachfolgenden wird das Anwendungsbeispiel von Goodman verändert und die Notation teilweise angepasst übernommen.

Gegeben sei eine Population, die anhand von zwei Merkmalen in einer Vierfeldertafel dargestellt werden kann. Beispielsweise lassen sich anhand von *Geschlecht* und *Berufstätigkeit* vier Gruppen aus einer Population der Arbeitsfähigen erkennen, die weiblichen Berufstätigen G_{WB}, die männlichen Berufstätigen G_{MB}, die weiblichen Arbeitslosen G_{WA} und die männlichen Arbeitslosen G_{MA}. In dem Fall definiert Goodman (1953) einen unbekannten Parameter β_1 als die durchschnittliche Wahrscheinlichkeit, dass eine weibliche Person berufstätig ist, beziehungsweise einen unbekannten Parameter β_2 als die durchschnittliche Wahrscheinlichkeit, dass eine weibliche Person arbeitslos ist (siehe Tabelle 3.1). Diese Parameter werden im Kontext der Wählerwanderung als die *Übergangswahrscheinlichkeiten* bezeichnet (Ambühl, 2003, S. 9).

	G_W	G_M	
G_B	β_1	$1-\beta_1$	X
G_A	β_2	$1-\beta_2$	$1-X$
	Y	$1-Y$	1

Tabelle 3.1: *Parameter des Ökologischen Regressionsmodells von Goodman (1953, S. 663 f.) in 2×2 Tabellenform, angepasst an die Notation in dieser Arbeit. Hinweis: Die zeilenweisen Randsummen der inneren Zellen besitzen nicht die Werte X und $1-X$, sondern eins.*

Betrachten wir eine Stichprobe i, die $g_{B,i}$ Individuen aus der Gruppe G_B und $g_{A,i}$ Individuen aus der Gruppe G_A enthält, dann wäre $X_i = g_{B,i}/(g_{B,i} + g_{A,i})$ der bekannte Anteil der Individuen aus der Gruppe G_B in der Stichprobe i und der (bekannte) erwartete Anteil der Individuen aus der Gruppe G_W wäre gleich

$$E(Y_i) = \beta_1 X_i + \beta_2(1 - X_i) \tag{3.1}$$

(Goodman, 1953, S. 664). Nach einer Umformung der Gleichung

$$
\begin{aligned}
E(Y_i) &= \beta_1 X_i + \beta_2(1 - X_i) \\
&= \beta_1 X_i + \beta_2 - \beta_2 X_i \\
&= \underbrace{\beta_2} + \underbrace{(\beta_1 - \beta_2)} X_i \\
&= \theta_0 \quad + \quad \theta_1 \cdot X_i, \qquad (3.2)
\end{aligned}
$$

$$
mit \quad \beta_2 = \theta_0 \quad und \quad \theta_1 = \beta_1 - \beta_2 = \beta_1 - \theta_0 \quad \Leftrightarrow \quad \beta_1 = \theta_0 + \theta_1,
$$

lassen sich gemäß dem Autor die unbekannten Parameter β_1 und β_2 anhand mehrerer Stichproben unverzerrt durch den kleinsten Quadrat Schätzer von θ_0 und θ_1 bestimmen (Goodman, 1953, S. 664).

Um das Identifikationsproblem bei dem Verfahren zu vermeiden und eine eindeutige Lösung zu finden (Cho und Manski, 2009, S. 9), setzt Goodman (1953, S. 664) die Annahme fest, dass die unbekannten Parameter β_1 und β_2 bei allen Stichproben konstant sind. Im Fall der Wählerwanderungsanalyse bedeutet dies, dass die Übergangswahrscheinlichkeiten für alle Wahlgebiete oder Wahlbezirke identisch sein sollen (Ambühl, 2003, S. 10). Goodman (1953) warnt, dass anhand seiner Methode Schätzwerte außerhalb des Intervalls $[0, 1]$ möglich sind. In dem Fall fordert er, die oben genannte Annahme zu überprüfen. Falls diese sich bestätigt, sollte nach seinem Vorschlag beispielsweise die negative Übergangswahrscheinlichkeit β_2 als 0 betrachtet werden. Davon ausgehend liefert die Lösung der Gleichung $E(Y_i) = X_i\beta_1$ die neue Schätzung für β_1 (Goodman, 1953, S. 664).

Obwohl die Durchführung und Interpretation des Ökologischen Regressionsmodells relativ einfach ist, werden die Annahmen gleicher Übergangswahrscheinlichkeiten in der Realität selten erfüllt (Ambühl, 2003, S. 31). Die Überprüfung dieser Annahme aus den Randsummen ist nach Ambühl (2003) vor allem nicht möglich, weswegen Goodman keine verlässliche Methode dazu bietet.

3.1.2 Ökologische Regression bei der Wählerwanderungsanalyse

Am Beispiel einer $R \times C$ Tabelle demonstrieren Klima et al. (2015, S. 3), wie sich die Methode von Goodman für die Wählerwanderungsanalyse erweitern lässt. In der Tabelle 3.2 stellen die R Zeilen die Parteien aus der ersten Wahl und die C Spalten die Parteien aus der zweiten Wahl dar. Die Ränder repräsentieren die relativen

Häufigkeiten, das heißt die Ergebnisse der ersten und der zweiten Wahl, während die β Koeffizienten die Übergangswahrscheinlichkeiten darstellen. Man kann beispielsweise $\beta_{NW,CSU}$ als die Wahrscheinlichkeit interpretieren, dass ein Wähler, der bei der ersten Wahl nicht gewählt hat, bei der zweiten Wahl seine Stimme der *CSU* gibt.

Partei	CSU_2	SPD_2	$\dots$	NW_2	
CSU_1	$\boldsymbol{\beta}_{\text{CSU,CSU}}$	$\boldsymbol{\beta}_{\text{CSU,SPD}}$	$\dots$	$\boldsymbol{\beta}_{\text{CSU,NW}}$	$P(\text{CSU}_{1,i})$
SPD_1	$\boldsymbol{\beta}_{\text{SPD,CSU}}$	$\boldsymbol{\beta}_{\text{SPD, SPD}}$	$\dots$	$\boldsymbol{\beta}_{\text{SPD, NW}}$	$P(\text{SPD}_{1,i})$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
NW_1	$\boldsymbol{\beta}_{\text{NW,CSU}}$	$\boldsymbol{\beta}_{\text{NW,SPD}}$	$\dots$	$\boldsymbol{\beta}_{\text{NW,NW}}$	$P(\text{NW}_{1,i})$
	$P(\text{CSU}_{2,i})$	$P(\text{SPD}_{2,i})$	$\dots$	$P(\text{NW}_{2,i})$	1

Tabelle 3.2: *Wahldaten zwischen zwei Wahlen für den Wahlbezirk i: Relative Häufigkeiten P und Übergangswahrscheinlichkeiten β (Klima et al., 2015, Tabelle 1, Gleichung 2).*

Um die β Koeffizienten zu schätzen, stellt man mittels einer Regression jede der C Parteien aus der zweiten Wahl ins Verhältnis zu allen Parteien aus der ersten Wahl. Zum Beispiel stellen Klima et al. (2015, S. 4) die Gleichung

$$\begin{aligned} P(\text{CSU}_{2,i}) = \beta_{\text{CSU,CSU}} \cdot P(\text{CSU}_{1,i}) \; + \; \beta_{\text{SPD,CSU}} \cdot P(\text{SPD}_{1,i}) \\ + \; \dots \; + \; \beta_{\text{NW,CSU}} \cdot P(\text{NW}_{1,i}). \end{aligned} \tag{3.3}$$

für die *CSU* auf (siehe die farbig markierte Zellen der Tabelle 3.2).

3.1.3 Duncan und Davis: Methode der Ränder

Im Jahr 1953 schlugen Duncan und Davis in ihrem Artikel *„An alternative to Ecological Correlation“* ein anderes Verfahren zur Lösung des Korrelationsproblems von Robinson vor. Ihre Idee besteht darin, die individuelle Korrelation über das kleinste Maximum und über das größte Minimum zu approximieren (Duncan und Davis, 1953, S. 666).

Für jede Stichprobe beziehungsweise für jeden Wahlkreis i lässt sich die Gleichung

$$Y_i = \beta_1^i X_i + \beta_2^i (1 - X_i) \tag{3.4}$$

(Gschwend, 2006, S. 228 f.) folgendermaßen umformulieren:

$$\Leftrightarrow \quad \beta_1^i X_i = Y_i - \beta_2^i (1 - X_i) \qquad \Leftrightarrow \quad \beta_2^i (1 - X_i) = Y_i - \beta_1^i X_i \tag{3.5}$$

$$\Leftrightarrow \quad \beta_1^i = \frac{Y_i}{X_i} - \frac{1 - X_i}{X_i} \beta_2^i \qquad \Leftrightarrow \quad \beta_2^i = \frac{Y_i}{1 - X_i} - \frac{X_i}{1 - X_i} \beta_1^i. \tag{3.6}$$

Die Übergangswahrscheinlichkeiten β_1^i und β_2^i liegen im Intervall $[0, 1]$, wobei β_1^i maximal wird, wenn $\beta_2^i = 0$ ist und minimal, wenn $\beta_2^i = 1$ ist (Gschwend, 2006, S. 230). Dementsprechend können die Grenzen von β_1^i und β_2^i aus den Gleichungen 3.6 durch

$$\beta_1^i \in \left[\max\left(0, \frac{Y_i - (1 - X_i)}{X_i}\right); \min\left(\frac{Y_i}{X_i}, 1\right) \right] \tag{3.7}$$

$$\beta_2^i \in \left[\max\left(0, \frac{Y_i - X_i}{1 - X_i}\right); \min\left(\frac{Y_i}{1 - X_i}, 1\right) \right] \tag{3.8}$$

für jede Stichprobe i bestimmt werden (Ambühl, 2003, S. 27; Cho und Manski, 2009, S. 7; Gschwend, 2006, S. 230). Weiterhin kann man die unteren und die oberen Grenzen für die gesamte Population, $\bar{\beta}_1^U$, $\bar{\beta}_1^O$, $\bar{\beta}_2^U$ und $\bar{\beta}_2^O$ durch die gewichtete Summe der unteren und der oberen Grenzen von β_1^i beziehungsweise von β_2^i folgendermaßen berechnen:

$$\bar{\beta}_1^U = \frac{1}{\sum_i N_i X_i} \sum_i \beta_1^{U,i} X_i N_i \qquad \bar{\beta}_1^O = \frac{1}{\sum_i N_i X_i} \sum_i \beta_1^{O,i} X_i N_i \tag{3.9}$$

$$\bar{\beta}_2^U = \frac{1}{\sum_i N_i (1 - X_i)} \sum_i \beta_2^{U,i} (1 - X_i) N_i \qquad \bar{\beta}_2^O = \frac{1}{\sum_i N_i (1 - X_i)} \sum_i \beta_2^{O,i} (1 - X_i) N_i \tag{3.10}$$

(Ambühl, 2003, S. 27). Dabei betonen Duncan und Davis (1953, S. 666), dass die Approximation umso genauer wird, je mehr Stichproben vorhanden sind. Konkret bedeutet dies bei der Wählerwanderungsanalyse, dass die Schätzung anhand von Wahlbezirken eine genauere Approximation liefert als die Schätzung anhand von Stadtbezirken, da ein Stadtbezirk die aggregierten Daten über mehrere Wahlbezirke beinhaltet.

Die Methode ist einfach, schränkt die Menge der möglichen Lösungen ein (Ambühl, 2003, S. 28) und setzt vor allem keine fragwürdigen Annahmen voraus (Cho und Manski, 2009, S. 7). Trotzdem wurde das Vorgehen oft kritisiert, da die geschätzten Grenzen oftmals zu breit und deswegen wenig informativ und präzise sind (Ambühl, 2003, S. 28; Klima et al., 2015, S. 3).

3.2 Entwicklung der hierarchischen Modelle

3.2.1 King: Das EI Modell

Ein neuer Ansatz von King im Jahr 1997 verbindet die beiden oben beschriebenen Methoden, die Ökologische Regression und die Methode der Ränder. Im Gegensatz zum Verfahren von Goodman (1953) findet die Schätzung der interessierenden Parameter beim EI Modell von King (1997) für jede Stichprobe i statt und erlaubt hierfür Abweichungen zwischen den Wahlkreisen. Außerdem erfolgt die Analyse auf zwei Ebenen und die Parameter β_1^i und β_2^i werden als zufällige Effekte betrachtet. Dabei werden β_1^i und β_2^i durch die Methode der Ränder für jeden Wahlbezirk beschränkt, um unrealistische Schätzungen, die außerhalb des Einheitsintervalls liegen, zu vermeiden. (King, 1997, S. 26; Gschwend, 2006, S. 230)

Ausgehend von der Gleichung 3.4 setzt King eine Trunkierte-Bivariate-Normal-Verteilung für die zufälligen Effekte β_1^i und β_2^i voraus. Das heißt, die Normal-Verteilung wird auf das Einheitsquadrat $[0,1] \times [0,1]$ reduziert. In dem ersten Schritt werden fünf Parameter μ_{β_1}, μ_{β_2}, $\sigma^2_{\beta_1}$, $\sigma_{\beta_1\beta_2}$, $\sigma_{\beta_2\beta_1}$ und $\sigma^2_{\beta_2}$ geschätzt, wobei für die Kovarianz $\sigma_{\beta_1\beta_2} = \sigma_{\beta_2\beta_1}$ gilt. Aus der geschätzten Verteilung werden im zweiten Schritt Bayesianische Simulationen durchgeführt, um die interessierenden Schätzwerte $\hat{\beta}_1^i$ und $\hat{\beta}_2^i$ zu erzeugen. (Ambühl, 2003, S. 34 f.; Gschwend, 2006, S. 231 f.)

Primäre Vorteile dieses Modells sind das Vermeiden unrealistischer Lösungen außerhalb des Einheitsintervalls (Ambühl, 2003, S. 35 f.) und die Abmilderung der selten zutreffenden Annahme der gleichen Übergangswahrscheinlichkeiten (Gschwend, 2006, S. 231). Dennoch betonen Cho und Manski (2009, S. 10), dass die Verteilungsannahme des EI Modells an sich die Annahme der ähnlichen Übergangswahrscheinlichkeiten umfasst. Deswegen unterscheidet sich diese Annahme, gemäß den Autoren, nicht wesentlich von den Annahmen des Modells von Goodman (1953). Letztendlich können die beiden Ansätze gleichermaßen zu falschen Ergebnissen führen, wenn die Annahmen nicht erfüllt sind (Cho und Manski, 2009, S. 10). Allerdings stellt das Verfahren von King (1997) einen Ausgangspunkt und die Motivation für die nachfolgenden hierarchischen Modelle dar, da die Schätzung der interessierenden Parameter auf zwei Ebenen durchgeführt wird (Klima et al., 2015, S. 5).

3.2.2 Rosen: Multinomial-Dirichlet-Modell

Aus dem vorherigen Modell entwickelten King et al. (1999) ein hierarchisches *Binomial-Beta-Modell* für 2×2 Tabellen. Nachfolgend haben Rosen et al. (2001) dieses erweitert, um die Analyse für $R \times C$ Tabellen zu ermöglichen. Das Modell wird hier mit einer allgemeinen Notation (siehe Tabelle 3.3) und anhand des Beispiels für die Analyse der Übergangswahrscheinlichkeiten zwischen zwei Wahlen (vergleiche Tabelle 3.2) präsentiert. Gegeben seien R Parteien aus der ersten Wahl und C Parteien aus der zweiten Wahl in einem Gebiet mit $i = 1, \ldots, p$ Wahlbezirke. Folglich stellen $Y_{1,i}, \ldots, Y_{C,i}$ die Anteile der Wähler dar, die in dem Wahlbezirk i die Partei c bei der zweiten Wahl gewählt haben. Analog stellen $X_{1,i}, \ldots, X_{R,i}$ die Anteile der Wähler dar, die in dem Wahlbezirk i ihre Stimme an die Partei r bei der ersten Wahl vergeben haben. Die unbekannten Parameter β_{rc}^i bezeichnen die Übergangswahrscheinlichkeiten von Partei r zur Partei c. Weiterhin werden die absoluten Häufigkeiten der zweiten Wahl als $Y_i^a = (Y_{1,i}^a, \ldots, Y_{C,i}^a)$ bezeichnet, wobei Index a für *absolut* steht.

		2. WAHL				
		$c = 1$	$c = 2$	$\ldots$	$c = C$	
1. WAHL	$r = 1$	$\boldsymbol{\beta}_{11}^i$	$\boldsymbol{\beta}_{12}^i$	$\ldots$	$1 - \sum_{c=1}^{C-1} \boldsymbol{\beta}_{1c}^i$	$X_{1,i}$
	$r = 2$	$\boldsymbol{\beta}_{21}^i$	$\boldsymbol{\beta}_{22}^i$	$\ldots$	$1 - \sum_{c=1}^{C-1} \boldsymbol{\beta}_{2c}^i$	$X_{2,i}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	$r = R$	$\boldsymbol{\beta}_{R1}^i$	$\boldsymbol{\beta}_{R2}^i$	$\ldots$	$1 - \sum_{c=1}^{C-1} \boldsymbol{\beta}_{Rc}^i$	$1 - \sum_{r=1}^{R-1} X_{r,i}$
		$Y_{1,i}$	$Y_{2,i}$	$\ldots$	$1 - \sum_{c=1}^{C-1} Y_{c,i}$	1

Tabelle 3.3: *$R \times C$ Tabelle der Wahldaten mit relativen Häufigkeiten (übernommen von Rosen et al. (2001, S. 137) und angepasst an die Wählerwanderungsanalyse und die Notation in dieser Arbeit).*

Auf der ersten Ebene des Verfahrens gehen Rosen et al. (2001) von einer Multinomial-Verteilung für Y_i^a aus (siehe Abbildung 3.1, Gleichung 3.13). Hierzu gilt

$$\theta_{c,i} = \beta_{1c}^i X_{1,i} + \beta_{2c}^i X_{2,i} + \cdots + \beta_{Rc}^i (1 - \sum_{r=1}^{R-1} X_{r,i}) = \sum_{r=1}^{R} \beta_{rc}^i X_{r,i} \qquad (3.11)$$

für jeden Parameter $\theta_{1,i}, \ldots, \theta_{C,i}$, wobei $\sum_{c=1}^{C} \theta_{c,i} = 1$ ist (Rosen et al., 2001, S. 137). Die Likelihood lässt sich bezüglich des Wahlbezirkes i folgendermaßen bestimmen:

$$\theta_{1,i}^{Y_{1,i}^a} \times \ldots \times \theta_{C-1,i}^{Y_{C-1,i}^a} \times (1 - \sum_{c=1}^{C-1} \theta_{c,i})^{N_i - \sum_{c=1}^{C-1} Y_{c,i}^a} \tag{3.12}$$

(Rosen et al., 2001, S. 137). Dabei bezeichnet $N_i = \sum_{c=1}^{C} Y_{c,i}^a$ die Anzahl aller Wähler im Wahlbezirk i. Auf der zweiten Ebene nehmen die Autoren eine unabhängige Dirichlet-Verteilung für β_r^i an (siehe Abbildung 3.1, Gleichung 3.14) und setzen die Parameter β_{rc}^i in Abhängigkeit einer Kovariable Z_i (Rosen et al., 2001, S. 137). Die Modellierung mit Kovariablen ist hier allerdings nicht von Interesse, weswegen im weiteren Verlauf ein Modell ohne Kovariablen betrachtet wird. Dieser Ansatz wurde von Lau et al. (2007, S. 46) in `R` (R Core Team, 2015) als Zusatzpaket `eiPack` (Lau et al., 2012) implementiert. Demnach ist auf der letzten Ebene eine Gamma-Hyperpriori-Verteilung für die Parameter α_{rc} angenommen (Abbildung 3.1, Gleichung 3.15), obwohl Rosen et al. (2001, S. 138) eine Exponential-Hyperpriori-Verteilung vorschlagen.

Multinomial-Dirichlet-Modell

ERSTE EBENE: $\boldsymbol{Y_i^a \sim Multinomial(N_i, \theta_i)}$ (3.13)

mit $Y_i^a = (Y_{1,i}^a, \ldots, Y_{C,i}^a)$ ▶ Anzahl der Wähler der Parteien $1, \ldots, C$

$\theta_i = (\theta_{1,i}, \ldots, \theta_{C,i})$ ▶ Anteil der Wähler der Parteien $1, \ldots, C$

$$\theta_{c,i} = \sum_{r=1}^{R} \beta_{rc}^i X_{r,i} \qquad \sum_{c=1}^{C} \theta_{c,i} = 1 \qquad E(Y_{c,i}^a) = N_i \theta_{c,i}$$

ZWEITE EBENE: $\boldsymbol{\beta_{r,i} \overset{iid}{\sim} Dirichlet(\alpha_{r1}, \ldots, \alpha_{rC})}$, für $r = 1, \ldots, R$ (3.14)

mit $\beta_{r,i} = (\beta_{r1}^i, \ldots, \beta_{rC}^i)$ ▶ Übergangswahrscheinlichkeiten

$$\sum_{c=1}^{C} \beta_{rc}^i = 1 \qquad E(\beta_{rc}^i) = \frac{\alpha_{rc}}{\sum_{c=1}^{C} \alpha_{rc}}$$

DRITTE EBENE: $\boldsymbol{\alpha_{rc} \overset{iid}{\sim} Gamma(\lambda_1, \lambda_2)}$, für $r = 1, \ldots, R,\ c = 1, \ldots, C$ (3.15)

mit $E(\alpha_{rc}) = \dfrac{\lambda_1}{\lambda_2}$

Abbildung 3.1: *Zusammenfassung der Verteilungen des ökologischen Multinomial-Dirichlet-Modells ohne Kovariablen (Rosen et al., 2001, S. 137; Lau et al., 2007, S. 46; Gelman et al., 2014, S. 576-579).*

Nach dem Bayes Theorem lässt sich eine Posteriori-Verteilung proportional zu

$$p(Daten|\beta_i, i = 1, \ldots, p) \times p(\beta_i, i = 1, \ldots, p|\alpha) \times p(\alpha) \tag{3.16}$$

$$= \prod_{i=1}^{p} \prod_{c=1}^{C} \theta_{c,i}^{Y_{c,i}^a} \tag{3.17}$$

$$\times \prod_{i=1}^{p} \prod_{r=1}^{R} \left\{ \frac{\Gamma(\sum_{c=1}^{C} \alpha_{rc})}{\prod_{c=1}^{C} \Gamma(\alpha_{rc})} \prod_{c=1}^{C} (\beta_{rc}^{i})^{\alpha_{rc}-1} \right\} \tag{3.18}$$

$$\times \prod_{r=1}^{R} \prod_{c=1}^{C} \frac{\lambda_2^{\lambda_1}}{\Gamma(\lambda_1)} \alpha_{rc}^{\lambda_1 - 1} \exp\{-\lambda_2 \alpha_{rc}\} \tag{3.19}$$

bestimmen (Rosen et al., 2001, S. 138). Da das Modell ohne Kovariablen betrachtet wird, ist die ursprüngliche Parametrisierung $d_r \exp(\gamma_{rc} + \delta_{Z_i})$ (siehe dazu Rosen et al., 2001, S. 137 f.) hier durch α_{rc} ersetzt und die Gamma- anstelle der Exponential-Verteilung dargelegt. Die Schätzung ist weder analytisch noch durch Integration möglich, weshalb die Inferenz mithilfe von Markov-Chain-Monte-Carlo-Verfahren durchgeführt wird. Die Autoren verwenden dazu einen Gibbs-Sampler. Die, für die Ziehungen benötigten, vollständig bedingten Dichten für β_{rc}^i und α_{rc} können jedoch nicht einer bekannten Verteilung zugeordnet werden, weshalb letztendlich ein Metropolis-Algorithmus angewendet wird (Rosen et al., 2001, S. 138 f.).

3.2.3 Greiner und Quinn: Multinomial-Log-Normal-Modell

Ein alternatives, hierarchisches Vorgehen für $R \times C$ Tafeln kam im Jahr 2009 von Greiner und Quinn. Gegenüber dem obigen Modell von Rosen et al. (2001), werden von den Autoren anstelle von Übergangswahrscheinlichkeiten β_{rc}^i die absoluten Häufigkeiten der inneren Zellen direkt ermittelt. Sie argumentieren, dass der Vorteil des Ansatzes die Gewichtung der Kreuztabellen einzelner Wahlbezirke proportional zur deren Größe ist. Zum Beispiel liefert eine 2×2 Tabelle mehr Information mit Randsummen von 400 und 600 als eine Tabelle mit Randsummen von 40 und 60. Hingegen werden, gemäß den Autoren, bei den Methoden, die die relative Häufigkeiten verwenden, beide Situationen gleich bewertet. Denn beiden hätten die Randsummen von 40 Prozent und 60 Prozent. Ein bedeutender Nachteil des Verfahrens, im Gegensatz zum Multinomial-Dirichlet-Modell von Rosen et al. (2001), ist die Gefahr von einem langsamen und schwerfälligen Modell-Fitting. (Greiner und Quinn, 2009, S. 68 f.; Greiner und Quinn, 2010, S. 1778 ff.)

		2. WAHL				
		$c = 1$	$c = 2$	$\dots$	$c = C$	
1. WAHL	$r = 1$	$\boldsymbol{N}^i_{11}$	$\boldsymbol{N}^i_{12}$	$\dots$	$\boldsymbol{N}^i_{1C}$	$X^a_{1,i}$
	$r = 2$	$\boldsymbol{N}^i_{21}$	$\boldsymbol{N}^i_{22}$	$\dots$	$\boldsymbol{N}^i_{2C}$	$X^a_{2,i}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	$r = R$	$\boldsymbol{N}^i_{R1}$	$\boldsymbol{N}^i_{R2}$	$\dots$	$\boldsymbol{N}^i_{RC}$	$X^a_{R,i}$
		$Y^a_{1,i}$	$Y^a_{2,i}$	$\dots$	$Y^a_{C,i}$	N_i

Tabelle 3.4: *$R \times C$ Tabelle der Wahldaten mit absoluten Häufigkeiten (übernommen von Greiner und Quinn (2009, S. 68), verallgemeinert und angepasst an die Wählerwanderungsanalyse und die Notation in dieser Arbeit).*

Um die Konsistenz bei der Notation beizubehalten (siehe Tabelle 3.4), repräsentieren nach wie vor $Y^a_{1i}, \dots, Y^a_{Ci}$ die Anzahl der Wähler, welche in dem Wahlbezirk i die Partei c bei der zweiten Wahl gewählt haben und analog stellen $X^a_{1i}, \dots, X^a_{Ri}$ die Anzahl der Wähler dar, welche in dem Wahlbezirk i ihre Stimme der Partei r bei der ersten Wahl gegeben haben. Die interessierenden unbekannten Parameter werden als N^i_{rc} bezeichnet und stellen die Anzahl der Wähler dar, die in dem Wahlbezirk i von Partei r zur Partei c gewandert sind. Die gesamte Anzahl der Wähler in einem Wahlbezirk i ist somit gleich

$$N_i = \sum_{r=1}^{R} X^a_{r,i} = \sum_{c=1}^{C} Y^a_{c,i} = \sum_{r=1}^{R} \sum_{c=1}^{C} N^i_{rc}. \tag{3.20}$$

Die Übergangswahrscheinlichkeiten können im Nachhinein einfach durch

$$\beta^i_{rc} = \frac{N^i_{rc}}{X^a_{r,i}} \tag{3.21}$$

bestimmt werden (Greiner und Quinn, 2009, S. 68; 2010, S. 1779).

Die Autoren beschreiben das Modell anhand eines Anwendungsbeispiels, bei dem die Schätzung der inneren Zellen in Abhängigkeit der Kovariable „Bevölkerungsgruppe gemäß Hautfarbe“ ermittelt werden soll (Greiner und Quinn, 2009, S. 68; 2010, S. 1775 f.). Dementsprechend stehen die nachfolgenden Annahmen des Modells ursprünglich im Verhältnis zum sogenannten *„racial block voting“* oder *„racially polarized voting“*, das einen Umstand bezeichnet, in dem die Individuen innerhalb einer Gruppe ähnliches und zwischen den Gruppen unterschiedliches Verhalten ausdrücken (Greiner und Quinn,

2010, S. 1775). Das kann sich durchaus auf den Fall der Wählerwanderung anwenden lassen, da die Wechselstimmen in der Regel die Tendenz haben, zur derjenigen Partei abzuwandern, die ähnliche Ansichten und soziale Werte vertritt, wie die zuvor gewählte Partei (Andreadis und Chadjipadelis, 2009, S. 207). In diesem Sinne gehen Greiner und Quinn (2009) davon aus, dass jeder Wähler die Wahrscheinlichkeit besitzt, eine der C Parteien bei der zweiten Wahl zu unterstützen, welche von seiner Wahlentscheidung bei der ersten Wahl und dem Wahlbezirk i abhängig ist. Sie nehmen die Randsummen für jeden Wahlbezirk i als fest an und betrachten die individuellen Wahlentscheidungen bei der zweiten Wahl unabhängig voneinander (Greiner und Quinn, 2009, S. 70).

Multinomial-Log-Normal-Modell

ERSTE EBENE:
$$(\boldsymbol{N}^i_{r1}, \ldots, \boldsymbol{N}^i_{rC}) \sim \boldsymbol{Multinom}(\boldsymbol{X}^a_{r,i}, \boldsymbol{\theta}_{r,i}) \tag{3.22}$$

mit $\theta_{r,i} = (\theta^i_{r1}, \ldots, \theta^i_{rC})$ ▶ Wahrscheinlichkeiten, die Parteien $1, \ldots, C$ zu wählen, falls bei der ersten Wahl Partei r gewählt wurde.

$$\sum_{c=1}^{C} N^i_{rc} = X^a_{r,i} \qquad \sum_{c=1}^{C} \theta^i_{rc} = 1 \qquad E(N^i_{rc}) = X^a_{r,i}\theta^i_{rc}$$

wobei $N^i_{11}, \ldots, N^i_{1C} \perp\!\!\!\perp N^i_{21}, \ldots, N^i_{2C} \perp\!\!\!\perp \ldots \perp\!\!\!\perp N^i_{R1}, \ldots, N^i_{RC}$

ZWEITE EBENE:
$$\boldsymbol{\omega}_i = (\boldsymbol{\omega}^T_{1,i}, \boldsymbol{\omega}^T_{2,i}, \ldots, \boldsymbol{\omega}^T_{R,i})^T \tag{3.23}$$

$$\sim \boldsymbol{N}_{Rx(C-1)} \left(\boldsymbol{\mu} = \begin{pmatrix} \mu^T_1 \\ \mu^T_2 \\ \ldots \\ \mu^T_R \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \Sigma_1 & \Sigma_{12} & \cdots & \Sigma_{1R} \\ \Sigma_{21} & \Sigma_2 & \cdots & \Sigma_{2R} \\ \vdots & \vdots & \vdots & \vdots \\ \Sigma_{R1} & \Sigma_{R2} & \cdots & \Sigma_R \end{pmatrix} \right)$$

mit $\omega^T_{r,i} = \left(log\left(\frac{\theta^i_{r1}}{\theta^i_{rC}}\right), \ldots, log\left(\frac{\theta^i_{rC-1}}{\theta^i_{rC}}\right) \right)$

DRITTE EBENE:
$$\boldsymbol{\mu} \sim \boldsymbol{N}(\boldsymbol{\mu}_0, \boldsymbol{K}_0) \tag{3.24}$$
$$\boldsymbol{\Sigma} \sim \boldsymbol{invWish}_{\nu_0}(\boldsymbol{\Psi}_0)$$

Abbildung 3.2: *Zusammenfassung der Verteilungen des ökologischen Multinomial-Log-Normal-Modells (Greiner und Quinn, 2009, S. 70 f.; Gelman et al., 2014, S. 576 ff.).*

Auf der ersten Ebene setzen Greiner und Quinn (2009, S. 70) eine unabhängige Multinomial-Verteilung für jede der R Zeilen voraus (siehe Abbildung 3.2, Glei-

chung 3.22). In Bezug auf die zeilenweise Unabhängigkeit lässt sich die Likelihood multiplikativ aus R multinomialverteilten Vektoren zusammensetzen. Auf der zweiten Ebene transformieren die Autoren logistisch die R, C-dimensionale, multinomiale Wahrscheinlichkeitsvektoren $\theta_{r,i}$ für jeden Wahlbezirk i. Hierfür betrachten sie die *Nichtwähler*, das heißt die Spalte C, als Referenzkategorie. Die transformierten R Vektoren $\omega_{r,i}$, jeweils mit einer reduzierten $(C-1)$ Dimension, nehmen sie als unabhängig und identisch $R(C-1)$-dimensional normalverteilt an (siehe Abbildung 3.2, Gleichung 3.23). Letztendlich setzen sie auf der letzten Ebene eine Normal- und eine Inverse-Wishart-Hyperpriori-Verteilung für die Parameter μ und Σ voraus (siehe Abbildung 3.2; Gleichung 3.24).

Um die gemeinsame Posteriori-Verteilung zu bestimmen, summieren Greiner und Quinn (2009) über die ersten $(C-1)+(R-1)$ unbekannten Zellen für jeden Wahlbezirk i und integrieren über den Parameter θ_{rc}. Die Spaltensummen betrachten sie hierbei als Funktionen der vollständigen Daten, welche unbeobachtet sind. Mit einer Matrix N_{beob}, deren i-te Zeile die Randsummen des i-ten Wahlbezirkes enthält und dem Parametervektor $\theta_i = (\theta_{1,i}^T, \theta_{2,i}^T, \ldots, \theta_{R,i}^T)$ bestimmen Greiner und Quinn (2009, S. 71 f.) die Posteriori-Verteilung folgendermaßen:

$$p(\mu, \Sigma | N_{beob}) \propto p(\mu, \Sigma) \prod_{i=1}^{p} \Bigg[\int \tag{3.25}$$

$$3.26 \sum_{N_{11}^i = UG_{N_{11}^i}}^{OG_{N_{11}^i}} \sum_{N_{12}^i = UG_{N_{12}^i}(N_{11}^i)}^{OG_{N_{12}^i}(N_{11}^i)} \cdots \sum_{N_{1C-1}^i = UG_{N_{1C-1}^i}(N_{11}^i,\ldots,N_{1C-2}^i)}^{OG_{N_{1C-1}^i}(N_{11}^i,\ldots,N_{1C-2}^i)}$$

$$\sum_{N_{21}^i = UG_{N_{21}^i}(N_{11}^i,\ldots,N_{1C-1}^i)}^{OG_{N_{21}^i}(N_{11}^i,\ldots,N_{1C-1}^i)} \cdots \cdots \sum_{N_{R-1,C-1}^i = UG_{N_{R-1,C-1}^i}(N_{11}^i,\cdots,N_{R-1,C-2}^i)}^{OG_{N_{R-1,C-1}^i}(N_{11}^i,\cdots,N_{R-1,C-2}^i)} \tag{3.26}$$

$$\binom{X_{1,i}^a}{N_{11}^i \; N_{12}^i \ldots N_{1C}^i} \binom{X_{2,i}^a}{N_{21}^i \; N_{22}^i \ldots N_{2C}^i} \cdots \binom{X_{R,i}^a}{N_{R1}^i \; N_{R2}^i \ldots N_{RC}^i} \tag{3.27}$$

$$\times \left(\theta_{11,i}^{N_{11,i}} \; \theta_{12,i}^{N_{12,i}} \ldots \theta_{1C,i}^{N_{1C,i}}\right) \left(\theta_{21,i}^{N_{21,i}} \theta_{22,i}^{N_{22,i}} \ldots \theta_{2C,i}^{N_{2C,i}}\right) \cdots$$

$$\cdot \left(\theta_{R1,i}^{N_{R1,i}} \theta_{R2,i}^{N_{R2,i}} \ldots \theta_{RC,i}^{N_{RC,i}}\right) \tag{3.28}$$

$$\times |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(\omega_i^* - \mu)^T \Sigma^{-1} (\omega_i^* - \mu)\right\} \tag{3.29}$$

$$\times\ (\theta^i_{11}\theta^i_{12}\dots\theta^i_{1C}\ \theta^i_{21}\theta^i_{22}\dots\theta^i_{2C}\ \cdots\ \theta^i_{R1}\theta^i_{R2}\dots\theta^i_{RC})^{-1} \tag{3.30}$$

$$\times\ I(N^i_{11}+N^i_{21}+\cdots+N^i_{R1}=Y^a_{1,i}) \tag{3.31}$$

$$\cdot\ I(N^i_{12}+N^i_{22}+\cdots+N^i_{R2}=Y^a_{2,i}) \tag{3.32}$$

$$\vdots$$

$$\cdot\ I(N^i_{1C}+N^i_{2C}+\cdots+N^i_{RC}=Y^a_{C,i}) \tag{3.33}$$

$$\times\ I(N^i_{11}+N^i_{12}+\cdots+N^i_{1C}=X^a_{1,i}) \tag{3.34}$$

$$\cdot\ I(N^i_{21}+N^i_{22}+\cdots+N^i_{2C}=X^a_{2,i}) \tag{3.35}$$

$$\vdots$$

$$\cdot\ I(N^i_{R1}+N^i_{R2}+\cdots+N^i_{RC}=X^a_{R,i}) \tag{3.36}$$

$$\times\ I(\theta^i_{11}+\theta^i_{12}+\cdots+\theta^i_{1C}=1) \tag{3.37}$$

$$\cdot\ I(\theta^i_{21}+\theta^i_{22}+\cdots+\theta^i_{2C}=1) \tag{3.38}$$

$$\vdots$$

$$\cdot\ I(\theta^i_{R1}+\theta^i_{R2}+\cdots+\theta^i_{RC}=1)\qquad d\theta_i\Big] \tag{3.39}$$

Die Notation wurde hierbei verallgemeinert. UG und OG in der Gleichung 3.26 kennzeichnen, dass die unteren und die oberen Grenzen bei der Summierung berücksichtigt werden. Die Notation in Klammern bedeutet, dass die Summierung jeweiliger Größe von allen vorher summierten Größen abhängt. Die Indikatorfunktionen in den Gleichungen 3.31 - 3.36 prüfen, dass die Zeilen- und Spaltensummen der unbeobachteten inneren Zellen den beobachteten Randsummen entsprechen. Die Gleichungen 3.37 - 3.39 bedingen eine zeilenweise Summierung der Parameter θ^i_{rc} auf den Wert eins. Die Schätzung der interessierenden Parameter N_{rc} findet mittels Gibbs-Sampler statt, wobei die Ränder deterministisch berücksichtigt werden (Greiner und Quinn, 2009, S. 72). Für nicht standardisierte bedingte Verteilungen (Gleichungen 3.27-3.29 und 3.36-3.38) wird der Metropolis-Hastings-Algorithmus angewendet (Greiner und Quinn, 2009, S. 80). Das Verfahren wurde von Autoren in `R` (R Core Team, 2015) als Zusatzpaket `RxCEcolInf` (Greiner et al., 2013) implementiert.

3.3 Hybridmodelle

3.3.1 Grundlage und Notation

Falls Individualdaten verfügbar sind, können diese fernerhin in die Analyse integriert werden. Die Modelle, die eine Kombination der Aggregat- und Individualdaten erfassen, werden *Hybridmodelle* genannt. Als Ergänzungen zur ökologischen Inferenz werden diese unter dem gleichen Kapitel beschrieben. Grundsätzlich gelten jedoch die Modelle, die Individualdaten in die Analyse einschließen, nicht mehr als Modelle der ökologischen Inferenz.

INDIVIDUALDATEN				AGGREGATDATEN		
	$Y=0$	$Y=1$		$Y=0$	$Y=1$	
$X=0$		$n_{0,i}$	$x^a_{0,i}$			$X^a_{0,i} - x^a_{0,i}$
$X=1$		$n_{1,i}$	$x^a_{1,i}$			$X^a_{1,i} - x^a_{1,i}$
	$n_i - y^a_i$	y^a_i	n_i	$N_i - Y^a_i - (n_i - y^a_i)$	$Y^a_i - y^a_i$	$N_i - n_i$

Tabelle 3.5: *Links:* 2×2 *Tabelle der Individualdaten mit absoluten Häufigkeiten. Rechts:* 2×2 *Tabelle der Aggregatdaten mit absoluten Häufigkeiten, adaptiert bezüglich der Individualdaten (übernommen von Wakefield (2004, S. 418) und angepasst an die Notation in dieser Arbeit).*

Ein Hybridmodell hat Wakefield im Jahr 2004 aus einem Vorschlagsverfahren für 2×2 Fälle entwickelt. Seine Notation wurde in der Tabelle 3.5 so angepasst, dass die absoluten Häufigkeiten der Individualdaten mit kleinen Buchstaben analog zu den großen Buchstaben der absoluten Häufigkeiten der Aggregatdaten bezeichnet sind. Die inneren Zellen der Tabelle der Individualdaten, $n_{0,i}$ und $n_{1,i}$, bilden an dieser Stelle die beobachteten Werte. Wakefield (2004, S. 419) adaptiert die Daten für die Analyse, indem er die Randsummen der Individualdaten von den entsprechenden Randsummen der Aggregatdaten subtrahiert. Er berichtet, dass die Verbesserung der Analyse bereits durch kleine Stichproben erreicht werden kann. Dennoch warnt er, dass die Stichproben repräsentativ sein sollen. Besonders anfällig für Verzerrungen können Nachwahlbefragungen sein, da die Befragten nicht immer ehrlich über ihre politischen Ansichten antworten (Wakefield, 2004, 420 f.).

Sein Verfahren hat andere Methodiker motiviert, die Individualdaten in ihre Methoden für $R \times C$ Fälle zu integrieren. Eine Erweiterung des Hybridmodells von Wakefield

(2004) auf das Multinomial-Dirichlet-Modell für $R \times C$ Fälle von Rosen et al. (2001) hat Schlesinger im Jahr 2013 in seiner Masterarbeit begründet. Schließlich implementierte er das Modell in `R` (R Core Team, 2015) als Zusatzpaket `eiwild` - *Ecological Inference with Individual level Data* (Schlesinger, 2014). Das Multinomial-Log-Normal-Modell von Greiner und Quinn (2009, 2010) geht bereits von einem *Individual-Level* aus und ermöglicht damit eine einfache Ergänzung des Modells durch die Individualdaten. Die hybride Version ihres Ansatzes ist im `RxCEcolInf` Paket (Greiner et al., 2013) in `R` (R Core Team, 2015) integriert.

		2. WAHL				
		$c = 1$	$c = 2$	$\dots$	$c = C$	
1. WAHL	$r = 1$	$\boldsymbol{n}^i_{11}$	$\boldsymbol{n}^i_{12}$	$\dots$	$\boldsymbol{n}^i_{1C}$	$x^a_{1,i}$
	$r = 2$	$\boldsymbol{n}^i_{21}$	$\boldsymbol{n}^i_{22}$	$\dots$	$\boldsymbol{n}^i_{2C}$	$x^a_{2,i}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	$r = R$	$\boldsymbol{n}^i_{R1}$	$\boldsymbol{n}^i_{R2}$	$\dots$	$\boldsymbol{n}^i_{RC}$	$x^a_{R,i}$
		$y^a_{1,i}$	$y^a_{2,i}$	$\dots$	$y^a_{C,i}$	n_i

Tabelle 3.6: *$R \times C$ Tabelle der Individualwahldaten mit absoluten Häufigkeiten (übernommen von Schlesinger (2013, S. 34) und angepasst an die Notation in dieser Arbeit).*

		2. WAHL				
		$c = 1$	$c = 2$	$\dots$	$c = C$	
1. WAHL	$r = 1$	$\boldsymbol{\beta}^i_{11}$	$\boldsymbol{\beta}^i_{12}$	$\dots$	$1 - \sum_{c=1}^{C-1} \boldsymbol{\beta}^i_{1c}$	$X^a_{1,i} - x^a_{1,i}$
	$r = 2$	$\boldsymbol{\beta}^i_{21}$	$\boldsymbol{\beta}^i_{22}$	$\dots$	$1 - \sum_{c=1}^{C-1} \boldsymbol{\beta}^i_{2c}$	$X^a_{2,i} - x^a_{2,i}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	$r = R$	$\boldsymbol{\beta}^i_{R1}$	$\boldsymbol{\beta}^i_{R2}$	$\dots$	$1 - \sum_{c=1}^{C-1} \boldsymbol{\beta}^i_{Rc}$	$X^a_{R,i} - x^a_{R,i}$
		$Y^a_{1,i} - y^a_{1,i}$	$Y^a_{2,i} - y^a_{2,i}$	$\dots$	$Y^a_{C,i} - y^a_{C,i}$	$N_i - n_i$

Tabelle 3.7: *$R \times C$ Tabelle der Aggregatwahldaten mit absoluten Häufigkeiten, die bezüglich der Individualwahldaten adaptiert werden (übernommen von Schlesinger (2013, S. 35) und angepasst an die Notation in dieser Arbeit).*

Zugunsten des Multinomial-Dirichlet-Hybridmodells bestätigt eine Simulationsstu-

die von Klima et al. (2016) die Ergebnisse von Wakefield (2004). Es hat sich wiederkehrend gezeigt, dass eine kleine Stichprobe die Schätzung verbessern kann. Um deren Hybridmodell zu evaluieren, haben Greiner und Quinn (2010) ebenfalls eine Simulationsstudie durchgeführt und kamen zu gleichen Ergebnissen. Sie schlagen vor, das Hybridmodell immer vor den reinen ökologischen Modellen zu bevorzugen, wenn Individualdaten vorhanden sind (Greiner und Quinn, 2010, S. 1785). Vorsicht ist geboten, falls bei der Stichprobe eine Bevölkerungsgruppe dominiert, warnen die Autoren. In Bezug auf die Wählerwanderungsanalyse sollten die Wähler einer Partei der ersten Wahl nicht eine Mehrheit der Befragten ausmachen. In den folgenden Unterabschnitten werden die beiden Hybridmodelle und ihre Annahmen kurz beschrieben.

