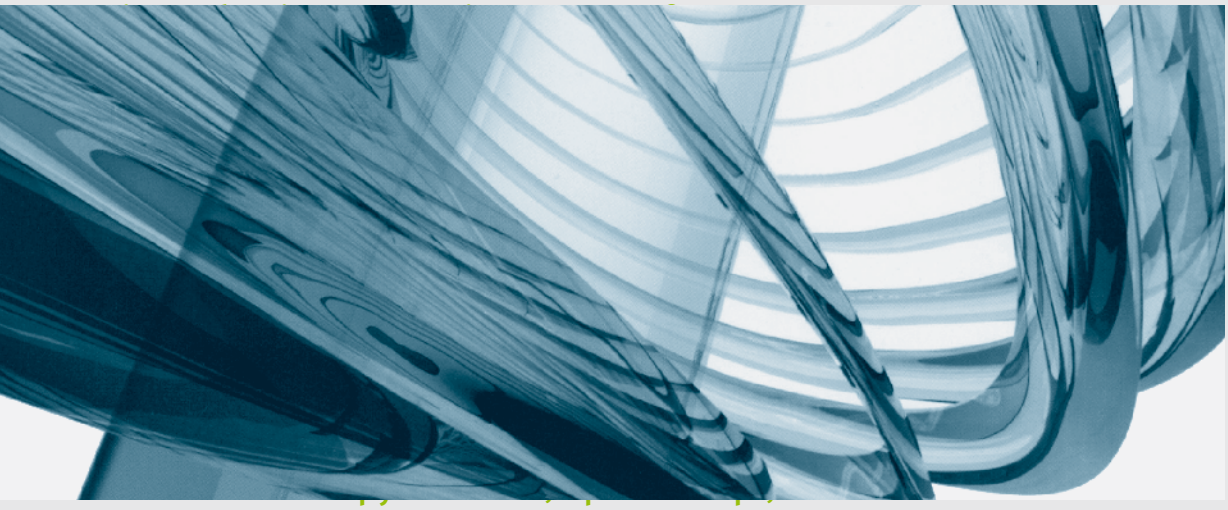
DESTATIS
wissen.nutzen.

STATISTIK UND WISSENSCHAFT

Ralf Münnich, Siegfried Gabler u.a.

Stichprobenoptimierung und Schätzung im Zensus 2011



Band 21

Statistisches Bundesamt

Bibliographische Information Der Deutschen Bibliothek
Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Daten sind im Internet über www.d-nb.de abrufbar.

Herausgeber: Statistisches Bundesamt, Wiesbaden

Internet: www.destatis.de

Ihr Kontakt zu uns:

www.destatis.de/kontakt

Informationen zu dieser Publikation unter

Tel.: +49 (0) 611 / 75 28 87

Statistischer Informationsservice

Tel.: +49 (0) 611 / 75 24 05
Fax: +49 (0) 611 / 75 33 30

Erschienen im Juli 2012

Print
Preis: EUR 24,80 [D]
Bestellnummer: 1030821-12900-1
ISBN: 978-3-8246-0992-5

Kostenfreier Download (PDF)
Artikelnummer: 1030821-12900-4

Vertriebspartner: IBRo Versandservice GmbH
Bereich Statistisches Bundesamt
Kastanienweg 1
18184 Roggentin
Deutschland
destatis@s-f-g.com
Tel.: + 49 (0) 3 82 04/ 6 65 43
Fax: + 49 (0) 3 82 04/ 6 69 19

Geleitwort

Das Forscherteam um Professor Dr. Münnich (Universität Trier) und PD Dr. Gabler (GESIS Mannheim) stellt mit dem vorliegenden Band 21 der Reihe *Statistik und Wissenschaft* die Ergebnisse und Empfehlungen des *Stichprobenforschungsprojekts zum deutschen Zensus 2011* vor. Damit stehen die Resultate eines mehr als 3½ Jahre dauernden Großprojekts zur methodischen Entwicklung der Stichprobenerhebung im Rahmen des registergestützten Zensus erstmals auch der wissenschaftlichen Öffentlichkeit zur Verfügung. In der Zensus-Haushaltebefragung wurden knapp 10 % der Bevölkerung Deutschlands befragt, um Informationen zu Über- und Untererfassungen im Melderegister, aber auch zu solchen Personenmerkmalen wie z. B. Bildung oder Erwerbstätigkeit zu erhalten, die aus vorhandenen Registern nicht gewonnen werden können. Die von den Forschern abgegebenen Empfehlungen basieren auf empirischen Forschungsergebnissen, die anhand umfangreicher Simulationsrechnungen auf einem möglichst realitätsgetreuen Datenbestand – einem „Modell" der realen Grundgesamtheit der Bevölkerung Deutschlands mit ihren zensusrelevanten Personenmerkmalen – gewonnen wurden. Das Gutachten diente bereits als methodisches Grundlagenpapier für die Ziehung der Stichprobe für die Haushaltebefragung. Jetzt werden die Forschungsergebnisse genutzt, um die Ergebnisse aus der Haushaltebefragung mittels des empfohlenen Regressionsschätzverfahrens hochzurechnen. Damit setzt die amtliche Statistik bei der Ermittlung der amtlichen Einwohnerzahlen die neuesten wissenschaftlichen Erkenntnisse um, die aus dem Projekt entstanden sind.

Ich bedanke mich bei allen, die im Rahmen dieses Forschungsprojekts – im Wissenschaftlerteam und in der amtlichen Statistik – zur Lösung dieser komplexen mathematisch-statistischen Fragestellungen beigetragen haben. Die Veröffentlichung in der Reihe *Statistik und Wissenschaft* möchte diese erfolgreiche Zusammenarbeit dokumentieren und anerkennen.

Wiesbaden, im Mai 2012

Roderich Egeler

Präsident des Statistischen Bundesamtes

Vorwort

Am 29. August 2006 wurde vom Bundeskabinett beschlossen, 2011 einen Register-gestützten Zensus durchzuführen. Das neue Zensusmodell beinhaltet u.a. eine statistische Korrektur von Melderegisterdaten um Über- und Untererfassungen (Karteileichen und Fehlbestände) sowie eine Erhebung zusätzlicher, nicht aus Registern verfügbarer Merkmale. Beides wird durch eine Haushaltebefragung auf Stichprobenbasis realisiert. Damit wurde ein Paradigmenwechsel in der Deutschen amtlichen Statistik herbeigeführt, der die Erforschung neuer statistischer Methoden erforderlich machte. Das Statistische Bundesamt vergab einen Forschungsauftrag, Stichprobendesign und Schätzmethodik für die Haushaltsstichprobe zu untersuchen – das Zensus-Stichprobenforschungsprojekt. Die vorliegende Monografie gibt einen Überblick über die Forschungsergebnisse dieses Projekts und insbesondere die daraus abgeleiteten Empfehlungen.

Die Autoren danken ausdrücklich den Statistischen Ämtern des Bundes und der Länder für den Forschungsauftrag und die konstruktive Zusammenarbeit. Ein besonderer Dank geht an die Kolleginnen und Kollegen der Projektgruppe 3, die auf Seiten des Auftraggebers für das Projekt und dessen Verlauf zuständig waren, für zahlreiche sehr inspirierende Diskussionen, die die Autoren immer wieder herausgefordert haben, Theorie und Anwendung bestmöglich in Einklang zu bringen. Insbesondere sind hier Herr Wolf Bihler vom Statistischen Bundesamt und Herr Josef Schäfer von IT.NRW zu nennen.

Ebenso danken wir der Zensuskommission für ihre Impulse und das stete Interesse an unserer Forschung und deren praktische Umsetzung.

Zu besonderem Dank sind wir unseren Beratern, Herrn Professor Partha Lahiri, PhD, Joint Programme of Survey Methodology und University of Maryland, sowie Herrn Professor Dr. Ulrich Rendtel, Freie Universität Berlin, verpflichtet. Mit zahlreichen Diskussionen und Hinweisen haben sie zu einer höheren Qualität unserer Ergebnisse sowie dieser Monografie beigetragen.

Danken möchten wir außerdem den Kolleginnen und Kollegen und Mitarbeiterinnen und Mitarbeitern beider Teams in Mannheim und Trier, insbesondere Frau Lucie Dostál und Herrn Dr. Martin Vogt, die ihre Dissertationen im Rahmen des Zensus-Projektes geschrieben haben.

Wir sehen dieses Projekt als exzellentes Beispiel der Forschungskooperation zwischen amtlicher und universitärer Statistik.

Trier und Mannheim, im April 2012

Ralf Münnich, Siegfried Gabler, Matthias Ganninger,
Jan Pablo Burgard und Jan-Philipp Kolb.

Inhalt

Abbildungsverzeichnis

Um die Ablesbarkeit zu erhöhen, finden sich ausgewählte Abbildungen in vergrößerter Darstellung im Kapitel 8 ab Seite 151.

Tabellenverzeichnis

Symbolverzeichnis

α	Präzisionsangabe
β	Regressions-Parametervektor
β_g	Regressions-Parametervektor für Gemeinde g
$\widehat{\beta}$	Schätzung des Regressions-Parametervektor
$\widehat{\beta}_g$	Schätzung des Regressions-Parametervektor in Gemeinde g
δ_d^2	Summe der reskalierten quadrierten Gewichte in Domain d
ε	Regressions-Fehlerterm
$e_{i,d}$	Residuen für Einheit i in Domain d
g	Gemeinde g
γ_d	Gewichte beim zusammengesetzten Schätzer in Domain d
γ_d^{opt}	Optimale Gewichtung für einen zusammengesetzten Schätzer in Domain d
M_h	Obergrenze der Box-Constraints in Schicht h
m_h	Untergrenze der Box-Constraints in Schicht h
MSE	Mean Squared Error
$MSE(*)$	(Design)-MSE des Schätzers $*$
$\widehat{MSE}(*)$	geschätzter (Design)-MSE eines Schätzers $*$
μ_d	Erwartungswertvektor in Domain d
N_{00}	Anzahl der Elemente im Durchschnitt der beiden Listen C und S
N_{11}	Anzahl der Elemente, die zu keiner der beiden Listen C und S gehören
N_C	Anzahl der Elementen in Liste C
N_F	Anzahl der Elementen nur in Liste S aber nicht in Liste C
N_K	Anzahl der Elementen nur in Liste C aber nicht in Liste S
N_S	Anzahl der Elementen in Liste S
N_h, $N_{h,A}$	Anzahl der Anschriften in Schicht h
$N_{h,d}$	Anzahl der Anschriften in Schicht h und Domain d
n_h, $n_{h,A}$	Anzahl der Anschriften in einer Stichprobe der Schicht h
$n_{h,prop}$	Stichprobenumfang in Schicht h bei proportionaler Allokation
$n_{h,opt}$	Stichprobenumfang in Schicht h bei optimaler Allokation
N_d	Umfang der Domain d
$\widehat{N}_d$	Geschätzter Umfang der Domain d

N_p	Umfang der Gesamtheit U
$\widehat{N}_P$	Petersen-Schätzer
$\widehat{N}_P^{DSE}$	Dual System-Schätzer
$\widehat{N}_P^{USA,2000}$	Dual System-Schätzer in USA aus dem Jahr 2000
$\mu_{X,d}$	Vektor der Erwartungswerte der Hilfsvariable X in Domain d
$\mu_{Y,d}$	Vektor der Erwartungswerte der Untersuchungsvariable Y in Domain d
$\widehat{\mu}_{Y,d}^{EBLUPA}$	EBLUPA-Schätzer für $\mu_{Y,d}$
$\widehat{\mu}_{Y,d}^{Synth}$	Naiver synthetischer Schätzer für $\mu_{Y,d}$
$\widehat{\mu}_{Y,d}^{SynthB}$	Synthetischer Schätzer für $\mu_{Y,d}$, Modell B
$\widehat{\mu}_{Y,d}^{YouRao}$	Schätzer von YOURAO für $\mu_{Y,d}$
p	Anteil einer Markmalsausprägung an der Gesamtzahl der Zensusbevölkerung
π_i	Inklusionswahrscheinlichkeit für Einheit i
p_{iR}	Wahrscheinlichkeit für Einheit i, der Registerbevölkerung anzugehören
p_{iZ}	Wahrscheinlichkeit für Einheit i, der Zensusbevölkerung anzugehören
p_K	Wahrscheinlichkeit zur Liste C aber nicht zur Liste S zu gehören
p_F	Wahrscheinlichkeit zur Liste S aber nicht zur Liste C zu gehören
p_{00}	Wahrscheinlichkeit zum Durchschnitt der beiden Listen C und S zu gehören
p_{11}	Wahrscheinlichkeit zu keiner der beiden Listen C und S zu gehören
$\widehat{\psi}$	Vektor der Varianzkomponenten
$RRMSE$	Relativer Root Mean Squared Error
$\widehat{\sigma}_u^2$	Geschätzte Modellvarianz
σ_u^2	Varianz der Effekte auf Domain-Ebene
σ_e^2	Varianz der individuellen Effekte
$s_{h,e,d}^2$	Stichprobenvarianz der Residuen e in Schicht h und Domain d
$s_{h,Y,d}^2$	Stichprobenvarianz des Untersuchungsmerkmals Y Schicht h und Domain d
$S_{h,A}^2$	Varianz der Anschriftengrößen Schicht h
$S_{h,Y}^2$	Varianz der y-Werte in Schicht h
$\widehat{t}_{h,Y}$	Schätzer für $\tau_{Y,h}$
τ_Z	Umfang der Zensusbevölkerung
$\tau_{h,Z}$	Umfang der Zensusbevölkerung in Schicht h
τ_F	Umfang der Fehlbestände

$\widehat{\tau}_F$	Geschätzter Umfang der Fehlbestände
τ_K	Umfang der Karteileichen
$\widehat{\tau}_K$	Geschätzter Umfang der Karteileichen
τ_R	Totalwert Registerbevölkerung
$\tau_{h,R}$	Umfang der Registerbevölkerung in Schicht h
$\tau_{R,dk}$	Umfang der Registerbevölkerung in Domain d und Klasse k
τ_Y	Summe der y-Werte in der Gesamtheit
$\widehat{\tau}_Y$	Geschätzte Summe der y-Werte in der Gesamtheit
$\tau_{Y,d}$	Summe der y-Werte in Domain d
$\tau_{Y<area>}$	Summe der y-Werte in Area $area$
$\widehat{\tau}_{Y<area>}$	Geschätzte Summe der y-Werte in Area $area$
$\widehat{\tau}_{Y,d}^{comp}$	Zusammengesetzter Schätzer für $\tau_{Y,d}$
$\widehat{\tau}_{Y,d}^{dir}$	Direkter Schätzer für $\tau_{Y,d}$
$\widehat{\tau}_{Y,d}^{GREG}$	GREG-Schätzer für $\tau_{Y,d}$
$\widehat{\tau}_{Y,d}^{HT}$	Horvitz-Thompson-Schätzer für $\tau_{Y,d}$
$\widehat{\tau}_{Y,d}^{SynthR}$	Regressions-synthetischer Schätzer für $\tau_{Y,d}$
τ_Z	Umfang der amtlichen Einwohnerzahl
$\widehat{\tau}_Z$	Geschätzter Umfang der amtlichen Einwohnerzahl
τ_Z	Umfang der amtlichen Einwohnerzahl in Domain d
$\widehat{\tau}_{Z,d}^{SPREE}$	Structure PREserving Estimator für $\tau_{Z,d}$
$\widehat{\tau}_{Z,k}^{CHAP-GER}$	Chapman-Schätzer für $\tau_{Z,d}$
$\widehat{\tau}_{Z<SDT>}$	Geschätzter Totalwert im Stadtteil SDT
$\widehat{\tau}_{Z<SDT,g>}$	Geschätzter Totalwert in Gemeinde g im Stadtteil SDT
θ	Erwarteter Anteil der registrierten Personen in der Stichprobe
u_d	Area Effekt in Domain d
U	Gesamtheit, Population
$\widehat{u}_{d\tilde{w}}$	Geschätzte Varianz in Domain d unter Berücksichtigung der Gewichte $\tilde{w}$
$V(\widehat{\tau}_Y)$	Varianz von $\widehat{\tau}_Y$
$V(\widehat{\tau}_{Y,d})$	Varianz von $\widehat{\tau}_{Y,d}$
$\widehat{V}(\widehat{\tau}_Y)$	Varianzschätzer für $V(\widehat{\tau}_Y)$
$\widehat{V}(\widehat{\tau}_{Y,d}^{HT})$	Schätzer für $V(\widehat{\tau}_{Y,d}^{HT})$

$V(\widehat{\tau}_{Y,prop})$	Varianz von $\widehat{\tau}_Y$ bei proportionaler Allokation
$V(\widehat{\tau}_{Y,opt})$	Varianz von $\widehat{\tau}_Y$ bei optimaler Allokation
$\widehat{V}(\widehat{\tau}_{Y,d}^{GREG})$	Varianzschätzer des GREG-Schätzers für $V(\widehat{\tau}_{Y,d})$
w_i	Designgewicht für Einheit i
$w_{h,i,d}$	Designgewicht für Einheit i in Schicht h und Domain d
$\tilde{w}_*$	Reskaliertes Gewicht für $*$
x	Stichproben-Matrix der unabhängigen Variablen in einer Regression
x_i	Stichproben-Vektor der unabhängigen Variablen der Einheit i in einer Regression
$\overline{X}_h$	Vektor der Mittelwerte der x-Werte in der Gesamtheit in Schicht h
$\overline{x}_h$	Vektor der Mittelwerte der x-Werte in der Stichprobe in Schicht h
Ξ	Varianz-Terme in der Gesamtheit bei geschichteter Auswahl
y	Stichproben-Vektor der abhängigen Variable in einer Regression
$\overline{y}_h$	Stichprobenmittel der y-Werte in Schicht h
$\overline{y}_{h,d}$	Stichprobenmittel der Stichprobenwerte in Schicht h der Domain d
ζ	Vorgegebene Schranke für den RRMSE

1 Einleitung

Im Rahmen des Zensus-Stichprobenforschungsprojektes wurden Stichprobendesigns und Schätzmethoden für die Haushaltsstichprobe des Register-gestützten Zensus 2011 erforscht. Dabei spielten zwei Ziele eine entscheidende Rolle: Einerseits sollte die amtliche Einwohnerzahl mit hinreichender Genauigkeit ermittelt werden (*Ziel 1*). Andererseits sollten Häufigkeitsverteilungen von Variablen geschätzt werden, die nicht in Registern enthalten sind (*Ziel 2*).

Die Gemeinden der Bundesrepublik Deutschland sind strukturell sehr unterschiedlich. Dies fällt insbesondere auf, wenn Gemeinden verschiedener Bundesländer miteinander verglichen werden. Deshalb galt es zunächst, Stichprobenbasiseinheiten geeignet zu definieren, auf denen Präzisionsvorgaben eingehalten werden können. Diese Stichprobenbasiseinheiten erlauben die Implementierung einer optimalen, geschichteten Zufallsstichprobe, die wiederum eine ganze Reihe von Bedingungen erfüllt.

Um die Effizienz des Stichprobendesigns und die Genauigkeit der Schätzverfahren zu analysieren, wurde eine realitätsnahe Simulationsgesamtheit erstellt, die schließlich über 85 Millionen Einträge umfasste. Diese hohe Zahl an Einträgen ist der Tatsache geschuldet, dass in der Simulationsgesamtheit Personen sowohl an ihrem Hauptwohnsitz als auch an ihrem Nebenwohnsitz erfasst sind. Das wesentliche Problem bei der Erzeugung dieser Simulationsgesamtheit bestand darin, dass die Verteilungen von zu implementierenden Variablen nicht für ganz Deutschland flächendeckend bekannt waren.

Die Forschungsarbeiten waren in drei Phasen aufgeteilt. In einer ersten Phase wurde zunächst ein Repertoire an Schätzmethoden auf ihre Einsetzbarkeit im Zensus hin untersucht. Unter anderem auf diesen Ergebnissen aufbauend wurden in der zweiten Phase ein optimales Stichprobendesign entwickelt sowie neuere Schätzmethoden auf ihre mögliche Eignung zum Einsatz im Zensus 2011 überprüft. Schließlich wurden in der dritten und letzten Phase konkrete Empfehlungen aus den Simulationen mit einem realitätsnahen Datenbestand abgeleitet.

Die vorliegende Monografie fasst die wesentlichen Ergebnisse der Forschungsarbeiten zusammen und gibt Empfehlungen, welche Methoden zum gegenwärtigen Zeitpunkt sinnvoll einsetzbar sind und welche möglicherweise in der Zukunft weitere Verbesserungen ermöglichen.

Das zweite Kapitel befasst sich mit der statistischen Methodik und ist in drei Teile unterteilt. In einem ersten Teil werden der Ablauf und die zwei zentralen Ziele des Stichprobenforschungsprojekts zum Zensus 2011 vorgestellt. Anschließend werden die notwendigen Grundlagen für die Ermittlung eines geeigneten Stichprobendesigns erläutert. Aus diesen Vorbetrachtungen wird schließlich das für den Zensus 2011 vorgeschlagene und inzwischen implementierte, geschichtete Stichprobendesign unter Box-Constraints abgeleitet (siehe auch Statistische Ämter des Bundes und der Länder 2011). Dieses Design wurde bereits in Münnich et al. (2010) eingeführt. Eine eingehende inhaltliche Erläuterung kann dem Bericht *Stichprobenverfahren und Allokation des Stichprobenumfangs für den Zensus 2011* entnommen werden (vgl. Statistisches Bundesamt 2010).

Ein zweiter Teil des zweiten Kapitels befasst sich anschließend mit den für den Zensus 2011 einsetzbaren Schätzverfahren. Diese teilen sich in sogenannte klassische Stichprobenverfahren, welche auf Basis von Stichprobendesign-Informationen Schätzungen ermöglichen, sowie Modell-basierte Schätzverfahren auf. Letztere umfassen insbesondere synthetische Methoden und Small AreaVerfahren. Diese Methoden ermöglichen aufgrund der Verwendung von Modellen in den Schätzungen einen Präzisionsgewinn, der aber mit Verzerrungen erkauft wird, welche aus dem Stichprobenziehungsprozess resultieren. Da diese Methoden bisher kaum in der amtlichen Statistik verwendet

werden, wird gelegentlich auch von einem *zweiten Paradigmenwechsel* in der deutschen amtlichen Statistik beim Zensus 2011 gesprochen. Im dritten Teil des zweiten Kapitels werden schließlich spezielle Schätzmethoden vorgestellt, welche sich insbesondere zur Untergliederung von Variablen oder kombinierten Verteilungen beziehungsweise auch tief untergliederten Tabellen eignen.