3.3.2 Multinomial-Dirichlet-Hybridmodell

Schlesinger (2013) beschreibt das Verfahren gleichermaßen auf dem Anwendungsbeispiel der Wählerwanderungsanalyse zwischen zwei Wahlen, jedoch folgt er annähernd der Notation von Wakefield (2004), die hier wiederum angeglichen wird. Für den Fall, dass die Daten einer Nachwahlbefragung zur Verfügung stehen, erweitert er die Tabelle 3.5 (links) auf eine $R \times C$ Tabelle, um die unbekannten Übergangswahrscheinlichkeiten β_{rc} zu schätzen (siehe Tabelle 3.6). Er nimmt an, dass n_{rc}^i für einige Wahlbezirke aus den Daten bekannt sind, wobei $(0 \leq i \leq p)$, $(0 \leq y_{c,i}^a \leq Y_{c,i}^a)$, $(0 \leq x_{r,i}^a \leq X_{r,i}^a)$ und $n_{rc}^i \in [0, \min\{y_{c,i}^a, x_{r,i}^a\}]$ gilt (Schlesinger, 2013, S. 34).

Entsprechend der zeilenweisen Binomial-Verteilung beim Ansatz von Wakefield (2004), setzt Schlesinger (2013, S. 34 f.) auf der ersten Ebene des Verfahrens eine unabhängige zeilenweise Multinomial-Verteilung für $n_{r1}^i, \ldots, n_{rC}^i$ voraus. Deren Parameter sind $\beta_{r1}^i, \ldots, \beta_{rC}^i$ und die Zeilensummen $x_{r,i}^a$, wobei nach wie vor $\sum_{c=1}^{C} \beta_{rc}^i = 1$ gilt (siehe die Abbildung 3.3, Gleichung 3.40). Hierbei nimmt er an, dass Individual- und Aggregatdaten die gleichen Übergangswahrscheinlichkeiten β_{rc}^i ergeben. Die Randsummen der Aggregatdaten werden, wie im Unterabschnitt 3.3.1 (Tabelle 3.5), gemäß den Informationen aus den Individualdaten je nach Wahlbezirk i angepasst (siehe Tabelle 3.7).

Alle drei Ebenen der Aggregatdaten folgen den gleichen Verteilungen wie das Multinomial-Dirichlet-Modell von Rosen et al. (2001), wobei auf der ersten Ebene die Spaltensummen modifiziert werden. Demnach definiert Schlesinger (2013, S. 35) an die-

ser Stelle eine Multinomial-Verteilung für $Y^a_{1,i} - y^a_{1,i}, \ldots, Y^a_{C,i} - y^a_{C,i}$ mit den Parametern $N_i - n_i$ und $\theta_{1,i}, \ldots, \theta_{C,i}$ (siehe Abbildung 3.3, Gleichung 3.41). Falls Vorwissen, wie zum Beispiel Wahlempfehlungen, vorhanden ist, kann dieses auf der dritten Ebene durch die *zellspezifischen* Hyperpriori-Verteilungen für α_{rc} verwendet werden (Schlesinger, 2013).

Multinomial-Dirichlet-Hybridmodell

INDIVIDUALDATEN

ERSTE EBENE:
$$(\boldsymbol{n}^i_{r1}, \ldots, \boldsymbol{n}^i_{rC}) \sim \boldsymbol{Multinomial}(\boldsymbol{x}^a_{r,i};\ \boldsymbol{\beta}^i_{r1}, \ldots \boldsymbol{\beta}^i_{rC}) \tag{3.40}$$

mit:
$$\sum_{c=1}^{C} \beta^i_{rc} = 1, \quad r = 1, \ldots, R$$

AGGREGATDATEN

ERSTE EBENE:
$$(\boldsymbol{Y}^a_{1,i} - \boldsymbol{y}^a_{1,i}, \ldots, \boldsymbol{Y}^a_{C,i} - \boldsymbol{y}^a_{C,i}) \sim \boldsymbol{Multinomial}(\boldsymbol{N}_i - \boldsymbol{n}_i;\ \boldsymbol{\theta}_{1,i}, \ldots, \boldsymbol{\theta}_{C,i}) \tag{3.41}$$

mit:
$$\sum_{c=1}^{C} \theta_{c,i} = 1, \quad \theta_{c,i} = \sum_{r=1}^{R} \beta^i_{rc} X_{r,i}, \quad X_{r,i} = \frac{X^a_{r,i} - x^a_{r,i}}{N_i - n_i}$$

ZWEITE EBENE:
$$(\boldsymbol{\beta}^i_{r1}, \ldots \boldsymbol{\beta}^i_{rC}) \sim \boldsymbol{Dirichlet}(\boldsymbol{\alpha}_{r1}, \ldots \boldsymbol{\alpha}_{rC}) \tag{3.42}$$

DRITTE EBENE:
$$\boldsymbol{\alpha}_{rc} \sim \boldsymbol{Gamma}(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) \tag{3.43}$$

oder zellspezifisch:
$$\boldsymbol{\alpha}_{rc} \sim \boldsymbol{Gamma}(\boldsymbol{\lambda}^{rc}_1, \boldsymbol{\lambda}^{rc}_2) \tag{3.44}$$

Abbildung 3.3: *Zusammenfassung der Verteilungen des Multinomial-Dirichlet-Hybridmodells (Schlesinger, 2013, S. 35 f.).*

Die gemeinsame Posteriori-Verteilung erweitert sich bei der Verwendung von Individualdaten und ergibt sich schließlich durch

$$f(\beta^i_{rc}, \alpha_{rc} | X_{r,i}, Y^a_{c,i}, n^i_{rc}, y^a_{c,i}, (\lambda_1, \lambda_2)) \propto \tag{3.45}$$

$$\times \left.\prod_{i=1}^{p} \prod_{r=1}^{R} \prod_{c=1}^{C} (\beta^i_{rc})^{n^i_{rc}} \right] \text{INDIVIDUALDATEN} \tag{3.46}$$

$$\times \prod_{i=1}^{p} \prod_{c=1}^{C} (\theta_{c,i})^{Y^a_{c,i} - y^a_{c,i}} \tag{3.47}$$

$$\times \prod_{i=1}^{p} \prod_{r=1}^{R} \left\{ \frac{\Gamma(\sum_{c=1}^{C} \alpha_{rc})}{\prod_{c=1}^{C} \Gamma(\alpha_{rc})} \prod_{c=1}^{C} (\beta^i_{rc})^{\alpha_{rc} - 1} \right\} \tag{3.48}$$

$$\times \prod_{r=1}^{R} \prod_{c=1}^{C} \frac{\lambda_2^{\lambda_1}}{\Gamma(\lambda_1)} \alpha_{rc}^{\lambda_1 - 1} \exp\{-\lambda_2 \alpha_{rc}\}. \tag{3.49}$$

Bei denjenigen Wahlbezirken i, für die keine Individualdaten vorhanden sind, reduziert sich die Posteriori-Verteilung auf die ohne Individualdaten (siehe Unterabschnitt 3.2.2, Gleichungen 3.17-3.19), da n_{rc}^i in der Gleichung 3.46 und $y_{c,i}^a$ in der Gleichung 3.47 in diesem Fall gleich null sind. Zur Schätzung werden Markov-Chain-Monte-Carlo-Methoden, speziell Metropolis-within-Gibbs-Sampler, durchgeführt. (Schlesinger, 2013, S. 36)

3.3.3 Multinomial-Log-Normal-Hybridmodell

Gemäß Greiner und Quinn (2009, 2010) impliziert die Annahme der festen Randsummen für jeden Wahlbezirk i, dass die Randsummen unabhängig vom Prozess der Datensammlung sind. Hierbei nehmen sie die Wahrscheinlichkeit, eine der C Parteien bei der zweiten Wahl zu unterstützen, abhängig von der Wahlentscheidung bei der ersten Wahl und dem Wahlbezirk i an. Die individuellen Wahlentscheidungen bei der zweiten Wahl betrachten sie als unabhängig voneinander. Diese Annahmen resultieren in den unabhängigen Multinomial-Verteilungen für die R Zeilen der unbekannten inneren Zellen bei den Aggregatdaten, die dem individuellen Wahlverhalten entsprechen (Greiner und Quinn, 2009, S. 70; Greiner und Quinn, 2010, S. 1781). Demnach können die Individualdaten ohne zusätzliche Annahmen ins Modell integriert werden.

Für eine Stichprobe S, die s aus p Wahlbezirken enthält, erweitern Greiner und Quinn (2009, S. 78) die Posteriori-Verteilung aus der Gleichungen 3.25 - 3.39, durch die Likelihood:

$$\left\{ \binom{N_i}{n_i}^{-1} \binom{n_i}{x_{1,i}^a \; x_{2,i}^a \; \cdots \; x_{R,i}^a} X_{1,i}^{x_{1,i}^a} X_{2,i}^{x_{2,i}^a} \cdots X_{R,i}^{x_{R,i}^a} \right. \tag{3.50}$$

$$\times \binom{x_{1,i}^a}{n_{11,i} \; n_{12,i} \; \cdots \; n_{1C,i}} \theta_{11}^{n_{11}} \theta_{12}^{n_{12}} \cdots \theta_{1C}^{n_{1C}} \tag{3.51}$$

$$\times \binom{x_{2,i}^a}{n_{21,i} \; n_{22,i} \; \cdots \; n_{2C,i}} \theta_{21}^{n_{21}} \theta_{22}^{n_{22}} \cdots \theta_{2C}^{n_{2C}} \tag{3.52}$$

$$\vdots$$

$$\times \binom{x^a_{R,i}}{n_{R1,i}\ n_{R2,i}\ \cdots\ n_{RC,i}} \theta_{R1}^{n_{R1}} \theta_{R2}^{n_{R2}} \cdots \theta_{RC}^{n_{RC}} \Bigg\}^{I(i \in S)} \tag{3.53}$$

Die Notation wurde hier bezüglich der Tabelle 3.6 angepasst. Der Indikator $i \in S$ weist darauf hin, dass ausschließlich die Wahlbezirke betrachtet werden, die in der Stichprobe vorhanden sind. In der Posteriori-Verteilung, bedingt auf den beobachteten Individualdaten, werden N^i_{rc} Parameter für jedes $i \in S$ adaptiert (Greiner und Quinn, 2009). Eine genauere Beschreibung wird im Jahr 2009 dennoch nicht gegeben. Ferner leiten sie an, dass nach dieser Ergänzung aus der gemeinsamen Posteriori-Verteilung die interessierenden Schätzwerte, in gleicher Weise wie bei der ökologischen Inferenz, durch Metropolis-within-Gibbs-Sampler gezogen werden können (Greiner und Quinn, 2009, S. 78). Im Jahr 2010 definieren Autoren die gemeinsame Posteriori-Verteilung des Hybridmodells in einer reduzierten Form proportional zu:

$$N(\mu|\mu_0, \kappa_0) \times Inv - Wish_{\nu_0}(\sigma|\Psi_0) \tag{3.54}$$

$$\times \prod_i \Bigg[\int \left(\sum_{M^i_{unbeob}} \left(\prod_{r,c} \frac{\theta^{i\ M^i_{rc}}_{rc}}{M^i_{rc}!} \right) \right)^{i \in S} \left(\sum_{N^i_{unbeob}} \left(\prod_{r,c} \frac{\theta^{i\ N^i_{rc}}_{rc}}{N^i_{rc}!} \right) \right)^{i \notin S} \tag{3.55}$$

$$\times \left(|\Sigma|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} (\omega_i - \mu)^T \Sigma^{-1} (\omega_i - \mu) \right\} \right) d\theta_i \Bigg] \tag{3.56}$$

(Greiner und Quinn, 2010, S. 1782). Hierbei definieren sie M^i_{rc} durch $N^i_{rc} - n^i_{rc}$. Die Gleichung 3.54 stellt die Hyperpriori-Verteilungen dar. Die weitere Erläuterung dieser Posteriori-Verteilung ist jedoch unklar, da die Autoren vier Zeilen beschreiben und nur drei darstellen. Die zweite und die dritte Zeile bezeichnen sie als Multinomial-Verteilung der inneren Zellen und die vierte Zeile als Multivariate-Normal-Verteilung. Es ist klar, dass sich die Beschreibung der vierten Zeile auf die Priori-Verteilung in der Gleichung 3.56 bezieht. Bei der Darstellung der Multinomial-Verteilung scheint eine Zeile zu fehlen. Demnach bleibt es unklar, wie genau die Posteriori-Verteilung durch die Individualdaten ergänzt wird.

4 Die Datenbasis

4.1 Datengrundlage und deskriptive Analyse

In diesem Abschnitt werden der Inhalt und der Ursprung der Aggregat- und Individualdaten beschrieben. Vor Beginn der Analyse liefern grafische Darstellungen einen ersten Überblick über das Wahlverhalten. Zur Erstellung der Diagramme kommen die folgenden R-Pakete zum Einsatz: `ggplot2` (Wickham, 2009), `ggthemes` (Arnold, 2016), `scales` (Wickham, 2016), `gridExtra` (Auguie, 2016) und `circlize` (Gu et al., 2014; Gu, 2015).

4.1.1 Amtliche Ergebnisse der betrachteten Wahlen

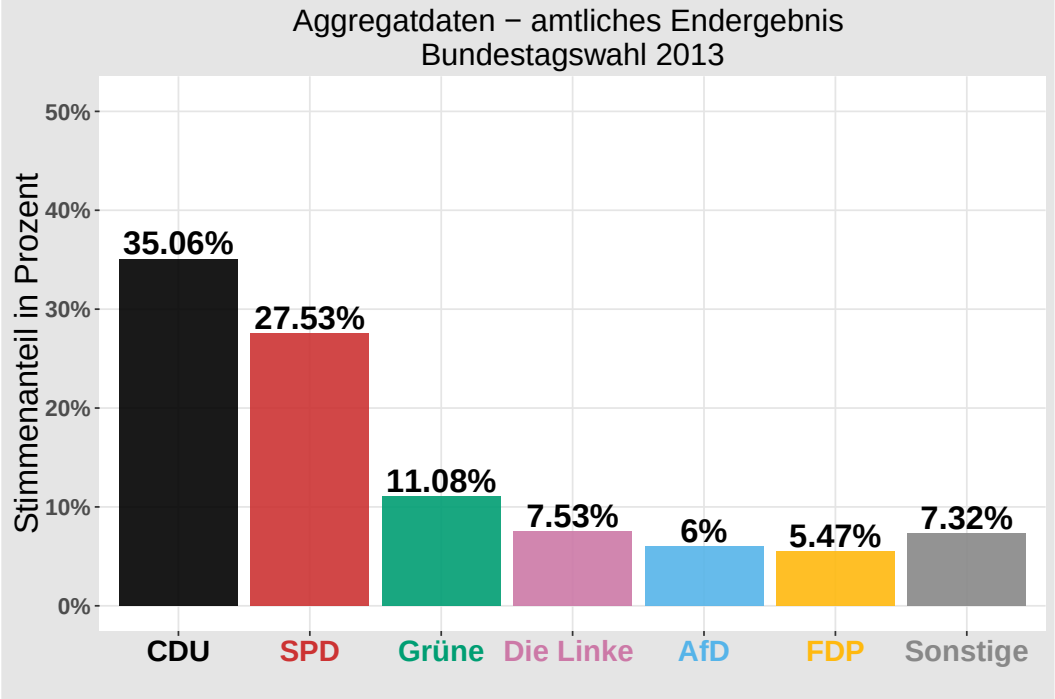

Abbildung 4.1: *Amtliches Endergebnis der Bundestagswahl im Jahr 2013. Quelle: Stadt Mannheim (2013).*

Die amtlichen Endergebnisse der **Bundestagswahl** im Jahr **2013** lassen sich von der offiziellen Internetseite der Stadt Mannheim (2013) herunterladen. In der Datei

`btw_wahlbezirke_neu` sind die Daten über die Erst- und Zweitstimmen der Bundestagswahl im `.xls` Format enthalten, wobei nur die Zweitstimmen in der Analyse betrachtet werden. Es sind die Ergebnisse der 137 796 Wähler von insgesamt 198 525 Wahlberechtigten für 20 Parteien zur Verfügung gestellt. Die Ergebnisse aller 135 744 gültigen Stimmen sind in der Abbildung 4.1 dargestellt. Auf der x-Achse liegen die sechs größten Parteien: *CDU*, *SPD*, *Grüne*, *Die Linke*, *AfD* und *FDP*, wie auch die Kategorie *Sonstige*, die alle kleinen Parteien umfasst. Die y-Achse zeigt wie viele Stimmen in Prozent die jeweilige Partei gewonnen hat. Mit 35.06 Prozent erreichte die *CDU* damals eine Mehrheit der Stimmen, während die *SPD* 27.53 Prozent erzielte. Darauf folgen die *Grünen* mit 11.08, *Die Linke* mit 7.53, *AfD* mit 6, *FDP* mit 5.46 und alle andere Parteien, die im Ganzen 7.32 Prozent erhielten. Das Prinzip und der Grund für die Zusammenfassung der kleinen Parteien in eine Kategorie wird in dem Unterabschnitt 4.2.1 auf der Seite 40 diskutiert. Im Anhang A.1.1 auf der Seite 85 ist die Liste aller Parteien zu finden.

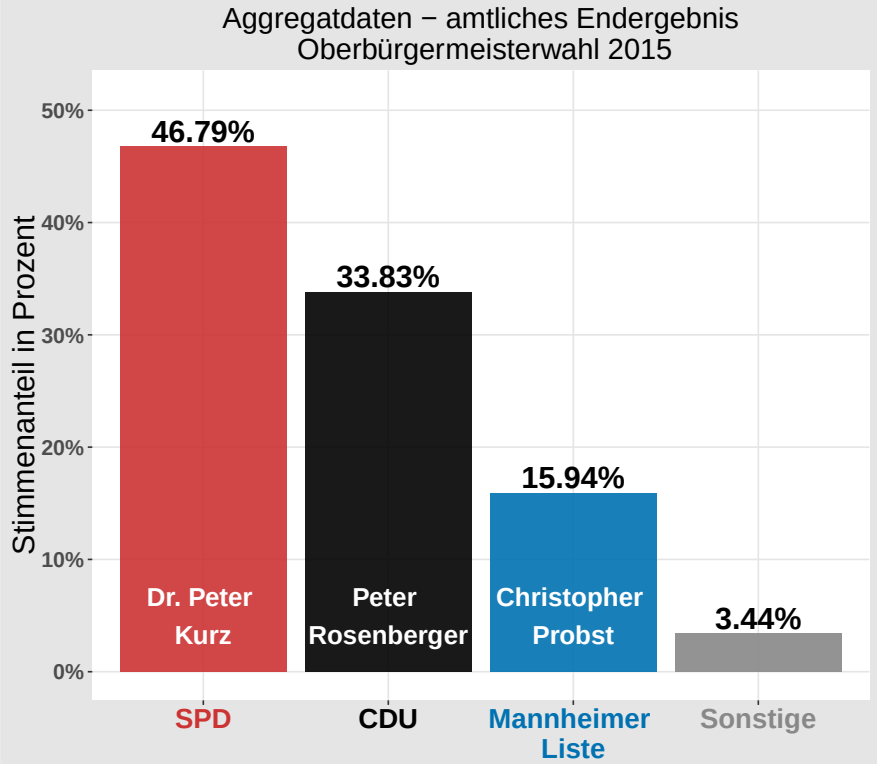

Abbildung 4.2: *Amtliches Endergebnis der Oberbürgermeisterwahl im Jahr 2015. Quelle: Stadt Mannheim (2015b).*

Amtliche Ergebnisse der **Oberbürgermeisterwahl** aus dem Jahr **2015** (Stadt Mannheim, 2015b) sind für den ersten Wahlkreis im Juni sowie für den zweiten Wahlkreis im Juli verfügbar. Hier wird die Datei

`obw2015_auswertungen_amtliches_endergebnis_fur_internet.xls` vom Juni für die Analyse verwendet, da die Nachwahlbefragung bereits im Juni im ersten Wahlkreis durchgeführt wurde. Obwohl die Anzahl der Wahlberechtigten im Jahr 2015 um 35 556 höher war als im Jahr 2013, sank 2015 die Anzahl der Wähler auf 71 866 und damit die Wahlbeteiligung von 69.4 Prozent auf nur 30.7 Prozent. Für den ersten Wahlkreis sind die Ergebnisse der vier stärksten Kandidaten, *Dr. Peter Kurz (SPD)*, *Christopher Probst (Mannheimer Liste)*, *Peter Rosenberger (CDU)* und *Christian Sommer (Die Partei)* freigegeben. Die Stimmen der anderen Kandidaten wurden unter der Kategorie *Andere Gewählte* bereits bei den amtlichen Ergebnissen zusammengezählt. Im weiteren Verlauf wird diese Kategorie als *Sonstige* bezeichnet, um die Darstellungen der beiden Wahlen abzugleichen. Der Kandidat *Christian Sommer (Die Partei)* wurde dieser Kategorie zugeteilt (siehe Unterabschnitt 4.2.1). Wie die gültigen Stimmen verteilt wurden zeigt die Abbildung 4.2. Auf der x-Achse sind die Kandidaten abgebildet und auf der y-Achse der Stimmenanteil der jeweiligen Kandidaten in Prozent. „Amtsinhaber Dr. Peter Kurz wird von den Mannheimer Kreisverbänden der SPD, der Grünen und der Linken unterstützt, Peter Rosenberger von der CDU und Christopher Probst von der Mannheimer Liste“ (Schredle, 2015). Da die *SPD*, die *Grünen* und *Die Linke* bei der Bundestagswahl 2013 insgesamt 46.14 Prozent erhielten, scheint der Gewinn von *Dr. Peter Kurz* mit 46.79 Prozent erwartungsgemäß. *Peter Rosenberger* erzielte 33.83 Prozent, genau 1.23 Prozent weniger als die *CDU* zwei Jahre vorher. Kandidat der *Mannheimer Liste*, *Christopher Probst*, bekam 15.94 Prozent der Stimmen und alle anderen Kandidaten sammelten insgesamt 3.44 Prozent aller Stimmen.

Das Stadtgebiet Mannheim besteht aus 17 Bezirken (siehe dazu Anhang A.1.2, Seite 86), die bei einer Wahl in Wahlgebäude und Wahlbezirke unterteilt werden. Das Prinzip der Zuordnung von Straßen zu den Wahlbezirken und Wahlgebäuden kann sich von Wahl zur Wahl ändern. Bei der Bundestagswahl im Jahr 2013 wurden hierfür 52 Wahlgebäude und 189 Wahlbezirke erstellt und bei der Oberbürgermeisterwahl im Jahr 2015 wurde das Stadtgebiet in 68 Wahlgebäude und 123 Wahlbezirke unterteilt. Eine Übersicht der elementaren Zahlen zum Populationsumfang und zur Bezirksunterteilung ist für das gesamte Stadtgebiet sowie für die Brief- und Urnenwähler getrennt, in der Tabelle 4.1 dargelegt.

BUNDESTAGSWAHL 2013			
	Gesamt	Urnenwähler	Briefwähler
Summe Wahlberechtigte	198 525		
Wahlberechtigte ohne Wahlschein	157 474		
Wahlberechtigte mit Wahlschein	41 051		
Wähler insgesamt	137 796	100 299	37 497
darunter mit Wahlschein	37 984	487	37 497
Ungültige Stimmen	2 052	1 706	346
Gültige Stimmen	135 744	98 593	37 151
Anzahl Stadtbezirke	17	17	17
Anzahl Wahlgebäude	52	51	1
Anzahl Wahlbezirke	189	150	39
OBERBÜRGERMEISTERWAHL 2015			
	Gesamt	Urnenwähler	Briefwähler
Summe Wahlberechtigte	234 081		
Wahlberechtigte ohne Wahlschein	210 953		
Wahlberechtigte mit Wahlschein	23 128		
Wähler insgesamt	71 866	50 995	20 871
darunter mit Wahlschein	21 110	239	20 871
Ungültige Stimmen	641	474	167
Gültige Stimmen	71 225	50 521	20 704
Anzahl Stadtbezirke	17	17	17
Anzahl Wahlgebäude	68	51	17
Anzahl Wahlbezirke	123	96	27

Tabelle 4.1: *Übersicht der wichtigsten Zahlen zum Populationsumfang und zur Bezirksunterteilung bei der Bundestagswahl 2013 (oben) und der Oberbürgermeisterwahl 2015 (unten).*

4.1.2 Nachwahlbefragung

Die Nachwahlbefragung (Felderer, 2015, persönliche Kommunikation) wurde im Rahmen der Lehrveranstaltung *Empirisches Forschungspraktikum* bei der Mannheimer Oberbürgermeisterwahl im Juni 2015 für fünf Wahlbezirke durchgeführt. Unter anderen sollten die Befragten zwei Fragen beantworten, die hier von Interesse sind: *„Für welchen Kandidaten haben Sie heute gestimmt?“*, *„Wenn Sie nun an die letzte Bundestagswahl im September 2013 denken: Welche Partei haben Sie damals mit Ihrer Zweitstimme gewählt?“*. Diese Daten ermöglichen einen ersten Einblick in das Individualwahlverhalten der Wähler.

Von insgesamt 1 575 Teilnehmern verweigerten 100 eine oder beide Fragen, 8 gaben

Fehlende Werte bei der Nachwahlbefragung

Angekreuzt bei der Bundestagswahl

Fehlende Werte Oberbürgermeisterwahl

	SPD	Grüne	Die Linke	CDU	AfD	FDP	Sonstige	Nichtwähler
Anzahl	2	2	0	6	0	2	0	1
Anteil	0.59 %	0.64 %	0 %	1.5 %	0 %	1.83 %	0 %	1.06 %

Angekreuzt bei der Oberbürgermeisterwahl

Fehlende Werte Bundestagswahl

	Dr. Peter Kurz	Peter Rosenberger	Christopher Probst	Sonstige	Nichtwähler
Anzahl	60	46	15	4	2
Anteil	8.06 %	9.09 %	8.06 %	5.71 %	18.18 %

Abbildung 4.3: *Fehlende Werte bei der Nachwahlbefragung. Anzahl bezeichnet die Häufigkeit der fehlenden Werte bei einer Wahl in Abhängigkeit von den Angaben bei der anderen Wahl. Der Anteil stellt das Verhältnis von Anzahl der fehlenden Werte bei einer Wahl zur Summe aller Angaben für die jeweilige Partei oder für den jeweiligen Kandidaten bei der anderen Wahl in Prozent dar.*

mehr als eine Antwort und 79 konnten sich nicht mehr erinnern, wem sie ihre Stimme im Jahr 2013 gegeben haben. Alle genannten Fälle wurden als fehlende Werte betrachtet und gelöscht, wodurch 185 Beobachtungen verloren gehen. In der Abbildung 4.3 lässt sich erkennen, dass deutlich mehr Werte bei der Angabe zur Bundestagswahl 2013 fehlen. Obwohl eine Mehrheit bei der Frage zur Oberbürgermeisterwahl (2015) den Kandidat *Dr. Peter Kurz* angekreuzt haben, zeigt das Verhältnis von Anzahl der fehlenden Werte zur Summe aller Angaben für die jeweiligen Kandidat, dass sich die Wähler aller drei großen Kandidaten angenähert gleichmäßig über ihre Wahl im Jahr 2013 nicht geäußert haben. Insgesamt lehnten 45 Probanden ab, eine Antwort auf beide Fragen zu geben. In der Abbildung 4.3 sind diese nicht dargestellt. Neben dem Problem der fehlenden Werte, muss noch ergänzt werden, dass der Anteil der *Nichtwähler* bei der Befragung, die nach der Wahl vor Ort stattfand, höchstwahrscheinlich nicht dem wirklichen Zustand der *Nichtwähler* entspricht. Denn, die Wahlberechtigten, die nicht gewählt haben, treten in der Regel auch nicht am Wahlort auf. Außerdem werden die Briefwähler bei der Befragung nicht betrachtet. Ferner wird im Unterabschnitt 4.2.4 (Seite 43) erklärt, warum das Ignorieren der Briefwähler bei der Wählerwanderungsanalyse problematisch sein kann.

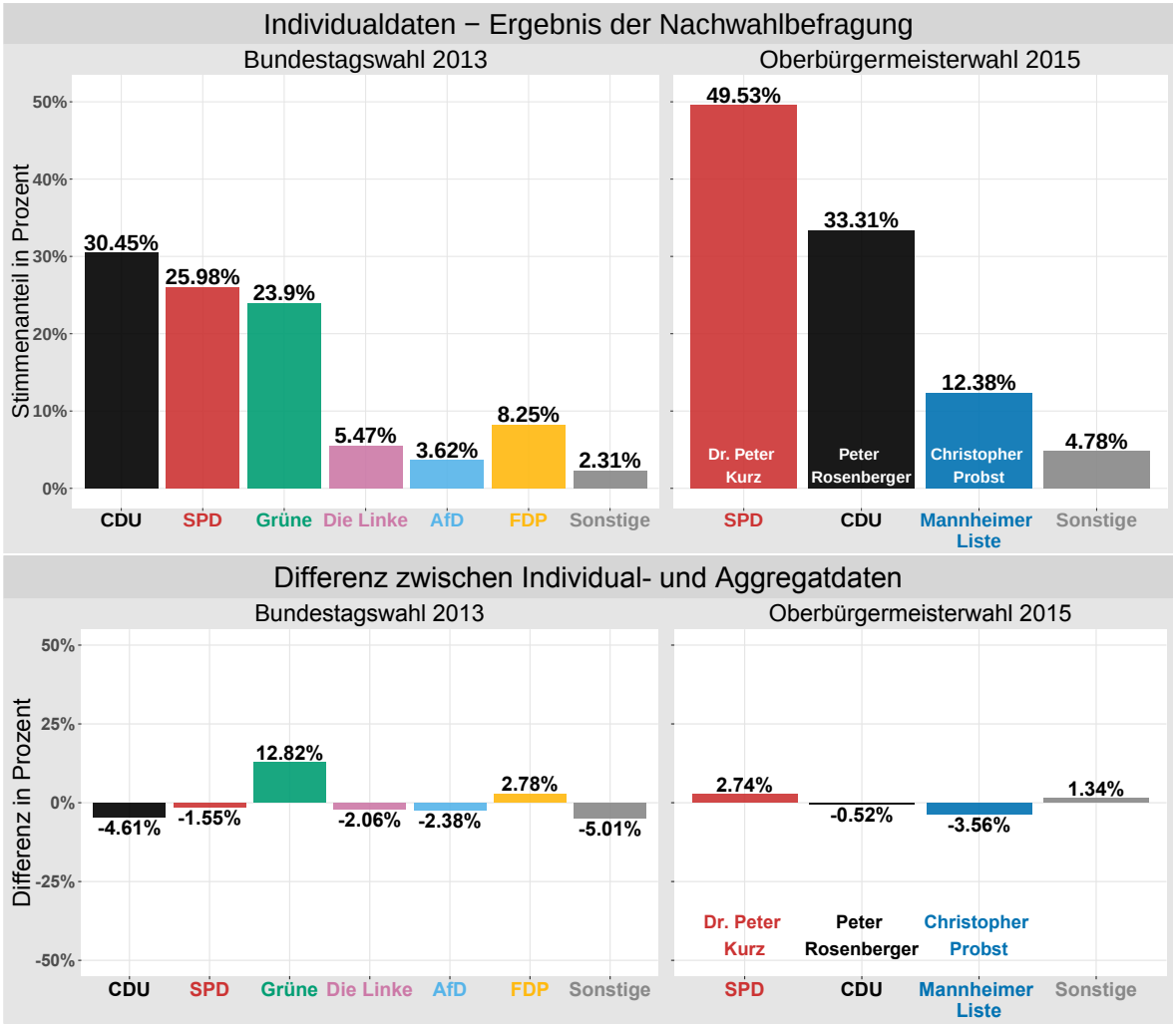

Abbildung 4.4: *Oben: Die Wahlergebnisse anhand der Nachwahlbefragung für die Bundestagswahl 2013 (links) und für die Oberbürgermeisterwahl 2015 (rechts). Unten: Die Differenz zwischen den Wahlergebnissen der Individual- und den Aggregatdaten für die Bundestagswahl 2013 und für die Oberbürgermeisterwahl 2015.*

Wie sich die Befragten bei dem Fragebogen geäußert haben ist in der Abbildung 4.4 (oben) dargestellt, wobei die *Nichtwähler* nicht berücksichtigt sind. In der unteren Grafik wird zusätzlich die Differenz zu den amtlichen Ergebnissen abgebildet. Die Differenz bei der Oberbürgermeisterwahl 2015 (unten rechts) zeigt eine bessere Übereinstimmung mit den amtlichen Ergebnissen. Die höchste Abweichung von 3.56 Prozent liegt beim Kandidaten *Christopher Probst* vor. Bei der Bundestagswahl 2013 (unten links) weist die Grafik generell etwas höhere Abweichungen auf. Unterdessen ist eine deutliche Überschätzung des Stimmenanteils der *Grünen* um 12.82 Prozent zu erkennen. Zur Untersuchung der möglichen Ursachen für diese Störung werden zusätzlich drei Grafiken erzeugt. Zuerst soll eine Darstellung der Ergebnisse nach Wahlbezirken in der Abbildung 4.5 zeigen, ob die Wahl der Bezirke bei der Durchführung der Nachwahlbefragung

die Ergebnisse beeinflussen könnte. Hierfür werden lediglich die fünf Wahlbezirke selektiert, die bei der Nachwahlbefragung betrachtet wurden. In den Abbildungen A.2 (Seite 87) und A.3 (Seite 88) im Anhang A.1.3 befinden sich zusätzlich die Ergebnisse aller Wahlbezirke. Für alle drei Grafiken werden die Wahlbezirke aggregiert, sodass die gleichen Ebenen bei der Bundestagswahl (2013) und bei der Oberbürgermeisterwahl (2015) mit der Nachwahlbefragung verglichen werden können. Im Unterabschnitt 4.2.3 auf der Seite 42 wird das Prinzip der Zusammensetzung der Wahlbezirke beschrieben. Aus den Grafiken lässt sich nicht erkennen, dass die amtlichen Ergebnisse der betrachteten Gebiete im Vergleich zu den Übrigen einen höheren Stimmenanteil für die *Grünen* aufweisen. Amtliches Ergebnis des Bezirkes 01251 weicht im Vergleich zu anderen nach oben ab. Dennoch wird der Stimmenanteil in der Nachwahlbefragung an dieser Stelle noch stärker überschätzt.

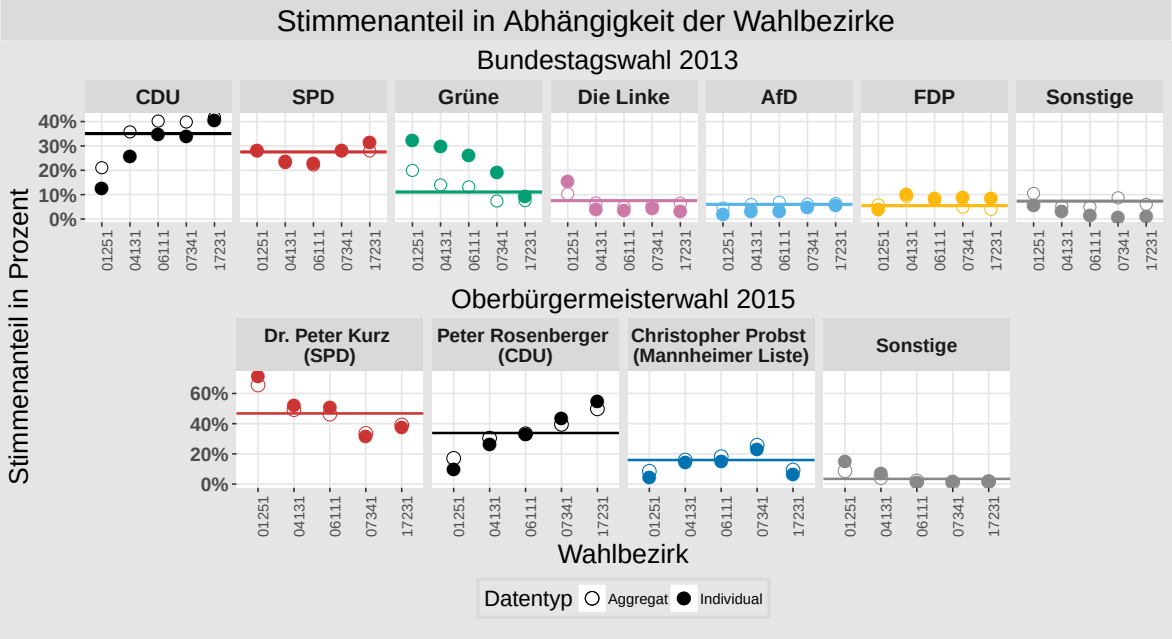

Abbildung 4.5: *Stimmenanteil in Abhängigkeit der Wahlbezirke bei der Bundestagswahl 2013 (oben) und der Oberbürgermeisterwahl 2015 (unten) für fünf Wahlbezirke, die bei der Nachwahlbefragung betrachtet wurden. Die dargestellten Wahlbezirke werden so aggregiert, dass alle Ebenen bei der Bundestagswahl (2013) und bei der Oberbürgermeisterwahl (2015) identisch sind und den Wahlbezirken bei der Nachwahlbefragung entsprechen.*

Obwohl rein grafische Beschreibungen nicht ausreichend sind um den Einfluss einer Variable festzustellen, so bietet eine visuelle Untersuchung der Strukturen von Alter und Bildungsabschluss der Befragten eine grobe Beschreibung der möglichen Einflüsse auf die Verzerrung. Die Grafik in der Abbildung 4.6 zeigt keine Indikatoren,

dass die Überschätzung der *Grünen* durch das Alter der Befragten verursacht wurde. Das Durchschnittsalter der Befragten, welche angeblich die *CDU*, *AfD* und *FDP* gewählt haben, ist etwas höher als das mittlere Alter aller Befragten. Die Befragten, die sich für die *Grünen* und *Die Linke* entschieden haben, sind im Durchschnitt ein wenig jünger. Diejenigen, die *Sonstige* angekreuzt haben, weisen ein um 9.13 Jahre niedrigeres mittleres Alter auf. Die Stadt Mannheim (2015c, S. 5, 15 f.) teilt mit, dass die Wahlbeteiligung bei der Oberbürgermeisterwahl 2015 in den älteren Altersgruppen generell höher war, insbesondere bei den 70-Jährigen und Älteren, wobei das Durchschnittsalter der Wahlberechtigten im Juni 48.4 Jahre betrug. Demzufolge entspricht das mittlere Alter der Befragten beinahe dem durchschnittlichen Alter in der Population der Wahlberechtigten.

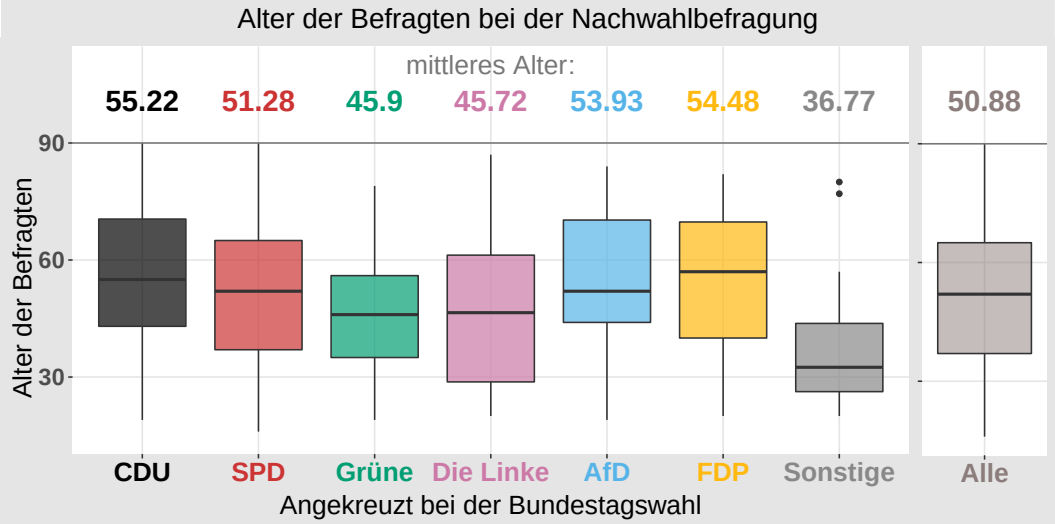

Abbildung 4.6: *Alter der Befragten bei der Nachwahlbefragung in Abhängigkeit der Angaben bei der Bundestagswahl (2013).*

Die Betrachtung der Bildungsabschlussquoten in der Tabelle 4.2 zeigt einen hohen Anteil von Personen mit Hochschulabschluss. Da die wahre Struktur des Bildungsabschlusses der Wahlberechtigten nicht bekanntgegeben wird, dienen die Informationen über den Bildungsabschluss der Bevölkerung Baden-Württembergs als ein Vergleichsmaß. Das Statistisches Bundesamt (2015, S. 65, 67) berichtet, dass 4.0 Prozent der Bevölkerung die Schule besuchen, 36.8 Prozent besitzen einen Volks- beziehungsweise Hauptschulabschluss, 26.7 Prozent einen Real- oder Mittelschulabschluss, 12.7 Prozent haben das Abitur oder Fachabitur und 17.3 Prozent Fachhochschul- oder Hochschulabschluss. Zu 0.5 Prozent sind die Angaben nicht bekannt und 2.0 Prozent haben keinen allgemeinen Schulabschluss. Dementsprechend erscheint in der Stichprobe einerseits ein

um 26.31 Prozent geringerer Anteil des Volks- oder Hauptschulabschlusses und ein um 5.63 Prozent geringerer Anteil des Real- oder Mittelschulabschlusses. Andererseits weist die Stichprobe einen um 8.52 Prozent höheren Anteil bei Abitur oder Fachabitur und einen um 39.44 Prozent höheren Anteil des Hochschulabschlusses auf. Die Befragten, die die *Grünen* angekreuzt haben, besitzen zu 61.04 Prozent einen Hochschulabschluss. Somit liegt hier starke Abweichung von der Hochschulabschlussquote der Bevölkerung vor und eine generell höhere Quote im Vergleich zur allen anderen Befragten. Da die Struktur des Bildungsabschlusses der Wahlberechtigten nicht bekannt gegeben wird, lässt sich allerdings keine zuverlässige Schlussfolgerung über die Quelle des Fehlers anhand von Bildungsabschlussquoten ziehen.

Bildungsabschluss der Befragten bei der Nachwahlbefragung

Bildungsabschluss	Angekreuzt bei der Bundestagswahl: CDU	SPD	Grüne	Die Linke	AfD	FDP	Sonstige	Alle
Noch Schüler	0 %	0.61 %	0 %	0 %	0 %	0 %	0 %	0.16 %
Volks-, Hauptschulabschluss	12.14 %	15.03 %	3.25 %	9.09 %	15.91 %	12.38 %	3.33 %	10.49 %
Real-, Mittelschulabschluss	26.12 %	24.85 %	12.34 %	15.15 %	31.82 %	20 %	6.67 %	21.07 %
Abitur / Fachabitur	19 %	19.02 %	23.38 %	30.3 %	15.91 %	20.95 %	40 %	21.22 %
Hochschulabschluss	42.74 %	39.26 %	61.04 %	45.45 %	36.36 %	46.67 %	50 %	46.74 %
Kein Abschluss	0 %	1.23 %	0 %	0 %	0 %	0 %	0 %	0.32 %
Summe	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %

Tabelle 4.2: *Bildungsabschluss der Befragten bei der Nachwahlbefragung in Abhängigkeit der Angaben bei der Bundestagswahl (2013).*

Die Ergebnisse der Wählerwanderung sind anhand von Individualdaten in Form der prozentualen Übergangsanteile in der Tabelle 4.3 gegeben und in der Abbildung 4.7 zusätzlich visuell dargestellt. Im Anhang A.1.4 auf der Seite 89 befindet sich eine gleichartige Darstellung ohne die Kategorie *Nichtwähler*. Eine Übergangstabelle zwischen zwei gleichartigen Wahlen, beispielsweise der Bundestagswahl 2009 und der Bundestagswahl 2013, wird normalerweise so erzeugt, dass die Übergangszellen der selben Parteien beider Wahlen, die sogenannten Loyalen oder Treuen, auf der Diagonale liegen. Die Übrigen, die Wechselnden, werden nichtdiagonal positioniert (Klima et al., 2015, S. 2). Hierfür werden die Treuen gemäß den oben erwähnten Wahlempfehlungen untereinander gestellt, das heißt als Loyale betrachtet. Demzufolge entspricht die

bisherige Aufstellung nicht der Reihenfolge der Parteien im weiteren Verlauf der Arbeit.

Nachwahlbefragung 2015						
		Oberbürgermeisterwahl 2015				
		Dr. Peter Kurz (SPD)	Peter Rosenberger (CDU)	Christopher Probst (Mannheimer Liste)	Sonstige	Nichtwähler
Bundestagswahl 2013	SPD	70.33 %	17.21 %	10.09 %	2.37 %	0 %
	Grüne	68.06 %	13.55 %	11.94 %	5.16 %	1.29 %
	Die Linke	56.34 %	11.27 %	11.27 %	19.72 %	1.41 %
	CDU	24.81 %	62.28 %	11.14 %	1.77 %	0 %
	AfD	12.77 %	48.94 %	36.17 %	0 %	2.13 %
	FDP	28.04 %	52.34 %	17.76 %	1.87 %	0 %
	Sonstige	43.33 %	0 %	16.67 %	40 %	0 %
	Nichtwähler	52.69 %	29.03 %	7.53 %	7.53 %	3.23 %

Tabelle 4.3: *Die Übergangstabelle zwischen der Bundestagswahl 2013 und der Oberbürgermeisterwahl 2015 anhand der Nachwahlbefragung.*

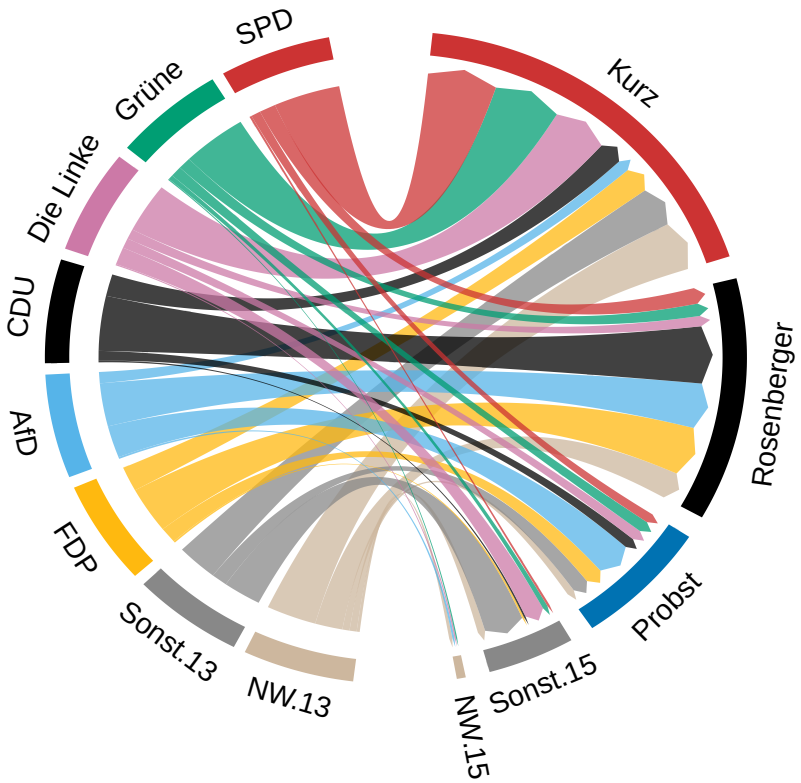

Abbildung 4.7: *Die Übergangswahrscheinlichkeiten zwischen der Bundestagswahl 2013 (links) und der Oberbürgermeisterwahl 2015 (rechts) anhand der Nachwahlbefragung. Die Breite jedes Pfeilendes drückt den Anteil an Stimmen aus, den der jeweilige Kandidat von verschiedenen Parteien gewonnen hat.*

Die Wurzeln der inneren Pfeile des Kreisdiagramms in der Abbildung 4.7, gefärbt

nach den dazugehörigen Parteien, fangen bei der Bundestagswahl 2013 an und enden bei der Oberbürgermeisterwahl 2015. Dabei drückt die Breite jedes Pfeilendes den Anteil an Stimmen aus, den der jeweilige Kandidat von verschiedenen Parteien gewonnen hat. Für eine bessere Übersicht werden die Namen verkürzt, wobei sich der ganze Name der Kandidaten aus der Tabelle 4.3 ablesen lässt. Die Abkürzung *NW* bezeichnet die *Nichtwähler*. Zusammen mit der Tabelle 4.3 zeigt die Grafik, dass die oben genannte Behauptung der Unterschätzung der *Nichtwähler* bei der Oberbürgermeisterwahl (2015) bestätigt werden kann.