Das dritte Kapitel bildet den eigentlichen Schwerpunkt dieses Buches. Es befasst sich mit den Ergebnissen der umfangreichen Simulationsstudie. Zunächst werden die Methoden der Datengenerierung beschrieben. Sie werden benötigt, um eine realitätsnahe Simulationsgesamtheit zu erzeugen, auf der die verschiedenen Stichprobendesigns und Schätzmethoden mittels Simulationsmethoden analysiert werden können. Mit Hilfe dieser Methoden werden zahlreiche benötigte Variablen unter Verwendung der Register-Informationen erzeugt und dargestellt. Ein besonderes Augenmerk wurde auf die Erzeugung von Karteileichen und Fehlbeständen gelegt. Diese haben einen maßgeblichen Einfluss auf die Qualität der *Ziel 1*-Schätzungen.

Anschließend folgen die Auswertungen der Simulationsstudie. Zunächst werden die Ergebnisse bezüglich *Ziel 1* vorgestellt. Hierbei werden die Ermittlung der amtlichen Einwohnerzahl sowie die Schätzung von Karteileichen und Fehlbeständen behandelt. Es folgen die Auswertungen zu *Ziel 2*. Hierauf folgen die Auswertungen zur Schätzung von hochdimensionalen Kreuztabellen, die auch Hypercubes genannt werden.

Einen weiteren Schwerpunkt des dritten Kapitels bilden Auswertungen zu Sonderproblemen. Diese umfassen diverse Fragestellungen, die zielübergreifend betrachtet werden beziehungsweise von besonderer Wichtigkeit sind, aber nicht unmittelbar auf die beiden zentralen Aufgaben ausgerichtet sind.

Eine kurze Zusammenfassung der Empfehlungen nebst Ausblick auf zukünftige Forschungsfelder eines Register-gestützten Zensus folgt im vierten Kapitel.

2 Methodische Grundlagen

2.1 Einführende Bemerkungen

Ein Zensus ist die Basis der Bevölkerungsstatistik und skizziert idealerweise ein Gesamtbild der Gesellschaft (vgl. Grohmann 2009). Im Jahr 2011 fand in Deutschland nach 1981 (ehemalige DDR) bzw. 1987 (früheres Bundesgebiet) wieder ein Zensus statt. Der Zensus 2011 wurde im Rahmen einer EU-weiten Volkszählungsrunde als Register-gestützter Zensus durchgeführt. Neben der Auswertung der Einwohnermelderegister wurde eine Stichprobe erhoben, die Informationen über mögliche Registerfehler sowie im Register nicht vorhandene Variablen liefern soll. Diese Art des Zensus ersetzte die traditionelle Volkszählung mit einer Befragung aller Haushalte durch Interviewer. Der Register-gestützte Zensus ist durch geringeren Erhebungsaufwand und eine erhebliche Reduktion der Befragungsbelastung der Bevölkerung gekennzeichnet.

Mit Hilfe des Zensus werden politikrelevante Größen ermittelt, die beispielsweise für den Länderfinanzausgleich oder den kommunalen Finanzausgleich von enormer Bedeutung sind (vgl. Wagner 2010). Speziell einwohnerzahlabhängige Bemessungsgrenzen haben für Gemeinden eine große Relevanz. Zudem gibt der Zensus u.a. Auskunft über Erwerbsleben, Unterhaltsquellen, Ausbildung und Pendlerverhalten der Personen und Haushalte in regionalen und inhaltlichen Untergliederungen.

Der Zensus 2011 ist Teil eines EU-weiten Verfahrens und soll die Vergleichbarkeit mit Volkszählungen anderer Länder, trotz der Anwendung unterschiedlicher Methoden, ermöglichen. Zu diesem Zweck wurde die EG-Verordnung über Volks- und Wohnungszählungen verabschiedet, die länderübergreifende Standards vorgibt.[1] Zur Vorbereitung des Zensus 2011 und zur Umsetzung der EG-Verordnung wurden in Deutschland das Zensusvorbereitungsgesetz 2011[2] und das Zensusgesetz 2011[3] verabschiedet.

Das Zensus Stichprobenforschungsprojekt ist in eine ganze Reihe von Untersuchungen und Prozessschritten eingebettet, die hauptsächlich von den Statistischen Ämter des Bundes und der Länder durchgeführt wurden oder noch durchgeführt werden. Unter anderem wurde mit dem Zensustest 2001 eine vorbereitende Erhebung realisiert, um das nun implementierte Verfahren erstmals zu testen (vgl. Statistische Ämter des Bundes und der Länder 2004 sowie Bierau 2001). Weiterhin wurde ein Anschriften- und Gebäuderegister aufgebaut (vgl. Kleber et al. 2009), welches als einer der zentralen Pfeiler des Zensus 2011 gilt.

Neben den strukturellen Anforderungen, wie der IT-Sicherheit und der Verarbeitung von erhobenen Daten, stellten sich auch zahlreiche inhaltliche Fragen, so zum Beispiel nach der Ausgestaltung des Fragebogens. Daneben mussten zahlreiche Entscheidungen getroffen werden, die Sonderproblematiken betrafen. Als Beispiel sei hier die Behandlung von Sonderanschriften (wie beispielsweise Studentenwohnheime oder Kasernen) angeführt (vgl. Berg und Bihler 2011, S. 319). Einer der letzten Schritte des gesamten Verfahrens wird die Haushaltegenerierung sein. Hierbei wird jede Person im Melderegister gemäß verschiedener deterministischer und statistischer Verfahren einem Haushalt zugeordnet (vgl. Fürnrohr und König 1999 sowie Fürnrohr et al. 2002).

---

1 Verordnung (EG) Nr. 763/2008 des Europäischen Parlaments und des Europäischen Rates vom 9. Juli 2008 über Volks- und Wohnungszählungen (Amtsblatt der EU Nr. L 218, S. 14).

2 Gesetz zur Vorbereitung eines Register-gestützten Zensus einschließlich einer Gebäude- und Wohnungszählung 2011 (Zensusvorbereitungsgesetz 2011 - ZensVorbG 2011) vom 8. Dezember 2007 (BGBl. I S. 2808).

3 Gesetz zur Anordnung des Zensus 2011 sowie zur Änderung von Statistikgesetzen vom 8. Juli 2009 (BGBl. S. 1781).

Beim Zensus 2011 wurden zunächst Informationen aus den Melderegistern zur Ermittlung der Bevölkerungszahl herangezogen. Zusätzlich wurde eine ergänzende Stichprobe gezogen, die hauptsächlich zwei Zielen dient. Zum einen müssen mögliche Registerfehler (Karteileichen und Fehlbestände) abgeschätzt werden, damit die aus den Melderegistern ermittelten Bevölkerungszahlen statistisch korrigiert werden können. Zum anderen werden mit der Stichprobe weitere interessierende personenbezogene Merkmale erhoben, wie etwa solche zur Ausbildung oder zum Erwerbsleben, die in den Registern nicht vorhanden sind.

Für die Durchführung des ersten Register-gestützten Zensus in Deutschland bedurfte es also gezielter Untersuchungen sowohl zur Stichprobenziehung als auch zur Schätzmethodik. Hierbei interessierte vor allem das Zusammenspiel dieser beiden Komponenten hinsichtlich ihrer Effizienz unter Berücksichtigung des Erhebungsaufwandes der Stichprobenerhebung für den Zensus 2011. Diese methodischen Fragen wurden im Rahmen des Zensus Stichprobenforschungsprojektes eingehend untersucht (siehe z.B. Münnich et al. 2011*a* oder Münnich et al. 2011*b*).

Dabei wurden in diesem Projekt insbesondere unterschiedliche Stichprobenziehungsmethoden und aktuell in der statistischen Forschung diskutierte Schätzmethodiken, mit Schwerpunkt auf klassischen Methoden und den Small Area-Verfahren, untersucht. Als Datengrundlage diente ein anonymisierter Melderegisterabzug, der vor 2010 zusammengestellt wurde. Damit ist er nicht deckungsgleich mit dem im Zensus 2011 verwendeten Abzug. Es wird angenommen, dass keine starken Strukturbrüche im Zwischenzeitraum stattgefunden haben. Unter dieser Annahme sind die Ergebnisse des Zensus Stichprobenforschungsprojektes weitestgehend auf die neuen Melderegisterabzüge übertragbar.

Für die Stichprobenziehung wurde im Projekt ein neues Verfahren entwickelt, das die verfügbaren Informationen optimal nutzt und die Erfüllung sowohl der statistischen als auch der gesetzlich festgelegten Anforderungen an den Zensus ermöglicht. Weiterhin wurden aktuelle Schätzer auf ihre Chancen (Verbesserung der Ergebnisse) und Risiken (mögliche Qualitätsverluste bei ungeeignetem Einsatz) hin untersucht. Um die Untersuchungen möglichst reliabel zu gestalten, wurde eine groß angelegte Simulationsstudie programmiert. Hierbei wurden verschiedene plausible Registerfehlermodelle implementiert, anhand derer gezeigt wurde, wie sich die verschiedenen Kombinationen von Schätzern, Stichprobendesigns und Fragestellungen bei unterschiedlichen Registerfehlerstrukturen verhalten. Weiterhin wurden in anderen Ländern existierende Ansätze auf ihre Anwendbarkeit im deutschen Zensus hin überprüft.

Weitere Informationen zu den verschiedenen Bestandteilen des Zensus sind auf der Homepage der Statistischen Ämter des Bundes und der Länder unter http://www.zensus2011.de zu finden. Aktuelle Links und Informationen zum Zensus Stichprobenforschungsprojekt können unter http://www.uni-trier.de/index.php?id=40374 abgerufen werden.

2.1.1 Ziele im Zensus 2011

Im Rahmen des Zensus 2011 werden in Deutschland zwei Ziele verfolgt. Die sogenannten Fragen von *Ziel 1* befassen sich mit dem Themenkomplex der amtlichen Einwohnerzahl. Hierzu gehört insbesondere die amtliche Einwohnerzahl selbst. Darüber hinaus haben die Untersuchungen zum Zensustest 2001 gezeigt, dass die Einwohnermelderegister von unterschiedlicher Qualität sind und Registerfehler aufweisen. Es wird hier zwischen Übererfassungen, sogenannten Karteileichen, und Untererfassungen, sogenannten Fehlbeständen, unterschieden. Im Rahmen der hier durchgeführten Untersuchungen soll davon ausgegangen werden, dass zunächst die Einwohnermelderegister ausgewertet werden. Anschließend werden mit Hilfe statistischer Verfahren, die im Folgenden genauer zu betrachten sind, die Zahl der Karteileichen und Fehlbestände geschätzt. Diese Schätzun-

gen werden zunächst zur Korrektur der Registerauswertung verwendet, woraus die amtliche Einwohnerzahl resultiert.

Das *Ziel 2* umfasst alle weiteren Fragestellungen und damit alle Variablen bzw. deren Proxies, die nicht durch Register abgedeckt sind. Dabei wird insbesondere auf folgende interessierende Punkte eingegangen:

- Variablen, die Auskunft über das Ausbildungsprofil geben, wie beispielsweise höchster allgemeiner Schulabschluss;
- Variablen, die Auskunft über das Erwerbsprofil geben;
- Stellung im Beruf;
- Überwiegender Lebensunterhalt;
- Daten der Bundesagentur für Arbeit;
- Zuzugsjahr von Ausländern (ZZJ).

Die im Rahmen des Zensus Stichprobenforschungsprojektes untersuchten Variablen stellen einen Komplex bedeutsamer Fragestellungen dar, die aufgrund ihrer Wichtigkeit für den Zensus beziehungsweise wegen der Bedeutung für die Schätzmethodik ausgewählt wurden. Im weiteren Verlauf der Ausführungen wird auf einzelne Aspekte der Daten und ihrer Besonderheiten genauer eingegangen.

Zur Ermittlung der amtlichen Einwohnerzahl τ_Z (Z für Zensus) werden neben der Anzahl, der in den Einwohnermeldeämtern registrierten Personen τ_R die Anzahl der Karteileichen τ_K (K als Index bzw. KAL im Text) und Anzahl der Fehlbestände τ_F (F bzw. FEB im Text) benötigt.

Somit ergeben sich als zentrale Kennwerte in *Ziel 1*

$$\tau_Z = \tau_R - \tau_K + \tau_F \quad . \tag{2.1.1}$$

Von diesen Gesamtwerten sind jedoch nur die Werte des Registers bekannt. Wie zuvor dargestellt, werden KAL und FEB geeignet geschätzt und zur Ermittlung der amtlichen Einwohnerzahl herangezogen. Man erhält:

$$\widehat{\tau}_Z = \tau_R - \widehat{\tau}_K + \widehat{\tau}_F \quad . \tag{2.1.2}$$

Dabei wird stets davon ausgegangen, dass die verschiedenen interessierenden Anzahlen auf unterschiedlichen Aggregationsebenen benötigt werden. Diese werden nach ihrer Einführung präzisiert.

Ziel 2 gestaltet sich prinzipiell einfacher, da keine Korrekturen von Registern herangezogen werden müssen. Es bezeichnet allgemein Y eine interessierende *Ziel 2*-Variable. Dann muss in Analogie zu *Ziel 1* die Summe der y-Werte geschätzt werden. Es interessiert also in der zu betrachtenden Aggregationsebene τ_Y, das durch $\widehat{\tau}_Y$ zu schätzen ist.

2.1.2 Präzisionsanforderungen

Wie im letzten Abschnitt dargestellt, werden verschiedene Zielgrößen geschätzt. Hierzu werden neben den Auswertungen der Einwohnermelderegister, die für Deutschland flächendeckend zur Verfügung stehen, Daten einer Stichprobenerhebung verwendet. Eine eingehende Darstellung der

Stichprobenerhebung folgt weiter unten. Wie in Stichprobenerhebungen üblich, dienen zur Quantifizierung der Qualität der Erhebung Präzisionsvorgaben, die sinnvollerweise vor der Untersuchung festgelegt werden.

Im Rahmen des Zensus 2011 wurden Vorgaben für die beiden Ziele unterschiedlich formuliert. Für das *Ziel 1*, die amtliche Einwohnerzahl, werden Gemeinden ab 10.000 Einwohnern sowie Stadtteile großer Städte berücksichtigt. Da die Einwohnerzahl in Gemeinden mit weniger als 10.000 Einwohnern nicht durch Schätzungen ermittelt wird, müssen in diesem Fall auch keine Präzisionsanforderungen für den Stichprobenfehler angegeben werden.

Es bezeichnet $< g >$ die g-te Gemeinde beziehungsweise SDT einen speziellen Stadtteil. Betrachtet man als Maß der Variabilität eines nicht notwendigerweise unverzerrten Schätzverfahrens den MSE (Mean Squared Error), dann wird die Präzisionsvorgabe für *Ziel 1*-Fragestellungen durch den RRMSE (relative Root Mean Square Error)[4]

$$\mathrm{RRMSE}(\widehat{\tau}_{Z<g>}) = \frac{\mathrm{RMSE}(\widehat{\tau}_{Z<g>})}{\tau_{Z<g>}} \leq 0,005 \tag{2.1.3}$$

beziehungsweise

$$\mathrm{RRMSE}(\widehat{\tau}_{Z<SDT,g>}) = \frac{\mathrm{RMSE}(\widehat{\tau}_{Z<SDT,g>})}{\tau_{Z<SDT,g>}} \leq 0,005 \tag{2.1.4}$$

angegeben. Der RRMSE eines Schätzers ist als Einheiten-unabhängiges Maß besonders geeignet, die relative Genauigkeit der Schätzung bezogen auf die Einwohnerzahl anzugeben.

Der RRMSE darf bei *Ziel 1* in den zu betrachtenden Gemeinden und Stadtteilen nicht über 0,5 % liegen.

Die Wahl der Präzisionsvorgaben gestaltet sich bezüglich *Ziel 2* wesentlich komplexer. Während in *Ziel 1* die Bezugsgröße im Nenner stets die Einwohnerzahl der zu betrachtenden Subpopulation ist, reduziert sich der Nenner in *Ziel 2* auf den Teil der Subpopulation, der die interessierende Eigenschaft aufweist. Bei seltenen Ereignissen kann die Verwendung relativer Maße zu Anforderungen führen, die kaum noch zu erfüllen sind. Andererseits wären absolute Maße aufgrund der insgesamt geringen Variabilität von seltenen Beobachtungen fast immer erfüllt. Dieser inhaltliche Zielkonflikt erfordert eine differenzierte Betrachtung der Präzisionsanforderungen.

Für eine *Ziel 2*-Variable sei der Anteil interessierender Beobachtungen an der Zensusbevölkerung gleich p, es gilt also $\tau_Y = p \cdot \tau_Z$. Im Falle $p \leq 1/15$ sind keine Präzisionsvorgaben vorgesehen. Im Falle $p \geq 1/15$ muss

$$\mathrm{RRMSE}(\widehat{\tau}_{Y<\mathrm{area}>}) \leq \frac{1}{p \cdot 100}$$

in der jeweiligen *Area*[5] erfüllt sein. Je höher der Anteil der interessierenden Merkmalsträger an der Gesamtbevölkerung ist, desto präziser müssen die Anforderungen eingehalten werden. Eine Übersicht konkreter Werte ist der Tabelle 2.1 zu entnehmen.

---

4 Der relative Root Mean Square Error wird in Abschnitt 6 definiert.

5 Area bezeichnet die geographische Einheit auf der die Schätzungen berechnet werden, dies kann zum Beispiel eine Gemeinde oder ein Kreis sein.

Tabelle 2.1: Anteil der Merkmalsausprägung und relativer Standardfehler

Anteil der interessierenden Merkmalsträger an der Population (in %)	Der relative Standardfehler für den geschätzten Totalwert der Merkmalsausprägung soll kleiner gleich. . . % sein
6,7	15
10	10
20	5
30	3,33
50	2
80	1,25

Für die *Ziel 2*-Fragestellungen wurden zusätzliche hierarchisch angeordnete Aggregationsebenen mit Präzisionsanforderungen belegt. So wurden speziell in Rheinland-Pfalz zusätzlich Präzisionsanforderungen für die Verbandsgemeinden eingeführt. Weiterhin ist die Erfüllung von Präzisionsanforderungen auch auf Kreisebene notwendig. Die Präzisionsanforderungen sind in § 7 Absatz 1 des Zensusgesetzes 2011 zu finden.[6]

Da im Allgemeinen der Anteil der Merkmalsträger an der Population p, der zur Konkretisierung der Präzisionsanforderung notwendig ist, nicht bekannt ist, muss dieser aus der Stichprobe geschätzt werden. Sofern das Merkmal kein seltenes Ereignis ist, was hier der Einfachheit halber mit $p < 1/15$ verbunden wird, führt dies nicht zu Problemen.

2.1.3 Datengrundlage und Stichprobeneinheiten

Im vorangegangenen Abschnitt wurden die Präzisionsanforderungen an die Schätzungen im Zensus 2011 dargestellt, die im Falle von *Ziel 2* einer hierarchischen Ordnung folgen. Im Rahmen der Beurteilung der Schätzverfahren müssen die Stichprobendesigns diese hierarchische Struktur berücksichtigen, damit die Qualitätsmessung a posteriori nicht von einer möglichen Abhängigkeit von zufällig realisierten Stichprobenumfängen in Teilpopulationen beeinflusst wird. Ebenso muss sichergestellt werden, dass kein Rahmenfehler (coverage error) entsteht.

Nachfolgend wird eine hierarchische Struktur von Zusammenfassungen regionaler Einheiten definiert, welche die einzelnen Ebenen der Präzisionsanforderungen berücksichtigt. Diese dient als Basis, um eine regionale Aufteilung des Gesamtstichprobenumfangs zu ermöglichen.

Definition 2.1.1. Stichprobenbasiseinheiten

Eine Stichprobenbasiseinheit ist als regionale Einheit definiert, aus der eine Teilstichprobe gezogen wird. Als Kurzbezeichnung für eine Stichprobenbasiseinheit wird nachfolgend der Begriff SMP (Sampling Point) verwendet. Diese Stichprobenbasiseinheiten werden in vier Typen nach folgendem Schema eingeteilt:

Typ 0 (SDT): Stadtteile ab 200.000 Einwohner (EW) aus Gemeinden mit mindestens 400.000 EW

Typ 1 (GEM): Gemeinden mit mindestens 10.000 EW, sofern sie nicht zum Typ 0 gehören

Typ 2 (VBG): Kleine Gemeinden (unter 10.000 EW) innerhalb eines Gemeindeverbands beziehungsweise einer Verbandsgemeinde werden zusammengefasst, sofern sie in der Summe mindestens 10.000 EW betragen

[6] Zensusvorbereitungsgesetz 2011 vom 8. Dezember 2007 (BGBl. I S. 2808), das durch Artikel 3 des Gesetzes vom 8. Juli 2009 (BGBl. I S. 1781) geändert worden ist.

Typ 3 (KRS): Zusammenfassung aller Gemeinden eines Kreises, die bis dahin noch keinem Typ zugeordnet wurden

Die Einteilung der SMPs in vier Typen erfolgt nach einem hierarchischen Schlüssel, bei dem ein Typ höherer Ordnung sich nur noch auf den Rest bezieht, der von dem Typ niederer Ordnung verblieben ist. Hierbei ist anzumerken, dass mit der Festlegung von Sampling Points keinerlei Festlegung der Teilstichprobenumfänge eines noch festzulegenden Stichprobendesigns erfolgt ist.

Insbesondere die Typen 2 und 3 garantieren eine flächendeckende Verteilung des Gesamtstichprobenumfangs, ohne den sich viele Fragen, etwa von Stadt- und Landverteilungen, kaum hätten vergleichend darstellen lassen. Ebenso garantiert dieser Schlüssel eine eindeutige Zerlegung Deutschlands in SMPs. Ferner werden neben den zu betrachtenden Einheiten, die in den Präzisionsanforderungen aufgeführt sind, sinnvolle Einheiten definiert, die im Gegensatz zu den einzelnen Gemeinden bezüglich ihrer Größe nicht mehr stark variieren.