Die Differenz zwischen den Randsummen der Nachwahlbefragung und den amtlichen Ergebnissen beweist, dass bei den Individualdaten ein Bias vorliegt. Dementsprechend ist die Analyse der Wählerwanderung ausschließlich anhand von Individualdaten in diesem Fall nicht zuverlässig. Neben den ergänzten möglichen Indizien, respektive den fehlenden Werten, der Bildungsstruktur der Befragten, der unterschätzte Anteil der *Nichtwähler* und das Ignorieren der Briefwähler, stellt die Teilnahmeverweigerung eine weitere Gefahr für die Störung in den Daten dar. Die bessere Übereinstimmung der Antworten mit den amtlichen Ergebnissen bei der Oberbürgermeisterwahl im Jahr 2015 als bei der Bundestagswahl im Jahr 2013 spricht dafür, dass der Zeitabstand auch eine negative Rolle spielen könnte.

4.2 Aufbereitung der Daten

Die Analyse der Wählerwanderung ist ein Sonderfall der ökologischen Inferenz in dem Sinne, dass eine höhere Rücksicht auf die Datenaufbereitung gerichtet werden sollte. In diesem Abschnitt erfolgt eine Angabe der primären Punkte und die Beschreibung der möglichen und der angewendeten Vorgehensweisen bei der Datenaufbereitung. Kellermann (2011), Ambühl (2003), Andreadis und Chadjipadelis (2009) beschrieben bereits einige Vorschritte, die durchgeführt werden können oder sogar müssen. Allerdings widmen Klima et al. (2015) dem Thema etwas mehr Aufmerksamkeit und zeigen in ihrem Artikel, dass das Vorgehen bei der Datenaufbereitung die Schätzungen bedeutend beeinflussen kann. Ferner evaluieren die Autoren anhand von Simulationsstudien, wie viel Einfluss auf die Qualität der Schätzung durch verschiedene Vorgehensweisen bei einigen Modellen genommen werden kann. Deren Ergebnisse stellen den Ausgangspunkt für einige Entscheidungen bei der Datenaufbereitung in dieser Arbeit dar.

4.2.1 Anzahl der Parteien

Der erste relevante Punkt repräsentiert die Anzahl der Parteien, die in der Analyse betrachtet werden sollten. Kellermann (2011, S. 34) weist darauf hin, dass eine möglichst starke Reduktion der Parteien erforderlich ist, um eine Verringerung der zu schätzenden Parameter zu schaffen. Die kleinen Parteien beeinflussen die Gesamtanzahl an Stimmen nur minimal, was bei der Schätzung ihrer Übergangswahrscheinlichkeiten meistens zu falschen Ergebnissen führt (Achen und Shively, 1995 in: Andreadis und Chadjipadelis, 2009, S. 206 f.; Kellermann, 2011, S. 34). Achen und Shively schlagen deshalb vor, die kleinen Parteien den großen, bezüglich der gemeinsamen Ideologie, zuzuordnen.

Derweilen kommt eine andere Vorgehensweise in dieser Arbeit zur Anwendung. Es können alternativ alle kleinen Parteien unter einer Kategorie zusammengefasst werden (Ambühl, 2003, S. 20; Andreadis und Chadjipadelis, 2009, S 209 f.; Kellermann, 2011, S. 34). Dementsprechend werden hier alle Parteien und Kandidaten, die weniger als 5% der gesamten Stimmenanzahl aufweisen, der Kategorie *Sonstige* unterstellt. In der Tabelle 4.4 sind alle dazugehörigen Parteien aufgelistet. Von den Kandidaten der Oberbürgermeisterwahl 2015 blieb alleinig *Christian Sommer (Die Partei)* mit 3.27 Prozent der Stimmen für die Zuordnung zur Kategorie *Andere Gewählte*, die letztendlich in *Sonstige* umbenannt wurde.

	Partei	**Anteil**
01	PIRATEN	3.2 %
02	NPD	1.2 %
03	TIER-SCHUTZ-PARTEI	1.0 %
04	REP	0.4 %
05	RENTNER	0.4 %
06	FREIE-WÄHLER	0.3 %
07	ÖDP	0.2 %
08	VOLKSABSTIMMUNG	0.2 %
09	PARTEI DER VERNUNFT	0.1 %
10	PRO-DEUTSCHLAND	0.1 %
11	BIG	0.1 %
12	BüSo	0.0 %
13	MLPD	0.1 %
14	PBC	0.1 %
Σ	`Sonstige`	≈ **7.4** %

Tabelle 4.4: *Kleine Parteien, die bei der Bundestagswahl (2013) der Kategorie Sonstige zugeordnet wurden.*

Zudem empfehlen Kellermann (2011, S. 34) und Ambühl (2003, S. 19) die Wahlberechtigten, die nicht oder ungültig gewählt haben, als zusätzliche Kategorie *Nichtwähler* zu betrachten. In dieser Arbeit wurde die neue Variable durch die Differenz zwischen der Variable `Summe_Wahlberechtigte` und `Gültige_Stimmen` erzeugt. Eine grafische Darstellung der wahlbezirkspezifischen amtlichen Ergebnisse, in der die *Nichtwähler* betrachtet werden, liegt im Anhang A.1.5 auf der Seite 90 für die Bundestagswahl 2013 (Abbildung A.5) und für die Oberbürgermeisterwahl 2015 (Abbildung A.6) vor.

4.2.2 Bevölkerungsänderung

Das zweite Problem erfasst die Veränderung der Population zwischen zwei betrachteten Wahlen, die durch Neuwähler, den Gestorbenen und den Umzug von Personen auftritt (Ambühl, 2003, S. 20; Kellermann, 2011, S. 34; Klima et al., 2015, S. 14 f.). Dadurch wird die Annahme gleicher Population verletzt und der Populationsumfang N sowie die Anzahl der Wähler für jeden Wahlbezirk N_i werden bei der ersten und der zweiten Wahl nicht identisch sein. Deshalb ist es unbedingt nötig, die Daten vor der Analyse anzupassen.

Sind die Daten zur Bevölkerungsänderung verfügbar, dann lassen sich die Veränderungen als neue Kategorien im Modell betrachten, indem die *noch nicht Wähler* als zusätzliche Variable bei der ersten Wahl und *nicht mehr Wähler* bei der zweiten Wahl betrachtet werden (Ambühl, 2003, S. 20; Klima et al., 2015, S. 15). Die inneren Zellen dieser Kategorien betragen 0 (siehe Tabelle 4.5), da keine Person, die nach der ersten Wahl gestorben oder weggezogen ist, in der zweiten Wahl ein Neuwähler oder ein Zugezogener sein kann (Klima et al., 2015, S. 15).

		2. WAHL					
		P1	P2	...	Gestorbene	Weggezogene	
1. WAHL	P1						
	P2						
	⋮						
	Neuwähler				0	0	
	Zugezogene				0	0	
							N_i

Tabelle 4.5: *Tabelle der Wahldaten mit zusätzlichen Kategorien zur Bevölkerungsänderung (übernommen von Klima et al. (2015, S. 15) und verallgemeinert).*

Bei der praktischen Umsetzung kommt es allerdings häufig vor, dass die Daten zur Bevölkerungsänderung nicht verfügbar sind. Zwei Vorgehen sind in diesem Fall möglich. Das erste wurde von Hawkes (1969, in: Kellermann, 2011, S. 35 f.) vorgeschlagen. Er nimmt an, dass die Bevölkerungsänderung keinen Einfluss auf das Wahlverhalten hat, d.h. die *neuen Wähler* vergeben ihre Stimmen ähnlich wie *nicht mehr Wähler*. Demzufolge kann die Differenz gemäß der Stimmvergabe verteilt werden. Beispielsweise, wenn bei der zweiten Wahl mehr wahlberechtigte Personen als bei der ersten vorhanden sind, wird die Differenz zur ersten Wahl proportional zur Stimmvergabe bei der zweiten Wahl zugerechnet (mehr zur Berechnung in Kellermann, 2011, S. 35 f.). Die zweite Möglichkeit wäre die Addition der Differenz von der Anzahl der Wahlberechtigten zwischen den beiden Wahlen und der *Nichtwähler* Kategorie. Keine der oben genannten Methoden ist anhand der Simulationsstudie von Klima et al. (2015) beim Vorgehen mit der Bevölkerungsänderung zu bevorzugen. Deshalb kommt hier die unkomplizierte Methode zur Anwendung. Die Differenz zwischen der *Summe der Wahlberechtigten* bei der Bundestagswahl (2013) und der *Summe der Wahlberechtigten* bei der Oberbürgermeisterwahl (2015) wurde zur *Nichtwähler* Kategorie bei der Bundestagswahl (2013) gerechnet.

4.2.3 Veränderung der Wahlbezirke

Ein weiteres Problem verursacht die Veränderung der Aufstellung von Wahlbezirken zwischen zwei Wahlen, die durch die Vereinigung oder die Aufteilung der Bezirke entstehen kann (Klima et al., 2015, S. 14). Da die Variable *Wahlbezirk* bei der Analyse als Identitätsvariable betrachtet wird, ist es notwendig, die konstanten Gebiete vor der Analyse zu definieren. Bei den hier betrachteten zwei Wahlen ist diese Veränderung besonders stark ausgefallen. Für die Urnenwähler wurden aus ursprünglich 150 Wahlbezirken aus der Bundestagswahl 2013 lediglich 96 bei der Oberbürgermeisterwahl 2015 gebildet (siehe auch Tabelle 4.1). Eine Zuordnung der Straßen zu den Wahlbezirken ist von dem Wahlbüro der Stadt Mannheim (2016, persönliche Kommunikation) für beide Wahlen bekannt gegeben worden, woraus hergeleitet werden konnte, wie die Wahlbezirke verknüpft worden sind. Eine Auflistung der Beziehungen zwischen umgeordneten Bezirken ist im Anhang A.1.6 auf der Seite 91 zu finden. Überall, wo ein Wahlbezirk der Bundestagswahl 2013 so zugeschnitten ist, dass bei der Oberbürgermeisterwahl 2015 einige Straßen einem und die übrigen dem anderen Wahlbezirk zugeordnet sind,

mussten alle auf einer Ebene aggregiert werden (alle grau gefärbten Wahlbezirke in der Tabelle A.3). Dadurch konnten letztendlich nur 67 konstante Ebenen für die Identitätsvariable erstellt werden. Klima et al. (2015, 2016) betonen, dass in dieser Maßnahme, die eine reduzierte Anzahl an Gebiete für die Analyse bereitstellt, eine weitere Fehlerquelle vorliegt, denn damit reduziert sich auch der Informationsumfang. Ihre Simulationsstudie zeigt, dass die Anzahl der Wahlbezirke einen bedeutenden Einfluss auf die Schätzung hat (Klima et al., 2015, S. 18).

4.2.4 Briefwähler

Jeder Wahlberechtigte hat die Möglichkeit, einen Wahlschein zu beantragen, um seine Stimme per Briefwahl zu vergeben. Diese Möglichkeit erzeugt aus zwei Gründe einen zusätzlichen Ursprung für Komplikationen bei der Datenaufbereitung. Erstens, das Wahlverhalten der Briefwähler unterscheidet sich in der Regel von dem Wahlverhalten der Urnenwähler. Aus diesem Grund dürfen die Briefwähler bei der Analyse nicht ignoriert werden, was übrigens die Simulationsstudie von Klima et al. (2015) bestätigt. Denn das Ausschließen der Briefwähler führt bei allen Modellen zu schlechteren Ergebnissen. Zweitens, die Ergebnisse der Briefwähler werden in spezifischen postalischen Wahlbezirken dargestellt, die üblicherweise nicht identisch mit den Wahlbezirken bei der Urnenwahl sind. Das bedeutet, dass zusätzliche Berechnungen notwendig sind, um Briefwähler in die Analyse einzuschließen. Die Simulationsstudie zeigt hierbei, dass die Schätzung desto genauer wird, je präziser die Aufteilung gemacht werden konnte. (Klima et al., 2015, S. 15 f., 20 ff.)

Die betrachteten Datensätze in dieser Arbeit enthalten die Variable `Wahlb._mit_Wahlschein`, die für jeden Wahlbezirk die Anzahl der Wahlberechtigten angibt, die einen Wahlschein beantragt haben. Diese Variable kann zur einigermaßen zuverlässigen Gewichtung bei der Addition der Briefwähler dienen, da die Wahlbeteiligung der *Wahlberechtigten mit Wahlschein* in der Regel sehr hoch ist (Klima et al., 2015, S. 15 f.). Im Hinblick auf die Tatsache, dass die Anzahl und die Bezeichnung der Stadtbezirke bei den Brief- und Urnenwählern übereinstimmen, kann der Anteil der *Wahlberechtigten mit Wahlschein* für jeden Wahlbezirk nach dem dazugehörigen Stadtbezirk berechnet werden. Dabei beträgt die Summe aller berechneten Anteile innerhalb eines Stadtbezirkes eins. Diese Anteile können weiterhin mit der Summe der Briefwähler je nach Stadtbezirk multipliziert werden, um die Anzahl

der Briefwähler separat pro Wahlbezirk zu bekommen. Letztendlich lässt sich die gewichtete Anzahl der Briefwähler zu den Urnenwähler addieren. Die damit erzeugten Dezimalziffern müssen noch in ganze Zahlen umgeformt werden. Durch die Rundung tritt jedoch ein Fehler auf, sprich in einigen Zellen zeigt sich eine Abweichung von ein bis zwei Stimmen. Da die wahren Randsummen je Stadtbezirk bekannt gegeben wurden, können diese zur Überprüfung und zur Korrektur des Rundungsfehlers verwendet werden. Die wahren spaltenweisen Randsummen lassen sich diesbezüglich durch das Aggregieren aller (nicht gewichteten) Brief- und Urnenwähler je Stadtbezirk berechnen. Da die *Nichtwähler* bei der Berechnung betrachtet werden, entspricht die Variable *Summe Wahlberechtigten* den wahren zeilenweisen Randsummen. Das Vorgehen bei der Korrektur ist, zuerst die ganzen Zahlen so zu generieren, dass alle Nachkommastellen weggeworfen werden. Danach wird ein Vektor mit den Ordnungsnummern der verworfenen Reste erzeugt. Dieser Vektor dient dazu, eine Stimme zuerst an der Stelle einzufügen, wo der größte Dezimalrest vorliegt, falls bei dieser Zelle die gewichteten Zeilen- und Spaltensummen mit den wahren Randsummen nicht übereinstimmen. Alle Werte, die noch addiert werden müssen um die Randsummen anzupassen, werden einer Nullmatrix zugeordnet, die letztendlich zum gewichteten Datensatz addiert wird. Dieser Prozess wiederholt sich so lang, bis alle Randsummen angepasst wurden.

Im Anhang A.1.7 auf der Seite 92 befindet sich eine Darstellung der Differenzen zwischen den Brief- und den Urnenwählern der beiden betrachteten Wahlen. In der Abbildung A.7 werden hierbei die Differenzen ohne die Kategorie *Nichtwähler* betrachtet und in der Abbildung A.8 ist diese Kategorie berücksichtigt. Der Programmcode zur Berechnung des gewichteten Datensatzes in `R` (R Core Team, 2015) wurde auf einem reduzierten Beispieldatensatz simuliert, der die amtlichen Ergebnisse der ersten zwei Wahlbezirke der Oberbürgermeisterwahl 2015 beinhaltet. Diese Simulation ist wegen des Umfangs im elektronischer Anhang E dargelegt.

4.2.5 Die Endform der Aggregat- und Individualdaten

Vor der Darstellung der Ergebnisse im Kapitel 6 wird im Kapitel 5 beschrieben, wie die Analyse anhand der betrachteten Modellen in `R` (R Core Team, 2015) jeweils mit und ohne Individualdaten durchgeführt werden kann. Beim letzten Schritt der Datenaufbereitung müssen die Daten in eine Form gebracht werden, die für die Analyse im `eiwild`

Paket von Schlesinger (2014) oder im `RxCEcolInf` Paket von Greiner et al. (2013) geeignet ist. Während sich das geforderte Format der Individualdaten unterscheidet, wird hingegen die Form der Aggregatdaten in beiden Paketen gleichartig definiert.

WBZ	P1	P2	$\cdots$	NW_13	K1	K2	$\cdots$	NW_15
1	$X^a_{1,1}$	$X^a_{2,1}$	$\cdots$	$X^a_{R,1}$	$Y^a_{1,1}$	$Y^a_{2,1}$	$\cdots$	$Y^a_{C,1}$
2	$X^a_{1,2}$	$X^a_{2,2}$	$\cdots$	$X^a_{R,2}$	$Y^a_{1,2}$	$Y^a_{2,2}$	$\cdots$	$Y^a_{C,2}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
67	$X^a_{1,67}$	$X^a_{2,67}$	$\cdots$	$X^a_{R,67}$	$Y^a_{1,67}$	$Y^a_{2,67}$	$\cdots$	$Y^a_{C,67}$

Tabelle 4.6: *Die Endform der Aggregatdaten zwischen einer Bundestagswahl und einer Oberbürgermeisterwahl für 67 Wahlbezirke.*

WGB	P1.K1	P1.K2	$\cdots$	P1.NW_15	P2.K1	$\cdots$ $\cdots$	NW_13.NW_15
5	n^5_{11}	n^5_{12}	$\cdots$	n^5_{1C}	n^5_{21}	$\cdots$ $\cdots$	n^5_{RC}
18	n^{18}_{11}	n^{18}_{12}	$\cdots$	n^{18}_{1C}	n^{18}_{21}	$\cdots$ $\cdots$	n^{18}_{RC}
24	n^{24}_{11}	n^{24}_{12}	$\cdots$	n^{24}_{1C}	n^{24}_{21}	$\cdots$ $\cdots$	n^{24}_{RC}
31	n^{31}_{11}	n^{31}_{12}	$\cdots$	n^{31}_{1C}	n^{31}_{21}	$\cdots$ $\cdots$	n^{31}_{RC}
64	n^{64}_{11}	n^{64}_{12}	$\cdots$	n^{64}_{1C}	n^{64}_{21}	$\cdots$ $\cdots$	n^{64}_{RC}

Tabelle 4.7: *Die Endform der Individualdaten zwischen einer Bundestagswahl und einer Oberbürgermeisterwahl für 5 fiktive Wahlbezirke beim Multinomial-Dirichlet-Hybridmodell im* `eiwild` *Paket (Schlesinger, 2014).*

Wie die Endform der Aggregatdaten zwischen einer Bundestagswahl und einer Oberbürgermeisterwahl für 67 Wahlbezirke gestaltet werden soll, zeigt die Beispieltabelle 4.6. Die beiden betrachteten Datensätze der amtlichen Ergebnisse werden anhand der vorher erzeugten konstanten Ebenen der Identitätsvariable vereinigt. Die Ergebnisse der ersten Wahl $(X^a_{1,i}, \cdots, X^a_{R,i})$ und der zweiten Wahl $(Y^a_{1,i}, \cdots, Y^a_{C,i})$ werden somit für jeden Wahlbezirk i in einem Datensatz mit insgesamt 67 beziehungsweise p Zeilen gespeichert. In der Tabelle 4.6 bezeichnen die Abkürzungen WBZ, P, K und NW den *Wahlbezirk*, die *Partei*, den *Kandidaten* und die *Nichtwähler*. Die gleich genannten Variablen beider Wahlen bekommen einen Suffix mit dem Jahr der dazugehörigen Wahl zur Differenzierung. Bei einer Wählerwanderungsanalyse zwischen zwei gleichartigen Wahlen, zum Beispiel zweier Bundestagswahlen, sollen alle gleichen Parteien mit einem Suffix bezeichnet werden.

Bei den Hybridmodellen werden die absoluten inneren Zellen der Kreuztabelle von den Individualdaten je beobachteten Wahlbezirk in einer Zeile erfasst. Die Zeilenanzahl des Datensatzes für die Analyse im `eiwild` Paket (Schlesinger, 2014, S. 10) ist

identisch zur Anzahl der betrachteten Wahlbezirke bei der Nachwahlbefragung. Eine Darstellung von fünf fiktiven Wahlbezirke befindet sich in der Tabelle 4.7. Für die Analyse im `RxCEcolInf` Paket (Greiner et al., 2013, S. 12 f.) müssen hingegen die Individualdaten in eine Matrix umgeformt werden. Die Zeilenanzahl dieser Matrix ist gleich der Anzahl der Wahlbezirke bei den Aggregatdaten. An der Stelle wo keine Individualdaten vorhanden sind, werden zeilenweise Nullvektoren eingefügt, sodass jede Zeile dieser Matrix den jeweiligen Zeilen in den Aggregatdaten entspricht. Die Spaltennamen besitzen zusätzlich einen Präfix „KK.“[2]. Analog zum Beispiel in der Tabelle 4.7 wird eine Mustermatrix in der Abbildung 4.8 hergestellt.

$$
\begin{array}{c|ccccccc}
 & KK.P1.K1 & KK.P1.K2 & \cdots & KK.P1.NW\_15 & KK.P2.K1 & \cdots\ \cdots & KK.NW\_13.NW\_15 \\
\hline
[1] & 0 & 0 & \cdots & 0 & 0 & \cdots\ \cdots & 0 \\
[2] & 0 & 0 & \cdots & 0 & 0 & \cdots\ \cdots & 0 \\
[3] & 0 & 0 & \cdots & 0 & 0 & \cdots\ \cdots & 0 \\
[4] & 0 & 0 & \cdots & 0 & 0 & \cdots\ \cdots & 0 \\
[5] & n_{11}^{5} & n_{12}^{5} & \cdots & n_{1C}^{5} & n_{21}^{5} & \cdots\ \cdots & n_{RC}^{5} \\
[6] & 0 & 0 & \cdots & 0 & 0 & \cdots\ \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
[17] & 0 & 0 & \cdots & 0 & 0 & \cdots\ \cdots & 0 \\
[18] & n_{11}^{18} & n_{12}^{18} & \cdots & n_{1C}^{18} & n_{21}^{18} & \cdots\ \cdots & n_{RC}^{18} \\
[19] & 0 & 0 & \cdots & 0 & 0 & \cdots\ \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
[23] & 0 & 0 & \cdots & 0 & 0 & \cdots\ \cdots & 0 \\
[24] & n_{11}^{24} & n_{12}^{24} & \cdots & n_{1C}^{24} & n_{21}^{24} & \cdots\ \cdots & n_{RC}^{24} \\
[25] & 0 & 0 & \cdots & 0 & 0 & \cdots\ \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
[29] & 0 & 0 & \cdots & 0 & 0 & \cdots\ \cdots & 0 \\
[31] & n_{11}^{31} & n_{12}^{31} & \cdots & n_{1C}^{31} & n_{21}^{31} & \cdots\ \cdots & n_{RC}^{31} \\
[32] & 0 & 0 & \cdots & 0 & 0 & \cdots\ \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
[63] & 0 & 0 & \cdots & 0 & 0 & \cdots\ \cdots & 0 \\
[64] & n_{11}^{64} & n_{12}^{64} & \cdots & n_{1C}^{64} & n_{21}^{64} & \cdots\ \cdots & n_{RC}^{64} \\
[65] & 0 & 0 & \cdots & 0 & 0 & \cdots\ \cdots & 0 \\
[66] & 0 & 0 & \cdots & 0 & 0 & \cdots\ \cdots & 0 \\
[67] & 0 & 0 & \cdots & 0 & 0 & \cdots\ \cdots & 0
\end{array}
$$

Abbildung 4.8: *Die Matrix-Endform der Individualdaten zwischen einer Bundestagswahl und einer Oberbürgermeisterwahl für 5 fiktive Wahlbezirke beim Multinomial-Log-Normal-Hybridmodell im `RxCEcolInf` Paket (Greiner et al., 2013).*

[2] Greiner und Quinn (2009, S. 78) bezeichnen die beobachtete Anzahl der inneren Zellen der Kreuztabelle mit K_{rc}^{i}.

5 Praktische Anwendung der Modelle in R

5.1 Multinomial-Dirichlet-Modell

Für die Analyse anhand des Multinomial-Dirichlet-Modells von Rosen et al. (2001) sind zwei Zusatzpakete in R verfügbar, das `eiPack` Paket von Lau et al. (2012) und das `eiwild` Paket von Schlesinger (2014). Da das Paket von Schlesinger die Kombination der Individual- und Aggregatdaten ermöglicht (siehe Unterabschnitt 3.3.2 des Kapitels 3), werden die Übergangswahrscheinlichkeiten in dieser Arbeit mithilfe von seinem Paket erzeugt.

Im `eiwild` Paket lässt sich die Analyse mittels der Funktion

```
indAggEi(form, aggr, indi=NULL, IDCols=c("ID"),                    1
         whichPriori="gamma", prioriPars=list(shape=4, rate=2),    2
         startValsAlpha=NULL, startValsBeta=NULL,                  3
         betaVars=NULL, alphaVars=NULL,                            4
         sample, burnin=0, thinning=1, verbose=1, ...)             5
```

durchführen. Die grau gefärbten Befehle werden bei der Analyse mit ihren Defaultwert verwendet. Mit dem Befehl `whichPriori` ist es alternativ möglich, die Exponential-Hyperpriori-Verteilung zu wählen und mit `startValsAlpha` und `startValsBeta` lassen sich in einer Matrixform die Startwerte für α und β festlegen. Bei dem Defaultwert `NULL` werden die Startwerte zufällig aus den entsprechenden Verteilungen gezogen. (Schlesinger, 2014, S. 9)

5.1.1 Die Datensätze

Durch den `aggr` und `indi` Befehle gibt man die Datensätze an, wobei für die Analyse ohne Individualdaten beim Argument `indi` der Defaultwert `NULL` unverändert bleiben soll. Der Name der Identitätsvariable kann durch den Befehl `IDCols` definiert werden, wenn diese anders als „ID“ genannt wird. Für die Analyse mit Individualdaten muss zusätzlich die Identitätsvariable des Individualdatensatzes angefügt werden, sodass der eingegebene Vektor der Länge zwei die entsprechenden Namen der beiden Identitätsvariablen enthält. Im Unterabschnitt 4.2.5 des Kapitels 4 wurde be-

reits beschrieben, in welcher Form die Datensätze vorliegen müssen, um die Analyse durchführen zu können. Als Beispiel hat der Autor ein Datenset integriert, der durch `data(topleveldat)` abrufbar ist. Dieser umfasst zwei Musterdatensätze. Die beiden, `aggr` und `indi`, Datensätze sind in einer reduzierten Form in der Abbildung 5.1 dargestellt. Neben der Eingabe der Daten muss noch das Verhältnis der Ergebnisse der zweiten Wahl zu den Ergebnissen der ersten Wahl durch eine Formel definiert werden. Für die Datensätze aus der Abbildung 5.1 kann diese beispielsweise durch

```
Form <- cbind(CSU_2, SPD_2, LINK_2, GRUN_2) ~ cbind(CSU_1, SPD_1, Link_1)        6
```

festgelegt werden. (Schlesinger, 2014, S. 9 ff.; 2013, S. 38)

Beispieldatensatz der Aggregatdaten (`aggr`)

	ID	CSU_1	SPD_1	Link_1	CSU_2	SPD_2	LINK_2	GRUN_2
1	1	2327	1447	194	1925	1015	274	754
2	2	883	674	78	742	405	129	359
3	3	8867	5946	684	7349	3855	1191	3102

Beispieldatensatz der Individualdaten (`indi`)

	ID	CSU_1.CSU_2	CSU_1.SPD_2	CSU_1.LINK_2	CSU_1.GRUN_2	...	Link_1.GRUN_2
1	13	11	56	26	22	...	30
2	18	37	39	59	56	...	5
3	5	16	55	41	61	...	20

Abbildung 5.1: *Beispieldatensätze der Individual- und Aggregatdaten aus dem* `eiwild` *Paket (Schlesinger, 2014, S. 19).*

5.1.2 *Sample, Burn-In* und *Thinning*

Die gewünschte Stichprobengröße der Kette kann mit dem Befehl `sample` eingefügt werden. Diese entspricht der Anzahl an Ziehungen, die nach dem *Burn-In* und *Thinning* gespeichert werden soll. Die gesamte Anzahl der durchgeführten Iterationen im `eiwild` Paket ist gleich dem eingegebenen `burnin` Wert addiert zum Produkt von `sample` und `thinning`. Mit den Defaultwerten `thinning = 1` und `burnin = 0` ist somit `sample` identisch zur gesamten Anzahl der Iterationen. Der Befehl `verbose` erlaubt das Monitoring während des Prozesses, indem, abhängig vom eingegebenem Wert, auf dem Bildschirm angezeigt wird, wie viele Iterationen bereits durchgeführt wurden. (Schlesinger, 2014, S. 9)

5.1.3 Varianz und Akzeptanzwahrscheinlichkeit

Die Befehle `betaVars` und `alphaVars` ermöglichen, die Varianzen der Vorschlagsdichten von β und α im Voraus festzulegen. Diese beeinflussen die Akzeptanzwahrscheinlichkeit, welche durch kleine Varianzen zu hoch werden kann und umgekehrt. Bei der kleinen Varianz und hohen Akzeptanzwahrscheinlichkeit macht die Markov-Kette viele kleine Schritte und erfasst damit nicht den ganzen Wertebereich. Hingegen wird bei der hohen Varianz und niedriger Akzeptanzwahrscheinlichkeit der breitere Wertebereich berücksichtigt, währenddessen die neuen Werte zu oft abgelehnt werden. Dadurch kann sich die Kette bei einem Wert zu lang halten. (Schlesinger, 2013, S. 43)

Um diese Umstände zu vermeiden, empfiehlt Schlesinger (2013) mithilfe der Funktion

```
tuneVars(form, aggr, indi=NULL,  IDCols=c("ID"),                              7
         whichPriori="gamma", prioriPars=list(shape=4, rate=2),               8
         accRat=c(0.4, 0.6), minProp=0.7, maxiter=20, sample=10000,           9
         verbose=10000, verboseTune=TRUE, improv=NULL,                        10
         betaVars=NULL, alphaVars=NULL,                                       11
         startValsAlpha=NULL, startValsBeta=NULL, ...)                        12
```

vor der Analyse die optimalen Varianzen zu finden. Alle Befehle bis auf `sample` und `verbose`, die identisch wie bei der `indAggEi()` Funktion sind, sollen hier in gleicher Weise definiert werden. Der Stichprobenumfang wurde hingegen bei der `tuneVars()` Funktion von dem Autor durch `sample` auf den Defaultwert von 10 000 gesetzt und muss nicht nach der Maßgabe des Stichprobenumfangs bei der `indAggEi()` bestimmt werden. Hierzu ist `verbose` beliebig zu wählen. Die `tuneVars()` Funktion hat noch zusätzlich fünf spezifische Befehle. Der Befehl `accRat` gibt den Bereich ein, in dem sich die Akzeptanzraten befinden sollen und `minProp` legt den Anteil der Parameter fest, die sich in diesem Bereich befinden sollen. Eine Stichprobe wird demnach entweder so lang wieder gezogen, bis das vorgegebene `minProp` erreicht wurde, oder bis zum Durchlauf aller Iterationen, deren Anzahl durch `maxiter` bestimmt wird. Für die Analyse in dieser Arbeit wird `minProp` auf 0.8 und `maxiter` auf 50 gesetzt. Zur Beschreibung der restlichen Befehle siehe die Literaturangabe. Nachdem der durch die `tuneVars()` erzeugte *Output* als Objekt gespeichert wurde, beispielsweise unter dem Name `tune`, können letztendlich die optimierten Varianzen mit `betaVars=tune[["betaVars"]]` und `alphaVars=tune[["alphaVars"]]` ins Modell integriert werden. (Schlesinger, 2014, S. 9; Schlesinger, 2013, S. 44 f.)

5.1.4 Hyperpriori-Parameter und Priori-Wissen

Nach der Untersuchung des Einflusses der Gamma-Hyperpriori-Verteilung von α_{rc} auf die Dirichlet-Priori-Verteilung von β^i_{rc}, nimmt Schlesinger (2013) für die Hyperpriori-Parameter die Defaultwerte $Gamma(\lambda_1 = 4, \lambda_2 = 2)$ aus dem `eiPack` Paket (Lau et al., 2012). Diese Wahl ermöglicht einen möglichst breiten Wertebereich für die Ziehungen von β^i_{rc} (Schlesinger, 2013, S. 55). Dennoch warnt der Autor, dass die Bestimmung einer vollständig nichtinformativen Priori-Verteilung für die Fälle mit $C > 2$ nicht möglich ist. Für die Schätzung der Übergangswahrscheinlichkeiten der Loyalen ist bei der Wählerwanderungsanalyse vor allem problematisch, dass bei mehreren betrachteten Parteien der zweiten Wahl hohe Werte von β^i_{rc} selten gezogen werden (Schlesinger, 2013; Klima et al., 2016). Gemäß Klima et al. (2016) wird dieser Effekt mit der Zunahme der betrachteten Parteien noch stärker. Die Ergänzung des Multinomial-Dirichlet-Modells im `eiwild` Paket (Schlesinger, 2014) ermöglicht die Hyperpriori-Parameter zellspezifisch zu definieren. Dadurch lässt sich Vorwissen ins Modell integrieren, respektive die informative Priori wird eingesetzt, um einen höheren Bereich für die Loyalen zu erzielen. Falls bekannt ist, dass einige Parteien eine hohe Unterstützung bekommen, erwarten Klima et al. (2016, S. 10) einen sinnvollen Verteilungsbereich zwischen 0.6 und 1 für die Loyalen. Anhand von Simulationsstudien berichten sie, dass eine Verbesserung der Schätzung durch eine informative Priori erzielt werden kann, wenn die Anzahl an Wahlbezirken oder Wahlkreisen niedrig ist.

Mithilfe der Funktion

```
prioriPlot(pars, which, cols, alphaSample=10000, betaSample=300, plot=TRUE)      13
```

kann im `eiwild` Paket der Einfluss der Hyperpriori-Verteilung von α_{rc} auf die Priori-Verteilung von β^i_{rc} grafisch dargestellt und untersucht werden. Die Funktion führt zuerst eine Simulation durch, um $\alpha_{r1}, \ldots, \alpha_{rC}$ aus Hyperpriori-Verteilung und $\beta^i_{r1}, \ldots, \beta^i_{rC}$ aus $Dir(\alpha_{r1}, \ldots, \alpha_{rC})$ zu ziehen. Falls nicht anders definiert, werden 10 000 Iterationen (`alphaSample`) durchgeführt und bei jeder Iteration werden $\beta^i_{r1}, \ldots, \beta^i_{rC}$ (`betaSample`) 300 mal gezogen. Damit werden letztendlich `alphaSample` $\times$ `betaSample` Ziehungen von β^i_{rc} gespeichert und grafisch dargestellt. Nach der visuellen Untersuchung lassen sich Hyperpriori-Parameter λ_1 und λ_2 durch den Befehl `prioriPars` in Funktionen `indAggEi()` und `tuneVars()` zellspezifisch definieren. Für die Datensätze aus der Abbildung 5.1 können beispielsweise mit einer Liste, die zwei Matrizen enthält, die Para-

meter wie folgt bestimmt werden:

```
PrioriPars <- list(shape =matrix(c(30, 4, 4, 4,                                    14
                                   4, 30, 4, 4,                                    15
                                   4, 4, 30, 4), nrow=3, ncol=4, byrow=TRUE),      16
                   rate =matrix(c(1, 2, 2, 2,                                      17
                                  2, 1, 2, 2,                                      18
                                  2, 2, 1, 2), nrow=3, ncol=4, byrow=TRUE))        19
```

Die Parameter λ_1^{rc} und λ_2^{rc} werden durch die Befehle `shape` und `rate` definiert. (Schlesinger, 2014, S. 11, 14; Schlesinger, 2013, S. 51 f.)

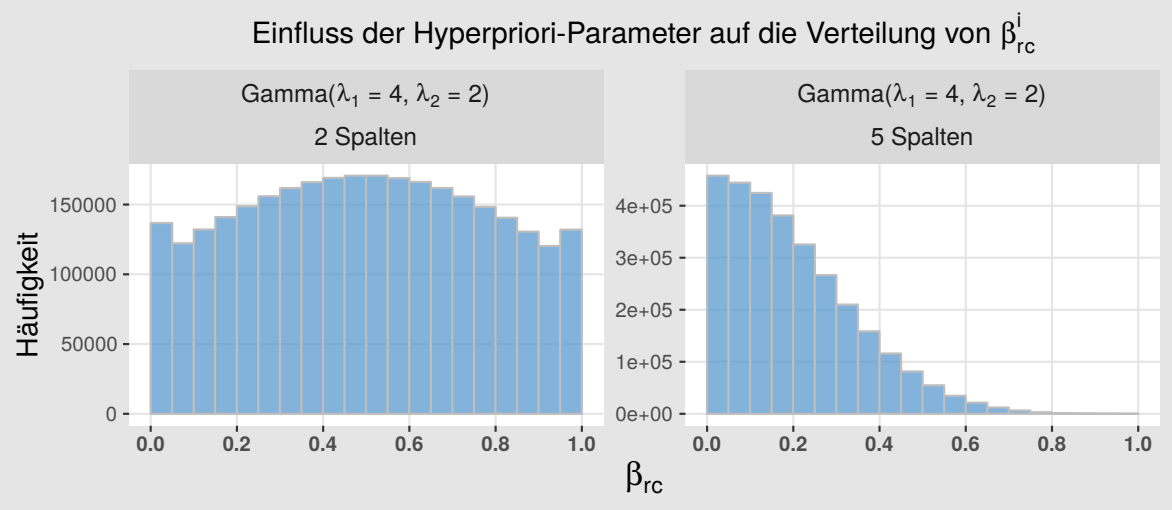

Abbildung 5.2: *Einfluss der Defaultwerte der Hyperpriori-Parameter* $Gamma(\lambda_1 = 4, \lambda_2 = 2)$*, auf die Verteilung von* β_{rc}^i *im* `eiwild` *Paket bei zwei Spalten (links) und bei fünf Spalten (rechts).*

Die Grafiken der Priori-Verteilung sind in dieser Arbeit mit dem `ggplot2` Paket (Wickham, 2009) hergestellt worden, wobei die Ziehungen mithilfe der `prioriPlot()` Funktion erzeugt wurden, um einen Datensatz für die `ggplot2` Grafik zu erstellen. In der Abbildung 5.2 ist ein Vergleich der Fälle $C = 2$ und $C = 5$ Spalten mit Defaultwerte $Gamma(\lambda_1 = 4, \lambda_2 = 2)$ dargestellt. Da die gezogenen Verteilungen für jede Spalte c und für eine beliebige Anzahl der Zeilen R gleich sind, wird in der Abbildung 5.2 lediglich eine Grafik für den Fall mit zwei Spalten (links) und eine Grafik für den Fall mit fünf Spalten (rechts) dargestellt. Die Anzahl der Zeilen R hat aufgrund der zeilenweisen Verteilungsannahme keinen Einfluss auf die Verteilung von β_{rc}^i (Schlesinger, 2013, S. 52). Die Grafik bestätigt, dass für $C = 2$ mit den Defaultwerten der Hyperpriori-Parameter eine gleichmäßige, nichtinformative Priori-Verteilung erzeugt wird. Die rechtsschiefe Verteilung für $C = 5$ zeigt, dass die Werte über 0.6 kaum gezogen werden. Die betrachteten Datensätze bestehen insgesamt aus fünf Kategorien für die zweite Wahl. Aufgrund

der geringen Anzahl an Wahlbezirken und hohen Anzahl an Kategorien könnten die Loyalen unterschätzt werden. Weiterhin wird in der Abbildung 5.3 visuell untersucht, ob für $C = 5$ die Werte von $Gamma(\lambda_1 = 30, \lambda_2 = 1)$ und $Gamm(\lambda_1 = 30, \lambda_2 = 2)$ für die Loyalen (links) die Verteilung von β^i_{rc} verbessern können. Für die anderen vier Spalten (rechts) sind dabei die Defaultwerte $Gammma(\lambda_1 = 4, \lambda_2 = 2)$ verwendet worden. Mittels der `prioriPlot()` Funktion wird für zellspezifisch definierte Parameter, für jede Spalte getrennt, eine Grafik erzeugt. Hier sind die gleichen Spalten $c = 2, 3, 4, 5$ wiederum in einer Grafik dargestellt. Eine Verbesserung des Verteilungsbereiches für die Loyalen und ein schmalerer Wertebereich für die anderen zeigt sich in beiden Fällen. Der Bereich zwischen 0.6 und 1, der von Klima et al. (2016) empfohlen wurde, wird durch die Werte $Gamma(\lambda_1 = 30, \lambda_2 = 1)$ annähernd erreicht. Eine Mehrheit der gezogenen Werte liegt dabei zwischen 0.7 und 0.9. Ein etwas breiterer Bereich ergibt sich durch $Gamma(\lambda_1 = 30, \lambda_2 = 2)$, wobei größtenteils die Werte zwischen 0.5 und 0.8 gezogen werden.

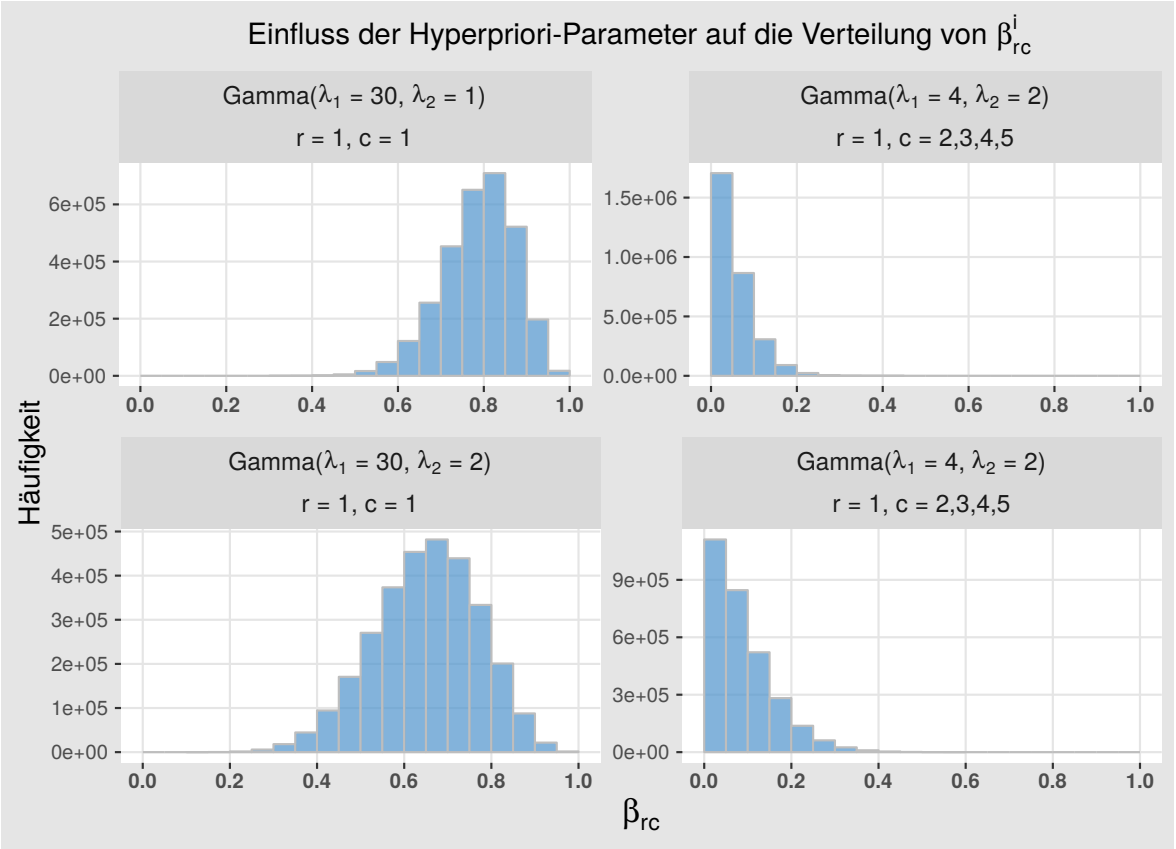

Abbildung 5.3: *Einfluss der Hyperpriori-Parameter auf die Verteilung von β^i_{rc} bei einer Zeile und fünf Spalten im* `eiwild` *Paket. Oben: Hyperpriori-Parameter $Gamma(\lambda_1 = 30, \lambda_2 = 1)$ (links) für eine Zelle und Defaultwerte für die übrigen Vier (rechts). Unten: Hyperpriori-Parameter $Gamma(\lambda_1 = 30, \lambda_2 = 2)$ (links) für eine Zelle und Defaultwerte für die übrigen Vier (rechts).*

5.2 Multinomial-Log-Normal-Modell

Für die Wählerwanderungsanalyse anhand des Multinomial-Log-Normal-Modells werden im `RxCEcolInf` Paket von Greiner et al. (2013) zwei verschiedene Funktionen für die Analyse mit und ohne Individualdaten verwendet. Demzufolge kann die ökologische Inferenz mithilfe der Funktion

```
Analyze(fstring, rho.vec, data =NULL,
        num.iters=1e+06, save.every =1000, burnin =10000,
        mu.vec.0=rep(log((0.45/(mu.dim -1))/0.55),  mu.dim),
        kappa=10, nu =(mu.dim+6), psi =mu.dim,
        mu.vec.cu=runif(mu.dim, -3,  0), NNs.start =NULL,
        THETAS.start=NULL, ...,
        print.every=10000)
```

durchgeführt werden. Die Funktion für die Analyse mit Individualdaten

```
AnalyzeWithExitPoll(fstring, rho.vec, exitpoll, data =NULL,
                    num.iters=1e+06, save.every =1000, burnin =10000,
                    ..., MMs.start=NULL, ...,
                    print.every=10000)
```

unterscheidet sich grundsätzlich nur in der Bezeichnung und in ein paar Befehle. Demnach beziehen sich die folgenden Beschreibungen auf die beiden Funktionen, falls nicht anders angegeben. Die Befehle `mu.vec.cu`, `NNs.start` und `THETAS.start` ermöglichen, in der gleichen Abfolge, die Startwerte für den Vektor μ, für die absoluten Häufigkeiten der inneren Zellen N_{rc}^i und für die Wahrscheinlichkeiten der inneren Zellen θ_{rc}^i zu bestimmen. Bei dem Hybridmodell lassen sich zusätzlich durch `MMs.start` die Startwerte für die unbeobachteten Werte, das heißt für die Differenzen $N_{rc}^i - n_{rc}^i$, bestimmen. Die Autoren empfehlen die Verwendung der Defaultwerte, wobei Startwerte zufällig gezogen werden. (Greiner et al., 2013, S. 3 f., 10 ff.)