Mit der Wahl des Namens Stichprobenbasiseinheiten soll vermieden werden, dass diese mit den Small Areas beziehungsweise den Schichten, die innerhalb dieser Einheiten gebildet werden, direkt in Verbindung gebracht werden. Eine Darstellung dieser SMPs erfolgt in Abbildung 2.1.[7]

Die Einfärbung wurde mit Hilfe der Klassifikation der SMPs gemacht (SDT: gelb; GEM: rot; VBG: grün; KRS: blau) diese sind durch feine schwarze Linien abgegrenzt. Die Grenzen der Stadtteile sind synthetisch erzeugt worden.Man erkennt, dass zwischen den Bundesländern sehr unterschiedliche Verteilungen der Typen von Stichprobenbasiseinheiten vorliegen. Die Extremfälle ergeben sich zwischen Nordrhein-Westfalen, Rheinland-Pfalz sowie den Stadtstaaten. Eine detaillierte Übersicht über die Anzahl der Stichprobenbasiseinheiten nach Bundesländern ist in Tabelle 2.2 gegeben.

Tabelle 2.2: Verteilung der SMP-Typen in den Bundesländern

	SMP-Typ (Anzahl)				
Bundesland	SDT	GEM	VBG	KRS	Summe
Baden-Württemberg	2	244	126	35	407
Bayern	8	216	30	71	325
Berlin	12	0	0	0	12
Brandenburg	0	71	5	14	90
Bremen	3	1	0	0	4
Hamburg	7	0	0	0	7
Hessen	3	168	0	21	192
Mecklenburg-Vorpommern	0	24	30	12	66
Niedersachsen	2	205	68	34	309
Nordrhein-Westfalen	12	339	0	17	368
Rheinland-Pfalz	0	46	122	20	188
Saarland	0	40	0	5	45
Sachsen	4	69	13	22	108
Sachsen-Anhalt	0	60	27	11	98
Schleswig-Holstein	0	53	52	11	116
Thüringen	0	33	6	17	56
Deutschland	53	1.569	479	290	2.391

[7] Die Abbildung stammt aus einer Forschungsarbeit des Forschungszentrums für Regional- und Umweltstatistik zum Thema *Wirkung der Verwendung von Verbandsgemeinden beim Zensus 2011 in Rheinland-Pfalz.* Den verwendeten Karten liegen die Vektordaten in den Verwaltungsgrenzen 1 : 250.000 des Bundesamtes für Kartographie und Geodäsie zu Grunde.

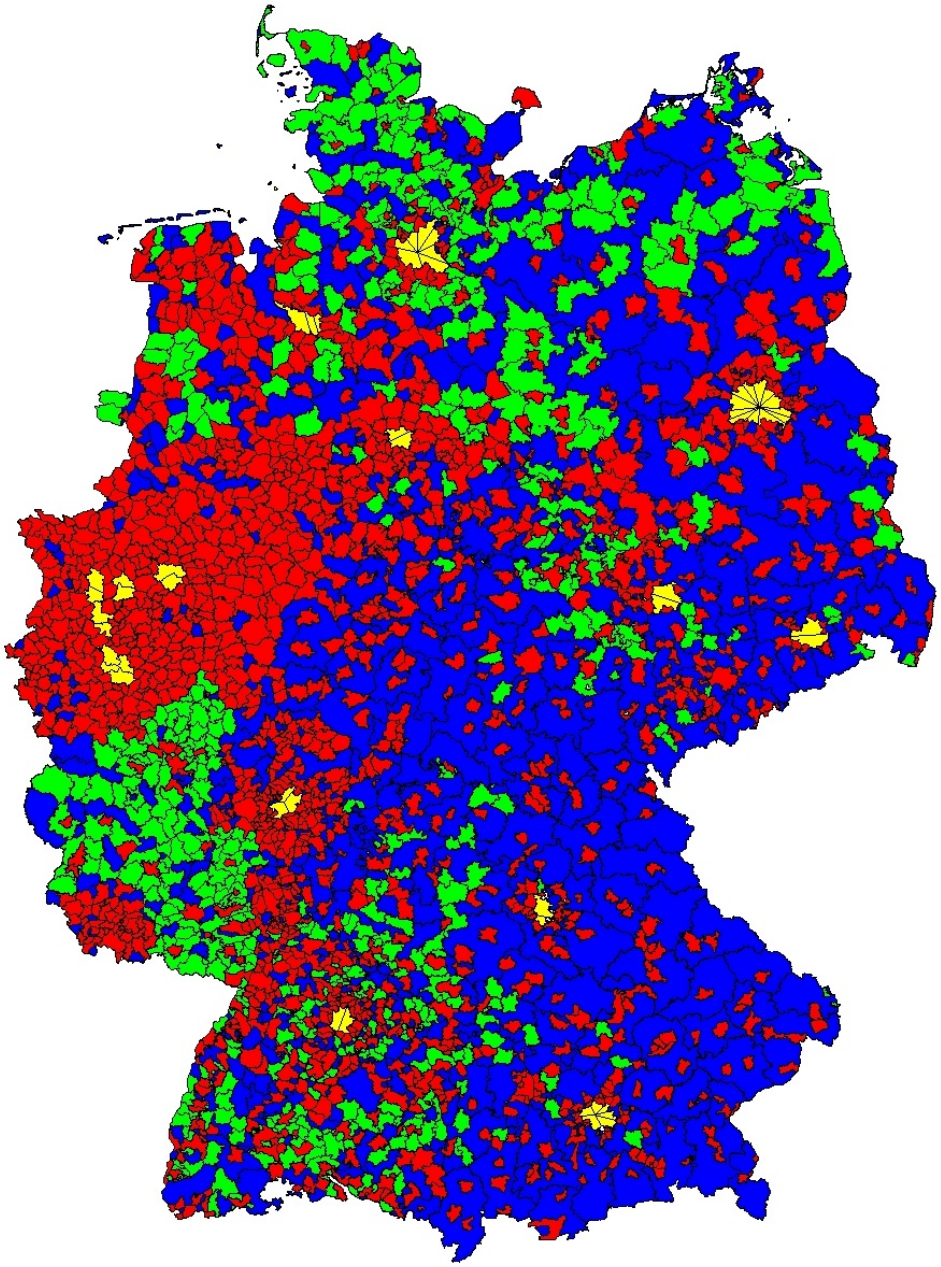


Abbildung 2.1: Darstellung der Stichprobenbasiseinheiten in Deutschland

Hierbei sei angemerkt, dass der Auswertung nicht die letzten aktuellen Registerauswertungen zugrunde liegen und damit geringfügige Abweichungen[8] auftreten können.

[8] Letztendlich hat sich die tatsächlich verwendete Anzahl an SMPs noch um 26 reduziert. Die tatsächlich verwendeten Zahlen sind in (Berg und Bihler 2011, S. 321) zu finden.

2.2 Das Stichprobendesign im Zensus 2011

2.2.1 Vorbemerkungen

Nach der Festlegung der Stichprobenbasiseinheiten muss ein *optimales* Stichprobendesign gefunden werden. Dabei entsteht jedoch unmittelbar die Frage nach der Definition des Begriffs der Optimalität bezogen auf die Stichprobenziehung des Zensus 2011. Bei der Lösung dieser Frage müssen zahlreiche Anforderungen berücksichtigt werden:

1. Erfüllung aller gestellten Präzisionsanforderungen,
2. Verwendung eines möglichst geringen Auswahlsatzes,
3. Begrenzung der Gesamtkosten sowie
4. keine zu große Variation der Auswahlwahrscheinlichkeiten.

Grundsätzlich sind Genauigkeit der Schätzung und Kosten beziehungsweise Aufwand der Erhebung gegeneinander gerichtete Ziele und lassen sich nicht gemeinsam und allgemein optimieren. Es handelt sich hierbei also um eine multikriterielle Betrachtung, bei der die verschiedenen Ziele beachtet werden müssen. Aufgrund der dargestellten Präzisionsvorgaben und der aus der Stichprobenziehung resultierenden Informationslage müssen bei einer Optimierung des Stichprobendesigns Kriterien der MSE-Minimierung herangezogen werden.

Ebenso sind die zuvor dargestellten Vorteile einer regionalen *Neugliederung* in Stichprobenbasiseinheiten bedeutend. Sie ermöglichen einen relativ einfachen und kohärenten Lösungsansatz für derartige Optimierungsfragen im Rahmen der Untersuchungen. Es kann gezeigt werden, dass das Aggregat automatisch die Präzissionsvorgaben erfüllt, wenn diese für untergeordnete Einheiten erfüllt sind. Dies ist dann auch bei den hierarchisch angelegten SMPs der Fall, bei denen Gemeinden zu Verbandsgemeinden und weiter zu Kreisen aggregiert werden.

Hat man H unabhängige Schätzer $\hat{t}_{1,Y}, \ldots, \hat{t}_{H,Y}$, die alle $V\left(\frac{\hat{t}_{h,Y}}{\tau_{h,Z}}\right) \leq \alpha$ erfüllen, so gilt

$$V\left(\frac{\sum_{h=1}^{H} \hat{t}_{h,Y}}{\tau_Z}\right) = \sum_{h=1}^{H} \left(\frac{\tau_{h,Z}}{\tau_Z}\right)^2 V\left(\frac{\hat{t}_{h,Y}}{\tau_{h,Z}}\right) \leq \alpha \sum_{h=1}^{H} \left(\frac{\tau_{h,Z}}{\tau_Z}\right)^2 \leq \alpha \left(\sum_{h=1}^{H} \frac{\tau_{h,Z}}{\tau_Z}\right)^2 = \alpha \quad .$$

Damit kann die Erfüllung von Präzisionsvorgaben auf untergeordneten Einheiten stets auch auf deren Aggregate übertragen werden.

Gesetzlich festgelegt ist die Ziehung von Anschriften. Der Rahmen der Erhebung basiert auf dem Gebäude- und Wohnungsregister, deren Vollständigkeit für die Optimierung vorausgesetzt wird. Die Ziehung der Anschriften aus dem Register erfolgt schließlich separat für alle SMPs.

2.2.2 Voruntersuchungen zum Stichprobendesign

Etwas komplexer gestaltet sich die Frage, welches Stichprobendesign verwendet werden soll. Aus der Stichprobenliteratur ist bekannt, dass mehrstufige Designs zwar aus erhebungstechnischen Gründen oft einfacher anwendbar sind und möglicherweise einen geringeren Erhebungsaufwand verursachen, allerdings schnell ineffizient sind und damit Präzisionsvorgaben verletzen. Daher eignen sich nur wenige Designs für eine eingehendere Betrachtung:

Unequal Probability Designs Diese Designs sind im Rahmen Design-basierter Schätzverfahren oft sehr effizient. Eine Umsetzung dieser Ziehungsverfahren auf großen Datenmengen gestaltet sich indes auch mit modernen Computern problematisch, da Algorithmen zur Zie-

hung sehr großer Stichproben mit bekannten und positiven Inklusionswahrscheinlichkeiten erster und zweiter Ordnung immer noch kaum verfügbar und außerordentlich rechenintensiv sind. Die Effizienz der Verfahren kann sich im Einsatz bei Modell-basierten Schätzverfahren umkehren. Ebenso kann die Optimierung bezüglich einzelner Variablen einen erheblichen Effizienzverlust anderer Variablen verursachen.

Balanced Sampling Dieses Verfahren nutzt verfügbare Randinformationen zur weiteren Optimierung des Designs. Dadurch erreicht der Horvitz-Thompson-Schätzer (siehe Abschnitt 2.3.2.1) bereits diejenige Effizienz, die von Regressionsschätzverfahren erreicht wird. Die Generierung von Stichproben in großen Grundgesamtheiten ist indes problematisch und hat sich in den Testsimulationen als zu aufwändig. Zudem ist der Effizienzgewinn kaum messbar.

Geschichtete Stichprobenverfahren Geschichtete Ziehungen lassen sich sehr einfach implementieren. Die Einbindung von Optimierungskalkülen ist sehr einfach möglich. Eine besondere Bedeutung hat hier die optimale Allokation des Stichprobenumfangs auf die Schichten nach Neyman (1934) und Tschuprov (1923).

Eine detaillierte Beschreibung der Designs findet man in der einschlägigen Stichprobenliteratur, wie etwa Cochran (1977), Särndal et al. (1992) oder Tillé (2006).

Im Rahmen der Voruntersuchungen zum optimalen Stichprobendesign für den Zensus 2011 wurden zahlreiche Ansätze untersucht und miteinander verglichen. Bei der Beurteilung der Genauigkeit und damit der Erfüllung von Präzisionsvorgaben müssen die zu verwendenden Schätzverfahren festgelegt werden. Prinzipiell spielen ebenso die interessierenden Zielvariablen eine wesentliche Rolle. Aufgrund der gesetzlichen Vorgaben muss aber *Ziel 1* bevorzugt berücksichtigt werden.

Als prioritäre Schätzverfahren wurden diejenigen Schätzverfahren ausgewählt, die im Rahmen der Forschungsprojekte DACSEIS[9] und EURAREA[10] für Small Area-Schätzungen herangezogen wurden. Hierzu gehören die klassischen Verfahren, wie der Horvitz-Thompson-Schätzer (HT) und der verallgemeinerte Regressionsschätzer (GREG), sowie die Modell-basierten Schätzer EBLUPA (vergleichbar mit dem Battese-Harter-Fuller *basic unit-level*-Modell) und EBLUPB (vergleichbar mit dem Fay-Herriot *basic area-level*-Modell). Eine eingehende Darstellung folgt in den Abschnitten 2.3.2 und 2.3.4.

Die Voruntersuchungen zu Stichprobendesigns und Schätzverfahren lieferten folgende Ergebnisse:

- In nahezu allen Schätzungen erzielte der GREG-Schätzer bessere Ergebnisse als der HT-Schätzer und zwar unabhängig vom Stichprobendesign (vgl. auch Cochran 1977 oder Münnich 1997). Insbesondere bei *Ziel 1* wurde ein erheblicher Effizienzgewinn beobachtet.
- Die Unequal Probability Designs erwiesen sich für die interessierende Fragestellung kaum besser als eine geeignet optimierte geschichtete Zufallsstichprobe, waren aber bezüglich der Wahl der Zielvariablen wesentlich weniger robust.
- Modell-basierte Verfahren zeigten auffällige Effizienzverluste bei Stichproben-Designs, welche in besonderer Weise Optimierungsmethoden verwenden. Dies betraf vor allem den EBLUPB. Hierbei spielen stark variierende Auswahlwahrscheinlichkeiten eine zentrale Rolle.

[9] www.dacseis.de, WP10.

[10] www.ons.gov.uk/ons/guide-method/method-quality/general-methodology/spatial-analysis-and-modelling/eurarea/index.html.

- Balancierte Designs brachten nicht die von ihnen erwarteten Verbesserungen. Überdies war die Implementierung bei der Verwendung vieler Kalibrierungsvariablen problematisch.

Sehr stark optimierte Designs erwiesen sich zwar in der Regel in ihrem Einsatzbereich als vorteilhaft, zeigten jedoch eine reduzierte Robustheit in Bezug auf Gegebenheiten (z.B. andere Zielvariablen), die nicht in die Optimierung eingingen. Insbesondere lieferten sie oft sehr stark variierende Designgewichte. Verfolgt man die Diskussion von Gelman (2007), dann muss man bei stark variierenden Designgewichten mit erheblichen Problemen bei der Schätzung von statistischen Modellen ausgehen (siehe dazu auch Burgard und Münnich 2010 oder Münnich und Burgard 2012). Dies spielt in zweierlei Hinsicht für die Schätzungen im Zensus 2011 eine wichtige Rolle:

1. Wie später zu sehen sein wird, können manche Fragen mit klassischen Schätzmethoden nicht mehr hinreichend akkurat gelöst werden. In diesen Fällen müssen Small Area-Methoden herangezogen werden. Da diesen Methoden statistische Modelle zugrunde liegen, dürfen die Designgewichte nicht so stark variieren.
2. Eine Verwendung der Daten durch Wissenschaftler ist nach dem Zensus noch nicht geklärt, womit sich geeignete statistische Analysen im Falle sehr stark variierender Gewichte fast ausschließen.

Meng et al. (2009) schlagen einen Gelman-Faktor von höchstens 10 vor. Der Gelman-Faktor ist der Quotient zwischen dem größten und dem kleinsten Designgewicht. Die Designgewichte sind wiederum als Reziprokwerte der Inklusionswahrscheinlichkeiten definiert. Sie weisen darauf hin, dass Gelman-Faktoren über 100 inakzeptabel sind. Es sei angemerkt, dass etwa bei Business-Statistiken Gelman-Faktoren im hohen 1000er-Bereich keine Seltenheit sind. Sehr hohe Gelman-Faktoren sind auch im Zensus-Test beobachtbar.

Die Untersuchungen legten die Verwendung geschichteter Zufallsstichproben nahe. Da zur geeigneten Ermittlung von Ziel 1-Schätzungen eine robuste Strategie herangezogen werden sollte, wurde als Benchmark-Schätzverfahren ein separat-kombinierter Regressionsschätzer vorgeschlagen. Dieser Regressionsschätzer verwendet Regressionsmethoden auf jedem SMP unabhängig von der Schichtung. Eine eingehende Darstellung des Schätzers erfolgt in Abschnitt 2.3.2.2.

2.2.3 Proportionale und optimale Allokation

Ausgangspunkt für die Bestimmung der Teilstichprobenumfänge für die SMPs ist die Schätzung von Totalwerten. Hier ist die Schätzung der amtlichen Einwohnerzahl von besonderer Bedeutung. Als Referenzschätzfunktion, die im Allgemeinen eine hinreichende Genauigkeit ermöglicht, wird der separat-kombinierte Regressionsschätzer verwendet. Allgemein ist dieser für H Schichten durch

$$\widehat{\tau}_Y = \sum_{h=1}^{H} N_h \cdot \left(\overline{y}_h + \left(\overline{\mathbf{X}}_h - \overline{\mathbf{x}}_h \right)' \cdot \widehat{\beta} \right) \tag{2.2.1}$$

für jeden einzelnen SMP definiert. Dabei ist

N_h	die Anzahl der Anschriften in h-ter Schicht
$\overline{y}_h$	das Stichprobenmittel der y-Werte in h-ter Schicht
$\overline{\mathbf{X}}_h$	der Vektor der Mittelwerte der x-Werte in der Gesamtheit in h-ter Schicht
$\overline{\mathbf{x}}_h$	der Vektor der Mittelwerte der x-Werte in der Stichprobe in h-ter Schicht und
$\widehat{\beta}$	die Schätzung des Regressionsparametervektors.

Aufgrund der separaten Vorgehensweise in jedem SMP, kann auf die Indizierung der SMPs an dieser Stelle verzichtet werden. Im Falle einer Hilfsvariablen (hier: Anschriftengröße = Zahl der registrierten Einwohner) ist die Varianz von $\widehat{\tau}_y$

$$\mathrm{V}\left(\widehat{\tau}_y\right) = \sum_{h=1}^{H} N_h^2 \cdot \frac{S_{h,Y}^2}{n_h} \cdot \left(1 - \frac{n_h}{N_h}\right) \cdot \left(1 - \rho^2\right) \tag{2.2.2}$$

wobei

n_h die Anzahl der Stichproben-Anschriften in der h-ten Schicht
$S_{h,Y}^2$ die Varianz der y-Werte in der h-ten Schicht und
ρ der Korrelationskoeffizient zwischen Anschriftengröße und Zahl der tatsächlich vorhandenen Einwohner bezeichnet.

Bei dieser theoretischen Varianz wird ein homogener Zusammenhang zwischen Untersuchungs- und Hilfsmerkmal unterstellt, welcher mit Hilfe des Korrelationskoeffizienten ρ ausgedrückt wird. $1 - \rho^2$ liefert den Effizienzgewinn durch Verwendung des Regressionsschätzers. Ziel der nachfolgenden Umformungen ist die Herleitung geeigneter Darstellungen der Varianz (2.2.2) bei unterschiedlichen Allokationen des Gesamtstichprobenumfangs, mit deren Hilfe dann die Mindeststichprobenumfänge abgeleitet werden können. Für die proportionale und optimale Allokation gilt (vgl. Cochran 1977 oder Münnich 1997):

$$n_{h,\mathrm{prop}} = \frac{N_h}{N} \cdot n \tag{2.2.3}$$

sowie

$$n_{h,\mathrm{opt}} = \frac{N_h \cdot S_{h,Y}}{\sum_{\ell=1}^{H} N_\ell \cdot S_{\ell,Y}} \cdot n \quad . \tag{2.2.4}$$

Für die proportionale Allokation erhält man durch Einsetzen von (2.2.3)

$$\mathrm{V}\left(\widehat{\tau}_{Y,\mathrm{prop}}\right) = \left(1 - \rho^2\right) \cdot \left(\frac{N}{n} - 1\right) \cdot \sum_{h=1}^{H} N_h \cdot S_{h,Y}^2 \quad . \tag{2.2.5}$$

Für die optimale Aufteilung erhält man mit Hilfe von (2.2.4)

$$\mathrm{V}\left(\widehat{\tau}_{Y,\mathrm{opt}}\right) = \left(1 - \rho^2\right) \cdot \left(\frac{1}{n} \cdot \left(\sum_{h=1}^{H} N_h \cdot S_{h,Y}\right)^2 - \sum_{h=1}^{H} N_h \cdot S_{h,Y}^2\right) \quad . \tag{2.2.6}$$

Definiert man nun

$$\Xi_{prop} := \sum_{h=1}^{H} N_h \cdot S_{h,Y}^2 \tag{2.2.7}$$

sowie

$$\Xi_{opt} := \Big(\sum_{h=1}^{H} N_h \cdot S_{h,Y} \Big)^2 \quad , \tag{2.2.8}$$

dann können die Varianzen für die proportionale und die optimale Allokation wie folgt dargestellt werden:

$$\mathsf{V}\left(\widehat{\tau}_{Y,prop}\right) = \left(1 - \rho^2\right) \cdot \left(\frac{N}{n} - 1\right) \cdot \Xi_{prop} \tag{2.2.9}$$

und

$$\mathsf{V}\left(\widehat{\tau}_{Y,opt}\right) = \left(1 - \rho^2\right) \cdot \left(\frac{1}{n} \cdot \Xi_{opt} - \Xi_{prop}\right) \quad . \tag{2.2.10}$$

Kostenoptimale Aufteilungen des Stichprobenumfangs wurden ebenfalls untersucht. Da sie letztendlich aber für den Einsatz im Zensus kaum geeignet eingesetzt werden kann, werden sie hier nicht weiter angeführt.

Zu beachten ist, dass für die Varianzformeln die üblichen Approximationen verwendet wurden, die im Extremfall von den tatsächlichen Varianzen abweichen können. Ein solcher Extremfall liegt etwa dann vor, wenn eine große Zahl Schichten mit kleinen Schichtumfängen vorliegt oder Annahmen, wie die zuvor erwähnte Homogenität, verletzt werden.