5.2.1 Die Datensätze

Die Eingabe der Daten erfolgt durch den Befehl `data` für die Aggregatdaten und `exitpoll` für die Individualdaten. Die Ergebnisse der Nachwahlbefragung müssen hierfür in einer bestimmten Form vorliegen. Die Endform wurde bereits im Unterabschnitt 4.2.5 des Kapitels 4 beschrieben und in der Abbildung 4.8 dargestellt. Zusätzlich werden in der Abbildung 5.4 die Daten, die im `RxCEcolInf` Paket simuliert wurden, in einer reduzierten Form präsentiert. Die Beispieldaten lassen sich durch die Funktion

`SimData <- gendata.ep()` generieren und mit dem Befehl `SimData$GQdata` für die Aggregatdaten oder `SimData$EPInv$returnmat.ep` für die Individualdaten abrufen (Greiner et al., 2013, S. 20). Die Identitätsvariable ist im `RxCEcolInf` Paket nicht von Relevanz, da die Zeilen der Individualdaten und deren Reihenfolge an die Zeilen der Aggregatdaten angepasst werden müssen.

Simulierte Aggregatdaten (`SimData$GQdata`)

	bla	whi	his	Dem	Rep	Abs
precinct1	279	723	36	170	201	667
precinct2	2	15	1016	155	211	667
precinct3	123	262	630	76	249	690
precinct4	105	69	849	231	255	537
precinct5	1	348	697	92	210	744
precinct6	17	1	1022	171	127	742
precinct7	64	920	7	69	190	732
precinct8	186	827	0	112	340	561
precinct9	546	400	82	284	81	663
precinct10	384	622	9	133	328	554

Simulierte Individualdaten (`SimData$EPInv$returnmat.ep`)

	KK.bla.Dem	KK.bla.Rep	KK.bla.Abs	KK.whi.Dem	...	KK.his.Abs
[1,]	0	0	0	0	...	0
[2,]	0	0	0	0	...	0
[3,]	0	0	0	0	...	0
[4,]	0	0	0	0	...	0
[5,]	0	0	0	0	...	0
[6,]	0	0	0	0	...	0
[7,]	0	0	0	0	...	0
[8,]	0	0	0	0	...	0
[9,]	19	3	21	5	...	2
[10,]	0	0	0	0	...	0

Abbildung 5.4: *Eine verkürzte Darstellung der simulierten Beispieldatensätze aus dem* `RxCEcolInf` *Paket (Greiner et al., 2013, S. 20).*

Das Verhältnis zwischen den Parteien oder den Kandidaten der ersten und der zweiten Wahl wird durch den Befehl `fstring` eingegeben. Für die simulierten Datensätze in der Abbildung 5.4 lässt sich diese beispielsweise durch

```
Fstring <- "Dem, Rep, Abs ~ bla, whi, his"                    31
```

bestimmen (Greiner et al., 2013, S. 14). Die Reihenfolge der eingegebenen Parteien oder Kandidaten spielt hierbei eine wesentliche Rolle, da die letzte Spalte als Referenzkategorie automatisch gewählt wird (Greiner et al., 2013, S. 6). Das heißt, man kann durch Veränderung der Reihenfolge eine andere Spalte als Referenzkategorie bestimmen. Es lohnt sich ferner hinzuweisen, dass bei der Eingabe kein zusätzlicher Abstand auftreten darf. Wird in einem *String Character* in `R` ein Zeilenbruch vorgenommen, so wird auto-

matisch \n eingefügt. Das heißt, die Formel muss innerhalb der Anführungszeichen in einer Zeile eingegeben werden. Im Fall von großen R und C kann das unpraktisch sein und zur Unübersichtlichkeit des Codes führen. Alternativ lässt sich die Formel in mehreren Zeilen eintragen und im Nachhinein korrigieren. Beispielsweise können mittels der Funktion

```
str_replace_all(Fstring, "\n", "")
```

aus dem `stringr` Paket (Wickham, 2015b) die Abstände aus dem gespeichertem *String* Objekt gelöscht werden. Greiner et al. (2013) thematisieren diese Problematik bei der Beschreibung des Paketes nicht. Sie stellen sogar einige Beispiele dar, deren Formeln in zwei Zeilen geschrieben sind (Greiner et al., 2013, S. 2 f., 9) und somit den Fehler

```
Error in `colnames<-`(`*tmp*`, value =c("Bosley", "Roberts", "Ribaudo", :
length of 'dimnames' [2] not equal to array extent
```

erzeugen.

5.2.2 *Sample, Burn-In* und *Thinning*

Mit dem Befehl `num.iters` wird im `RxCEcolInf` Paket die Anzahl aller Iterationen definiert. *Thinning* kann durch `save.every` und *Burn-In* durch `burnin` bestimmt werden. Dabei muss die Anzahl der Iterationen nach dem *Burn-In* (`num.iters-burnin`) durch den Wert von *Thinning* (`save.every`) teilbar sein. Die endgültige Stichprobengröße oder *Sample* lässt sich schließlich durch `(num.iters-burnin)/save.every` berechnen. Die Defaultwerte `num.iters = 1000000`, `save.every = 1000` und `burnin = 10000` liefern somit eine Stichprobe der Größe 990. Mit dem Befehl `print.every` kann bestimmt werden, wie viele Iterationen des Prozesses auf dem Bildschirm gezeigt werden. (Greiner et al., 2013, S. 4 f., 9)

5.2.3 Varianz und Akzeptanzwahrscheinlichkeit

Für die ökologische Inferenz wird mittels der Funktion

```
Tune(fstring, data=NULL,
     num.runs=12, num.iters=10000,
     rho.vec=rep(0.05, ntables),
     kappa=10, nu=(mu.dim+6), psi=mu.dim,
     mu.vec.0=rep(log((.45/(mu.dim-1))/.55), mu.dim),
     mu.vec.cu=runif(mu.dim, -3, 0), ...)
```

und für das Hybridmodell mittels der Funktion

```
TuneWithExitPoll(fstring, exitpoll, data =NULL, num.runs=12,                    39
                 num.iters=10000, rho.vec =rep(0.05,  ntables), ...)            40
```

ein Vektor generiert, der mit der Kovarianzmatrix Σ multipliziert werden kann, um die Akzeptanzwahrscheinlichkeiten zu optimieren. Die Eingabe von `fstring`, `data` und `exitpoll` erfolgt analog zur Funktion `Analyse()` oder `AnalyseWithExitPoll()`. Der Befehl `num.runs` gibt die Anzahl der Wiederholungen an und der Befehl `num.iters` bestimmt die Anzahl der Iterationen. Für die Analyse in dieser Arbeit wird `num.runs` auf 50 Wiederholungen gesetzt und die vorgegebene Anzahl der Iterationen von 10 000 wird verwendet. In die Funktion `Tune()` oder `TuneWithExitPoll()` lassen sich durch `rho.vec` die Startwerte dieses Vektors bestimmen. Diese werden dann mittels der Funktion angepasst, um die Akzeptanzwahrscheinlichkeiten zwischen 0.2 und 0.5 für die Ziehungen von θ^i_{rc} zu erlangen. Ist der *Output* der Funktion beispielsweise unter dem Namen `Tune_LN` gespeichert, so lässt sich der optimierte Vektor durch die Eingabe von `rho.vec = Tune_LN$rhos` in die Funktion `Analyse()` oder `AnalyseWithExitPoll()` integrieren. (Greiner et al., 2013, S. 22, 25)

5.2.4 Hyperpriori-Parameter

Greiner et al. (2013) ermöglichen dem Benutzer die vorbestimmten Hyperpriori-Parameter bei der Analyse und bei der Varianzanpassung (*Tuning*) zu ändern. Das Einbeziehen des Vorwissens durch zellspezifische Bestimmung ist dennoch nicht möglich. Deswegen werden bei der Analyse anhand des Multinomial-Log-Normal-Modells die Defaultwerte akzeptiert. Für die Normal-Hyperpriori-Verteilung von dem Priori-Parameter μ bestimmt der skalare Wert `kappa = 10` (κ) die Diagonale der Kovarianzmatrix und der Vektor `mu.vec.0 = rep(log((0.45/(mu.dim - 1))/0.55), mu.dim)` (μ_0) die Mittelwerte. Für die Inverse-Wishart-Hyperpriori-Verteilung von dem Priori-Parameter Σ werden die Freiheitsgrade mit `nu = (mu.dim + 6)` (ν_0) und die Diagonale der Matrixparameter durch den skalaren Wert `psi = mu.dim` (ψ) definiert.

6 Ergebnisse

In diesem Kapitel erfolgt die Konvergenzdiagnose der erzeugten Ketten, Vergleich der Ketten und Modelle und letztendlich die Darstellung der Ergebnisse eines gewählten Modells. Die Ketten der beiden betrachteten Modelle werden zuerst mit zehn Millionen Iterationen für jede Version des Modells berechnet. Damit kann untersucht werden, ob das Modell konvergiert und welches *Thinning* und *Burn-In* für ein *Sample* von 1 000 Ziehungen geeignet ist. Dem Grunde nach werden danach für jedes Modell drei verdünnte Ketten erzeugt. Deren Vergleich soll erkennen lassen, ob bei mehreren Durchführungen die gleichen Ergebnisse erzeugt werden. Die Güte eines Modells lässt sich nicht überprüfen, da die wahren Übergangswahrscheinlichkeiten nicht bekannt sind. Deswegen kann ein Modell nur auf Grund der Konvergenzdiagnose und des Vergleichs der Ergebnisse innerhalb und zwischen den Modellen gewählt werden. Die subjektive Auswertung spielt dabei auch eine Rolle und kann nicht vermieden werden.

Die Konvergenz der Ketten wird visuell untersucht. Dafür sind die grafische Darstellungen der *Density* (Dichte) und *Trace Plots* der gezogenen absoluten Häufigkeiten der inneren Zellen nützlich. Die absoluten Häufigkeiten werden im `eiwild` (Schlesinger, 2014) mit `counts` und im `RxCEcolInf` (Greiner et al., 2013) mit `NNs` bezeichnet. Für die Erstellung der Grafiken in diesem Kapitel kommen die folgenden R-Pakete zum Einsatz: `ggplot2` (Wickham, 2009), `ggthemes` (Arnold, 2016), `scales` (Wickham, 2016), `gridExtra` (Auguie, 2016), `grid` (R Core Team, 2015), `RColorBrewer` (Neuwirth, 2014), `circlize` (Gu et al., 2014; Gu, 2015), `stringr` (Wickham, 2015b) und `reshape2` (Wickham, 2015a). Da sich einige Grafiken über ganze Seiten erstrecken, lassen sich diese nicht passend in den Text integrieren. Deswegen werden alle Grafiken in folgenden Abschnitten am Ende des dazugehörigen Abschnitts oder Unterabschnitts dargelegt.

Aus den erzeugten Ketten können die Posteriori-Mittelwerte der absoluten Häufigkeiten für jede Zelle der Wählerwanderungstabelle berechnet werden. Je nach Interesse lassen sich danach die relativen Häufigkeiten oder Übergangswahrscheinlichkeiten bestimmen. Um die Differenzen zwischen den Ergeb-

nissen von verschiedenen Modellen oder Ketten darstellen zu können, werden zwei Vergleichsmaße verwendet. Der Hauptunterschied liegt darin, ob die Distanzen anhand von Übergangswahrscheinlichkeiten oder von relativen Häufigkeiten der inneren Zellen berechnet werden und wie sich die Differenzen letztendlich interpretieren lassen. Für zwei Übergangstabellen T^A_{ueber} und T^B_{ueber} schlägt Schlesinger (2013, S. 59) den *Mean Absolut Error* vor, der folgendermaßen definiert wird:

$$MAE(T^A_{ueber}, T^B_{ueber}) = \frac{1}{R \times C} \sum_{r=1}^{R} \sum_{c=1}^{C} |\beta_{rc}^{T^A_{ueber}} - \beta_{rc}^{T^B_{ueber}}| \tag{6.1}$$

Der *Mean Absolut Error* oder MAE kann Werte zwischen 0 und 1 annehmen und wird als durchschnittliche Differenz pro Zelle interpretiert. Ein alternatives Vergleichsmaß stellt die *Absolute Distanz* oder AD dar, die für zwei Wählerwanderungstabellen mit relativen Häufigkeiten, T^A_{rel} und T^B_{rel}, durch

$$AD(T^A_{rel}, T^B_{rel}) = \sum_{r=1}^{R} \sum_{c=1}^{C} |T^A_{rel}(r, c) - T^B_{rel}(r, c)| \tag{6.2}$$

berechnet werden kann (Klima et al., 2015, S. 9). Dieses Maß kann Werte zwischen 0 und 2 annehmen. Nach Halbierung lässt sich $AD/2$ als Anteil der Stimmen interpretieren, die innerhalb einer Tabelle umverteilt werden müssen, um zwei identische Tabellen zu erhalten. Hierbei bezeichnet *Absolut* in der Bezeichnung des Vergleichsmaßes, dass negative Differenzen in positive Werte transformiert werden. Das Ignorieren des Vorzeichens ist bei der Berechnung von beiden Vergleichsmaßen nötig, da sich die negativen und die positiven Werte ansonsten bei der Summierung gegenseitig entwerten würden. Über die Differenzen zwischen den Ergebnissen lassen sich im Prinzip auf Basis der beiden Vergleichsmaße ähnliche Schlussfolgerungen ziehen. Dadurch dass die Distanzen anhand von Übergangswahrscheinlichkeiten berechnet werden, betrachtet MAE alle Zeilen gleichwertig. Hingegen weisen die kleinen Parteien der ersten Wahl bei der Berechnung des AD niedrigere Differenzen auf als die großen Parteien.

6.1 Multinomial-Dirichlet-Modell

Die Simulationsstudie von Klima et al. (2016, S. 19) zeigt eine Verbesserung der Schätzung durch die Verwendung einer informative Priori bei kleiner Anzahl an Wahlbezirken. Zudem ist bekannt, dass *Dr. Peter Kurz* von der *SPD*, *Grünen* und *Die Linken* unterstützt wurde und dass die *CDU* Herrn *Peter Rosenberger* empfohlen hat

(Schredle, 2015). Dementsprechend werden bei der Schätzung anhand des Multinomial-Dirichlet-Modells (Rosen et al., 2001) die zellspezifischen Hyperpriori-Parameter der Gamma-Verteilung definiert (Schlesinger, 2013). Die Wähler, die bei der ersten Wahl eine Partei gewählt haben, die einen bestimmten Kandidaten unterstützt und diesem ihre Stimme in der zweiten Wahl gegeben haben, werden hierbei als Loyale betrachtet. Die Idee ist neben einem Modell mit Defaultwerten $Gamma(\lambda_1 = 4,\ \lambda_2 = 2)$ noch zwei Modelle mit Vorwissen, jeweils mit und ohne Individualdaten, zu berechnen und zu vergleichen. Für die Loyalen wird $Gamma(\lambda_1 = 30,\ \lambda_2 = 1)$ und $Gamma(\lambda_1 = 30,\ \lambda_2 = 2)$ verwendet, während für die übrigen Zellen die Defaultwerte unverändert bleiben. Der Einfluss dieser Parameter auf die Verteilung von β^i_{rc} ist bereits in der Abbildung 5.3 (Seite 52) des Kapitels 5 dargestellt. Zur Überprüfung der Konvergenz im Unterabschnitt 6.1.1 und zum Vergleich der Ketten im Unterabschnitt 6.1.2 werden neben den vier Zellen von Loyalen, `SPD.Kurz`, `Güne.Kurz`, `Linke.Kurz` und `CDU.Rosenberger`, noch `Sonstige_13.Sonstige_15` und `Nichtwähler_13.Nichtwähler_15` bei der grafischen Darstellung betrachtet.

6.1.1 Konvergenzdiagnose

In der Abbildung 6.1 (links) auf der Seite 61 befindet sich die Darstellung der *Trace of Counts* für die gewählten inneren Zellen des Multinomial-Dirichlet-Modells ohne Vorwissen. Von insgesamt zehn Millionen Iterationen wird lediglich jede tausendste abgebildet, um die Größe der zu speichernden Grafik zu reduzieren. Dabei wird die Nummer der durchgeführten und nicht der abgebildeten Iteration auf der x-Achse der Grafik dargestellt. Da die Berechnung der Mittelwerte die Speichergröße der Grafik nicht beeinflusst, werden diese aus allen gezogenen Werten jeweiliger Zellen berechnet und mit den waagerechten weißen Linien in der Grafik markiert. Die rechte Grafik zeigt die Dichten und die Mittelwerte der *Counts* der zweiten und der letzten Million aller Iterationen. Anstelle der ersten Million wird hier die Zweite dargestellt, um den Einfluss der Startwerte auf die Zellen `Sonstige_13.Sonstige_15` und `Nichtwähler_13.Nichtwähler_15` zu vermeiden. Diese drücken die Dichten auf die Seite, wodurch die Übersichtlichkeit der Dichten begrenzt ist. Das Ziel hierbei ist die Kettenteile am Anfang und am Ende zu vergleichen, um untersuchen zu können, ob die Kette stationär ist.

Aus den Grafiken ist visuell zu erkennen, dass das ökologische Multinomial-

Dirichlet-Modell ohne Vorwissen eine stationäre Verteilung wahrscheinlich erreicht hat. Die Zellen `Güne.Kurz` und `Linke.Kurz` weisen eine seht gute Übereinstimmung am Anfang und am Ende der Kette auf. Eine minimale Abweichung im Mittelwert zeigt sich am Anfang der Kette im Vergleich zu den wenig niedrigeren Werte am Ende der Kette bei der Zelle `SPD.Kurz`. Etwas höhere Werte werden am Anfang der Kette bei der Zelle `CDU.Rosenberger` gezogen. Bei der Zelle `Sonstige_13.Sonstige_15` tritt ein Abfallen im Verlauf der Kette auf. Solche Störungen sind jedoch bei kleinen Kategorien, bei denen wenig Daten vorhanden sind, zu erwarten. Der Mittelwert scheint jedoch nicht sehr stark davon beeinflusst zu sein. Die Kette der `Nichtwähler_13.Nichtwähler_15` Zelle ist wegen des Startwertes nach oben gedrückt. Dadurch lässt sich nicht genau erkennen, ob irgendwelche Störungen erscheinen. Die Dichte zeigt hierbei, dass die Werte bei der letzten Million etwas höher sind als bei der Zweiten. Obwohl die Startwerte bei `Sonstige_13.Sonstige_15` und `Nichtwähler_13.Nichtwähler_15` von dem Rest der Kette stark abweichen, wurde die Konvergenz bei allen Ketten schnell erreicht. Bereits ein *Burn-In* von 100 000, der mit senkrechten roten Linien in die Grafik gezeichnet wird, sollte hier reichen, um den Einfluss der Startwerte zu unterdrücken.

Um das geeignete *Thinning* zu bestimmen, wird ferner die Autokorrelation untersucht. In der Abbildung 6.2 (links) auf der Seite 62 zeigt sich nach dem *Burn-In* von 100 000 ohne *Thinning* für die Stichprobe von 1 000 Ziehungen eine hohe Autokorrelation bei allen Zellen. Obwohl einige Autoren dem *Thinning* kritisch gegenüber stehen (siehe Unterabschnitt 2.2.2 des Kapitels 2 auf der Seite 9), verwenden Klima et al. (2015, S. 9 f.) ein hohes *Thinning* bei Modellen der ökologischen Inferenz zur Behebung von starker Autokorrelation bei der Wählerwanderungsanalyse. Dementsprechend wird nach der Untersuchung ein *Thinning* von 2 000 angewendet, da diese die Korrelation ausreichend verringert (siehe Abbildung 6.2, rechts). Die übrigen Versionen des Multinomial-Dirichlet-Modells weisen ähnliche Entwicklungen und Merkmale bei der Konvergenz und bei der Autokorrelation auf. Demzufolge werden für alle Modelle identisches *Thinning* und *Burn-In* verwendet. Die gleichartige grafische Darstellungen der Ketten und der Autokorrelation sind für alle Versionen des Modells im Anhang A.2.1 (ab der Seite 93) und A.2.2 (ab der Seite 98) zu finden.

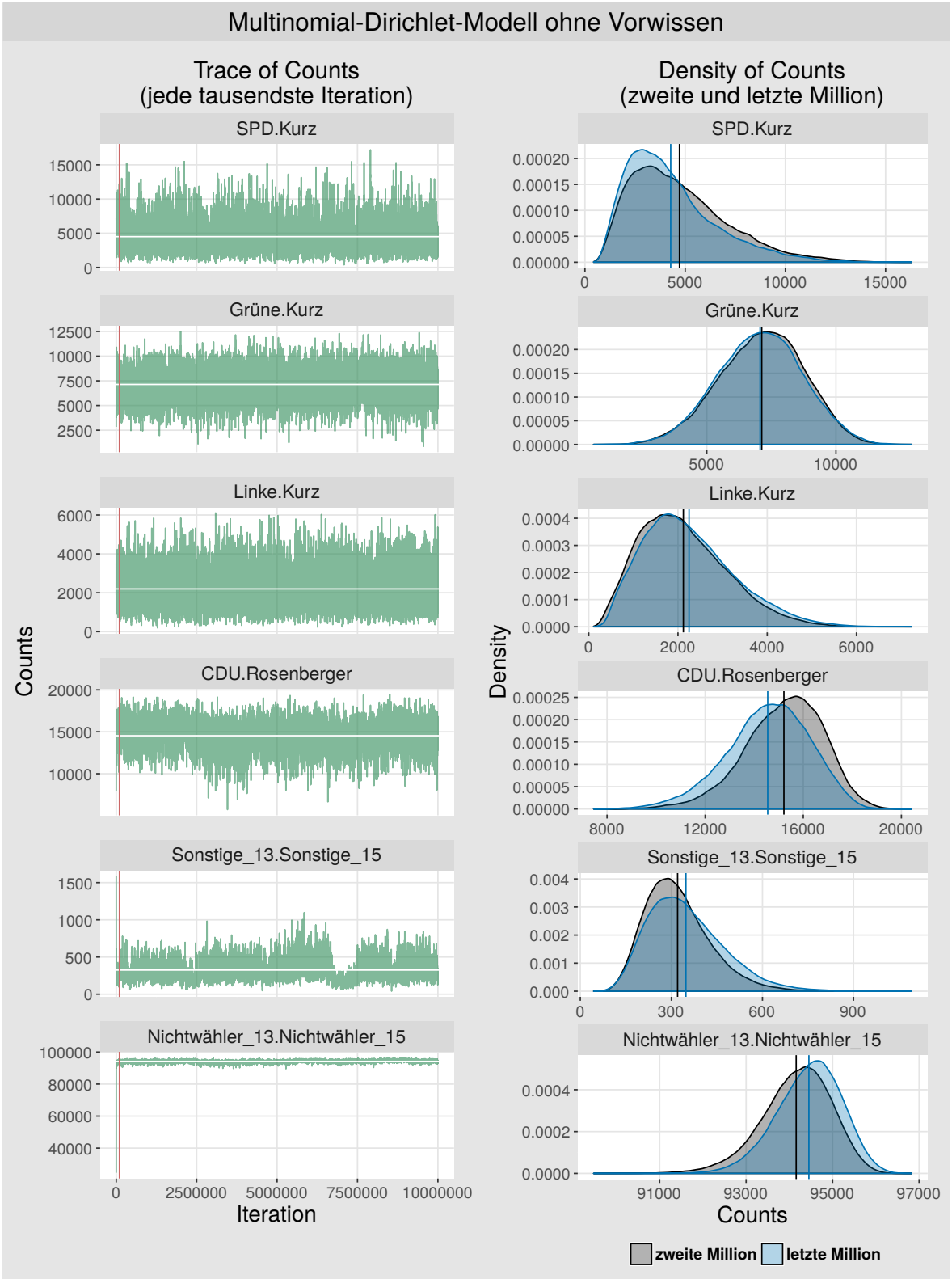

Abbildung 6.1: *Die Ketten (links) und die Dichten (rechts) der gezogenen absoluten Häufigkeiten (Counts) von fünf gewählten inneren Zellen des ökologischen Multinomial-Dirichlet-Modells ohne Vorwissen. Links: Von zehn Millionen Iterationen wird jede tausendste dargestellt. Die senkrechten roten Linien kennzeichnen die* 100 000*-ste von zehn Millionen Iterationen. Die waagerechten weißen Linien zeigen die Mittelwerte aller gezogenen Werte. Rechts: Die Dichten der zweiten und der letzten Million aller Iterationen und die dazugehörigen Mittelwerte (senkrechte Linien).*

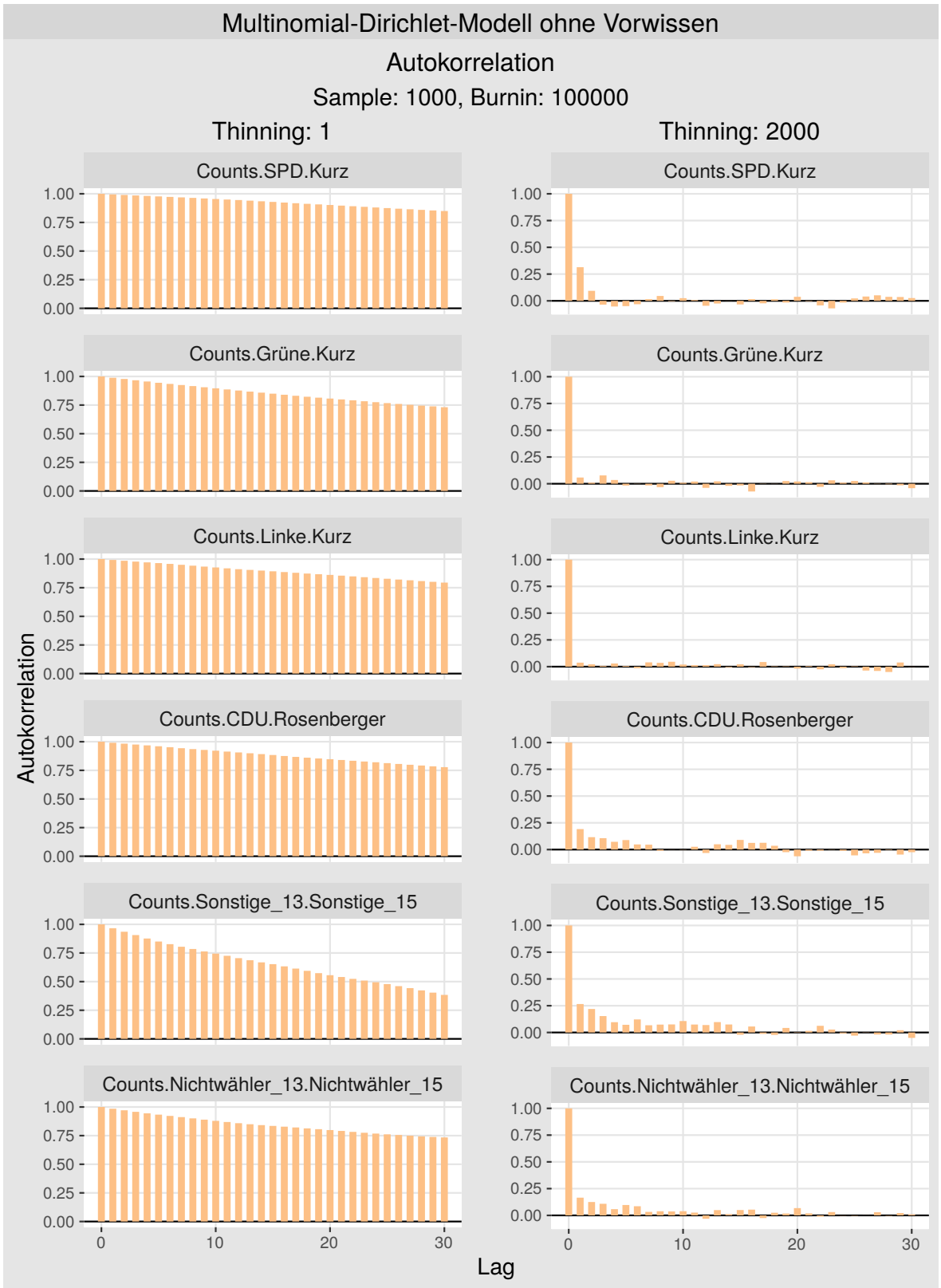

Abbildung 6.2: *Die Autokorrelationen der gezogenen absoluten Häufigkeiten (Counts) von fünf gewählten inneren Zellen des ökologischen Multinomial-Dirichlet-Modells ohne Vorwissen, anhand einer Stichprobe mit* 1 000 *Ziehungen nach dem Burn-In von* 100 000*. Links: Ohne Thinning. Rechts: Thinning von* 2 000*.*

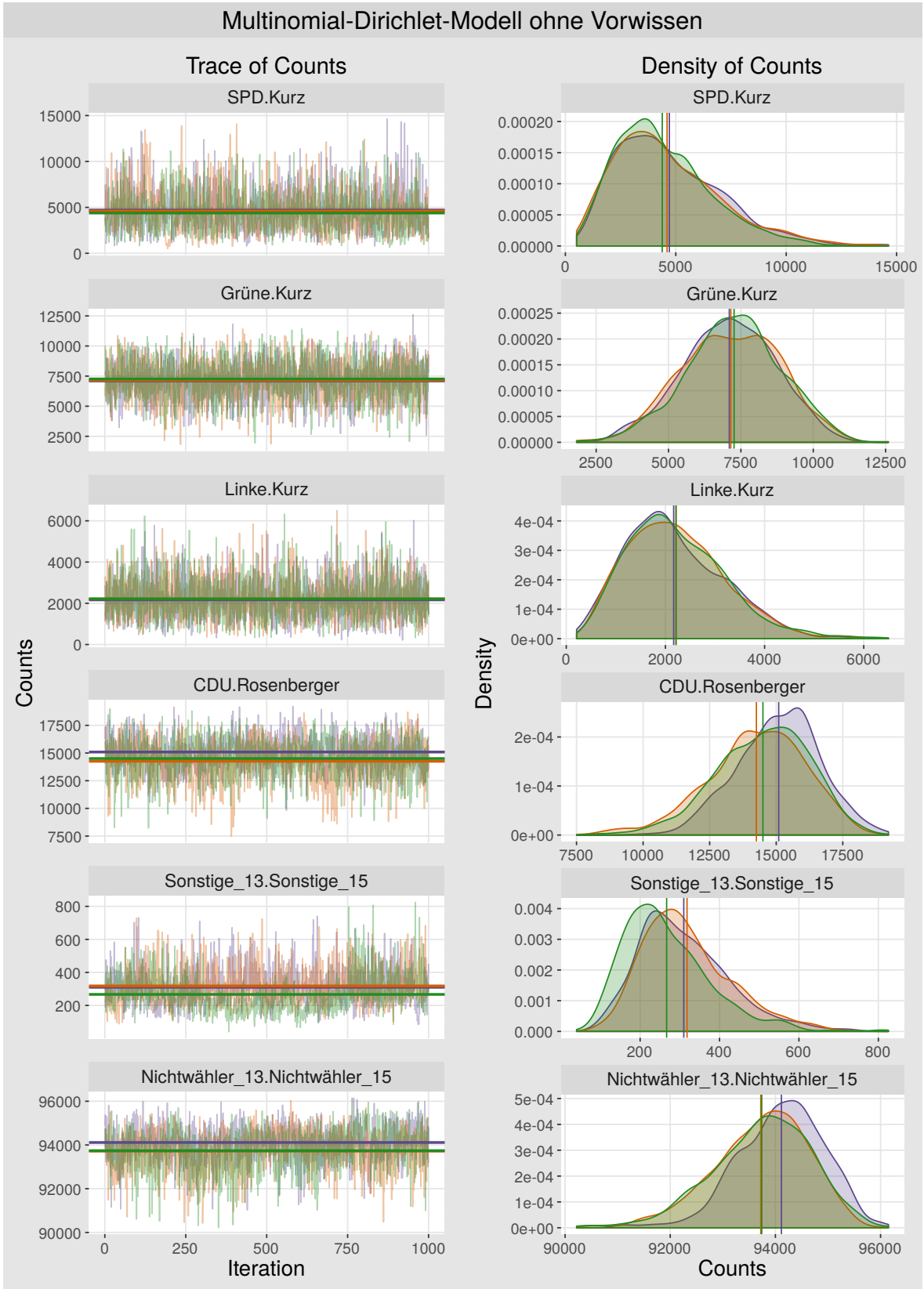

Abbildung 6.3: *Drei verdünnte Ketten der gezogenen absoluten Häufigkeiten (Counts) von fünf gewählten inneren Zellen des ökologischen Multinomial-Dirichlet-Modells ohne Vorwissen. Sample:* 1 000, *Burn-In:* 100 000 *und Thinning:* 2 000. *Links: Trace of Counts der drei Ketten und die dazugehörigen Mittelwerte (waagerechte Linien). Rechts: Dichten der verdünnten Ketten und die gleichen Mittelwerte senkrecht dargestellt.*

6.1.2 Ketten- und Modellvergleich

Drei Ketten, die nach dem *Burn-In* und *Thinning* für das Multinomial-Dirichlet-Modell ohne Vorwissen erzeugt sind, werden in der Abbildung 6.3 auf der Seite 63 dargestellt. Links in der Grafik befinden sich die Ketten, die sich entsprechend der Farbe unterscheiden lassen. Hierbei wird auf der x-Achse die Nummer der Iteration in der Stichprobe dargestellt. Rechts in der Grafik werden die Dichten der jeweiligen Ketten und deren Mittelwerte präsentiert und mit entsprechenden Farben differenziert.

Die Zellen `CDU.Rosenberger`, `Sonstige_13.Sonstige_15` und `Nichtwähler_13.Nichtwähler_15` zeigen etwas höhere Unterschiede zwischen den drei erzeugten Ketten als die übrigen dargestellten Zellen. Bei der Zelle `Sonstige_13.Sonstige_15` weist die grüne Kette im Mittel einen Abfallen im Vergleich zur orangen und violetten Kette auf. Bei einigen Versionen des Modells scheinen die Ketten, die hier eine Abweichung untereinander aufweisen, stabiler zu sein. Deren Darstellung befindet sich im Anhang A.2.3 (ab der Seite 103). Zusätzlich werden die Grafiken erzeugt, welche die drei verdünnten Ketten aller Zellen in einer Matrixform darstellen. Diese sind aus Gründen der Übersichtlichkeit im elektronischen Anhang dargelegt. Die Auflistung der Dateien, die in digitaler Version der Arbeit beigelegt werden, ist im Anhang E auf der Seite 117 zu finden.

Nach einer visuellen Untersuchung soll ferner der Vergleich anhand von absoluten Distanzen (AD) für jede Version des Modells präziser zeigen, wie hoch die Unterschiede zwischen den Ergebnissen über alle Zellen innerhalb der gleichen Modelle sowie zwischen verschiedenen Versionen des Modells sind. Zusätzlich werden die durchschnittlichen absoluten Differenzen pro Zelle (MAE) im Anhang A.2.4 auf der Seite 108 präsentiert. In der Abbildung 6.4 auf der Seite 66 sind die absoluten Distanzen zwischen den Ergebnissen von drei erzeugten Ketten innerhalb der Modelle in Prozentpunkten dargestellt. Die AD Werte sind hierbei über die Diagonale identisch beziehungsweise symmetrisch. Es zeigt sich, dass das Multinomial-Dirichlet-Hybridmodell bei allen Hyperpriori-Parameter die niedrigeren Distanzen zwischen den Ergebnissen der Ketten aufweist als das ökologische Multinomial-Dirichlet-Modell. Die Summe aller Distanzen der drei ökologischen Modelle beträgt hierbei 13.68 Prozent. Deutlich geringerer ist die Summe aller Distanzen der Hybridmodelle mit dem Wert von 4.5 Prozent. Von allen betrachteten Modellen zeigen sich bei den Hybridmodellen mit Vorwissen die geringsten Unterschiede zwischen den Ketten. Es ist anzumerken, dass eine Abnahme

von AD zwischen den Ketten durch die Integration des Vorwissens erreicht wird, allerdings nicht so stark wie durch das Verwenden der Individualdaten. Die Abnahme durch die Integration des Vorwissens ist jedoch innerhalb des ökologischen Modells mittels MAE nicht zu erkennen (siehe Abbildung A.24 im Anhang A.2.4 auf der Seite 108). Beim ökologischen Modell mit Vorwissen unterscheidet sich die erste Kette (violett) bei $Gamma(\lambda_1 = 30, \lambda_1 = 1)$ von der Zweiten und der Dritten stärker. Dieser Unterschied wird bei der AD unterdrückt, da ein starkes Abfallen der ersten Ketten bei der Zelle `AfD.Probst` auftritt. Außerdem liegt die erste Kette bei der Zelle `AfD.Nichtwähler` etwas höher als die Zweite und die Dritte. Anhand von AD werden die Distanzen bei kleinen Parteien, in diesem Fall *AfD*, niedriger als anhand von MAE. Beim ökologischen Modell mit Vorwissen $Gamma(\lambda_1 = 30, \lambda_1 = 2)$ unterscheidet sich die dritte Kette von der Ersten und der Zweiten stärker. Hier tritt eines Abfallen der dritten Kette (grün) bei der Zelle `Sonstige_13.Rosenberger` auf. Diese Störungen lassen sich visuell mittels der Darstellungen der Ketten aller Zellen erkennen, welche im elektronischen Anhang aufgeführt werden.

Die Abbildung 6.5 auf der Seite 66 präsentiert die absoluten Distanzen zwischen den Ergebnissen von verschiedenen Versionen des Modells. Analog ist die Matrix mit den AD Werte symmetrisch. Zum Vergleich wurde hierbei die erste der drei erzeugten Ketten für jede Version des Modells genommen. Wiederum zeigt sich, dass das Hybridmodell über alle Versionen stabiler ist als das ökologische Modell. Den größten Unterschied zu allen anderen Modellen weist das ökologische Modell ohne Vorwissen auf. Die Modelle mit Vorwissen weisen im Allgemeinen die niedrigsten absoluten Distanzen auf.

Weiterhin werden die zellspezifischen absoluten Differenzen der Modelle zur Nachwahlbefragung dargestellt. Um den Lesefluss dieser Arbeit nicht zu stören befindet sich die Abbildung 6.12 am Ende des Unterabschnittes 6.2 auf der Seite 74. In der Grafik sind die zellspezifischen Differenzen für jede Version des Modells in einem Box-Plot dargestellt. Unten sind die durchschnittlichen absoluten Differenzen (*Mean Absolut Error*) gegeben und mit „MAE“ in den abgebildeten Box-Plots eingezeichnet. Diese Grafik dient lediglich zum Vergleich und liefert keine Information über die Qualität der Schätzung. Erwartungsgemäß weisen die ökologischen Modelle fast doppelt so große Differenzen zur Nachwahlbefragung auf als die Hybridmodelle. Die geringsten zellspezifischen Differenzen und MAE zeigen sich bei dem Hybridmodell ohne Vorwissen.

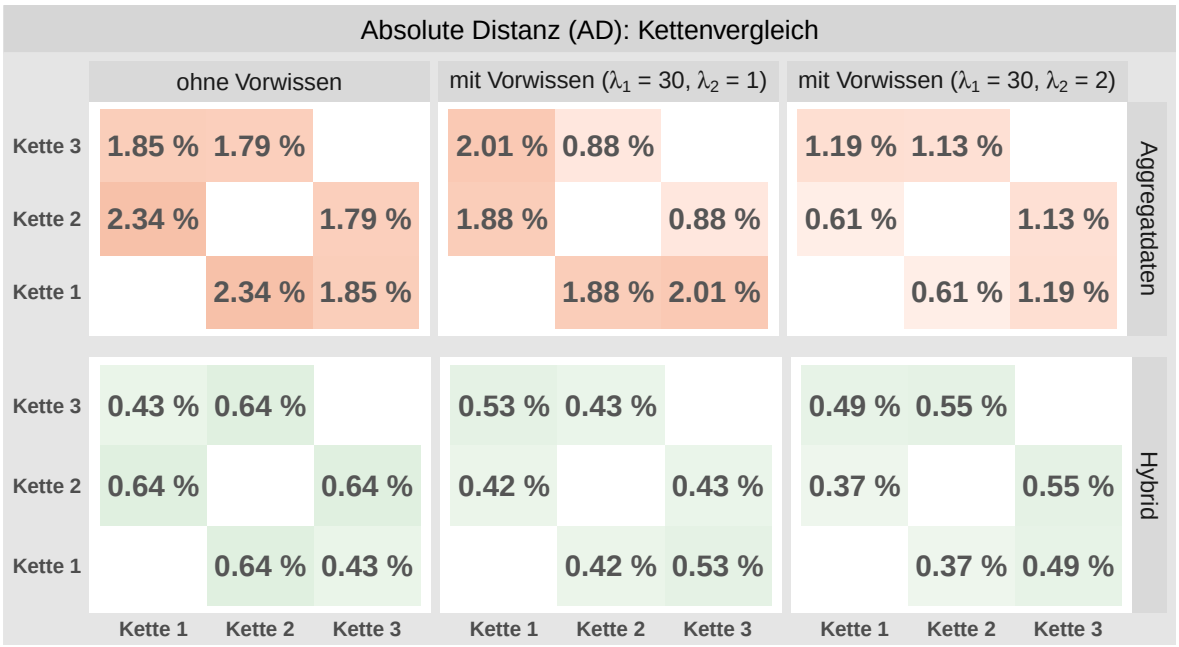

Abbildung 6.4: *Absolute Distanzen (AD) in Prozentpunkten zwischen den Ergebnissen der drei verdünnten Ketten für jede Version des ökologischen (oben) und des hybriden (unten) Multinomial-Dirichlet-Modells. Die Werte sind je nach Modell symmetrisch über die Diagonale.*

Absolute Distanz (AD): Modellvergleich

	Aggregatdaten	Aggregatdaten mit Vorwissen (30, 1)	Aggregatdaten mit Vorwissen (30, 2)	Hybrid	Hybrid mit Vorwissen (30, 1)	Hybrid mit Vorwissen (30, 2)
Aggregatdaten		23.6 %	20.26 %	18.85 %	22.42 %	21.27 %
Aggregatdaten mit Vorwissen (30, 1)	23.6 %		5.3 %	17.85 %	12.62 %	13.45 %
Aggregatdaten mit Vorwissen (30, 2)	20.26 %	5.3 %		14.35 %	10.59 %	10.8 %
Hybrid	18.85 %	17.85 %	14.35 %		7.56 %	6.04 %
Hybrid mit Vorwissen (30, 1)	22.42 %	12.62 %	10.59 %	7.56 %		1.68 %
Hybrid mit Vorwissen (30, 2)	21.27 %	13.45 %	10.8 %	6.04 %	1.68 %	

Abbildung 6.5: *Absolute Distanzen (AD) in Prozentpunkten zwischen den Ergebnissen von verschiedenen Versionen des Multinomial-Dirichlet-Modells (symmetrisch über die Diagonale). Zum Vergleich wurde die erste der drei verdünnten Ketten für jede Version des Modells verwendet.*

6.2 Multinomial-Log-Normal-Modell

Bei der Schätzung anhand des Multinomial-Log-Normal-Modells kann das Vorwissen nicht durch zellspezifische Parameter der Hyperpriori-Verteilung ins Modell integriert werden. Es sollte jedoch untersucht werden, ob die Wahl einer anderen Referenzkategorie die Schätzung beeinflusst. Infolgedessen werden hier insgesamt vier Versionen des Modells berechnet, zwei mit Referenzkategorie `Nichtwähler_15` und zwei mit Referenzkategorie `Kurz`, jeweils mit und ohne Individualdaten. Eine Kette mit zehn Millionen Iterationen, ohne *Burn-In* und *Thinning*, konnte mithilfe des `RxCEcolInf` (Greiner et al., 2013) Paketes nicht erzeugt werden, da der Fehler

```
Finished MCMC routine. Processing output...

Error in cbind(mu, Sigma, NNs, LAMBDA, TURNOUT, GAMMA, BETA):
long vectors not supported yet: bind.c:1304
Calls: Analyze -> cbind
```

aufgetreten ist. Deswegen wurde bereits bei der ersten Berechnung ein *Thinning* von 100 ohne *Burn-In* verwendet. Dadurch werden von den erzeugten zehn Millionen Iterationen schließlich 100 000 gespeichert. Im folgenden Unterabschnitt werden bei den grafischen Darstellungen die gleichen Zellen betrachtet wie beim Multinomial-Dirichlet-Modell.

6.2.1 Konvergenzdiagnose

In der Abbildung 6.6 wird im *Trace Plot* (links) von insgesamt 100 000 gespeicherten Iterationen lediglich jede zehnte aufgezeichnet, womit die gleiche Anzahl an Iterationen wie beim Multinomial-Dirichlet-Modell dargestellt wird. Auch wenn nur jede hundertste Iteration gespeichert werden konnte, wurden letztendlich 10 000 000 Iterationen durchgeführt. Dementsprechend werden nach wie vor die Nummern der durchgeführten Iterationen auf der x-Achse der Grafik angegeben. Die Mittelwerte konnten nur auf Basis der 100 000 gespeicherten Iterationen berechnet werden und sind mit waagerechten schwarzen Linien markiert. Die rechte Grafik zeigt die Dichten der absoluten Häufigkeiten (*Counts*) der zweiten und der letzten Million Iterationen, wobei wiederum nur jede hundertste Iteration betrachtet werden kann. Beim dargestellten ökologischen Multinomial-Log-Normal-Modell mit automatisch gewählter Referenzkategorie `Nichtwähler_15` wurde die Konvergenz scheinbar nicht erreicht. Eine stationäre Verteilung lässt sich über alle Zellen nicht erkennen. Dies gilt auch für die übrigen Ver-

sionen des Modells, welche im Anhang A.2.5 (ab der Seite 109) zu finden sind. Obwohl eine weitere Untersuchung nicht notwendig ist, werden die gleichen Grafiken wie beim Multinomial-Dirichlet-Modell erzeugt und kurz beschrieben. Die senkrechten roten Linien im *Trace Plot* (Abbildung 6.6 links) kennzeichnen 2 000 000 Iterationen (von zehn Millionen durchgeführten Iterationen), die im weiteren Verlauf verworfen werden. Die Autokorrelation ist für das ökologische Multinomial-Log-Normal-Modell mit Referenzkategorie `Nichtwähler_15` in der Abbildung 6.7 und für die anderen Versionen im Anhang A.2.6 (ab der Seite 112) dargestellt. Die hohe Korrelation verringert sich nach dem *Burn-In* und *Thinning* von 2 000 bei allen Modellen kaum. Bei einigen Zellen wird sie sogar höher. Ein Modell mit *Thinning*, welches die Korrelation ausreichend verringert, wäre hier praktisch nicht berechenbar.

6.2.2 Ketten- und Modellvergleich

Mit dem *Burn-In* von 2 000 000 und dem *Thinning* von 2 000 werden ferner drei Ketten mit Referenzkategorie `Nichtwähler_15` und eine Kette mit Referenzkategorie `Kurz` erzeugt. Die Grafik *Trace of Counts* und die Dichten (*Density*) der Ketten sind in der Abbildung 6.8 auf der Seite 72 für das ökologische Log-Normal-Modell und im Anhang A.2.7 auf der Seite 115 für das Multinomial-Log-Normal-Hybridmodell dargestellt. Alle dargelegten Zellen weisen ziemlich hohe Differenzen zwischen den Ketten auf. Bei der Zelle `Nichtwähler_13.Nichtwähler_15` zeigen sich hierbei etwas niedrigere Differenzen als bei den Restlichen.

Aus der Grafik lässt sich kaum erkennen, ob die Schätzung durch die Veränderung der Referenzkategorie beeinflusst wird. Dafür liefert der Vergleich von Ergebnissen über alle Zellen mithilfe von AD in der Abbildung 6.9 auf der Seite 73 eine bessere Übersicht. Das heißt, hier wird beim Kettenvergleich, neben den Ketten innerhalb der gleichen Modelle, zusätzlich eine Kette mit nicht automatisch gewählter Referenzkategorie betrachtet. Mit und ohne Individualdaten zeigt sich, dass die Ergebnisse der Ketten mit der Referenzkategorie `Kurz` höhere absolute Distanzen aufweisen als die Restlichen untereinander. Allgemein und im Vergleich zum Multinomial-Dirichlet-Modell sind die Distanzen zwischen den Ketten sehr hoch, wobei etwas niedrigeren absoluten Distanzen beim Hybridmodell erkennbar sind als beim ökologischen Modell.

Der Modellvergleich wird in der Abbildung 6.10 auf der Seite 73 dargestellt. Hierfür wird die erste Kette mit Referenzkategorie `Nichtwähler_15` und wiederum die Kette

mit Referenzkategorie `Kurz`, jeweils mit und ohne Individualdaten, genommen. Demnach wiederholen sich beim Modellvergleich einige Zellen, die bereits beim Kettenvergleich dargestellt wurden. Der Modellvergleich zeigt allerdings, dass die Schätzung durch die Individualdaten stärker beeinflusst wird als durch die Veränderung der Referenzkategorie. Hierbei wird die Distanz zwischen dem ökologischen und dem hybriden Modell durch die Wahl der Referenzkategorie `Kurz` um 15.44 Prozent geringer. Der Ketten- und Modellvergleich anhand von MAE wird im Anhang A.2.8 auf der Seite 108 präsentiert.

Die zellspezifischen absoluten Differenzen der Modelle zur Nachwahlbefragung werden in der Abbildung 6.11 auf der Seite 74 dargestellt. Wie beim Multinomial-Dirichlet-Modell weisen die ökologischen Modelle doppelt so großen Differenzen zur Nachwahlbefragung auf als die Hybridmodelle. Beim Hybridmodell mit Referenzkategorie `Kurz` zeigen sich die geringsten zellspezifischen Differenzen mit MAE von 7.72 Prozent. Bei den anderen Versionen des Modells sind die zellspezifischen Differenzen und MAE allgemein höher als beim Multinomial-Dirichlet-Modell.

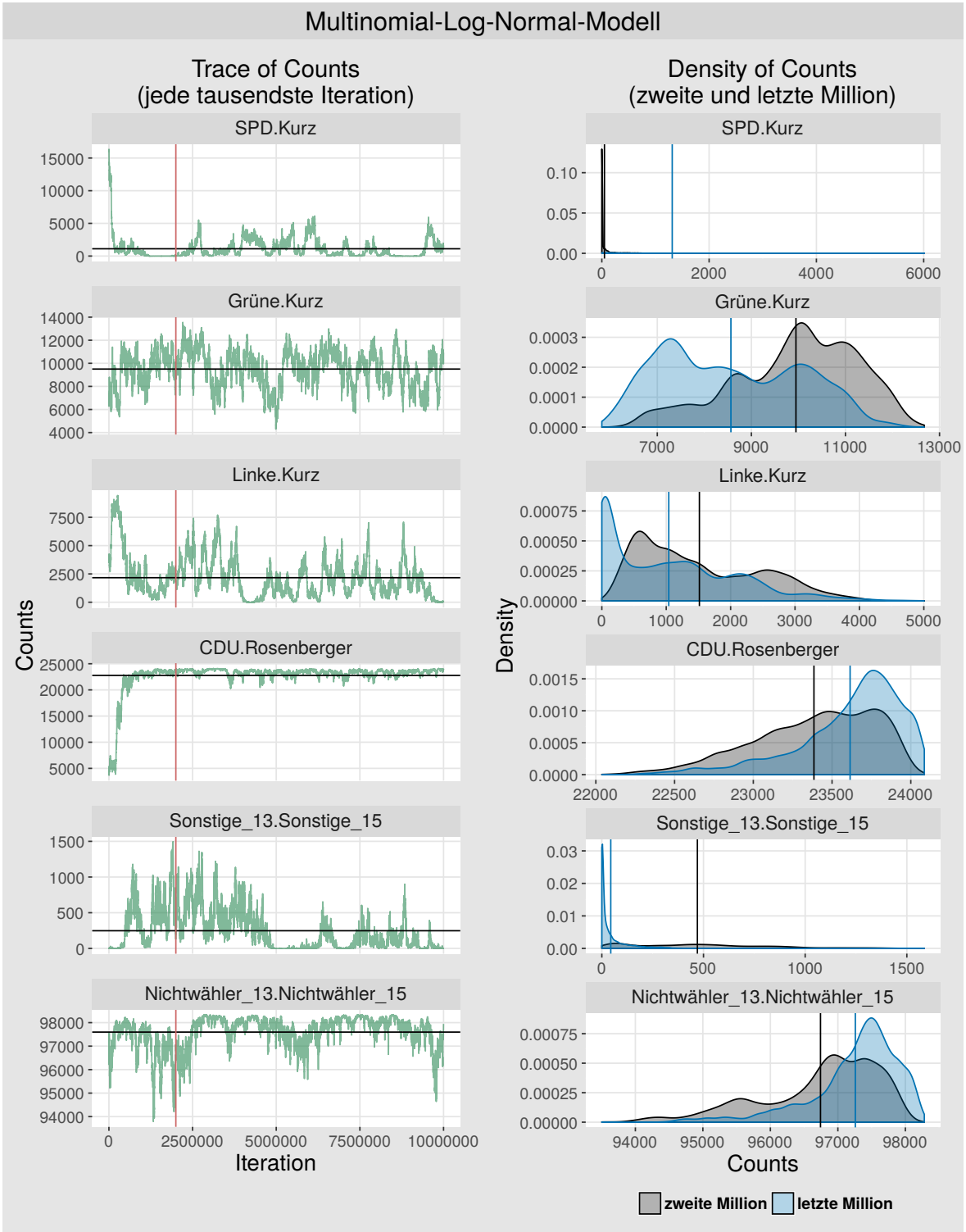

Abbildung 6.6: *Die Ketten (links) und die Dichten (rechts) der gezogenen absoluten Häufigkeiten (Counts) von fünf gewählten inneren Zellen des ökologischen Multinomial-Log-Normal-Modells mit automatisch gewählter Referenzkategorie* `Nichtwähler_15`*. Von zehn Millionen durchgeführten Iterationen konnte jede hundertste gespeichert werden. Links wird von zehn Millionen Iterationen jede tausendste dargestellt. Die senkrechten roten Linien kennzeichnen die* 2 000 000*-ste von zehn Millionen Iterationen. Die waagerechten schwarzen Linien zeigen die Mittelwerte von* 100 000 *gespeicherten Werten. Rechts: Die Dichten der zweiten und der letzten Million (jede hundertste Iteration betrachtet) und die dazugehörigen Mittelwerte (senkrechte Linien).*

Abbildung 6.7: *Die Autokorrelationen der gezogenen absoluten Häufigkeiten (Counts) von fünf gewählten inneren Zellen des ökologischen Multinomial-Log-Normal-Modells mit automatisch gewählter Referenzkategorie* `Nichtwähler_15` *anhand einer Stichprobe mit* 1 000 *Ziehungen nach dem Burn-In von* 2 000 000*. Links: Thinning von* 100*. Rechts: Thinning von* 2 000*.*

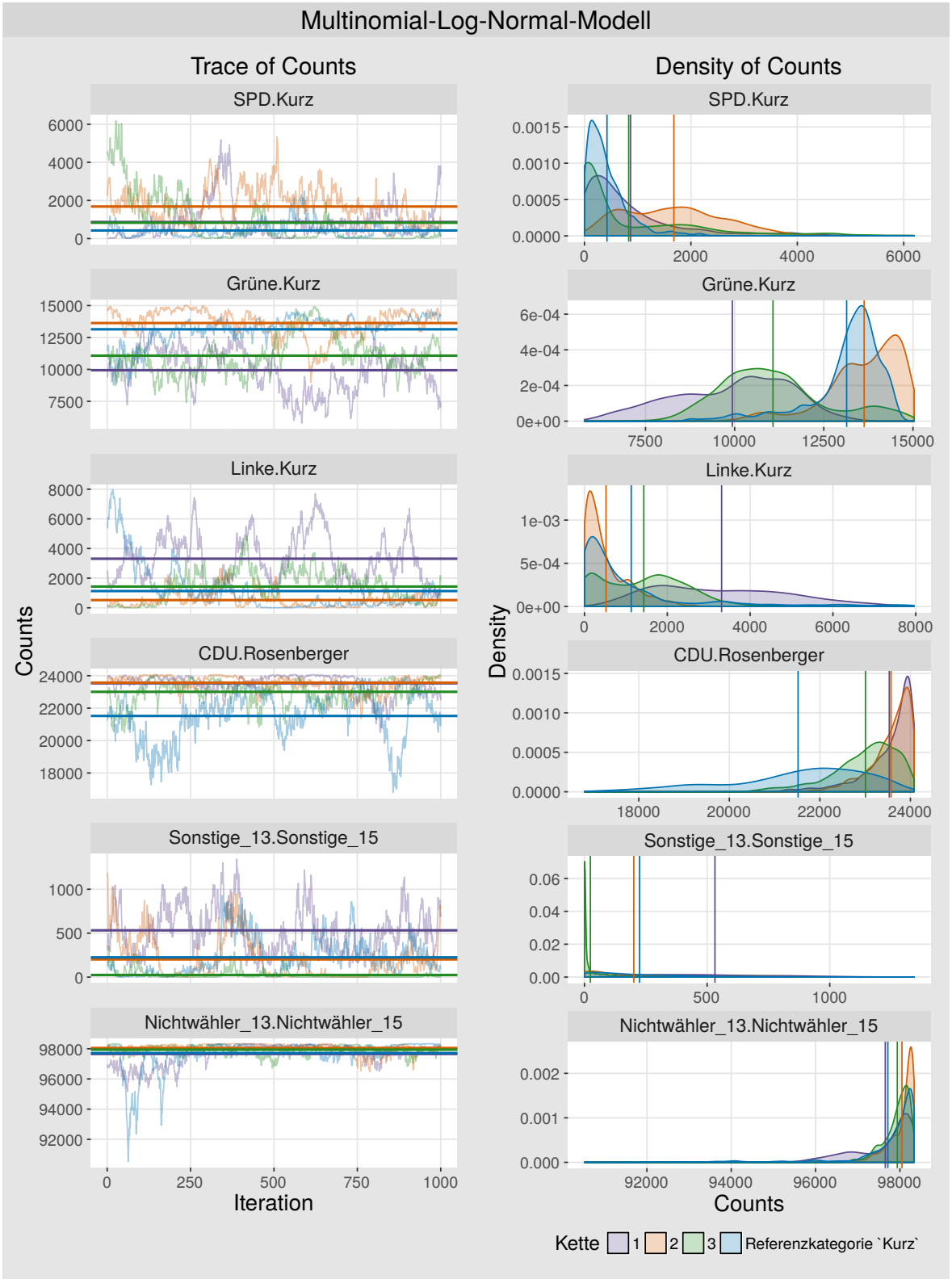

Abbildung 6.8: *Drei verdünnte Ketten der gezogenen absoluten Häufigkeiten (Counts) von fünf gewählten inneren Zellen des ökologischen Multinomial-Log-Normal-Modells mit automatisch gewählter Referenzkategorie* `Nichtwähler_15` *und eine verdünnte Kette mit Referenzkategorie* `Kurz`*. Sample:* 1 000*, Burn-In:* 2 000 000 *und Thinning:* 2 000*. Links: Trace of Counts der vier Ketten und die dazugehörigen Mittelwerte (waagerechte Linien). Rechts: Dichten der Ketten und die gleichen Mittelwerte senkrecht dargestellt.*

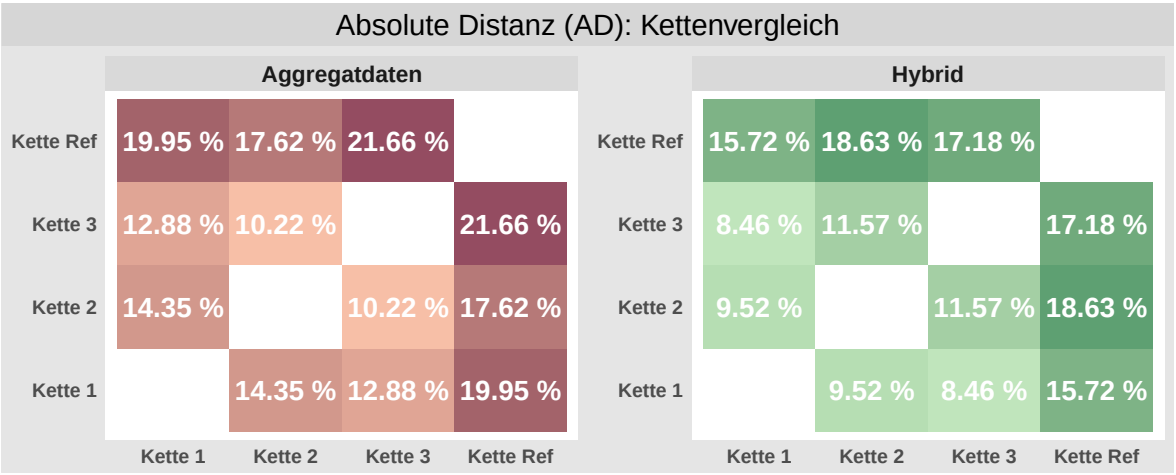

Abbildung 6.9: *Absolute Distanzen (AD) in Prozentpunkten zwischen den Ergebnissen der drei verdünnten Ketten mit automatisch gewählter Referenzkategorie* `Nichtwähler_15` *und einer Kette mit Referenzkategorie* `Kurz` *bei dem ökologischen (links) und bei dem hybriden (rechts) Multinomial-Log-Normal-Modell. Die Werte sind je nach Modell symmetrisch über die Diagonale.*

Absolute Distanz (AD): Modellvergleich

	Aggregatdaten	Aggregatdaten Referenzkategorie ‘Kurz‘	Hybrid	Hybrid Referenzkategorie ‘Kurz‘
Aggregatdaten		19.95 %	39.26 %	34.35 %
Aggregatdaten Referenzkategorie ‘Kurz‘	19.95 %		34.01 %	23.82 %
Hybrid	39.26 %	34.01 %		15.72 %
Hybrid Referenzkategorie ‘Kurz‘	34.35 %	23.82 %	15.72 %	

Abbildung 6.10: *Absolute Distanzen (AD) in Prozentpunkten zwischen den Ergebnissen von verschiedenen Versionen des Multinomial-Log-Normal-Modells (symmetrisch über die Diagonale). Zum Vergleich wurde die erste der drei verdünnten Ketten mit Referenzkategorie* `Nichtwähler_15` *und die Kette mit Referenzkategorie* `Kurz`*, jeweils für die Version mit und ohne Individualdaten, verwendet.*

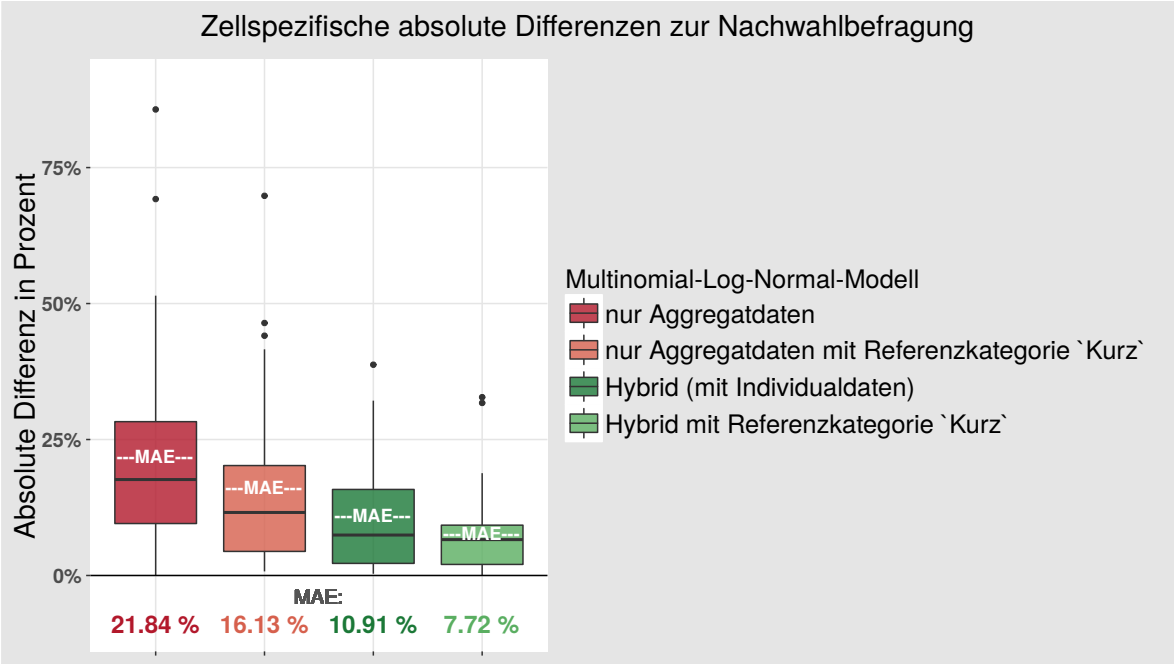

Abbildung 6.11: *Zellspezifische absolute Differenzen der Ergebnisse von verschiedenen Versionen des Multinomial-Log-Normal-Modells zur Nachwahlbefragung. Unten: Durchschnittliche absolute Differenzen pro Zelle (Mean Absolut Error).*

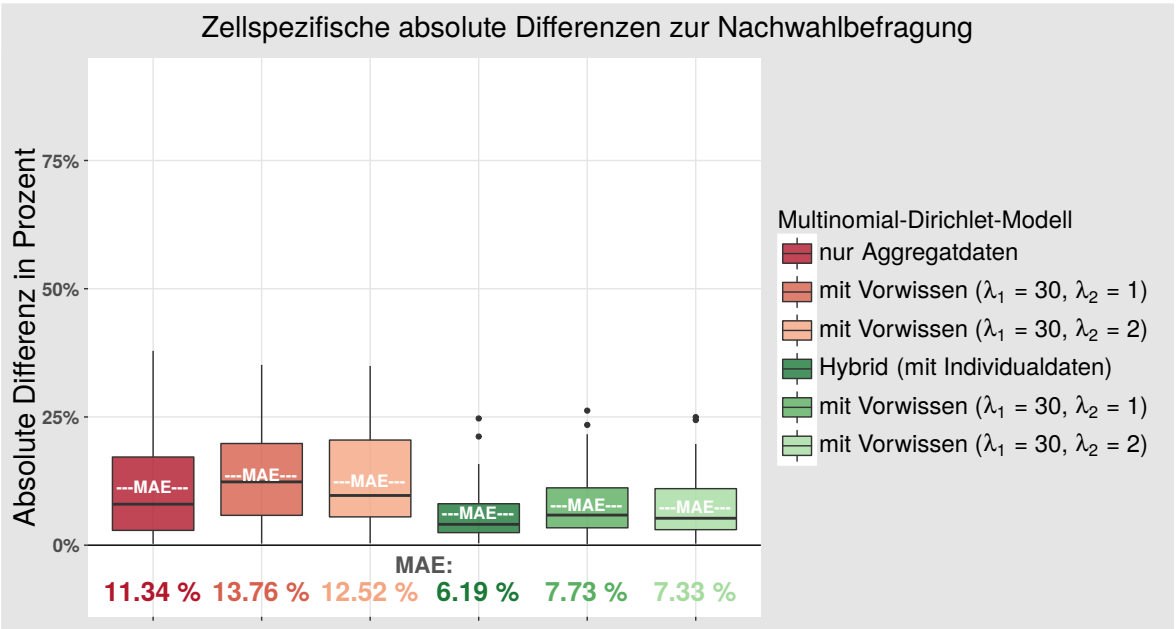

Abbildung 6.12: *Zellspezifische absolute Differenzen der Ergebnisse von verschiedenen Versionen des Multinomial-Dirichlet-Modells zur Nachwahlbefragung. Unten: Durchschnittliche absolute Differenzen pro Zelle (Mean Absolut Error).*

6.3 Modellwahl und Darstellung der Ergebnisse

Welches Modell der Wahrheit am nächsten liegt, lässt sich nicht testen. Nach der Methode der Elimination können zuerst alle Versionen des Multinomial-Log-Normal-Modells ausgeschlossen werden, da deren Ketten eine stationäre Verteilung nicht erreicht haben. Aus diesem Grund bleibt die Wahl einer Version des Multinomial-Dirichlet-Modells. Als Kriterium werden dafür die absoluten Distanzen zwischen den Ketten und Modellen verwendet. Es gibt hierbei keine Garantie, dass das Modell mit geringsten Differenzen zwischen den Ketten der Wahrheit entspricht. Allerdings können die Modelle, die bei jeder Durchführung unterschiedliche Ergebnisse erzeugen, nicht präzise und zuverlässig sein. Demzufolge werden die betrachteten Modelle mithilfe von AD solange eliminiert, bis ein Modell bleibt, das präziser und zuverlässiger als die Anderen ist. Der nächste Schritt ist demnach das Ausschließen aller Versionen des ökologischen Multinomial-Dirichlet-Modells. Alle verbleibenden drei Versionen des Hybridmodells weisen niedrige Distanzen zwischen den drei Ketten auf. Die zwei Versionen des Hybridmodells mit dem Vorwissen zeigen hierbei die niedrigste Distanz zwischen sich im Vergleich zu allen anderen Distanzen zwischen den verschiedenen Versionen des Modells. Die Wahl eines der Multinomial-Dirichlet-Hybridmodelle mit Vorwissen wird von den Ergebnissen der Simulationsstudie von Klima et al. (2016, S. 15 f., 19) unterstützt. Denn die Studie zeigt, dass die Analyse nur anhand von Aggregatdaten durch das Verwenden der Individualdaten verbessert wird, auch im Fall wenn die Nachwahlbefragung einen Bias aufweist. Außerdem wurde bei einer kleinen Anzahl an Wahlbezirken eine Verbesserung der Schätzung durch die Verwendung einer informativen Priori bemerkt. Zwischen den zwei übrigen Modellen wird letztendlich das Hybridmodell mit dem Hyperpriori-Parameter $Gamma(\lambda_1 = 30, \lambda_2 = 1)$ gewählt. Diese Entscheidung wird zufällig in `R` generiert, da die beiden verbleibenden Modelle nach Kriterium der Differenzen zwischen den drei Ketten sehr ähnlich sind.

Die geschätzten Übergangswahrscheinlichkeiten des gewählten Modells sind in der Tabelle 6.1 angegeben und in der Abbildung 6.13 visuell dargestellt. Der Gewinner der Oberbürgermeisterwahl 2015 ist die Kategorie *Nichtwähler*. Sogar 96.16 Prozent der Wahlberechtigten, die bei der Bundestagswahl im Jahr 2013 nicht gewählt haben, verzichten auch bei der Oberbürgermeisterwahl im Jahr 2015 auf ihr Recht zu wählen. Das heißt, keiner der Kandidaten konnte einen wesentlichen Anteil der *Nichtwähler* für sich erlangen. Nebenbei entschieden sich auch viele, die bei der Bundestagswahl

2013 eine der Parteien gewählt haben, bei der Oberbürgermeisterwahl 2015 keinen der Kandidaten zu unterstützen. So haben 79.71 Prozent der Wähler der kleinen Parteien, 61.8 Prozent der Wähler von *AfD*, 54.71 Prozent der Wähler von *SPD*, 52.48 Prozent der Wähler von *CDU* und 47.83 Prozent der Wähler von *Die Linken* bei der Oberbürgermeisterwahl 2015 nicht gewählt. Mit einer Nichtwählerquote von 31.49 Prozent weisen die Wähler von *FDP* eine etwas höhere Beteiligungsquote im Vergleich zu anderen auf. Der geringste Anteil der *Nichtwähler* von 14.66 Prozent zeigt sich bei den ehemaligen Wähler der *Grünen*. Der echte Gewinner, *Dr. Peter Kurz*, profitierte scheinbar gut von der Wahlempfehlungen. Mindestens von denen, die sich an der Wahl beteiligten, bekam er eine Mehrheit der Stimmen. Obwohl nur 45.29 Prozent der *SPD* Wähler und 52.17 Prozent der Wähler von *Die Linken* bei der Oberbürgermeisterwahl 2015 gewählt haben, gaben ihm 33.19 Prozent der *SPD* Wähler und 37.47 Prozent der Wähler von *Die Linken* ihre Stimme. Von den Wählern der *Grünen* erlangte *Dr. Peter Kurz* sogar 66.54 Prozent. Überzeugt hat er auch 22.01 Prozent der Wähler der *FDP*. Der zweite Kandidat *Peter Rosenberger* erwarb 32.76 Prozent der Stimmen der *CDU* Wähler, was dem größten Anteil von diejenigen entspricht, die bei der Oberbürgermeisterwahl 2015 gewählt haben. Er schaffte es, etwas mehr *FDP* Wähler als *Dr. Peter Kurz* zu gewinnen, sprich 26.51 Prozent. Mit 15.01 Prozent haben ihm doppelt so viele Wähler der *AfD* ihre Stimme gegeben. Nicht wesentlich weniger *AfD* Wähler (14.02 Prozent) unterstützte den dritten Kandidaten *Christopher Probst*. Von den Wählern der *FDP* entschieden sich 16.15 Prozent für ihn. Andere gewählte Kandidaten unter der Kategorie *Sonstige* bekamen insgesamt eine niedrige Unterstützung aller Wähler. Der Anteil lag zwischen 0.23 und 3.85 Prozent.

Der hohe Anteil der *Nichtwähler* und deren Aufteilung bietet möglicherweise eine Erklärung für den Bias bei den Kategorien der Bundestagswahl (2013) in der Nachwahlbefragung. Denn es werden einerseits die Parteien unterschätzt, deren Wähler ein hohen Anteil der *Nichtwähler* bei der Oberbürgermeisterwahl 2015 ausmachen. Anderseits werden die *Grünen* und die *FDP*, deren Wähler eine höhere Wahlbeteiligung im Vergleich zu den Übrigen aufweisen, überschätzt. Ein höherer Anteil der ehemaligen Wähler von *Grünen* in der Population der Wahlbeteiligten bei der Oberbürgermeisterwahl 2015 ist in diesem Fall zu erwarten. Dementsprechend darf die Rolle der *Nichtwähler* bei der Wählerwanderungsanalyse mittels einer Befragung nicht ignoriert werden.

Multinomial-Dirichlet-Hybridmodell mit Vorwissen Gamma(λ_1 = 30, λ_2 = 1)						
		Oberbürgermeisterwahl 2015				
		Dr. Peter Kurz (SPD)	Peter Rosenberger (CDU)	Christopher Probst (Mannheimer Liste)	Sonstige	Nichtwähler
Bundestagswahl 2013	SPD	33.19 %	5.11 %	6.17 %	0.82 %	54.71 %
	Grüne	66.54 %	8.19 %	6.92 %	3.69 %	14.66 %
	Die Linke	37.47 %	5.34 %	5.73 %	3.63 %	47.83 %
	CDU	6.97 %	32.76 %	7.2 %	0.59 %	52.48 %
	AfD	7.29 %	15.01 %	14.02 %	1.88 %	61.8 %
	FDP	22.01 %	26.51 %	16.15 %	3.85 %	31.49 %
	Sonstige	7.43 %	3.57 %	5.93 %	3.36 %	79.71 %
	Nichtwähler	1.08 %	1.4 %	1.13 %	0.23 %	96.16 %

Tabelle 6.1: *Die Übergangstabelle zwischen der Bundestagswahl 2013 und der Oberbürgermeisterwahl 2015 anhand des Multinomial-Dirichlet-Hybridmodells mit Hyperpriori-Parameter Gamma*($\lambda_1 = 30, \lambda_2 = 1$) *für die Zellen der Loyalen und Defaultwerte für die Restlichen.*

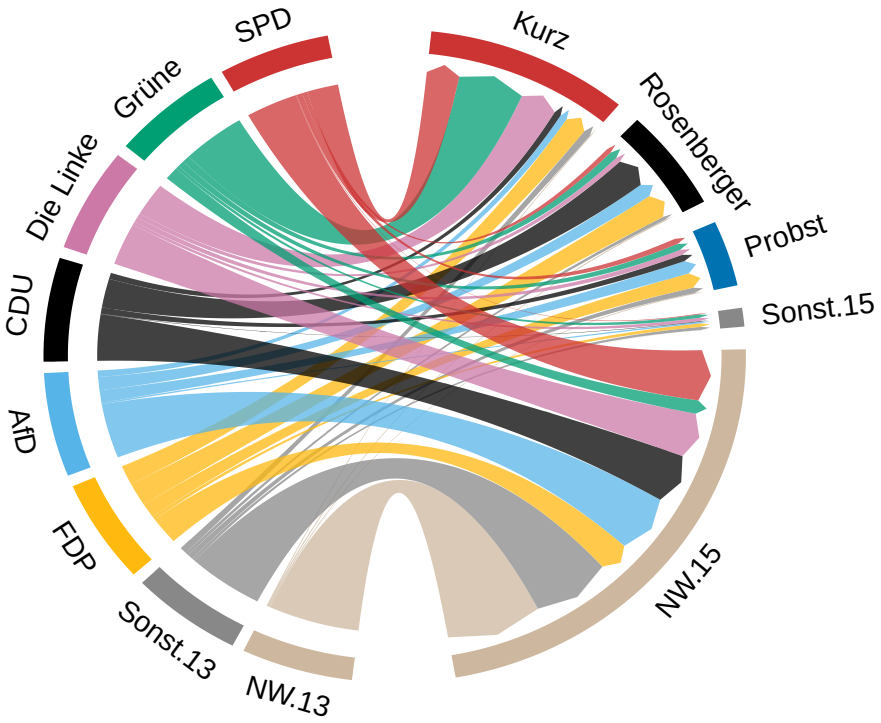

Abbildung 6.13: *Die Übergangswahrscheinlichkeiten zwischen der Bundestagswahl 2013 (links) und der Oberbürgermeisterwahl 2015 (rechts) anhand des Multinomial-Dirichlet-Hybridmodells mit Hyperpriori-Parameter Gamma*($\lambda_1 = 30, \lambda_2 = 1$) *für die Zellen der Loyalen und Defaultwerte für die Restlichen. Die Breite jedes Pfeilendes drückt den Anteil an Stimmen aus, den der jeweilige Kandidat von verschiedenen Parteien gewonnen hat.*

7 Fazit

Ziel der vorliegenden Arbeit war die Schätzung der Wählerwanderung zwischen der Bundestagswahl im Jahr 2013 und der Oberbürgermeisterwahl im Jahr 2015 in Mannheim. Zu diesem Zweck waren die amtlichen Ergebnisse und eine Nachwahlbefragung verfügbar. Das individuelle Wahlverhalten konnte aus der Nachwahlbefragung einfach und schnell herausgezogen und dargestellt werden. Ein Vergleich der Randsummen mit den amtlichen Ergebnissen wies jedoch darauf hin, dass die Individualdaten einen Bias aufweisen. Dementsprechend ist die Analyse der Übergänge zwischen zwei Wahlen nur anhand von vorhandenen Individualdaten unzuverlässig und unsicher. Die Analyse anhand von Aggregatdaten lässt sich mithilfe der ökologischen Inferenz durchziehen. Die neu entwickelten Hybridmodelle kombinieren dabei die ökologische Inferenz und die Individualdaten, um die Stärken der beiden Ansätze nutzen zu können. Demnach wurden für die Wählerwanderungsanalyse zwei ökologische hierarchische Modelle und deren hybriden Versionen verwendet. Das Multinomial-Dirichlet-Modell von Rosen et al. (2001) und dessen hybride Version von Schlesinger (2013) wurden mit dem `eiwild` Paket (Schlesinger, 2014) berechnet. Die Analyse anhand des ökologischen und des hybriden Multinomial-Log-Normal-Modells von Greiner und Quinn (2009, 2010) wurde mit dem `RxCEcolInf` Paket (Greiner et al., 2013) durchgeführt.

Vor der Analyse ist die Datenaufbereitung ein relevanter Schritt. Als Erstes ist die Anzahl der Parteien oder Kandidaten zu reduzieren, um die Anzahl der Parameter zu verringern. Dabei wurden alle kleine Parteien beziehungsweise Kandidaten einer Kategorie zugeordnet. Zweitens ändert sich die Population der Wähler mit dem Zeitabstand zwischen zwei Wahlen. Diese Differenz wurde zur *Nichtwähler* Kategorie bei der Bundestagswahl (2013) gerechnet. Außerdem ändert sich die Aufstellung der Wahlbezirke zwischen zwei Wahlen. Durch die Anpassung der Wahlbezirke entstand hier eine Reduktion. Dieser Informationsverlust kann beim Multinomial-Dirichlet-Modell durch Integration von Vorwissen kompensiert werden. Die Simulationsstudie von Klima et al. (2016) zeigt eine Verbesserung der Schätzung durch die Verwendung von informativen Priori-Verteilung bei einer kleinen Anzahl an Wahlbezirken. Weiterhin werden

die Ergebnisse der Briefwähler in spezifischen postalischen Wahlbezirken dargestellt, die nicht identisch mit den Wahlbezirke der Urnenwählern sind. Demzufolge mussten die Briefwähler durch zusätzliche Berechnungen den Urnenwähler zugerechnet werden. Schließlich wurden die Datensätze der Individual- und Aggregatdaten in die notwendige Form für die Analyse gebracht.

Bei der Schätzung anhand des Multinomial-Dirichlet-Modells können die zellspezifischen Hyperpriori-Parameter der Gamma-Verteilung definiert werden (Schlesinger, 2013). Dementsprechend wurde das Vorwissen über die Wahlempfehlungen benutzt, um die Unterschätzung der Zellen der Loyalen zu verhindern. Es wurden $Gamma(\lambda_1 = 30,\ \lambda_2 = 1)$ und $Gamma(\lambda_1 = 30,\ \lambda_2 = 2)$ als alternative Hyperpriori-Parameter für die Zellen der Loyalen überprüft. Somit wurden insgesamt sechs Versionen des Modells berechnet, eine mit Defaultwerte für alle Zellen, zwei mit erwähnten Hyperpriori-Parameter für die Zellen der Loyalen und alle drei einmal in ökologischer Version und einmal in hybrider Version. Bei der Schätzung anhand des Multinomial-Log-Normal-Modells wurden das ökologische Modell und das Hybridmodell einmal mit automatisch gewählter Referenzkategorie und einmal mit der Referenzkategorie `Kurz` berechnet.

Da sich die Güte der Modelle nicht testen lässt, wurde für die Modellwahl die Konvergenzdiagnose der erzeugten Ketten sowie die absoluten Distanzen (AD) zwischen den Ketten und Modellen als Kriterium verwendet. Von den zwei Hauptmodellen konnte zuerst das Multinomial-Log-Normal ausgeschlossen werden, da keine dessen Versionen konvergiert. Hierbei verringerte das Verwenden der Individualdaten die AD Werte zwischen den Ketten bei beiden Hauptmodellen. Demzufolge wurde als nächstes das ökologische Multinomial-Dirichlet-Modell ausgeschlossen. Die geringsten AD Werte zeigen sich bei den zwei Hybridmodellen mit Vorwissen. Da sich die beiden Versionen anhand der verwendeten Kriterien kaum unterschieden, wurde letztendlich das Hybridmodell mit dem Hyperpriori-Parameter $Gamma(\lambda_1 = 30, \lambda_2 = 1)$ zufällig in `R` gezogen.

Die Ergebnisse des gewählten Modells zeigen eine sehr hohe Wanderung von allen Parteien zu den *Nichtwähler* mit der Ausnahme der Wähler der *Grünen* und der *FDP*. Da gerade diese Kategorien bei der Nachwahlbefragung überschätzt wurden, stellt sich die Frage, ob das der Grund für den Bias darstellt. Wenn ja, sollte überlegt werden, ob die Nachwahlbefragung am Ausgang der Wahllokale eine zuverlässige Methode für

die Wählerwanderungsanalyse ist. Allerdings berichten viele Autoren, dass so erhobene Daten hilfreich sind, um die ökologische Inferenz zu verbessern (Greiner und Quinn, 2010; Wakefield, 2004), auch wenn ein Bias in den Daten vorliegt (Klima et al., 2016). Als eine eigenständige Methode oder als die Unterstützung im Hybridmodell ist die telefonische Befragung zusehen (Klima et al., 2016), denn somit können die *Nichtwähler* sowie die Briefwähler in der Analyse betrachtet werden. Welche Aussagen liefern diese Ergebnissen für die Politiker? Vor allem kann empfohlen werden, erforderliche Maßnahmen zu ergreifen, um die Bürgerinnen und Bürger zu animieren, ihr Wahlrecht zu nutzen. Da die Analyse der politischen Situation nicht Teil dieser Arbeit ist, können keine weiteren Hinweise über die Art und Weise der Maßnahmen gegeben werden.

Zielsetzung dieser Arbeit war es nicht zu beurteilen, welche der betrachteten Modelle im Allgemeinen besser ist. Vor allem deswegen, weil anhand von den hier durchgeführten Analysen dies nicht möglich wäre. Dennoch lassen sich einige Vor- und Nachteile nennen, die durch die Anwendung bemerkt wurden. Das Multinomial-Log-Normal-Modell ist in der praktischen Anwendung mithilfe des `RxCEcolInf` Paketes (Greiner et al., 2013) komplizierter. Diese Kritik umfasst zuerst die erforderliche Formatierung der Individualdaten (Unterabschnitt 4.2.5, S. 45 und 46) und die etwas komplizierte Angabe von *Sample*, *Thinning* und *Burn-In* (Unterabschnitt 5.2.2, S. 55). Zusätzlich wird die Angabe von dem Verhältnis zwischen den Kategorien der ersten und der zweiten Wahl durch einen *String Character* umständlich (Unterabschnitt 5.2.1, S, 54 und 55). Hierbei könnte die Bestimmung der Referenzkategorie besser gelöst werden als durch das aktuell notwendige Umschichten der Reihenfolge der eingegebenen Kategorien. Letztendlich erwähnen die Autoren selbst, dass das Modell-Fitting langsamer ist. Dazukommend lässt sich im `eiwild` Paket (Schlesinger, 2013) Vorwissen durch zellspezifische Hyperpriori-Parameter ins Modell integrieren (Unterabschnitt 5.1.4 ab Seite 50). Aus den erwähnten Gründen wird die Analyse schließlich mit dem `eiwild` Paket anhand des Multinomial-Dirichlet-Modells empfohlen, falls die Analyse auf ein Modell begrenzt werden muss. Insbesondere wenn Vorwissen vorhanden ist und die Anzahl der Wahlbezirke klein ausfällt. Ansonsten bietet sich an, die beiden Modelle durchzuführen, zu vergleichen und die Ergebnisse aus dem Modell mit stabilsten Ketten zu berechnen.

Literatur

Ambühl, M. (2003). *Methoden zur Rekonstruktion von Wählerströmen aus Aggregatdaten.* Bundesamt für Statistik (BFS). (17 Politik)

Andreadis, I. und Chadjipadelis, T. (2009). A Method for the Estimation of Voter Transition Rates. *Journal of Elections, Public Opinion and Parties*, *19* (2), 203–218.

Arnold, J. B. (2016). `ggthemes`: Extra Themes, Scales and Geoms for `ggplot2` [Software-Handbuch]. Zugriff auf `https://CRAN.R-project.org/package=ggthemes` (R package version 3.0.3 (2016-04-09))

Auguie, B. (2016). `gridExtra`: Miscellaneous Functions for „Grid“Graphics [Software-Handbuch]. Zugriff auf `https://CRAN.R-project.org/package=gridExtra` (R package version 2.2.1 (2016-02-29))

Cho, W. K. T. und Manski, C. F. (2009). Chross-Level/Ecological Inference. In J. M. Box-Steffensmeier, H. E. Brady und D. Collier (Hrsg.), *Oxford Handbook of Political Methodology.* Oxford University Press. Zugriff auf `http://cho.pol.illinois.edu/wendy/papers/eiexp.pdf`

Cowles, M. K. und Carlin, B. P. (1996). Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. *Journal of the American Statistical Assotiation*, *91* (434), 883–904. Zugriff auf `http//www.jstor.org/stable/2291683`

Duncan, O. D. und Davis, B. (1953). An Alternative to Ecological Correlation. *American Sociological Review*, *18* (6), 665–666.

Felderer, B. (2015). *E-Mail Kommunikation: Wahltagsbefragung - OBW 2015.* Universität Mannheim.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. und Rubin, D. B. (2014). *Bayesian Data Analysis.* Chapman & Hall, CRC Press.

Gelman, A. und Shirley, K. (2011). Inference from Simulations and Monitoring Convergence. In S. Brooks, A. Gelman, G. L. Jones und X.-L. Meng (Hrsg.), *Handbook of Markov Chain Monte Carlo.* Chapman & Hall, CRC Press. Zugriff auf `http://www.mcmchandbook.net/HandbookChapter6.pdf`

Geyer, C. J. (2011). Introduction to MCMC. In S. Brooks, A. Gelman, G. L. Jones und X.-L. Meng (Hrsg.), *Handbook of Markov Chain Monte Carlo.* Chapman & Hall, CRC Press. Zugriff auf `http://www.mcmchandbook.net/HandbookChapter1.pdf`

Goodman, L. A. (1953). Ecological Regressions and Behavior of Individuals. *American Sociological Review*, *18* (6), 663-664.

Greiner, D. J., Baines, P. und Quinn, K. M. (2013). `RxCEcolInf`: R × C Ecological Inference With Optional Incorporation of Survey Information [Software-Handbuch]. Zugriff auf `https://CRAN.R-project.org/package=RxCEcolInf` (R package version 0.1-3 (2013-07-24))

Greiner, D. J. und Quinn, K. M. (2009). R × C Ecological Inference: Bounds, Correlations, Flexibility and Transparency of Assumptions. *Journal of the Royal Statistical Society A*, *172* (1), 67–81. Zugriff auf `https://www.law.berkeley.edu/files/GQ-JRSSA.pdf`

Greiner, D. J. und Quinn, K. M. (2010). Exit Pooling and Racial Block Voting: Combining Individual-Level and R × C Ecological Data. *The Annals of Applied Statistics*, *4* (4), 1774–1796.

Greiner, D. J. und Quinn, K. M. (2012). Long Live the Exit Poll. , *141* (4), 9-22.