Ferner darf nicht vergessen werden, dass sich die Optimalität auf theoretische Stichprobenumfänge bezieht, die die Ganzzahligkeitsbedingung nicht erfüllen. Die in der Realität durchzuführenden Stichprobenziehungen verwenden jedoch nur ganzzahlige Teilstichprobenumfänge. Insofern sind die tatsächlichen Stichprobenumfänge *Näherungen* für die theoretischen Lösungen. Die Effekte dieses Ganzzahligkeitsproblems spielen jedoch erst bei großer Schichtanzahl eine Rolle. Für mögliche Rundungsverfahren verweisen wir auf Abschnitt 3.6.9.

Bei der optimalen Allokation kommt es in der Praxis vor, dass die $S^2_{h,Y}$ in einigen Schichten 0 oder sehr nahe bei 0 sind. Dies würde beliebig große Designgewichte hervorbringen. In solchen Fällen wird daher ein Mindeststichprobenumfang pro Schicht verwendet. Umgekehrt kann es bei der optimalen Allokation vorkommen, dass für einige Schichten $n_h > N_h$ ist. In diesem Fall könnte Lemma 1 aus Stenger und Gabler (2005) angewendet werden, das eine optimale Aufteilung des Stichprobenumfangs n unter den linearen Nebenbedingungen $0 < n_h \leq N_h$ für alle Schichten h und $\sum n_h \leq n$ gewährleistet.

Das Ziel ist es, eine simultane Optimierung der Stichprobenumfänge auf allen SMPs im Sinne der Zielvorstellungen aus Kapitel 2.2 zu realisieren. Dabei ist zu beachten, dass die zu ermittelnden optimalen Stichprobenumfänge Restriktionen unterliegen. Diese Restriktionen können etwa durch einen vorgegebenen Gesamtstichprobenumfang entstehen.

Verwendet man den RRMSE für den asymptotisch unverzerrten kombinierten Regressionsschätzer, dann erhält man aus

$$\mathrm{RRMSE}(\widehat{\tau}) = \frac{\sqrt{\mathsf{V}\left(\widehat{\tau}\right)}}{\tau_Z \cdot p} \leq \zeta \tag{2.2.11}$$

unter Verwendung der Gleichungen (2.2.9) und (2.2.10) für die proportionale Allokation

$$n \geq \frac{N \cdot \Xi_{prop} \cdot \left(1-\rho^2\right)}{\left(\tau_Z \cdot p \cdot \zeta\right)^2 + \Xi_{prop} \cdot \left(1-\rho^2\right)} \tag{2.2.12}$$

und für die optimale Allokation

$$n \geq \frac{\Xi_{opt} \cdot \left(1-\rho^2\right)}{\left(\tau_Z \cdot p \cdot \zeta\right)^2 + \Xi_{prop} \cdot \left(1-\rho^2\right)} \quad . \tag{2.2.13}$$

Man erkennt, dass die Mindeststichprobenumfänge in dieser Darstellung Funktionen der Vorgaben ζ, ρ und p bei gegebenen Termen Ξ_{prop} und Ξ_{opt} sowie τ_Z sind. Die notwendigen Informationen können aus den Anschriftenstrukturen der gelieferten Daten ermittelt oder gegebenenfalls approximiert werden. Diese Approximation wird durch $S^2_{h,Y} = S^2_{h,A}$ erzielt, wobei $S^2_{h,A}$ die Varianz der Anschriftengrößen in der h-ten Schicht ist. Darüber hinaus soll darauf hingewiesen werden, dass für $\rho = 0$ der Horvitz-Thompson-Schätzer für die geschichtete Zufallsstichprobe in den Betrachtungen berücksichtigt wird. In der Praxis wird dieser Fall jedoch so gut wie nie auftreten. Dies verdeutlicht die Abbildung 2.2.

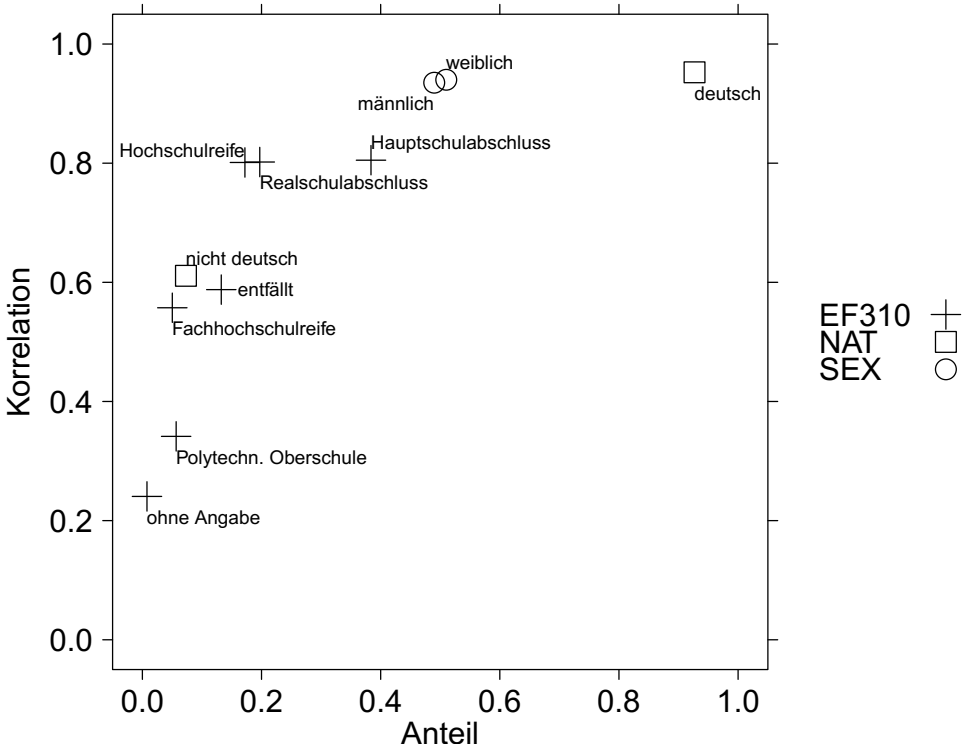


Abbildung 2.2: Korrelationen der Anzahl der Ausprägungen der Variablen EF310 (höchster Bildungsabschluss), NAT (Nationalität) und SEX (Geschlecht) auf Anschriften-Ebene mit der Anschriftengröße

Man erkennt, dass der Korrelationskoeffizient erst bei sehr kleinen Anteilen erheblich kleinere Werte aufweist. Hieraus kann abgeleitet werden, dass die Proportionalität des Vorkommens der interessierenden *Ziel 2*-Variablen zu den Anschriftengrößen weitgehend erhalten bleibt und damit die Optimierung bezüglich *Ziel 1* vielfach auch für *Ziel 2* zumindest akzeptabel ist.

Mit Hilfe dieser Darstellung wurden anschließend geeignete Teilstichprobenumfänge ermittelt. Dabei waren zudem folgende Punkte zu berücksichtigen:

- Es soll ein Gesamtstichprobenumfang von 8 % der Personen auf die SMPs aufgeteilt werden;
- Innerhalb der SMPs soll eine Schichtung nach Anschriftengröße durchgeführt werden (dies schließt eine weitergehende, eher aus inhaltlichen Gründen motivierte Schichtung nicht aus);
- Die zuvor gestellten Präzisionsanforderungen an *Ziel 1* und *2* müssen erfüllt sein;
- Um später möglicherweise interessierende Untersuchungen nicht unnötig zu erschweren – dies betrifft insbesondere die Verwendung von Small Area-Verfahren – sollte die Variabilität der Designgewichte nicht zu groß ausfallen.

Die Untersuchungen ergaben, dass bei Anwendung der optimalen Allokation in *Ziel 1* die Präzisionsvorgaben erfüllt wurden. Bei *Ziel 2* können einzelne sehr ungünstige heterogene Verteilungen von *Ziel 2*-Merkmalen auf die SMPs möglicherweise zu einer Nichterfüllung von *Ziels 2* führen, was angesichts der Fülle möglicher *Ziel 2*-Variablen aber nicht überraschend ist. Wie zuvor bereits motiviert, sollte bei wichtigen Variablen auch *Ziel 2* den zugehörigen Präzisionsanforderungen genügen. Wenig erfreulich ist die Tatsache, dass je nach Art der Schichtung der Gelman-Faktor über 3.000 liegen kann und damit unakzeptabel hoch ist.

Bei Verwendung der proportionalen Allokation entsteht das Problem, dass der Aufwand der Erhebung durch recht hohe Teilstichprobenumfänge vergleichsweise hoch liegen würde, wenn man den Präzisionsanforderungen genügen wollte. Somit schließt sich aus Effizienzgründen eine Verwendung der proportionalen Allokation aus. Damit entsteht die Notwendigkeit, einen Kompromiss einzugehen, der mit Hilfe einer optimalen Allokation unter Nebenbedingungen gelöst wird.

2.2.4 Optimale Allokation unter Box-Nebenbedingungen

Eine Allokation, die für alle Modelle geeignet ist, muss folgende Anforderungen erfüllen:

- Einhaltung der Präzisionsanforderungen, insbesondere bezüglich *Ziel 1*;
- Robustheit der Allokation gegenüber den interessierenden Fragestellungen und Ziele;
- Möglichkeit einer späteren Anwendung Modell-basierter Schätzverfahren.

Gerade der letzte Punkt referiert auf ein bisher kaum untersuchtes Ziel. Bekannt ist, dass die Modell-basierten Verfahren bei sehr unterschiedlichen Auswahlwahrscheinlichkeiten einen *Sample Selection Bias* aufweisen können. Daher ist es durchaus ratsam, auf sehr unterschiedliche Auswahlwahrscheinlichkeiten zu verzichten. Im letzten Abschnitt konnte man jedoch bereits erkennen, dass gerade bei der optimalen Allokation, und insbesondere bei schiefen Verteilungen, sehr unterschiedliche Designgewichte entstehen. Mögliche Auswirkungen derartiger Designgewichte wurden in Burgard und Münnich (2010) sowie Münnich und Burgard (2012) untersucht.

Schranken für Designgewichte müssen nach oben und unten gesetzt werden. Eine Beschränkung nach unten entspricht einer oberen Grenze des Auswahlsatzes. Obere Schranken müssen gesetzt werden, damit einzelne Gruppen nicht stark unterrepräsentiert sind. Eine Beschränkung nach oben

entspricht einem Mindestauswahlsatz. Sinnvoll erscheint es, dass der Zensus in jedem Falle besser sein soll als der Mikrozensus, für den der Auswahlsatz bei einem Prozent liegt. Eine Einschränkung der Variabilität der Gewichte bewirkt auch eine weniger ungleiche Behandlung der Bürger. Allerdings resultieren auch weniger effiziente Schätzungen im Design-basierten Kontext.

Vor der Durchführung des Zensus ist lediglich die Anzahl der relevanten Anschriften sowie die Anschriftengrößen bekannt. Wie bereits weiter oben angeführt, sind neben der tatsächlichen Anzahl an Personen, die an einer Anschrift wohnen, auch weitere Variablen mit der Anschriftengröße relativ hoch korreliert. Da mit dem Register-gestützten Zensus die erforderlichen Daten mit deutlich weniger Belastung für die Befragten als bei einer traditionellen Volkszählung gewonnen werden sollen, wird der erwartete Anteil θ, an Personen in der Stichprobe, begrenzt. Es stellt sich die Frage, wie sich verschiedene Variablen hinsichtlich der Genauigkeitsanforderungen verhalten, wenn die Auswahl an Anschriften bevölkerungsproportional oder bevölkerungsoptimal vorgenommen wird.