Gschwend, T. (2006). Ökologische Inferenz. In J. Benke, T. Gschwend, D. Schindler und K.-U. Schnapp (Hrsg.), *Methoden der Politikwissenschaft: neuere qualitative und quantitative Analyseverfahren* (S. 227-237). Nomos Verl.-Ges. Zugriff auf `http://www.ssoar.info/ssoar/handle/document/25840`

Gu, Z. (2015). Visualize Relations by Chord Diagram [Software-Handbuch]. Zugriff auf `https://cran.r-project.org/web/packages/circlize/vignettes/visualize_relations_by_chord_diagram.pdf`

Gu, Z., Gu, L., Eils, R., Schlesner, M. und Brors, B. (2014). `circlize` implements and enhances circular visualization in R. *Bioinformatics*, *30*, 2811-2812. (R package version 0.3.5 (2016-03-28))

Held, L. und Bové, D. S. (2014). *Applied Statistical Inference: Likelihood and Bayes*. Springer-Verlag.

Himmelwelt, H. T., Biberian, M. J. und Stockdale, J. (1978). Memory for past Vote: Implication of a Study of Bias in Recall. , *8* (3), 365–375.

Kellermann, T. (2011). Vom Wahlergebnis zur Wählerwanderung: Welche Wähler wechseln wie ihre Wahlentscheidung? *Stadtforschung und Statistik* (1), 34–40.

King, G. (1997). *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton University Press. Zugriff auf `http://gking.harvard.edu/files/gking/files/part1.pdf?m=1360039169`

King, G., Rosen, O. und Tanner, M. A. (1999). Binomial-Beta Hierarchical Models for Ecological Inference. *Sociological methods and research*, *28* (1), 61–90.

Klima, A., Schlesinger, T., Küchenhoff, H. und Thurner, P. W. (2016). Combining Aggregate Data and Exit-Polls for the Estimation of Voter Transitions [Unveröffentlichte Artikel].

Klima, A., Thurner, P. W., Molnar, C., Schlesinger, T. und Küchenhoff, H. (2015). Estimation of Voter Transitions Based on Ecological Inference: An Empirical Assessment of Different Approaches. *AStA Advances in Statistical Analysis*, 1–27.

Lau, O., Moore, R. T. und Kellermann, M. (2007). `eiPack`: R × C Ecological Inference and Higher-Dimension Data Management. *R News*, *7* (2), 43–47.

Lau, O., Moore, R. T. und Kellermann, M. (2012). `eiPack`: Ecological Inference and Higher-Dimension Data Management [Software-Handbuch]. Zugriff auf `https://CRAN.R-project.org/package=eiPack` (R package version 0.1-7 (2012-01-13))

Link, W. A. und Eaton, M. J. (2012). On Thinning of Chains in MCMC. *Methods in Ecology and Evolution*, *3* (1), 112–115. Zugriff auf `http://dx.doi.org/10.1111/j.2041-210X.2011.00131.x`

Neuwirth, E. (2014). `RColorBrewer`: ColorBrewer Palettes [Software-Handbuch]. Zugriff auf `https://CRAN.R-project.org/package=RColorBrewer` (R package version 1.1-2 (2014-12-07))

Payne, C., Brown, P. und Hanna, V. (1986). By-election Exit Polls. , *5* (3), 277-287.

R Core Team. (2015). R: A Language and Environment for Statistical Computing [Software-Handbuch]. Vienna, Austria. Zugriff auf `https://www.R-project.org/` (R version 3.2.3 (2015-12-10))

Robert, C. P. (2007). *The Bayesian Choice*. Springer Science+Bussines Media.

Robinson, W. S. (1950). Ecological Correlations and the Behavior of Individuals. *American Sociological Review*, *15* (3), 351–357.

Rosen, O., Jiang, W., King, G. und Tanner, M. A. (2001). Bayesian and Frequentist Inference for Ecological Inference: The R × C Case. *Statistica Neerlandica*, *55* (2), 134–156. Zugriff auf `http://gking.harvard.edu/files/gking/files/em.pdf?m=1360038990`

Schlesinger, T. (2013). *Kombination von Aggregat- und Individualdaten bei der Analyse von R × C-Tafeln: Neue Implementierung in R* (Unveröffentlichte Masterarbeit). Ludwig-Maximilians-Universität, München.

Schlesinger, T. (2014). `eiwild`: Ecological Inference with Individual and Aggregate Data [Software-Handbuch]. Zugriff auf `https://CRAN.R-project.org/package=eiwild` (R package version 0.6.7 (2014-03-06))

Schredle, M. (2015). *RheinneckarBlog.* Zugriff am 19. Juni 2016 auf http://www.rheinneckarblog.de/05/fdp-mannheim-keine-ob-wahlempfehlung/67829.html

Stadt Mannheim. (2013). *Ergebnisse der Bundestagswahl (Archiv): Wahlbezirksergebnis (XLS).* Zugriff am 22. März 2016 auf https://www.mannheim.de/stadt-gestalten/ergebnisse-bundestagswahl-archiv

Stadt Mannheim. (2015a). *17 Stadtbezirke.* Zugriff am 20. Juni 2016 auf https://www.mannheim.de/stadt-gestalten/stadtgebiet-und-flaechennutzung

Stadt Mannheim. (2015b). *Ergebnisse der Oberbürgermeisterwahl (Archiv): Wahlergebnisse aller Ebenen (XLS).* Zugriff am 22. März 2016 auf https://www.mannheim.de/stadt-gestalten/ergebnisse-oberbuergermeisterwahl-archiv

Stadt Mannheim. (2015c). *Wahlbeteiligung bei den Mannheimer Oberbürgermeisterwahlen im Jahr 2015* (Statistischer Bericht Mannheim Nr. 8). Zugriff am 21. Juni 2016 auf https://mannheim.de/stadt-gestalten/ergebnisse-oberbuergermeisterwahl-archiv

Statistisches Bundesamt. (2015). *Bildungsstand der Bevölkerung.* Zugriff am 21. Juni 2016 auf https://www.destatis.de/DE/Publikationen/Thematisch/BildungForschungKultur/Bildungsstand/BildungsstandBevoelkerung.html

Wahlbüro der Stadt Mannheim. (2016). *E-Mail Kommunikation: Straßenverzeichnissen - Zuordnung der Straßen zu den Wahlbezirken.*

Wakefield, J. (2004). Ecological Inference for 2 × 2 Tables. *Royal Statistical Society*, *167* (3), 385–445.

Wickham, H. (2009). ggplot2*: Elegant Graphics for Data Analysis.* Springer-Verlag New York. Zugriff auf http://ggplot2.org (R package version 2.1.0 (2016-03-01))

Wickham, H. (2015a). reshape2: Flexibly Reshape Data: A Reboot of the Reshape Package [Software-Handbuch]. Zugriff auf https://cran.r-project.org/web/packages/reshape2/index.html (R package version 1.4.1 (2014-12-06))

Wickham, H. (2015b). stringr: Simple, Consistent Wrappers for Common String Operations [Software-Handbuch]. Zugriff auf https://CRAN.R-project.org/package=stringr (R package version 1.0.0 (2015-04-30))

Wickham, H. (2016). scales: Scale Functions for Visualization [Software-Handbuch]. Zugriff auf https://CRAN.R-project.org/package=scales (R package version 0.4.0 (2016-02-26))

A Anhang

A.1 Die Datenbasis

A.1.1 Parteien der Bundestagswahl 2013

BUNDESTAGSWAHL 2013			
	Partei		Partei
01	CDU	11	ÖDP
02	SPD	12	PBC
03	FDP	13	VOLKSABSTIMMUNG
04	GRÜNE	14	MLPD
05	DIE LINKE	15	BüSo
06	AfD	16	BIG
07	PIRATEN	17	PRO-DEUTSCHLAND
08	NPD	18	FREIE-WÄHLER
09	REP	19	PARTEI DER VERNUNFT
10	TIER-SCHUTZ-PARTEI	20	RENTNER

Tabelle A.1: *Die Liste aller Parteien aus dem Datensatz der amtlichen Endergebnisse der Bundestagswahl im Jahr 2013 (Stadt Mannheim, 2013).*

A.1.2 Stadtbezirke Mannheim

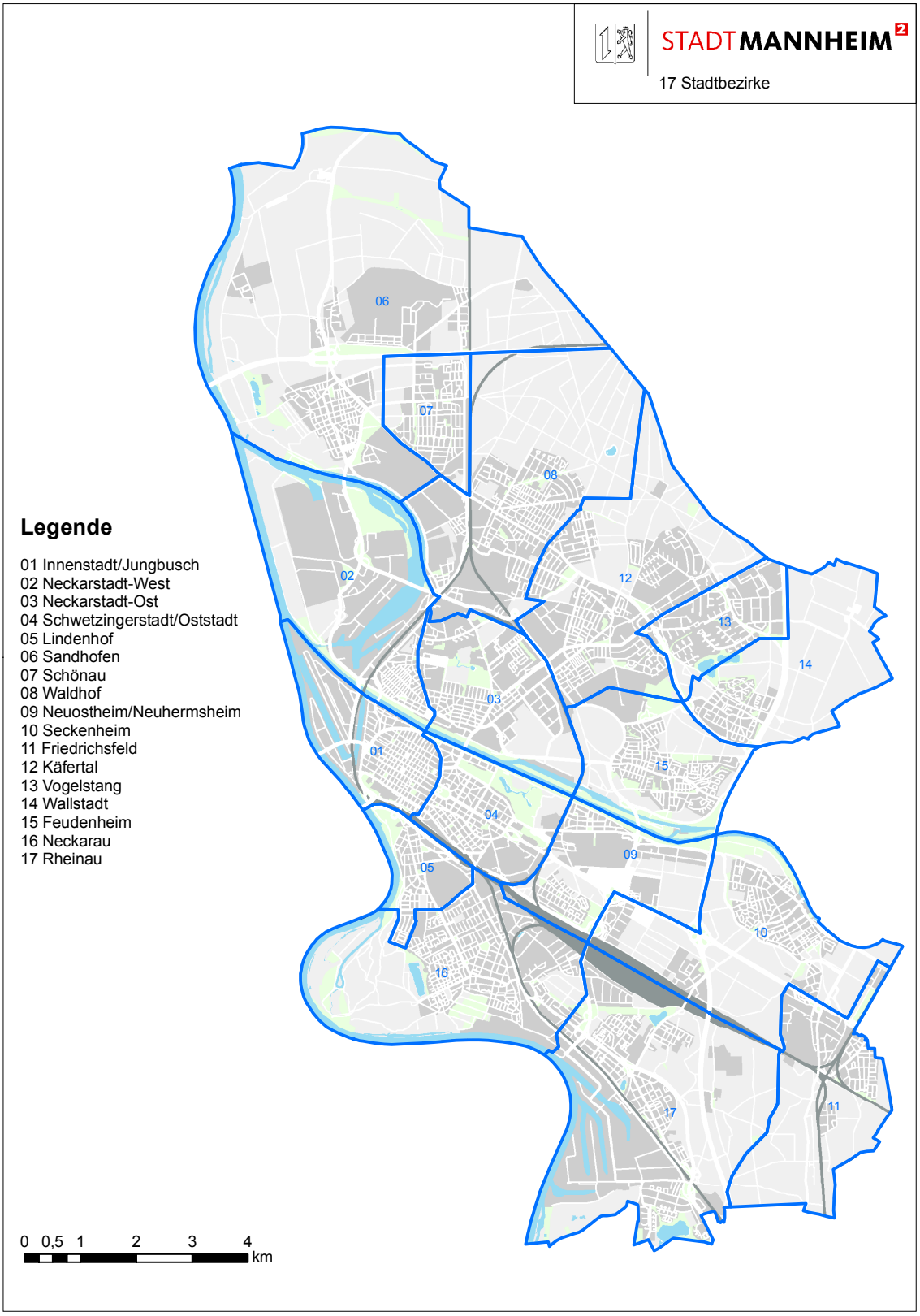

Abbildung A.1: *Mannheim: Aufteilung der Stadtbezirke, übernommen von Stadt Mannheim (2015a).*

A.1.3 Die Ergebnisse beider Wahlen nach Wahlbezirke

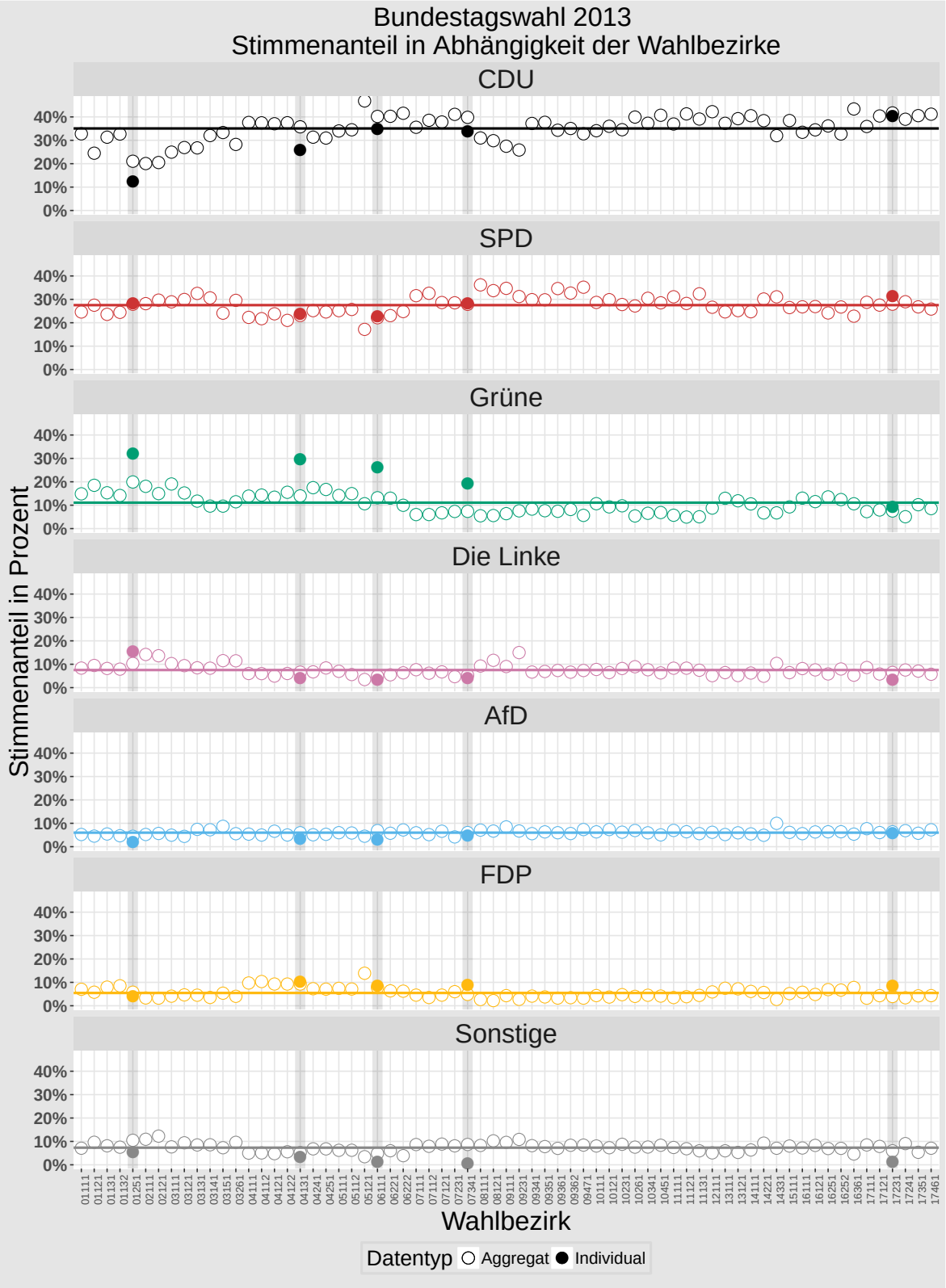

Abbildung A.2: *Amtliches Ergebnis der Bundestagswahl 2013 in Abhängigkeit der Wahlbezirke und die Ergebnisse der Nachwahlbefragung für fünf betrachtete Wahlbezirke. Die dargestellten Wahlbezirke werden so aggregiert, dass alle Ebenen gleich sind wie bei der Oberbürgermeisterwahl 2015.*

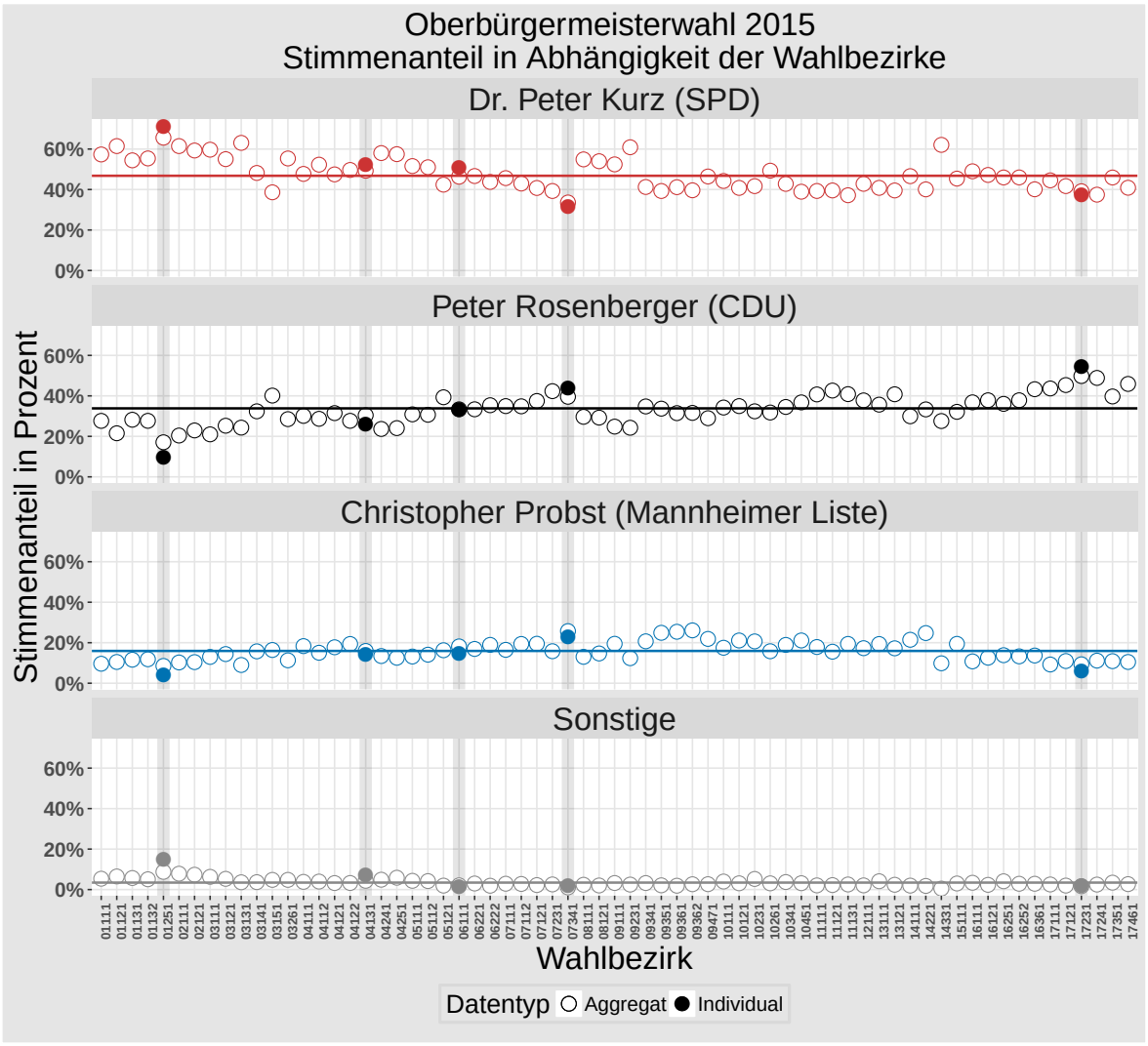

Abbildung A.3: *Amtliches Ergebnis der Oberbürgermeisterwahl 2015 in Abhängigkeit der Wahlbezirke und die Ergebnisse der Nachwahlbefragung für fünf betrachtete Wahlbezirke. Die dargestellten Wahlbezirke werden so aggregiert, dass alle Ebenen gleich sind wie bei der Bundestagswahl 2013.*

A.1.4 Die Ergebnisse der Wählerwanderung anhand von Individualdaten ohne *Nichtwähler* bei der Oberbürgermeisterwahl

Nachwahlbefragung 2015					
		Oberbürgermeisterwahl 2015			
		Dr. Peter Kurz (SPD)	Peter Rosenberger (CDU)	Christopher Probst (Mannheimer Liste)	Sonstige
Bundestagswahl 2013	SPD	70.33 %	17.21 %	10.09 %	2.37 %
	Grüne	68.95 %	13.73 %	12.09 %	5.23 %
	Die Linke	57.14 %	11.43 %	11.43 %	20 %
	CDU	24.81 %	62.28 %	11.14 %	1.77 %
	AfD	13.04 %	50 %	36.96 %	0 %
	FDP	28.04 %	52.34 %	17.76 %	1.87 %
	Sonstige	43.33 %	0 %	16.67 %	40 %
	Nichtwähler	54.44 %	30 %	7.78 %	7.78 %

Tabelle A.2: *Die Übergangstabelle zwischen der Bundestagswahl 2013 und der Oberbürgermeisterwahl 2015 anhand der Nachwahlbefragung, ohne „Nichtwähler“ bei der Oberbürgermeisterwahl 2015.*

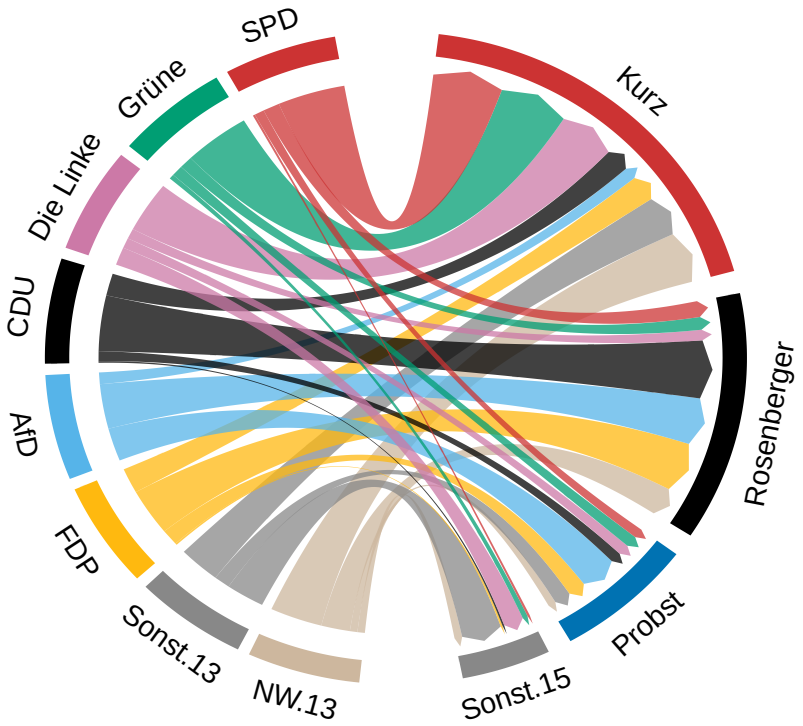

Abbildung A.4: *Die Übergangswahrscheinlichkeiten zwischen der Bundestagswahl 2013 (links) und der Oberbürgermeisterwahl 2015 (rechts) anhand der Nachwahlbefragung, ohne „Nichtwähler“ bei der Oberbürgermeisterwahl 2015. Die Breite jedes Pfeilendes drückt den Anteil an Stimmen aus, den der jeweilige Kandidat von verschiedenen Parteien gewonnen hat.*

A.1.5 Aggregatdaten - amtliches Ergebnis mit *Nichtwähler*

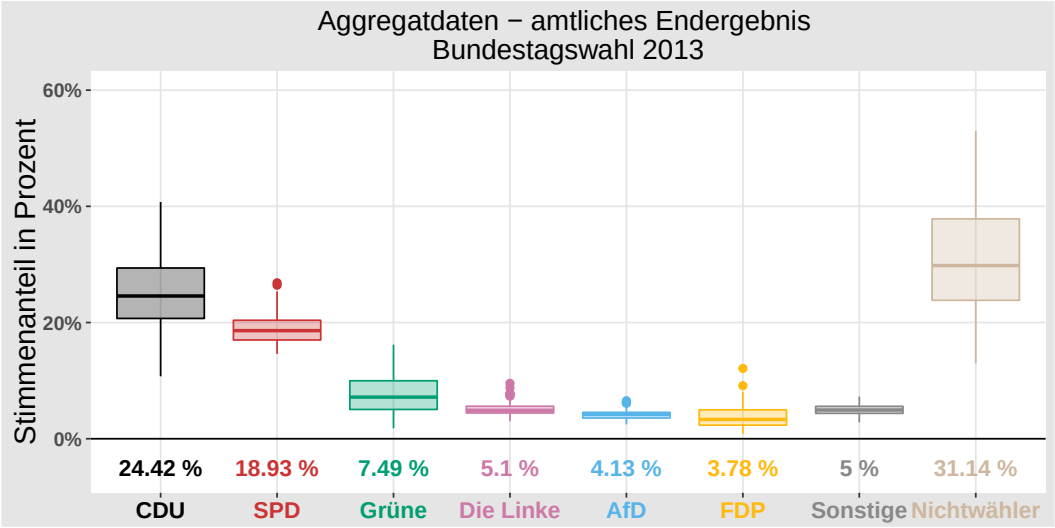

Abbildung A.5: *Wahlbezirkspezifische amtliche Endergebnisse der Bundestagswahl 2013 inklusive „Nichtwähler“. Unten: Der durchschnittliche Stimmenanteil über alle Wahlbezirke in Prozent. Quelle: Stadt Mannheim (2013).*

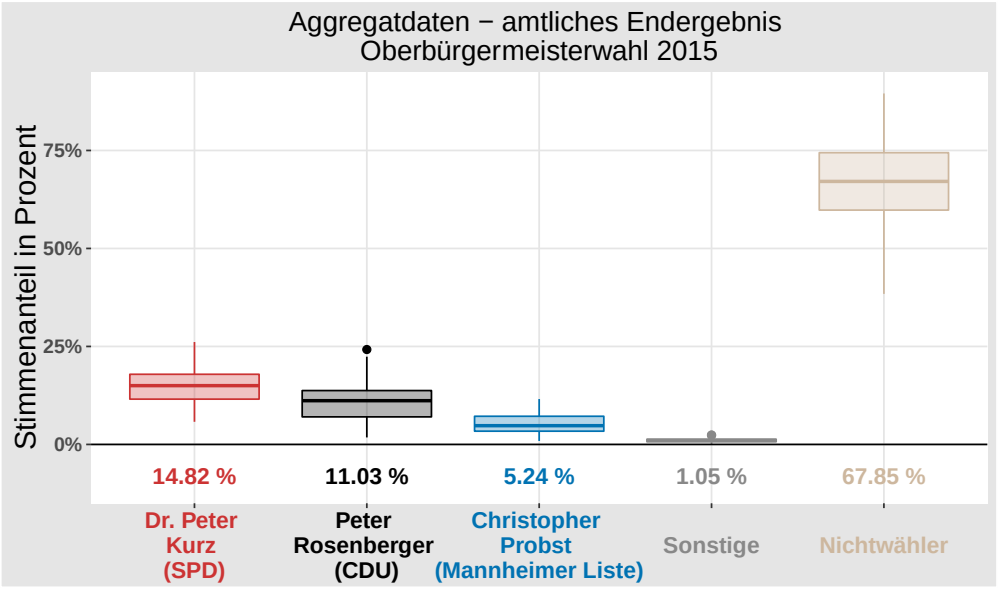

Abbildung A.6: *Wahlbezirkspezifische amtliche Endergebnisse der Oberbürgermeisterwahl 2015 inklusive „Nichtwähler“. Unten: Der durchschnittliche Stimmenanteil über alle Wahlbezirke in Prozent. Quelle: Stadt Mannheim (2015b).*

A.1.6 Vereinigung der Wahlbezirke

Nr.	BTW	OBW
1	01111	01111
	01112	
2	01121	01121
	01122	
	01122	01122
	01123	
3	01131	→ 01131
4	01132	→ 01132
5	01251	01251
	01252	
6	02111	02111
	02112	
	02112	02112
	02113	
7	02121	02121
	02122	
	02122	02122
	02123	
8	03111	03111
	03112	
	03112	03112
	03113	
	03113	03113
	03114	
9	03121	03121
	03122	
10	03131	→ 03131
11	03141	03141
	03142	
	03142	03142
	03143	
12	03151	→ 03151
13	03261	03261
	03262	
14	04111	→ 04111
15	04112	→ 04112
16	04121	→ 04121
17	04122	→ 04122
18	04131	04131
	04132	
	04132	04132
	04133	
19	04241	04241
	04242	
20	04251	04251
	04252	
	04252	04252
	04253	
21	05111	05111
	05112	
	05112	05112
	05113	
	.....	
22	05114	05113
	05115	
	05115	05114
	05116	
23	05121	→ 05121
24	06111	06111
	06112	
25	06221	→ 06221
26	06222	→ 06222
27	07111	07111
	07112	
	.....	
28	07113	07112
	07114	
29	07121	07121
	07122	
30	07231	→ 07231
31	07341	→ 07341
32	08111	08111
	08112	
	08112	08112
	08113	
33	08121	08121
	08122	
	08123	
34	09111	→ 09111
35	09231	→ 09231
36	09341	09341
	09342	
37	09351	09351
	09352	
38	09361	09361
	09362	
	.....	
39	09363	09362
	09364	
40	09471	09471
	09472	
	09472	09472
	09473	
41	10111	10111
	10112	
42	10121	10121
	10122	
	10122	10122
	10123	
43	10231	10231
	10232	
44	10261	10261
	10262	
45	10341	10341
	10342	
46	10451	10451
	10452	
47	11111	11111
	11112	
	11112	11112
	11113	
48	11121	11121
	11122	
49	11131	11131
	11132	
50	12111	12111
	12112	
	12112	12112
	12113	
	12114	12113
	12114	
	12115	
51	13111	13111
	13112	
	13112	13112
	13113	
52	13121	13121
	13122	
	13122	13122
	13123	
	13124	13123
	13124	
	13125	
53	14111	14111
	14112	
	14112	14112
	14113	
	14114	14113
	14114	
	14115	14114
	14115	
	14116	
54	14221	→ 14221
55	14331	→ 14331
56	15111	15111
	15112	
	15112	15112
	15113	
57	16111	16111
	16112	
	16112	16112
	16113	
	16114	16113
	16114	
	16115	
58	16121	16121
	16122	
59	16251	16251
	16252	
	.....	
60	16253	16252
	16254	
61	16361	16361
	16362	
	16362	16362
	16363	
	16364	
62	17111	17111
	17112	
	17112	17112
	17113	
63	17121	17121
	17122	
64	17231	17231
	17232	
	17232	17232
	17233	
65	17241	→ 17241
66	17351	→ 17351
67	17461	17461
	17462	
	17462	17462
	17463	

Tabelle A.3: *Vereinigung der Wahlbezirke zwischen der Bundestagswahl 2013 und der Oberbürgermeisterwahl 2015, wodurch 67 konstante Ebenen resultieren. Quelle: Wahlbüro der Stadt Mannheim (2016).*

A.1.7 Differenz des Stimmenanteils zwischen den Brief- und Urnenwählern

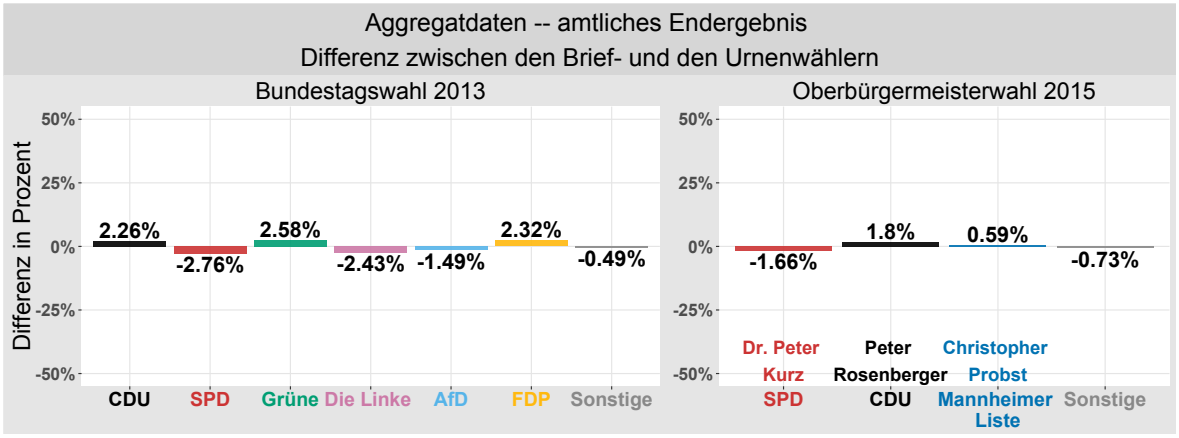

Abbildung A.7: *Differenz der Stimmenanteile zwischen den Brief- und den Urnenwählern bei der Bundestagswahl 2013 (links) und bei der Oberbürgermeisterwahl 2015 (rechts). Quelle: Stadt Mannheim (2013, 2015b).*

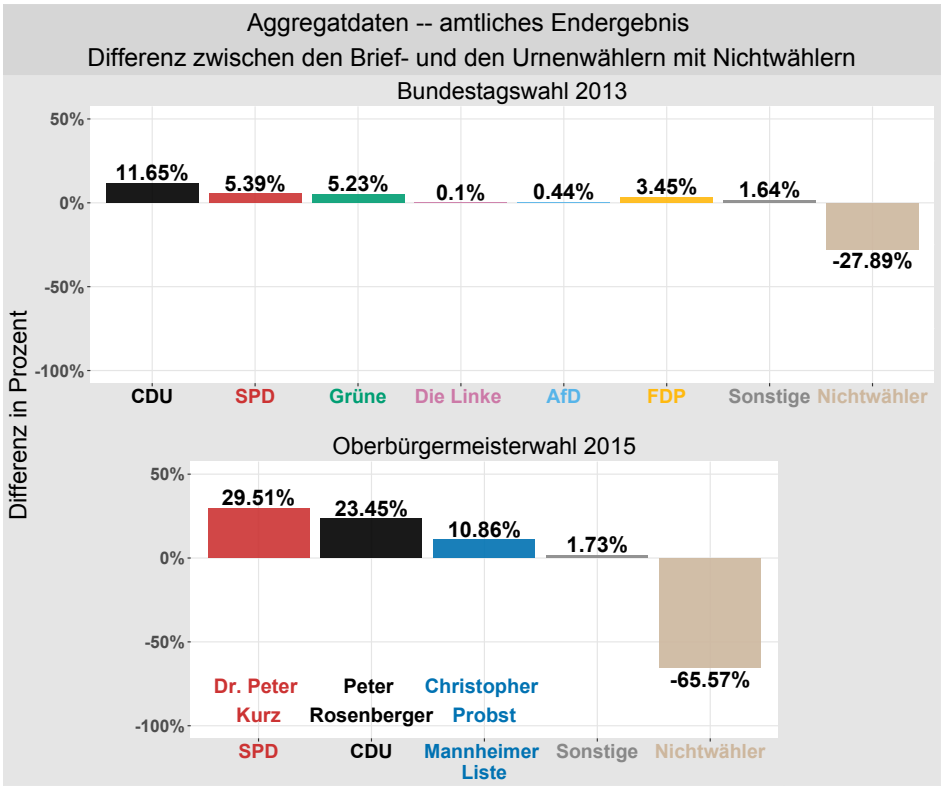

Abbildung A.8: *Differenz der Stimmenanteile zwischen den Brief- und den Urnenwählern bei der Bundestagswahl 2013 (oben) und bei der Oberbürgermeisterwahl 2015 (unten) inklusive „Nichtwähler“. Quelle: Stadt Mannheim (2013, 2015b).*

A.2 Konvergenzdiagnose, Ketten- und Modellvergleich

A.2.1 Multinomial-Dirichlet-Modell: *Trace-* und *Density of Counts*

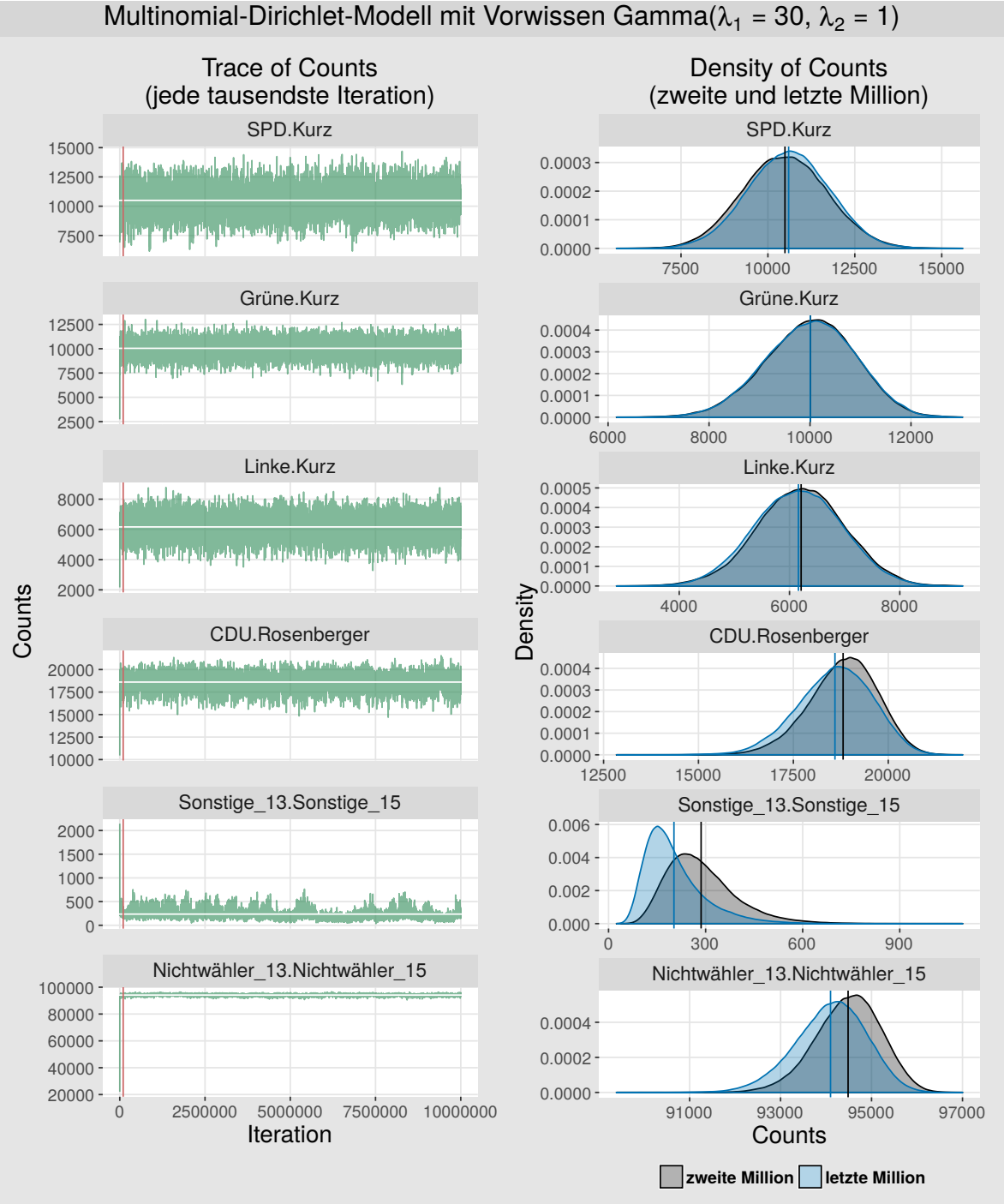

Abbildung A.9: *Die Ketten (links) und die Dichten (rechts) der gezogenen absoluten Häufigkeiten (Counts) von fünf gewählten inneren Zellen des ökologischen Multinomial-Dirichlet-Modells mit Hyperpriori-Parameter Gamma*$(\lambda_1 = 30, \lambda_2 = 1)$ *für die Zellen der Loyalen. Links: Von zehn Millionen Iterationen wird jede tausendste dargestellt. Die senkrechten roten Linien kennzeichnen die* 100 000*-ste von zehn Millionen Iterationen. Die waagerechten weißen Linien zeigen die Mittelwerte aller gezogenen Werte. Rechts: Die Dichten der zweiten und der letzten Million aller Iterationen und die dazugehörigen Mittelwerte (senkrechte Linien).*

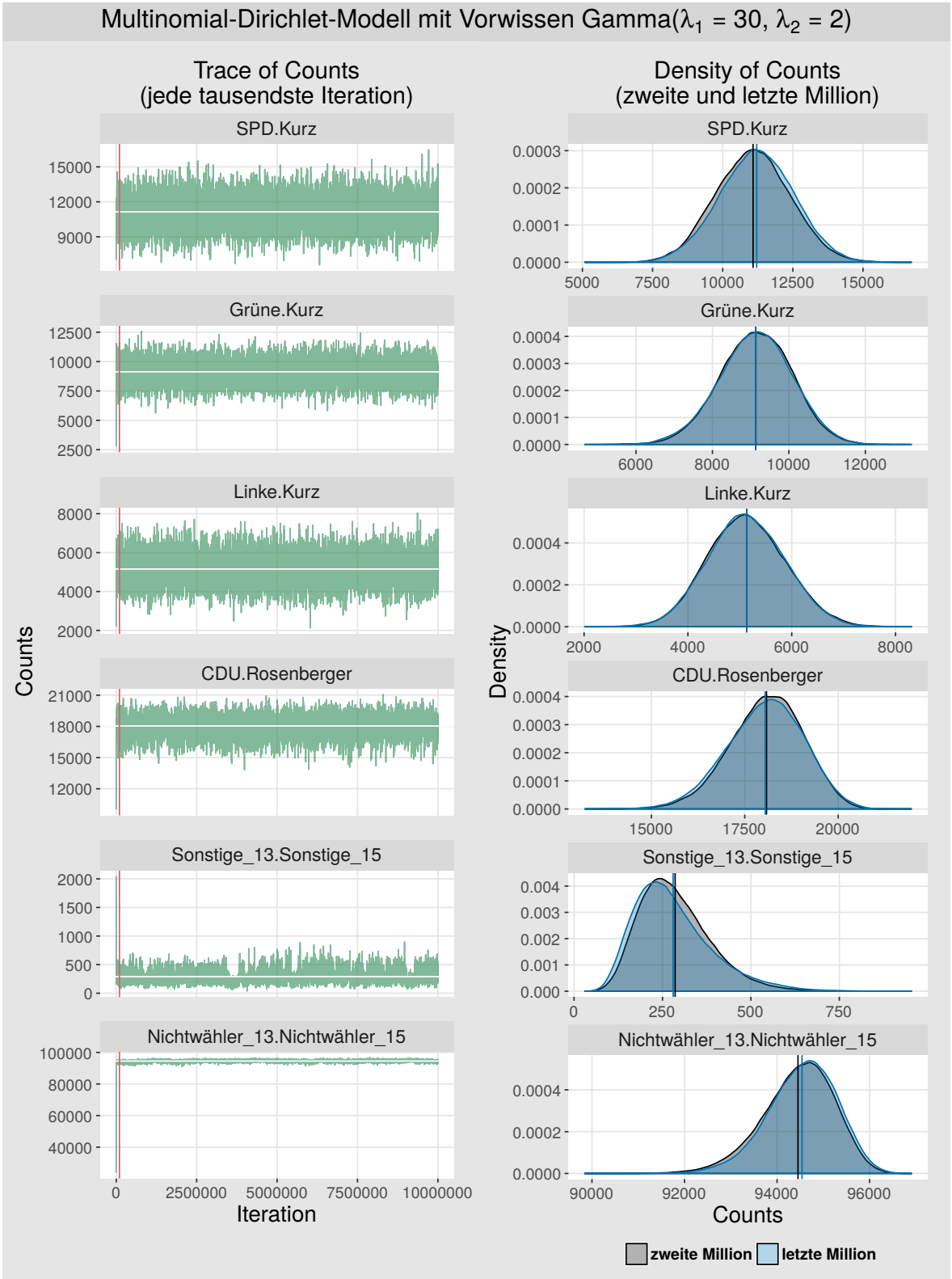

Abbildung A.10: *Die Ketten (links) und die Dichten (rechts) der gezogenen absoluten Häufigkeiten (Counts) von fünf gewählten inneren Zellen des ökologischen Multinomial-Dirichlet-Modells mit Hyperpriori-Parameter Gamma*$(\lambda_1 = 30, \lambda_2 = 2)$ *für die Zellen der Loyalen. Links: Von zehn Millionen Iterationen wird jede tausendste dargestellt. Die senkrechten roten Linien kennzeichnen die* 100 000*-ste von zehn Millionen Iterationen. Die waagerechten weißen Linien zeigen die Mittelwerte aller gezogenen Werte. Rechts: Die Dichten der zweiten und der letzten Million aller Iterationen und die dazugehörigen Mittelwerte (senkrechte Linien).*

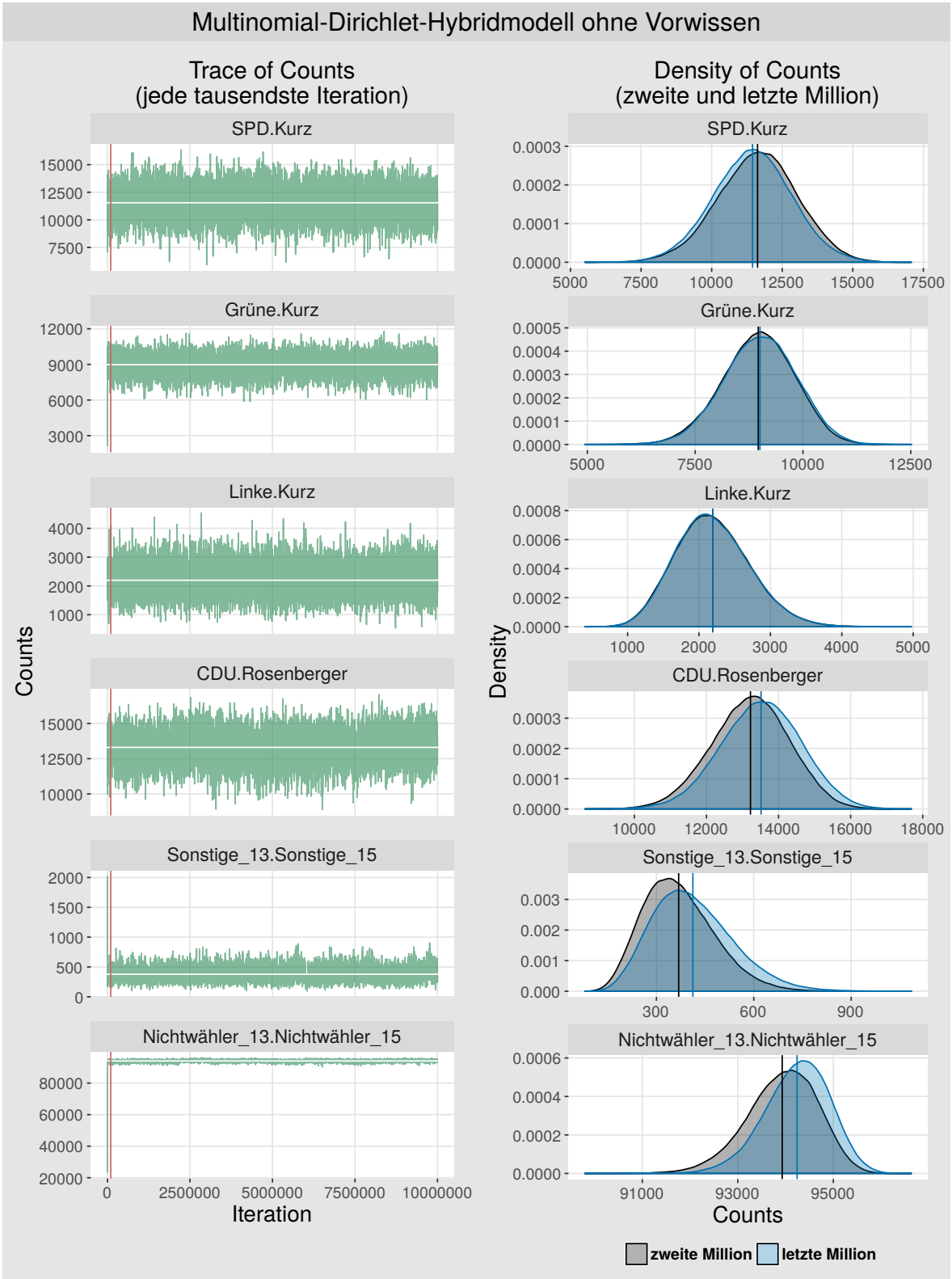

Abbildung A.11: *Die Ketten (links) und die Dichten (rechts) der gezogenen absoluten Häufigkeiten (Counts) von fünf gewählten inneren Zellen des Multinomial-Dirichlet-Hybridmodells ohne Vorwissen. Links: Von zehn Millionen Iterationen wird jede tausendste dargestellt. Die senkrechten roten Linien kennzeichnen die* 100 000*-ste von zehn Millionen Iterationen. Die waagerechten weißen Linien zeigen die Mittelwerte aller gezogenen Werte. Rechts: Die Dichten der zweiten und der letzten Million aller Iterationen und die dazugehörigen Mittelwerte (senkrechte Linien).*

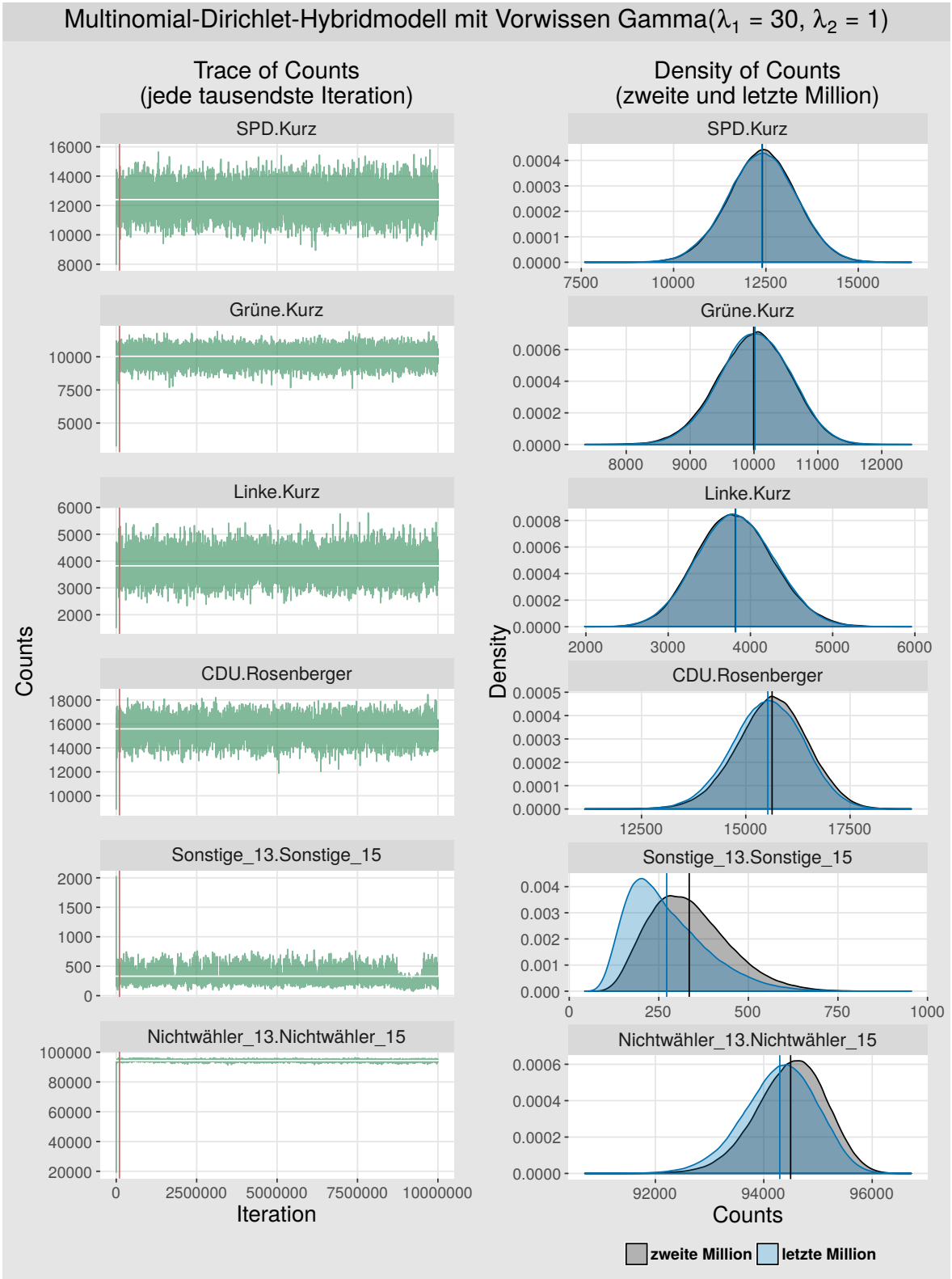

Abbildung A.12: *Die Ketten (links) und die Dichten (rechts) der gezogenen absoluten Häufigkeiten (Counts) von fünf gewählten inneren Zellen des Multinomial-Dirichlet-Hybridmodells mit Hyperpriori-Parameter Gamma*$(\lambda_1 = 30, \lambda_2 = 1)$ *für die Zellen der Loyalen. Links: Von zehn Millionen Iterationen wird jede tausendste dargestellt. Die senkrechten roten Linien kennzeichnen die* 100 000*-ste von zehn Millionen Iterationen. Die waagerechten weißen Linien zeigen die Mittelwerte aller gezogenen Werte. Rechts: Die Dichten der zweiten und der letzten Million aller Iterationen und die dazugehörigen Mittelwerte (senkrechte Linien).*

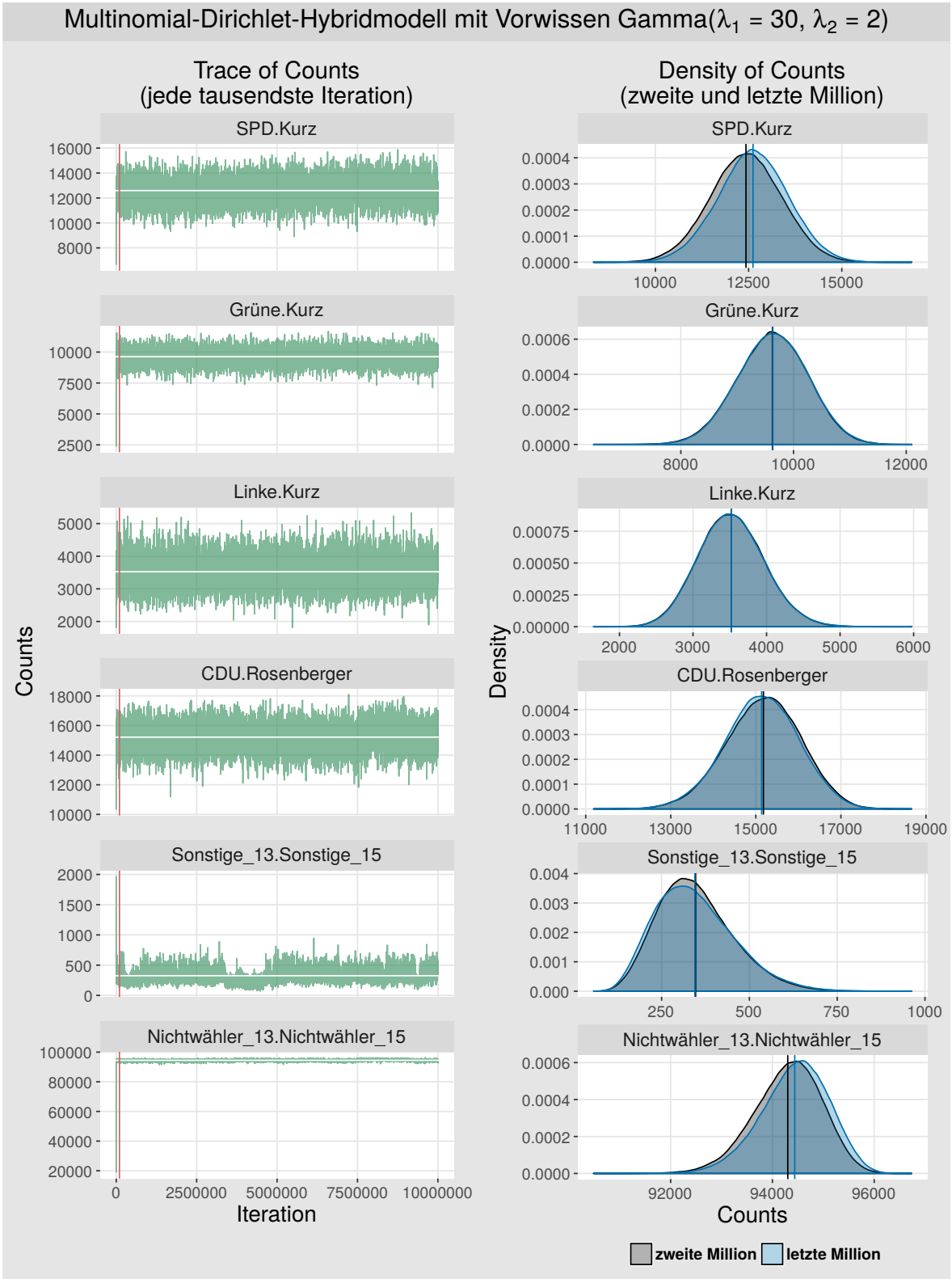

Abbildung A.13: *Die Ketten (links) und die Dichten (rechts) der gezogenen absoluten Häufigkeiten (Counts) von fünf gewählten inneren Zellen des Multinomial-Dirichlet-Hybridmodells mit Hyperpriori-Parameter Gamma*$(\lambda_1 = 30, \lambda_2 = 2)$ *für die Zellen der Loyalen. Links: Von zehn Millionen Iterationen wird jede tausendste dargestellt. Die senkrechten roten Linien kennzeichnen die* 100 000*-ste von zehn Millionen Iterationen. Die waagerechten weißen Linien zeigen die Mittelwerte aller gezogenen Werte. Rechts: Die Dichten der zweiten und der letzten Million aller Iterationen und die dazugehörigen Mittelwerte (senkrechte Linien).*

A.2.2 Multinomial-Dirichlet-Modell: Autokorrelationen

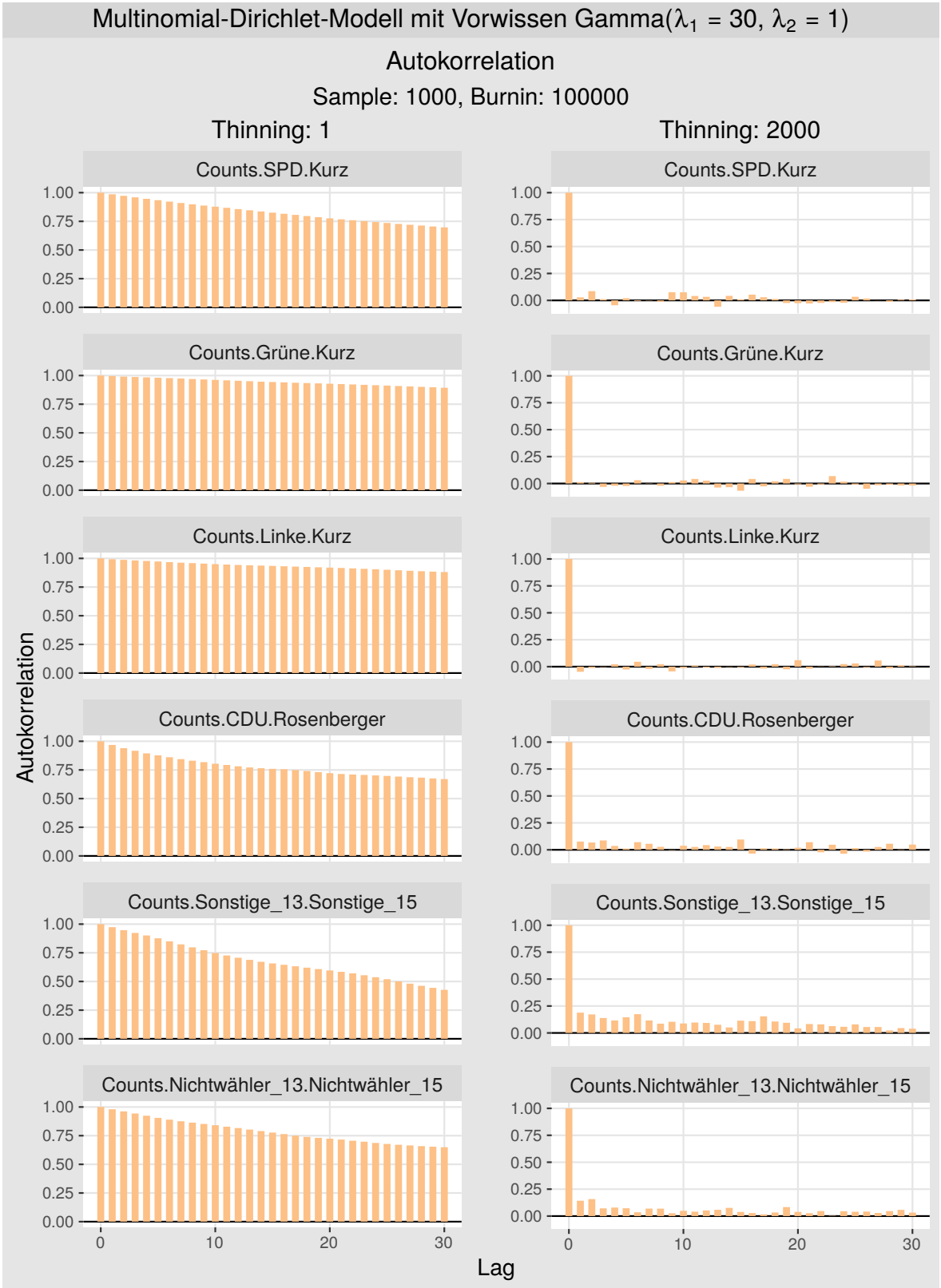

Abbildung A.14: *Die Autokorrelationen der gezogenen absoluten Häufigkeiten (Counts) von fünf gewählten inneren Zellen des ökologischen Multinomial-Dirichlet-Modells mit Hyperpriori-Parameter Gamma*$(\lambda_1 = 30, \lambda_2 = 1)$ *für die Zellen der Loyalen, anhand einer Stichprobe mit* 1 000 *Ziehungen nach dem Burn-In von* 100 000*. Links: Ohne Thinning. Rechts: Thinning von* 2 000*.*

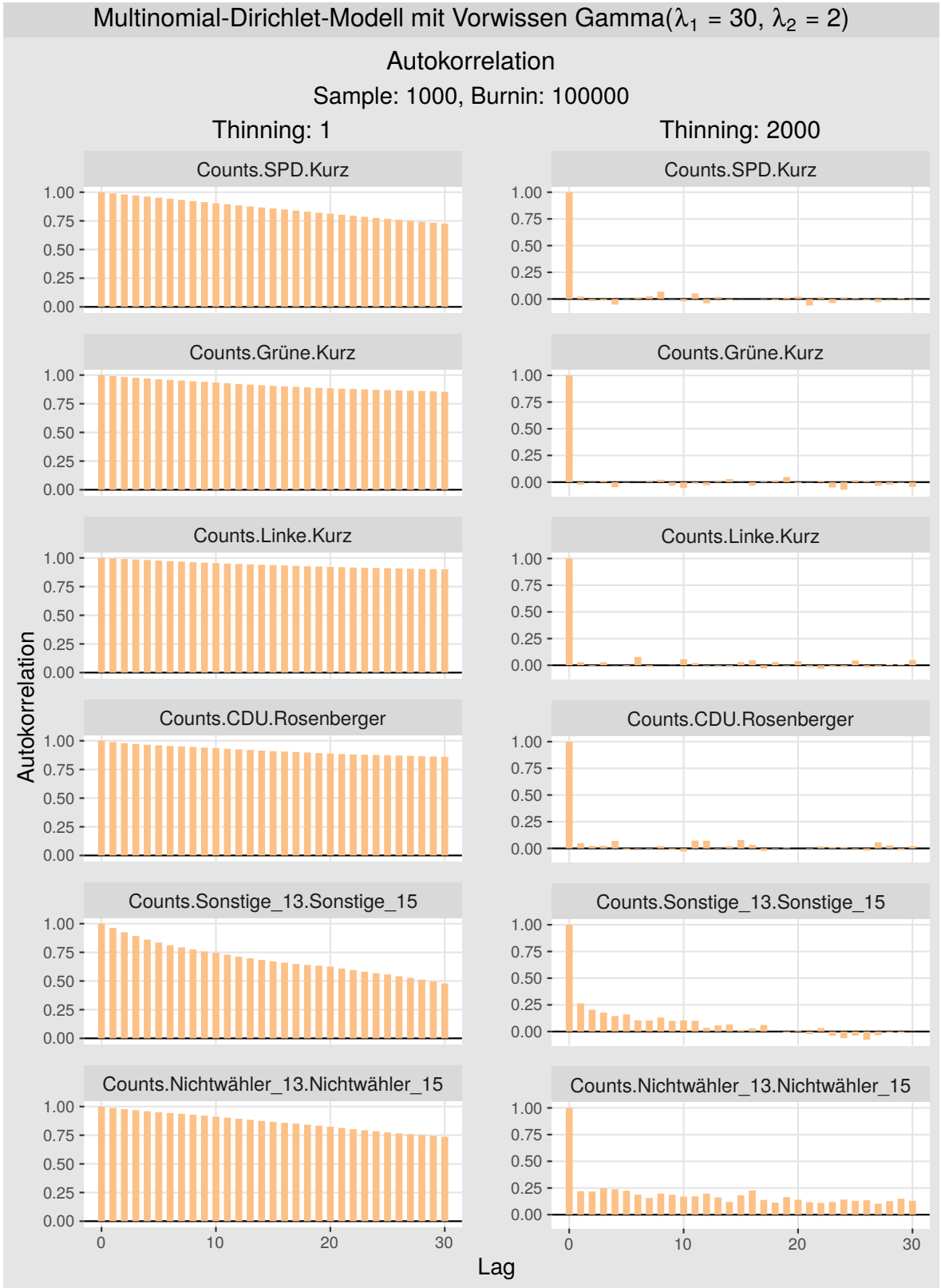

Abbildung A.15: *Die Autokorrelationen der gezogenen absoluten Häufigkeiten (Counts) von fünf gewählten inneren Zellen des ökologischen Multinomial-Dirichlet-Modells mit Hyperpriori-Parameter Gamma*$(\lambda_1 = 30, \lambda_2 = 2)$ *für die Zellen der Loyalen, anhand einer Stichprobe mit* 1 000 *Ziehungen nach dem Burn-In von* 100 000*. Links: Ohne Thinning. Rechts: Thinning von* 2 000*.*

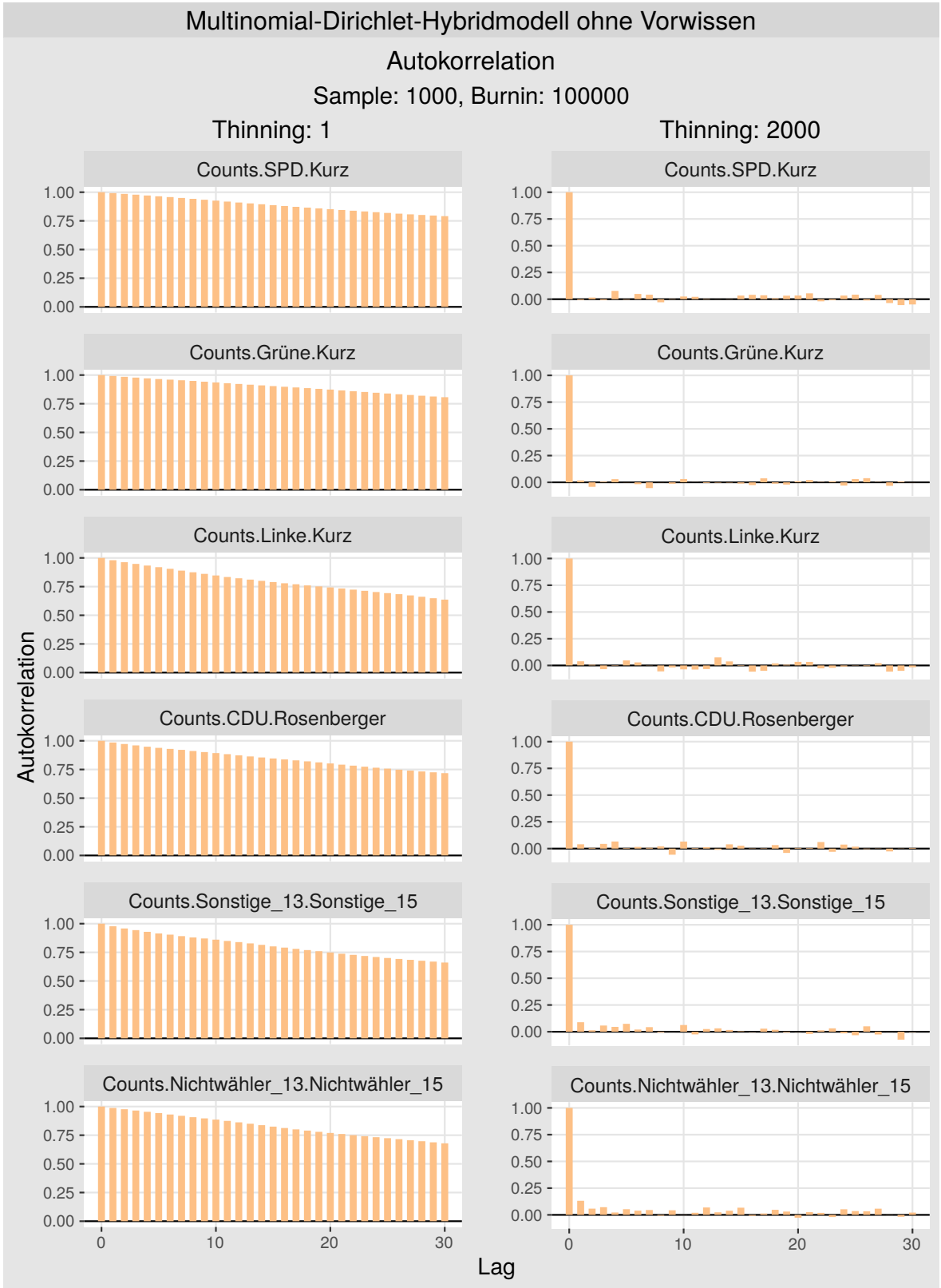

Abbildung A.16: *Die Autokorrelationen der gezogenen absoluten Häufigkeiten (Counts) von fünf gewählten inneren Zellen des Multinomial-Dirichlet-Hybridmodells ohne Vorwissen, anhand einer Stichprobe mit* 1 000 *Ziehungen nach dem Burn-In von* 100 000*. Links: Ohne Thinning. Rechts: Thinning von* 2 000*.*

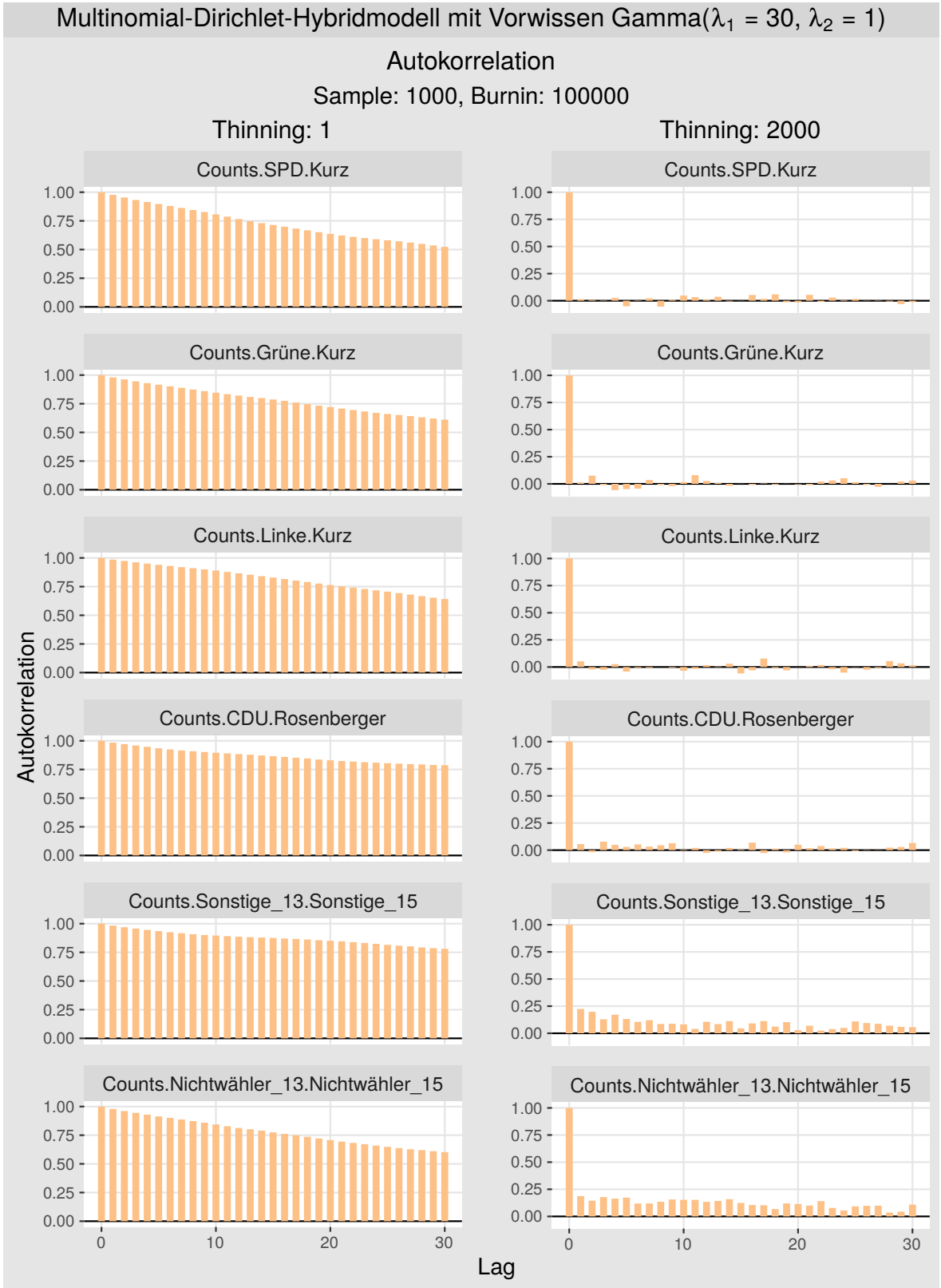

Abbildung A.17: *Die Autokorrelationen der gezogenen absoluten Häufigkeiten (Counts) von fünf gewählten inneren Zellen des Multinomial-Dirichlet-Hybridmodells mit Hyperpriori-Parameter Gamma*($\lambda_1 = 30, \lambda_2 = 1$) *für die Zellen der Loyalen, anhand einer Stichprobe mit* 1 000 *Ziehungen nach dem Burn-In von* 100 000. *Links: Ohne Thinning. Rechts: Thinning von* 2 000.

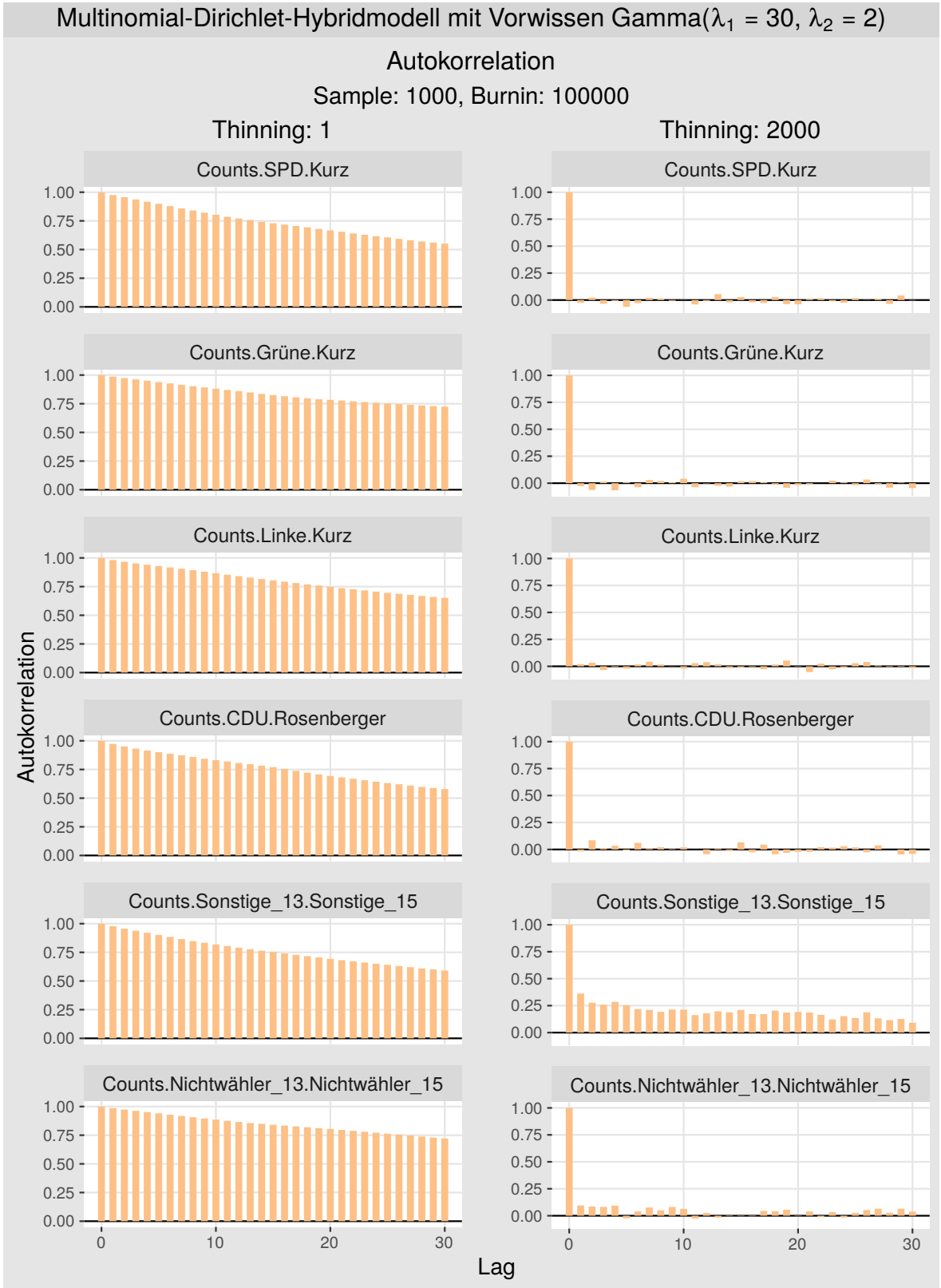

Abbildung A.18: *Die Autokorrelationen der gezogenen absoluten Häufigkeiten (Counts) von fünf gewählten inneren Zellen des Multinomial-Dirichlet-Hybridmodells mit Hyperpriori-Parameter Gamma*$(\lambda_1 = 30, \lambda_2 = 2)$ *für die Zellen der Loyalen, anhand einer Stichprobe mit* 1 000 *Ziehungen nach dem Burn-In von* 100 000*. Links: Ohne Thinning. Rechts: Thinning von* 2 000*.*

A.2.3 Multinomial-Dirichlet-Modell:

Trace of Counts* nach *Burn-In* und *Thinning

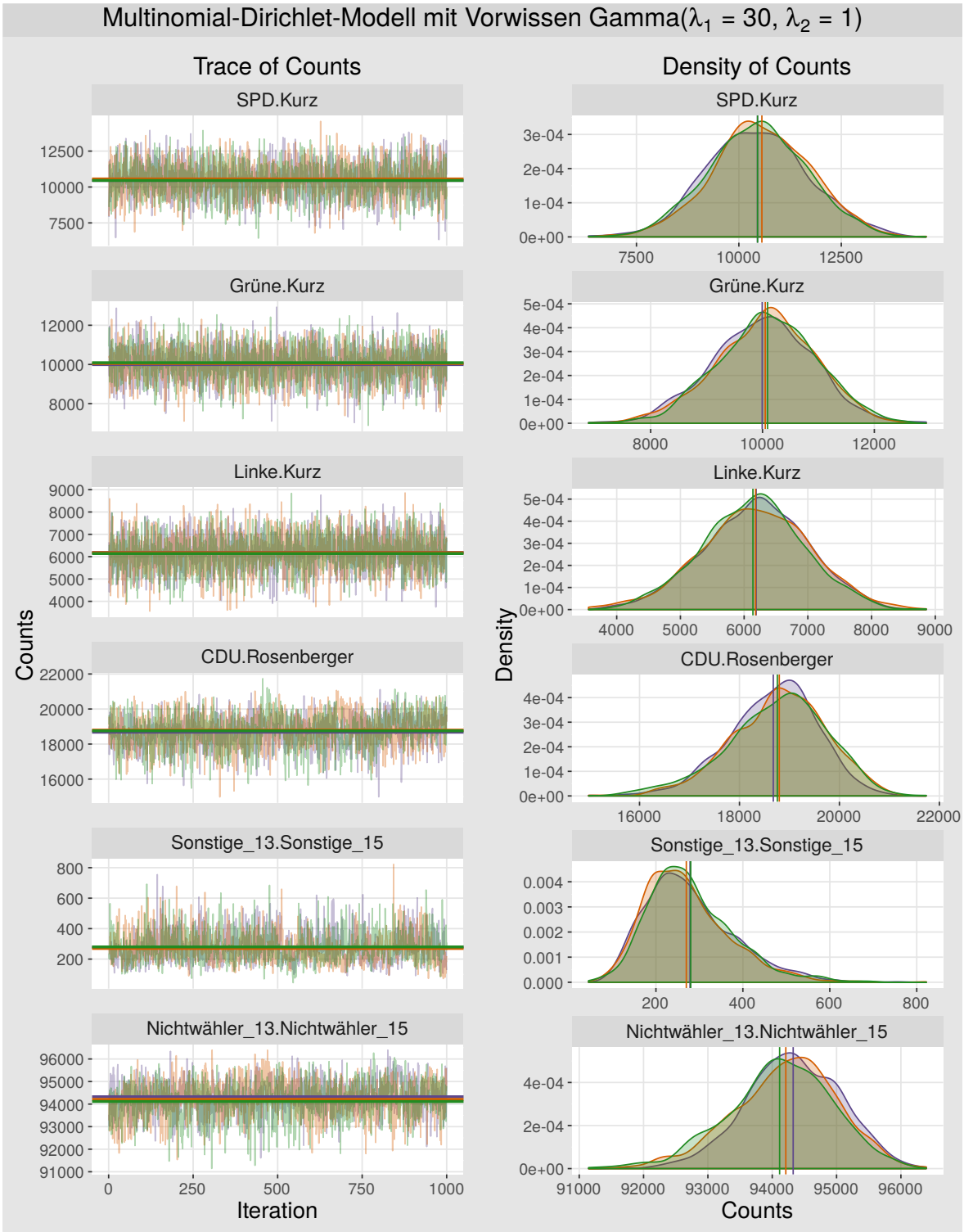

Abbildung A.19: *Drei verdünnte Ketten der gezogenen absoluten Häufigkeiten (Counts) von fünf gewählten inneren Zellen des ökologischen Multinomial-Dirichlet-Modells mit Hyperpriori-Parameter Gamma*$(\lambda_1 = 30, \lambda_2 = 1)$ *für die Zellen der Loyalen. Sample:* 1 000, *Burn-In:* 100 000 *und Thinning:* 2 000. *Links: Trace of Counts der drei Ketten und die dazugehörigen Mittelwerte (waagerechte Linien). Rechts: Dichten der verdünnten Ketten und die gleichen Mittelwerte senkrecht dargestellt.*

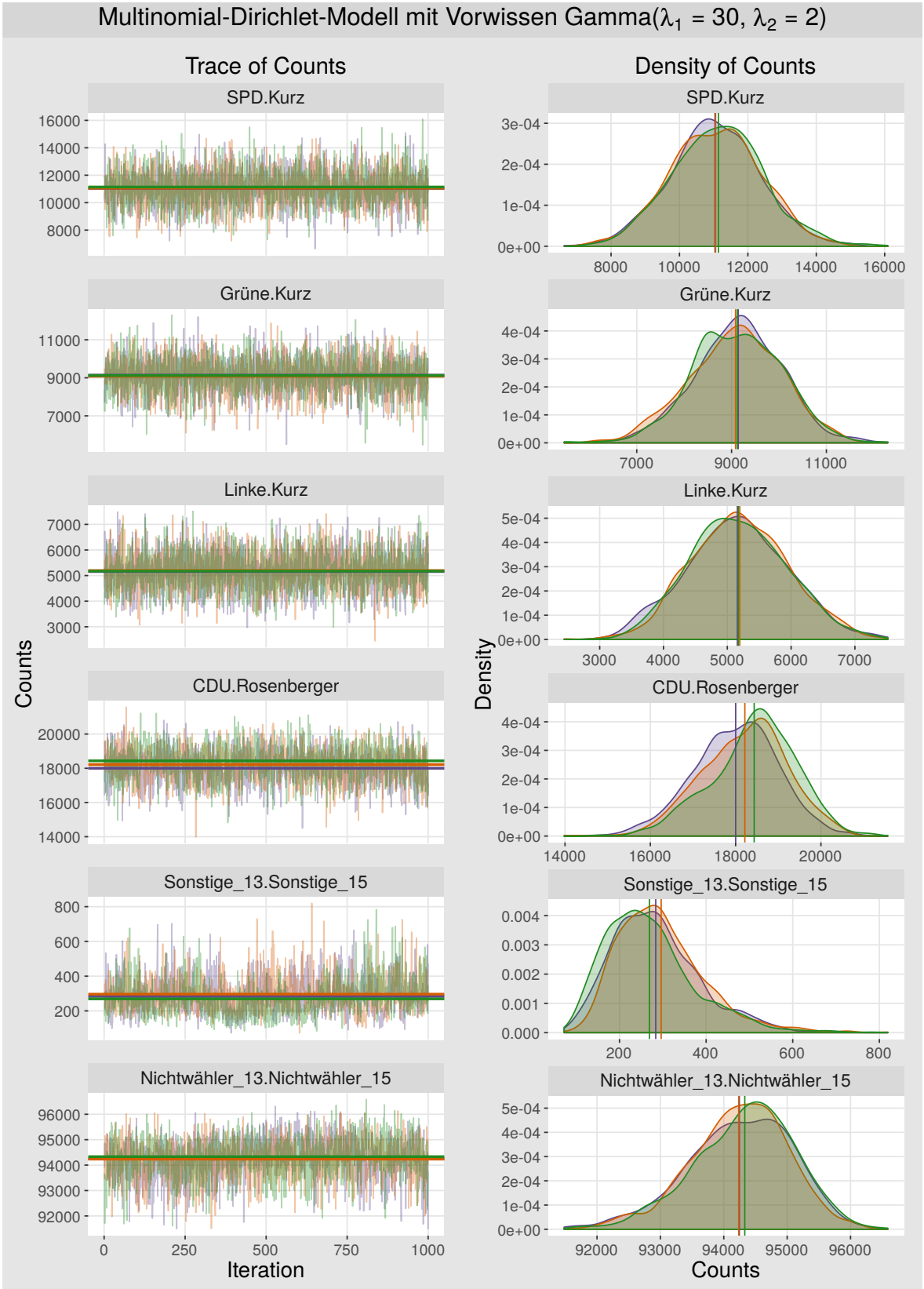

Abbildung A.20: *Drei verdünnte Ketten der gezogenen absoluten Häufigkeiten (Counts) von fünf gewählten inneren Zellen des ökologischen Multinomial-Dirichlet-Modells mit Hyperpriori-Parameter Gamma*$(\lambda_1 = 30, \lambda_2 = 2)$ *für die Zellen der Loyalen. Sample:* 1 000, *Burn-In:* 100 000 *und Thinning:* 2 000. *Links: Trace of Counts der drei Ketten und die dazugehörigen Mittelwerte (waagerechte Linien). Rechts: Dichten der verdünnten Ketten und die gleichen Mittelwerte senkrecht dargestellt.*

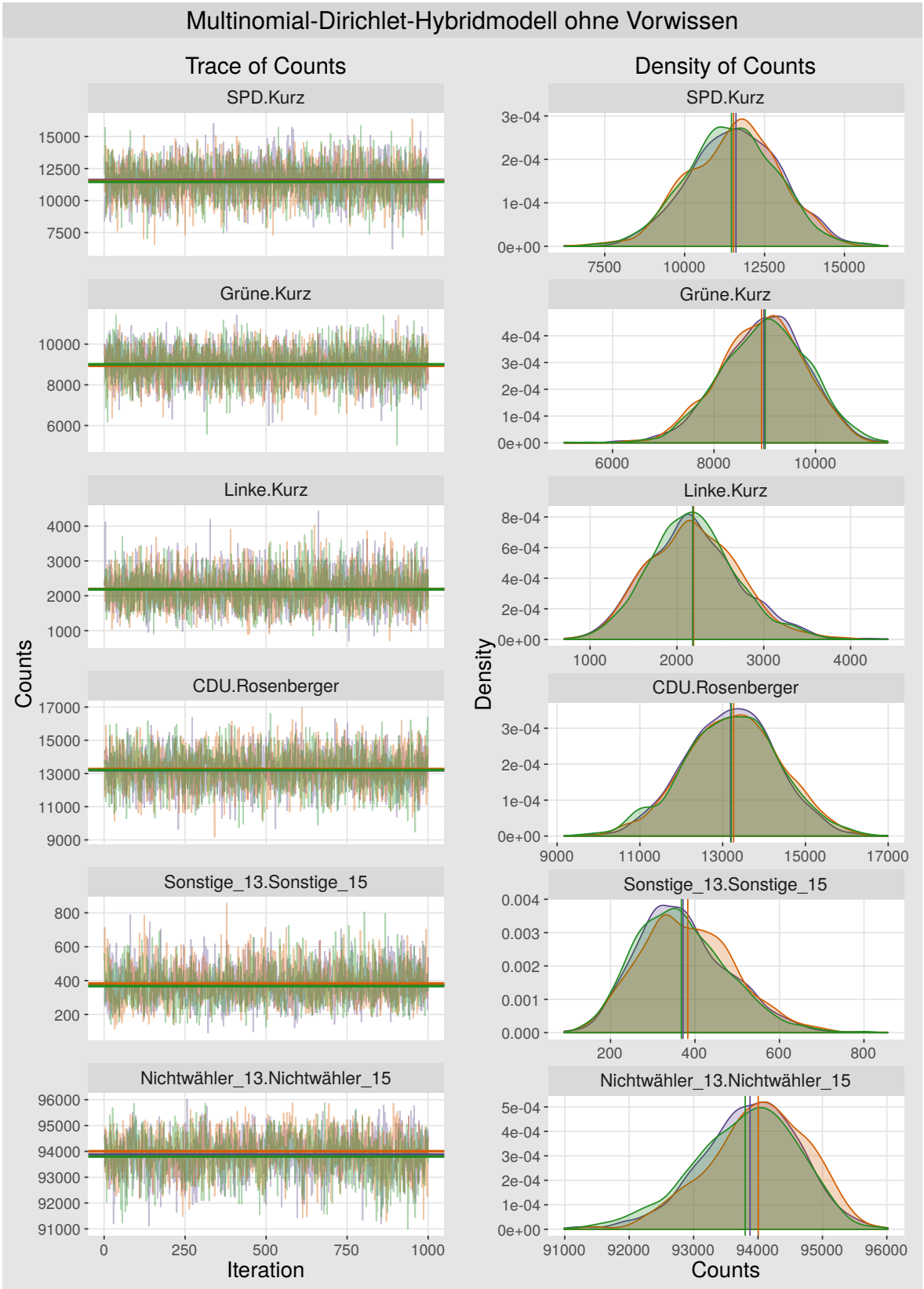

Abbildung A.21: *Drei verdünnte Ketten der gezogenen absoluten Häufigkeiten (Counts) von fünf gewählten inneren Zellen des Multinomial-Dirichlet-Hybridmodells ohne Vorwissen. Sample:* 1 000, *Burn-In:* 100 000 *und Thinning:* 2 000. *Links: Trace of Counts der drei Ketten und die dazugehörigen Mittelwerte (waagerechte Linien). Rechts: Dichten der verdünnten Ketten und die gleichen Mittelwerte senkrecht dargestellt.*

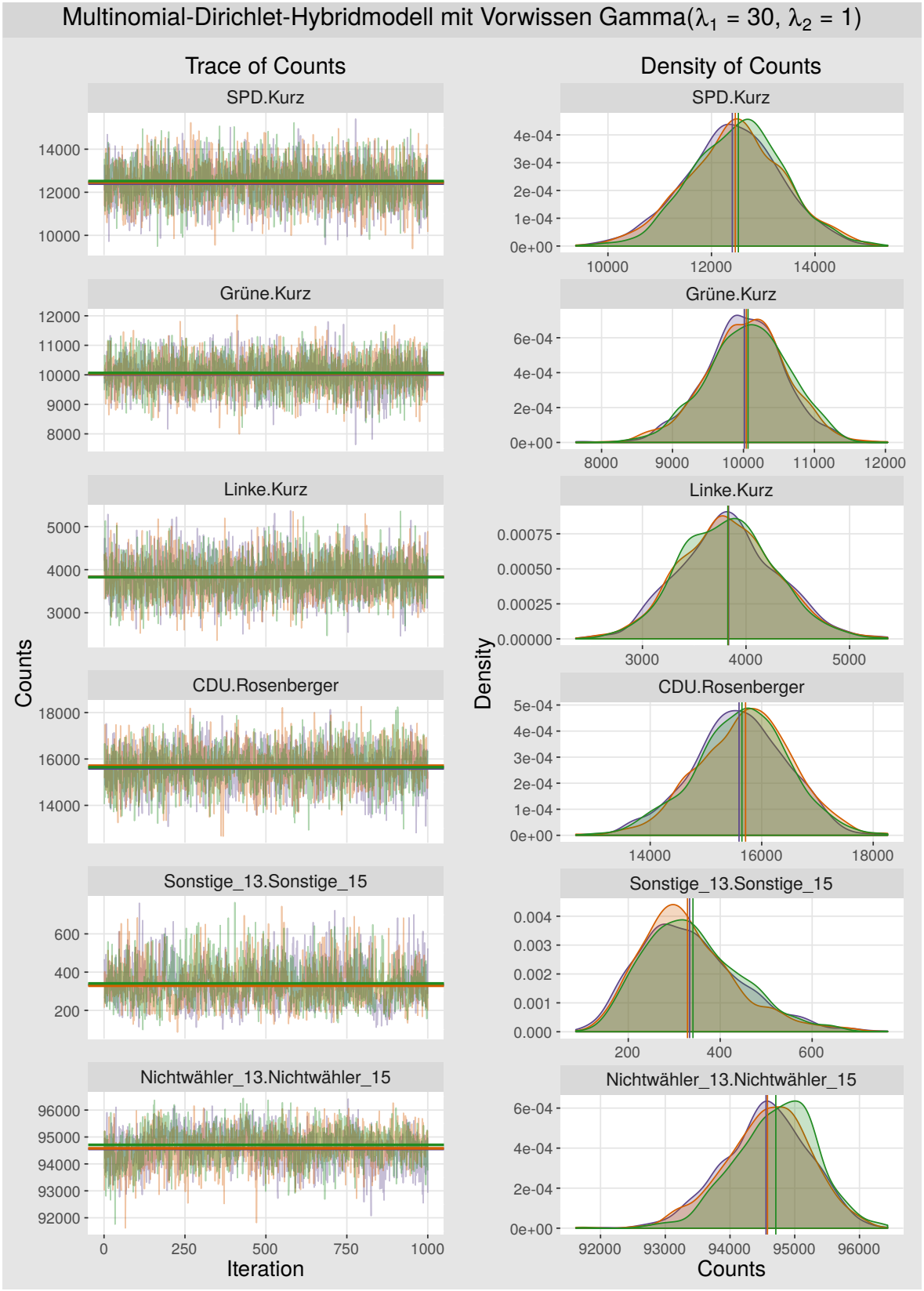

Abbildung A.22: *Drei verdünnte Ketten der gezogenen absoluten Häufigkeiten (Counts) von fünf gewählten inneren Zellen des Multinomial-Dirichlet-Hybridmodells mit Hyperpriori-Parameter Gamma*$(\lambda_1 = 30, \lambda_2 = 1)$ *für die Zellen der Loyalen. Sample:* 1 000, *Burn-In:* 100 000 *und Thinning:* 2 000. *Links: Trace of Counts der drei Ketten und die dazugehörigen Mittelwerte (waagerechte Linien). Rechts: Dichten der verdünnten Ketten und die gleichen Mittelwerte senkrecht dargestellt.*

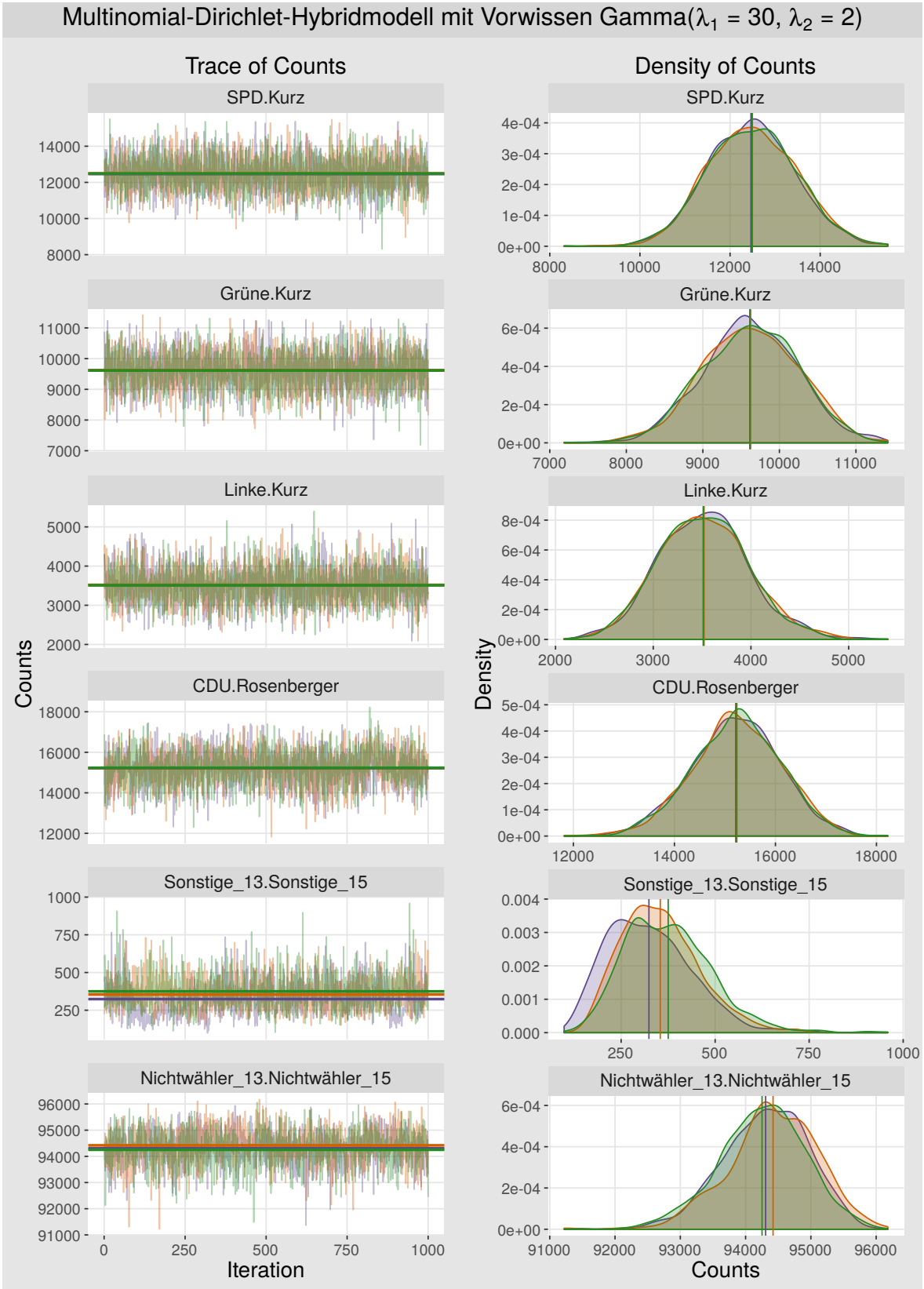

Abbildung A.23: *Drei verdünnte Ketten der gezogenen absoluten Häufigkeiten (Counts) von fünf gewählten inneren Zellen des Multinomial-Dirichlet-Hybridmodells mit Hyperpriori-Parameter Gamma*$(\lambda_1 = 30, \lambda_2 = 2)$ *für die Zellen der Loyalen. Sample:* 1 000, *Burn-In:* 100 000 *und Thinning:* 2 000. *Links: Trace of Counts der drei Ketten und die dazugehörigen Mittelwerte (waagerechte Linien). Rechts: Dichten der verdünnten Ketten und die gleichen Mittelwerte senkrecht dargestellt.*

A.2.4 Multinomial-Dirichlet-Modell: Ketten- und Modellvergleich mittels MAE

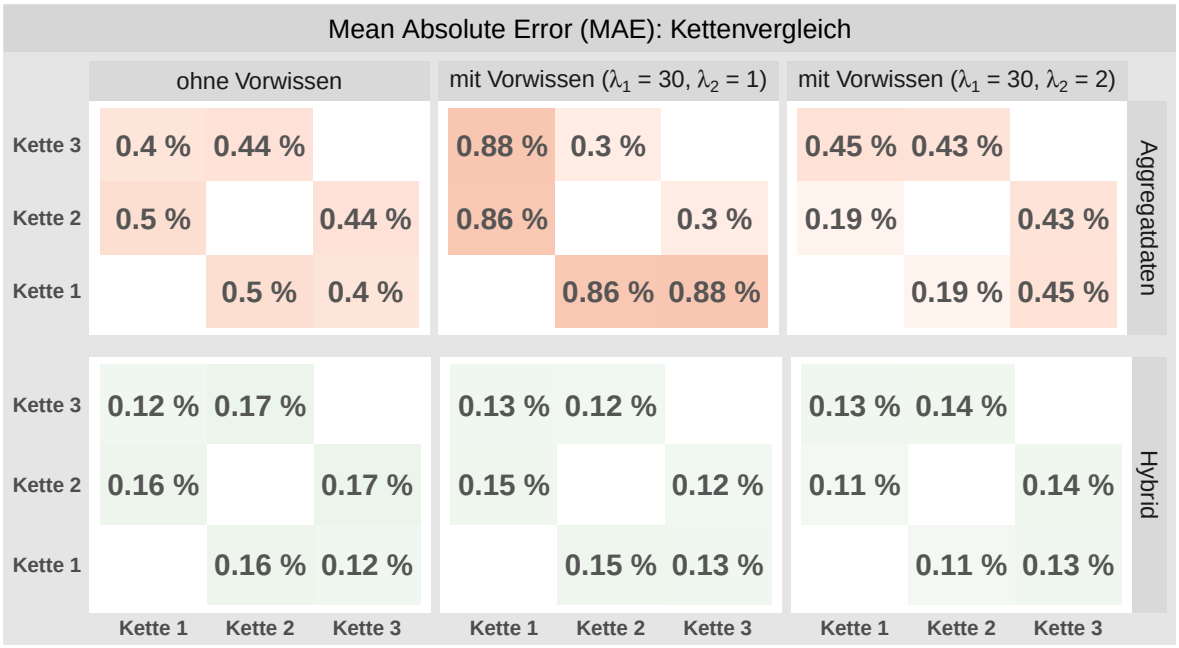

Abbildung A.24: *Mean Absolut Error (MAE) in Prozentpunkten zwischen den Ergebnissen von drei verdünnten Ketten für jede Version des ökologischen Multinomial-Dirichlet-Modells (oben) und des Multinomial-Dirichlet-Hybridmodells (unten). Die Werte sind je nach Modell symmetrisch über die Diagonale.*

Mean Absolute Error (MAE): Modellvergleich

	Aggregatdaten	Aggregatdaten mit Vorwissen (30, 1)	Aggregatdaten mit Vorwissen (30, 2)	Hybrid	Hybrid mit Vorwissen (30, 1)	Hybrid mit Vorwissen (30, 2)
Aggregatdaten		8.28 %	6.34 %	5.82 %	7.44 %	6.97 %
Aggregatdaten mit Vorwissen (30, 1)	8.28 %		2.26 %	5.96 %	4.27 %	4.56 %
Aggregatdaten mit Vorwissen (30, 2)	6.34 %	2.26 %		4.71 %	3.68 %	3.73 %
Hybrid	5.82 %	5.96 %	4.71 %		2.57 %	2.03 %
Hybrid mit Vorwissen (30, 1)	7.44 %	4.27 %	3.68 %	2.57 %		0.6 %
Hybrid mit Vorwissen (30, 2)	6.97 %	4.56 %	3.73 %	2.03 %	0.6 %	

Abbildung A.25: *Mean Absolut Error (MAE) in Prozentpunkten zwischen den Ergebnissen von verschiedenen Versionen des Multinomial-Dirichlet-Modells (symmetrisch über die Diagonale). Zum Vergleich wurde die erste der drei verdünnten Ketten für jede Version des Modells verwendet.*

A.2.5 Multinomial-Log-Normal-Modell: *Trace-* und *Density of Counts*

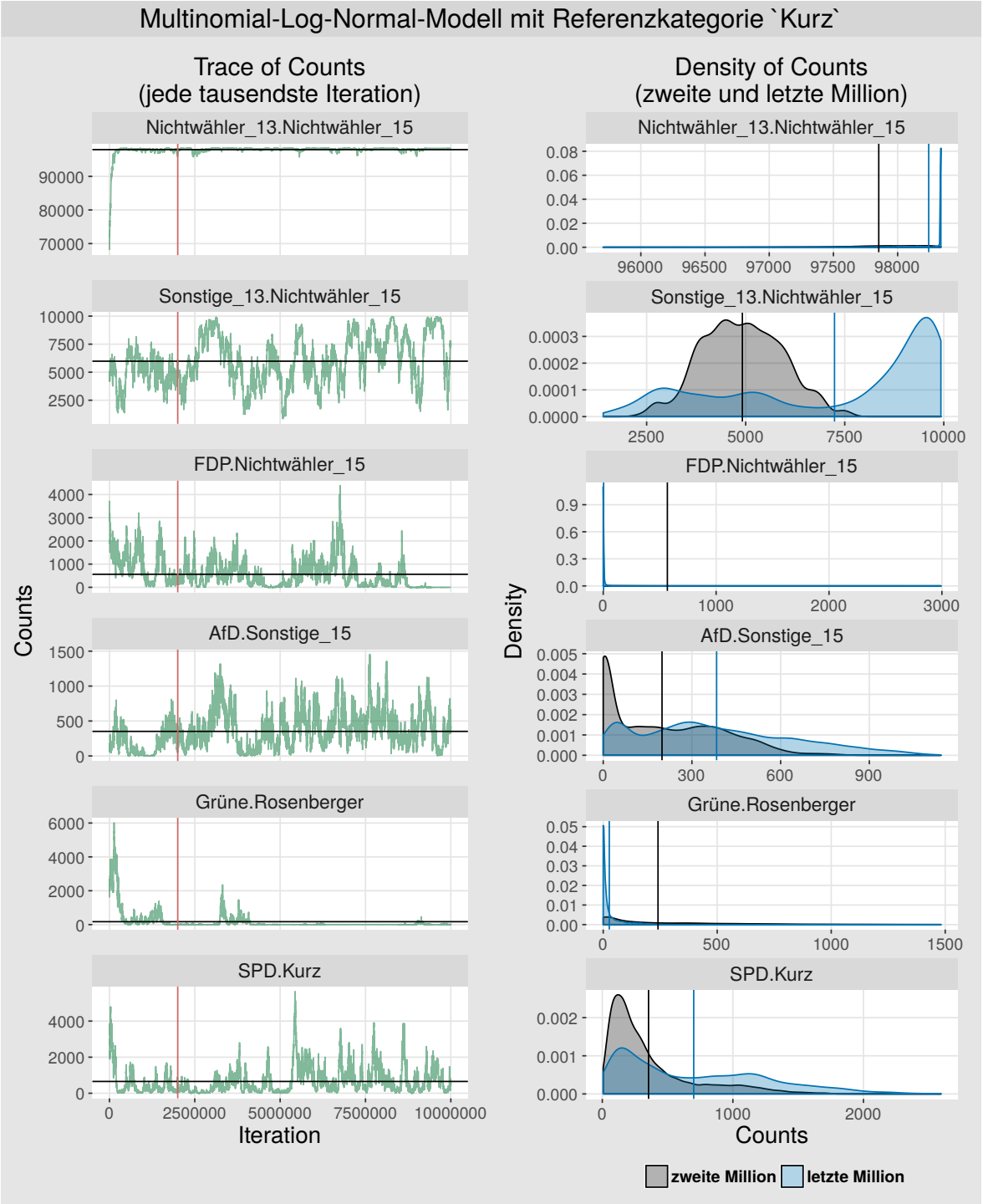

Abbildung A.26: *Die Ketten (links) und die Dichten (rechts) der gezogenen absoluten Häufigkeiten (Counts) von fünf gewählten inneren Zellen des ökologischen Multinomial-Log-Normal-Modells mit Referenzkategorie* `Kurz`*. Von zehn Millionen durchgeführten Iterationen konnte jede hundertste gespeichert werden. Links wird von zehn Millionen Iterationen jede tausendste dargestellt. Die senkrechten roten Linien kennzeichnen die 2 000 000-ste von zehn Millionen Iterationen. Die waagerechten schwarzen Linien zeigen die Mittelwerte von 100 000 gespeicherten Werten. Rechts: Die Dichten der zweiten und der letzten Million (jede hundertste Iteration betrachtet) und die dazugehörigen Mittelwerte (senkrechte Linien).*

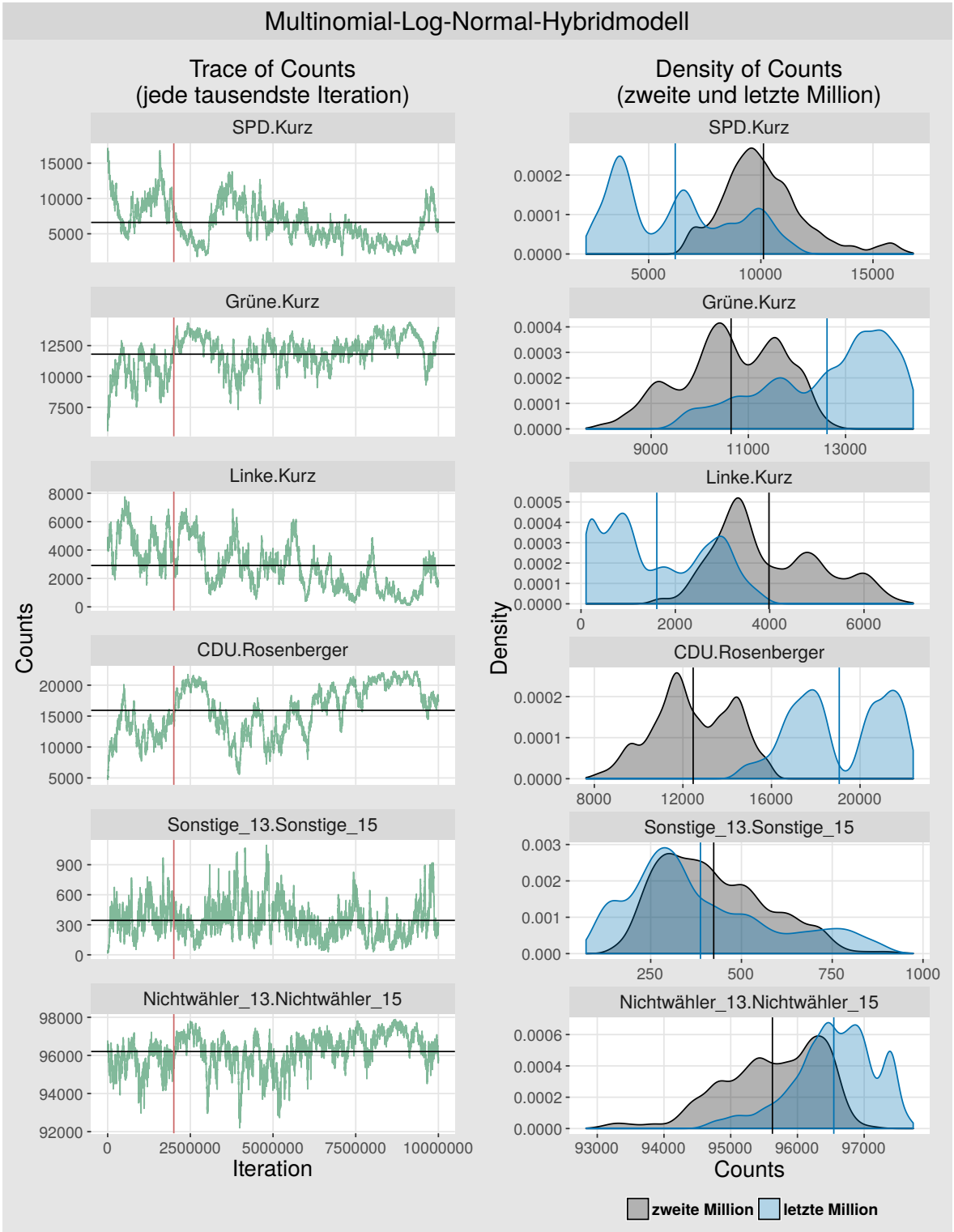

Abbildung A.27: *Die Ketten (links) und die Dichten (rechts) der gezogenen absoluten Häufigkeiten (Counts) von fünf gewählten inneren Zellen des Multinomial-Log-Normal-Hybridmodells mit automatisch gewählter Referenzkategorie* `Nichtwähler_15`*. Von zehn Millionen durchgeführten Iterationen konnte jede hundertste gespeichert werden. Links wird von zehn Millionen Iterationen jede tausendste dargestellt. Die senkrechten roten Linien kennzeichnen die 2 000 000-ste von zehn Millionen Iterationen. Die waagerechten schwarzen Linien zeigen die Mittelwerte von 100 000 gespeicherten Werten. Rechts: Die Dichten der zweiten und der letzten Million (jede hundertste Iteration betrachtet) und die dazugehörigen Mittelwerte (senkrechte Linien).*

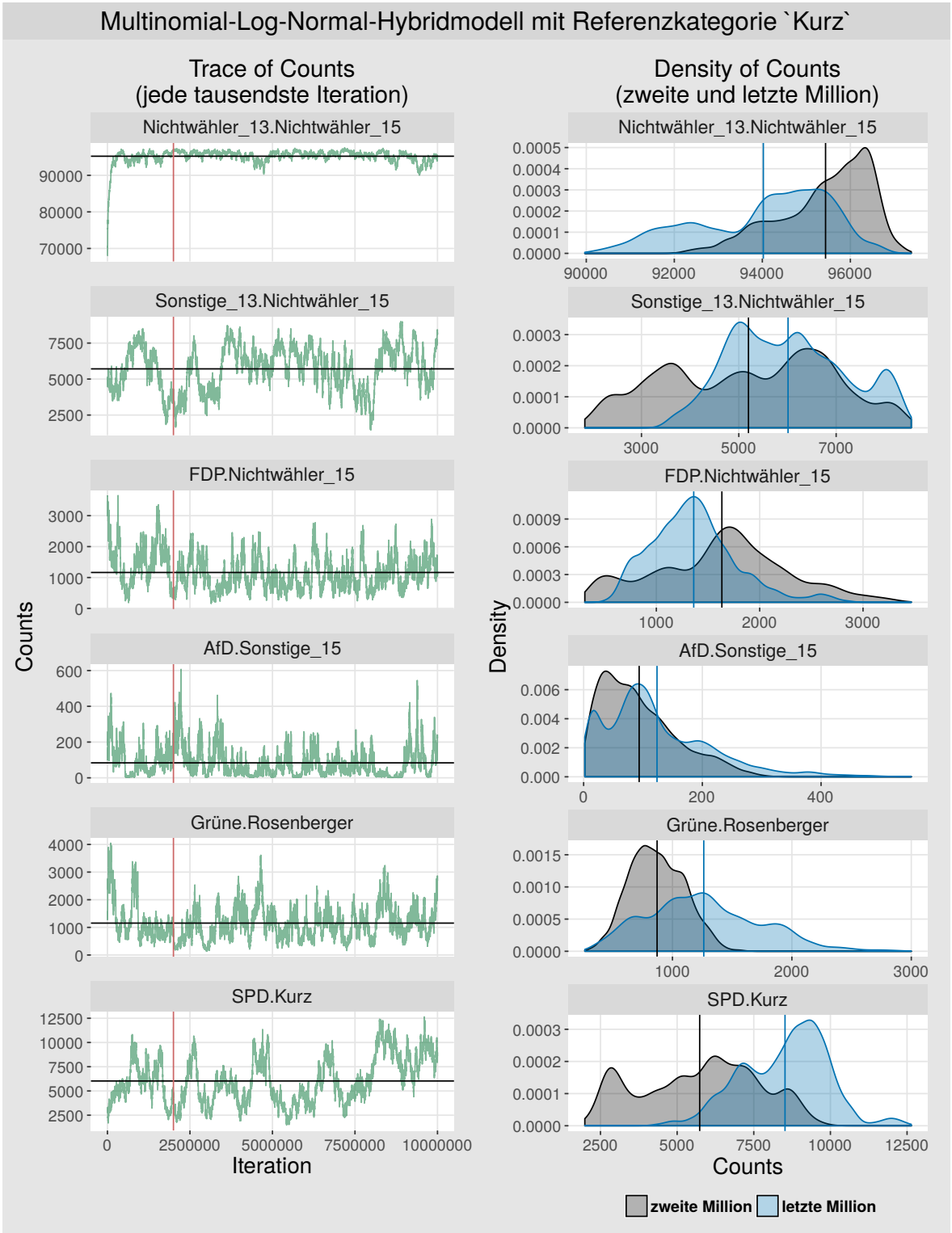

Abbildung A.28: *Die Ketten (links) und die Dichten (rechts) der gezogenen absoluten Häufigkeiten (Counts) von fünf gewählten inneren Zellen des Multinomial-Log-Normal-Hybridmodells mit Referenzkategorie* `Kurz`*. Von zehn Millionen durchgeführten Iterationen konnte jede hundertste gespeichert werden. Links wird von zehn Millionen Iterationen jede tausendste dargestellt. Die senkrechten roten Linien kennzeichnen die 2 000 000-ste von zehn Millionen Iterationen. Die waagerechten schwarzen Linien zeigen die Mittelwerte von 100 000 gespeicherten Werten. Rechts: Die Dichten der zweiten und der letzten Million (jede hundertste Iteration betrachtet) und die dazugehörigen Mittelwerte (senkrechte Linien).*

A.2.6 Multinomial-Log-Normal-Modell: Autokorrelationen

Abbildung A.29: *Die Autokorrelationen der gezogenen absoluten Häufigkeiten (Counts) von fünf gewählten inneren Zellen des ökologischen Multinomial-Log-Normal-Modells mit Referenzkategorie* `Kurz` *anhand einer Stichprobe mit* 1 000 *Ziehungen nach dem Burn-In von* 2 000 000*. Links: Thinning von* 100*. Rechts: Thinning von* 2 000*.*

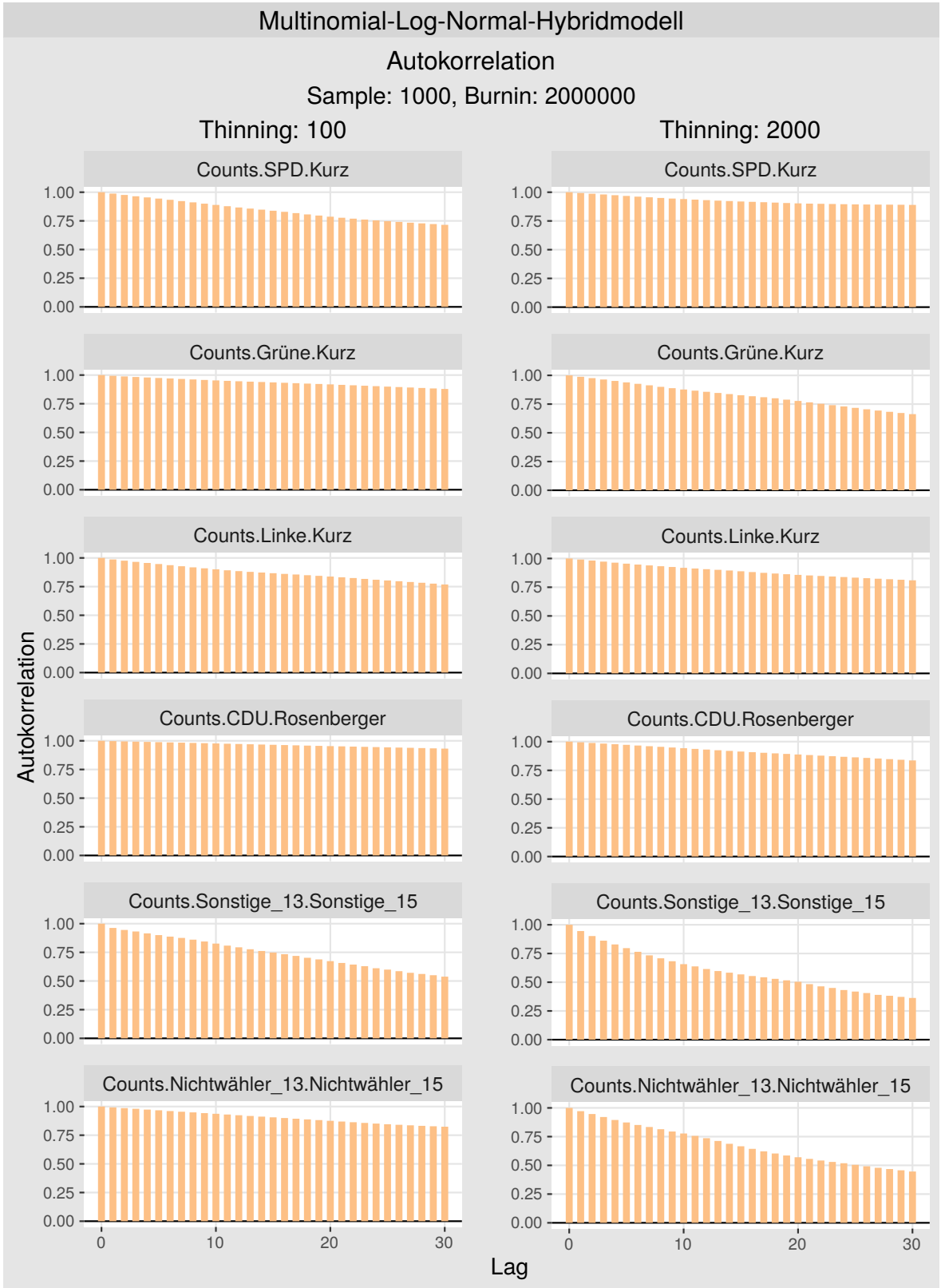

Abbildung A.30: *Die Autokorrelationen der gezogenen absoluten Häufigkeiten (Counts) von fünf gewählten inneren Zellen des Multinomial-Log-Normal-Hybridmodells mit automatisch gewählter Referenzkategorie* `Nichtwähler_15` *anhand einer Stichprobe mit* 1 000 *Ziehungen nach dem Burn-In von* 2 000 000*. Thinning von* 100*. Rechts: Thinning von* 2 000*.*

Abbildung A.31: *Die Autokorrelationen der gezogenen absoluten Häufigkeiten (Counts) von fünf gewählten inneren Zellen des Multinomial-Log-Normal-Hybridmodells mit Referenzkategorie* `Kurz` *anhand einer Stichprobe mit* 1 000 *Ziehungen nach dem Burn-In von* 2 000 000*. Thinning von* 100*. Rechts: Thinning von* 2 000*.*

A.2.7 Multinomial-Log-Normal-Modell:

Trace of Counts* nach *Burn-In* und *Thinning

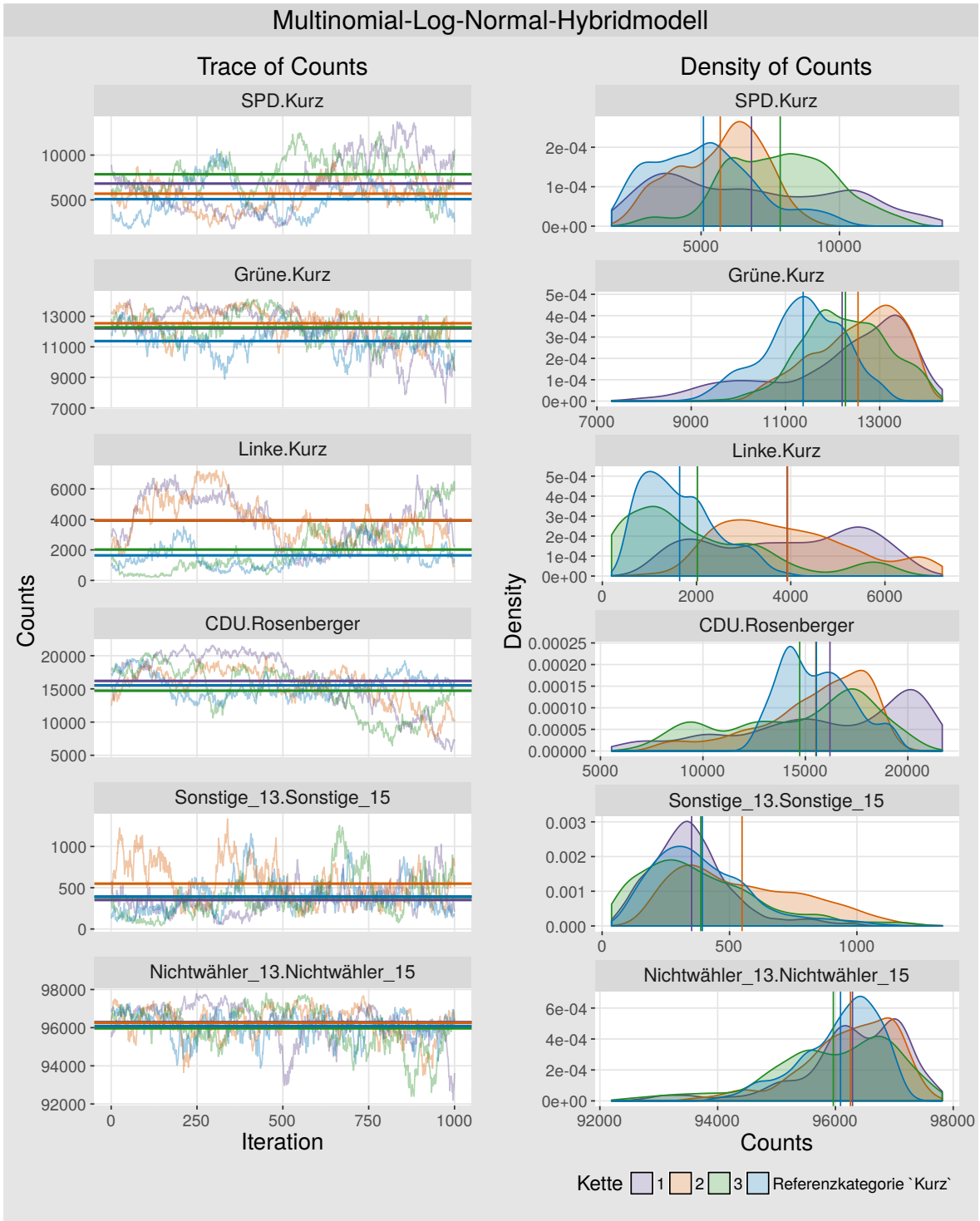

Abbildung A.32: *Drei verdünnte Ketten der gezogenen absoluten Häufigkeiten (Counts) von fünf gewählten inneren Zellen des Multinomial-Log-Normal-Hybridmodells mit automatisch gewählter Referenzkategorie* `Nichtwähler_15` *und eine verdünnte Kette mit Referenzkategorie* `Kurz`*. Sample:* 1 000*, Burn-In:* 2 000 000 *und Thinning:* 2 000*. Links: Trace of Counts der vier Ketten und die dazugehörigen Mittelwerte (waagerechte Linien). Rechts: Dichten der Ketten und die gleichen Mittelwerte senkrecht dargestellt.*

A.2.8 Multinomial-Log-Normal-Modell: Ketten- und Modellvergleich mittels MAE

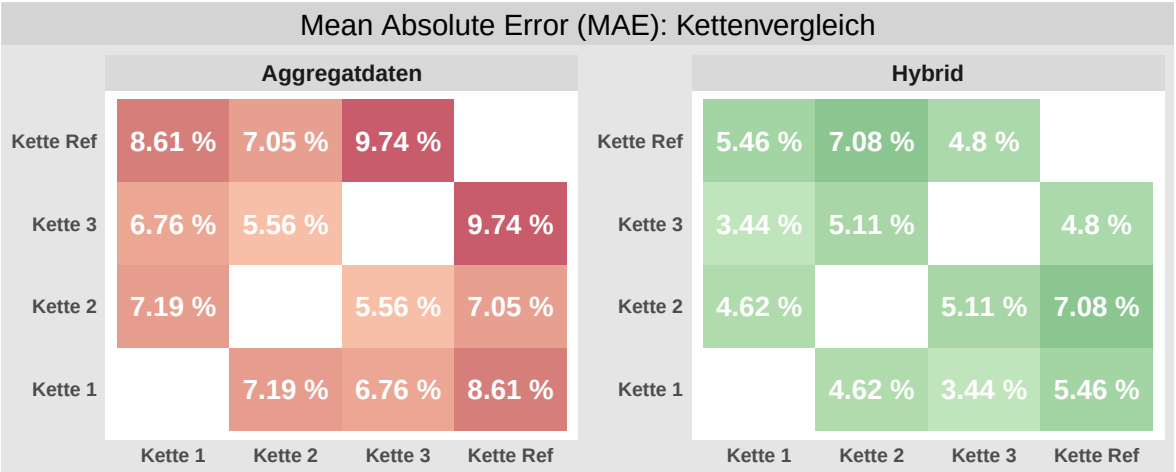

Abbildung A.33: *Mean Absolut Error (MAE) in Prozentpunkten zwischen den Ergebnissen von drei verdünnten Ketten mit automatisch gewählter Referenzkategorie* `Nichtwähler_15` *und einer Kette mit Referenzkategorie* `Kurz` *bei dem ökologischen Multinomial-Log-Normal-Modell (links) und bei dem Multinomial-Log-Normal-Hybridmodell (rechts). Die Werte sind je nach Modell symmetrisch über die Diagonale.*

Mean Absolute Error (MAE): Modellvergleich

	Aggregatdaten	Aggregatdaten Referenzkategorie 'Kurz'	Hybrid	Hybrid Referenzkategorie 'Kurz'
Aggregatdaten		8.61 %	13.8 %	12.66 %
Aggregatdaten Referenzkategorie 'Kurz'	8.61 %		11.6 %	8.79 %
Hybrid	13.8 %	11.6 %		5.46 %
Hybrid Referenzkategorie 'Kurz'	12.66 %	8.79 %	5.46 %	

Abbildung A.34: *Mean Absolut Error (MAE) in Prozentpunkten zwischen den Ergebnissen von verschiedenen Versionen des Multinomial-Log-Normal-Modells (symmetrisch über die Diagonale). Zum Vergleich wurde die erste der drei verdünnten Ketten mit Referenzkategorie* `Nichtwähler_15` *und die Kette mit Referenzkategorie* `Kurz`*, jeweils für die Version mit und ohne Individualdaten, verwendet.*

E Elektronischer Anhang

Der Inhalt der beigelegten CD ist in der Abbildung E.1 aufgelistet. Auf der CD ist ebenfalls diese Arbeit in digitaler Version unter dem Name `"MA_Kopecki.pdf"` vorhanden.

Im Ordner `Briefwaehler_Gewichtung_Bs` befindet sich der Beispielcode zur Addition der Briefwähler und die Erstellung des gewichteten Datensatzes in `R`, welcher mittels der amtlichen Ergebnisse der ersten zwei Wahlbezirke der Oberbürgermeisterwahl 2015 simuliert wird. Die benötigten Daten werden dem Ordner beigelegt.

Der Ordner `Grafiken_Ketten_Matrix` umfasst die Grafiken, welche die drei verdünnten Ketten aller Zellen in einer Matrixform für alle betrachteten Modelle darstellen und eine `README.txt` Datei mit der Beschreibung der Grafiken.

Im Ordner `R_Code` sind alle Programmcodes vorhanden, welche für die Datenaufbereitung, die Erstellung der Grafiken und für die Analyse verwendet wurden. Die erzeugten Dateien werden automatisch in den vier dazugehörigen Ordnern gespeichert und im weiteren Verlauf geladen. Das heißt, dieser Ordner und alle R Dateien müssen dem „Working Directory“ beigelegt werden. Im Ordner `Daten` liegen die nötigen rohen Datensätze vor. Aus Datenschutzgründen dürfen die Daten für das Erstellen der Grafiken zum Alter und zur Bildung der Befragten nicht beiliegen. Die Grafiken, die mithilfe des Codes nicht hergestellt werden können, befinden sich im Ordner `Grafiken/Deskriptive_Analyse`. Die R Dateien wurden für die Berechnungen auf dem Server vorbereitet. Die Nummern der Dateien geben die Reihenfolge der Durchführung an. Das heißt, die Datei unter der Nummer 02 kann erst dann durchgeführt werden, wenn der Durchlauf der Datei unter der Nummer 01 fertig ist. Die Dateien, die die gleichen Nummern besitzen (beispielsweise 03a bis 03j), können auf dem Server parallel berechnet werden. Eine Auflistung aller Pakete, die vor der Analyse installiert werden müssen, befindet sich in der Datei `00_Pakete.R`. Falls die Berechnungen auf einem privaten Rechner durchgeführt werden möchten, sind die Beschreibungen am Anfang des Codes zu beachten.

Die Datei `README.txt` enthält die vorliegende Beschreibung des CD Inhaltes.

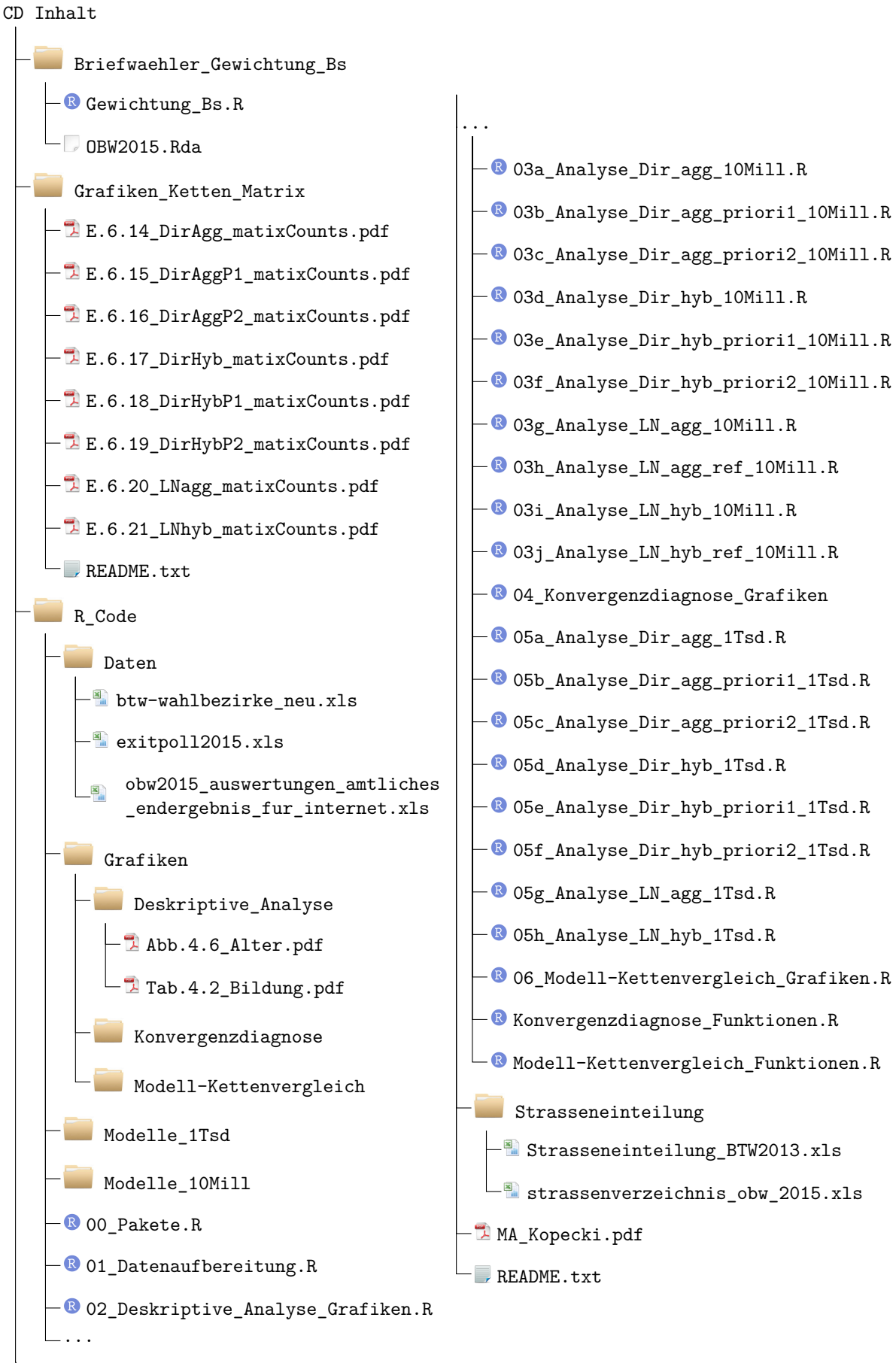

Abbildung E.1: *Inhalt der beigelegten CD*

Erklärung

Hiermit versichere ich, dass ich die vorliegende Masterarbeit selbständig angefertigt und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Die, aus fremden veröffentlichten oder nicht veröffentlichten Quellen, wörtlich oder sinngemäß übernommenen Gedanken sind als solche kenntlich gemacht. Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

München, den 17. August 2016

(Ort/Datum)

(Unterschrift)