Formal lässt sich die Anforderung an die zu bestimmenden Stichprobenumfänge in den Schichten durch die folgenden Box-Constraints darstellen:

$$m_{<g>h} \leq n_{<g>h} \leq M_{<g>h} \tag{2.2.14}$$

mit vorgegebenen Untergrenzen $m_{<g>h}$ und Obergrenzen $M_{<g>h}$ in Gemeinde g und Schicht h.

Darüber hinaus darf ein vorab festgelegter Gesamtumfang an Personen $\tau_Z \cdot \theta$ in der Stichprobe nicht überschritten werden, wobei τ_Z wieder die Zahl der tatsächlich (am Hauptwohnsitz) lebenden Personen bezeichnet. Aufgrund der Tatsache, dass in einer Schicht Anschriften von verschiedener Größe zusammengefasst sind, kann die Personenzahl in einer Schicht von Stichprobe zu Stichprobe variieren. Daher kann auch nur eine erwartete Personenzahl bei gegebenem $n_{<g>h,A}$ angegeben werden. Die erwartete Personenzahl in Gemeinde g und Schicht h ist gegeben durch $n_{<g>h,A} \cdot \frac{\tau_{<g>h,R}}{N_{<g>h,A}}$. Soll insgesamt der Anteil θ von (registrierten) Personen in Deutschland ausgewählt werden, erhalten wir als zusätzliche Nebenbedingung

$$\sum_{g=1}^{G} \sum_{h=1}^{H} n_{<g>h,A} \cdot \frac{\tau_{<g>h,R}}{N_{<g>h,A}} = \tau_R \cdot \theta \qquad . \tag{2.2.15}$$

Bei den obigen Betrachtungen gilt es zu bedenken, dass statt der Zahl der in einer Gemeinde g und Schicht h registrierten Personen $\tau_{<g>h,R}$ eigentlich die Zahl der tatsächlich in einer Gemeinde g und Schicht h lebenden Personen, also $\tau_{<g>h,Z}$, für die Berechnungen berücksichtigt werden sollte. Diese Zahl ist jedoch in der Planungsphase nicht verfügbar und es wird daher $\tau_{<g>h,R}$ verwendet.

Wie bereits erwähnt, können diese Anforderungen von der naiven Neyman-Tschuprov-Allokation im Rahmen geschichteter Zufallsstichproben nicht mehr ohne weiteres erfüllt werden. Da alle SMPs in der Optimierung berücksichtigt werden sollen, wird die gewichtete 2-Norm des Vektors aller RRMSEs der SMPs

$$||\mathbf{RRMSE}(\widehat{\tau}_Z)||_2 = \sqrt[2]{\sum_{g=1}^{G} w_{<g>} \mathrm{RRMSE}(\widehat{\tau}_{<g>,Z})^2} \tag{2.2.16}$$

unter den Nebenbedingungen (2.2.14) und (2.2.15) minimiert mit $w_{<g>} \propto \left(\sum_{h=1}^{H} \tau_{<g>h,Z} \right)^2 = \tau^2_{<g>,Z}$. Der RRMSE in Gemeinde g ist definiert als

$$\text{RRMSE}(\widehat{\tau}_{<g>}) = \frac{\sqrt{V\left(\sum_{g=1}^{G} \sum_{h=1}^{H} \widehat{\tau}_{<g>h,Z} \right)}}{\tau_{<g>,Z}} \quad . \tag{2.2.17}$$

Wegen (2.2.2) ist die Minimierung von (2.2.16) als Funktion von $n_{<g>h}$ identisch mit der Minimierung von $\sum_{g=1}^{G} \sum_{h=1}^{H} N^2_{<g>h} \cdot \frac{S^2_{<g>h,Z}}{n_{<g>h}}$. Da in der Planungsphase $S^2_{<g>h,Z}$ nicht verfügbar ist, wird in der vorangegangenen Formel $S^2_{<g>h,R}$ verwendet.

Zusammenfassend ist daher folgendes Minimierungsproblem zu lösen: Minimiere als Funktion von $n_{<g>h}$

$$\sum_{g=1}^{G} \sum_{h=1}^{H} N^2_{<g>h} \cdot \frac{S^2_{<g>h,R}}{n_{<g>h}}$$

unter den Nebenbedingungen

$$0 < m_{<g>h} \leq n_{<g>h} \leq M_{<g>h}$$

$$\sum_{g=1}^{G} \sum_{h=1}^{H} n_{<g>h,A} \cdot \frac{\tau_{<g>h,R}}{N_{<g>h,A}} = \tau_R \cdot \theta \quad .$$

Das so gestellte Problem einer nicht-linearen Optimierung unter Nebenbedingungen kann gelöst werden, wie Gabler et al. (2012) zeigen. Ein einfacher, dort vorgeschlagener Algorithmus ermöglicht die optimale Aufteilung eines Gesamtstichprobenumfangs n auf H Schichten, wobei sowohl untere als auch obere Grenzen für die Stichprobenumfänge in den Schichten eingehalten und die Nebenbedingung in (2.2.15) erfüllt werden. Dabei macht sich der Algorithmus die Tatsache zunutze, dass die Schichten exakt drei Klassen zugeordnet werden können. In Schichten, die der ersten Klasse U_1 angehören, wird der Stichprobenumfang exakt auf die untere Schranke m_h gesetzt, in Schichten der zweiten Klasse U_2 wird n_h exakt auf die obere Schranke M_h gesetzt und in der dritten Klasse U_3 wird der verbleibende Stichprobenumfang $n - \sum_{h \in U_1} m_h - \sum_{h \in U_2} M_h$ optimal im Sinne von Neyman-Tschuprov aufgeteilt. Das Problem besteht somit darin, diejenige Zusammensetzung der Klassen zu finden, für die insgesamt eine bestimmte Zielfunktion minimiert wird. Der Algorithmus löst dieses Problem dadurch, dass zunächst zwei geordnete Reihen gebildet werden, in denen die Schichten entsprechend ihrer Ausprägung auf $N_h \cdot S_{h,Y}$ in aufsteigender, beziehungsweise absteigender Reihenfolge angeordnet sind. Anschließend werden die Kombinationen dieser Ordnungen Schritt für Schritt abgearbeitet. Die erste Lösung, bei der alle Elemente aus U_3 die Nebenbedingungen erfüllen, ist die angestrebte Lösung. Durch eine effiziente Programmierung lässt sich diese optimale Aufteilung des Stichprobenumfangs unter Box-Constraints auch bei tausenden von Schichten auf Grund der linearen Komplexität des resultierenden Algorithmus in weniger als einer Sekunde ermitteln. Ein Vergleich verschiedener numerischer Verfahren dieses Problems wird in Münnich et al. (2011*b*) durchgeführt.

Zusammenfassend lässt sich also sagen, dass das vorgestellte Stichprobendesign eine neue Variante einer optimalen Allokation unter Nebenbedingungen ist. Hierdurch werden verschiedene Zielsetzungen realisiert. Auf der einen Seite können Mindeststichprobenumfänge formuliert werden, auf der anderen Seite kann die Variabilität von Designgewichten eingeschränkt werden, wodurch der statistische Modellbau erleichtert wird, der bei der Verwendung von Small Area-Modellen benötigt wird. Außerdem verhindern die Box-Constraints eine allzu ungleiche Befragungswahrscheinlichkeit der Bevölkerung.

Die Optimierungsroutinen wurden bei den Untersuchungen speziell auf das *Ziel 1* ausgerichtet, da nur hierfür Daten in der hinreichenden Qualität zur Verfügung stehen. Die nachfolgenden Grafiken wurden in analoger Weise aufgebaut. Die Zeilen werden durch unterschiedliche Entnahmeanteile gekennzeichnet (grüner Bereich), die Spalten durch verschiedene Schichtungen (hellbrauner Bereich). Die Typen stellen in diesem Abschnitt die vollständigen RRMSE der vier Kategorien dar, wobei bei Typ 2 (Verbandsgemeinden) eine proportionale Aufteilung von SMP-Typ 3-Stichprobenumfängen auf die beteiligten Verbandsgemeinden erfolgte. In diesen Fällen erfolgt in realiter eine zufällige Aufteilung des Stichprobenumfangs, da übergeordnet in Kreisresten gezogen wird. In allen Fällen ist die durchgezogen rote Linie der Benchmark, der aus den anfangs dargestellten Qualitätsanforderungen resultiert. Die gestrichelte blaue Linie kennzeichnet den mittleren RRMSE über alle im jeweiligen Szenario abgebildeten Fragestellungen.

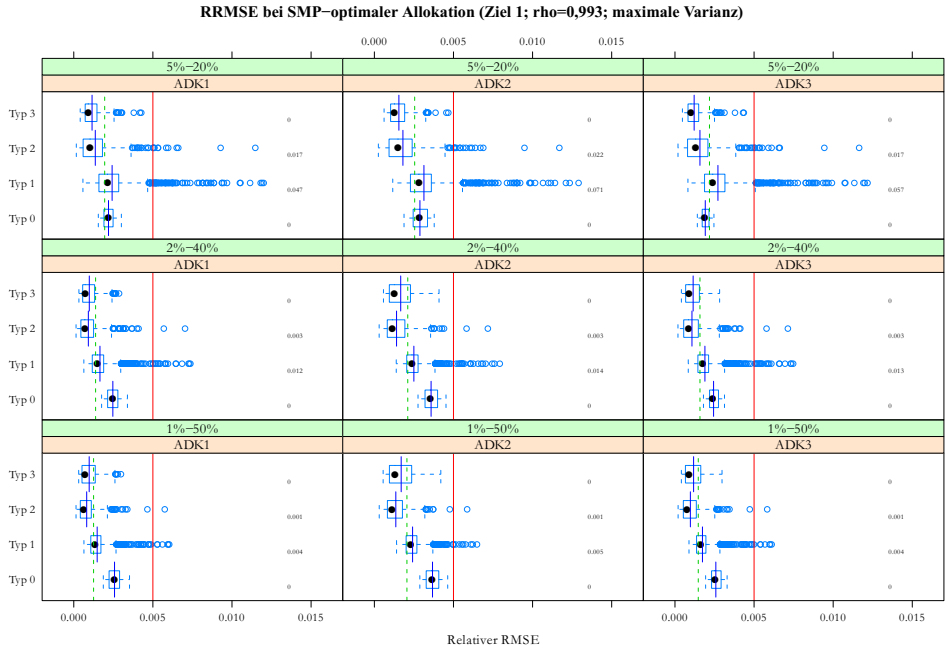


Abbildung 2.3: RRMSE für *Ziel 1* bei $\rho = 0{,}993$ bei drei Schichtungen und drei Entnahmeanteilsvariationen bei SMP-optimaler Allokation

In Abbildung 2.3 erkennt man, dass einzelne Gemeinden bzw. Verbandsgemeinden den Anforderungen nicht genügen. Im Durchschnitt aller Fragestellungen auf allen drei Ebenen ergeben sich indes keine Verletzungen der Qualitätsanforderungen. Es sei nochmals darauf hingewiesen, dass

bei *Ziel 1* Schätzungen auf Ebene der Verbandsgemeinden und Kreise nicht erfolgen und nur aus Vergleichsgründen ausgewiesen sind.

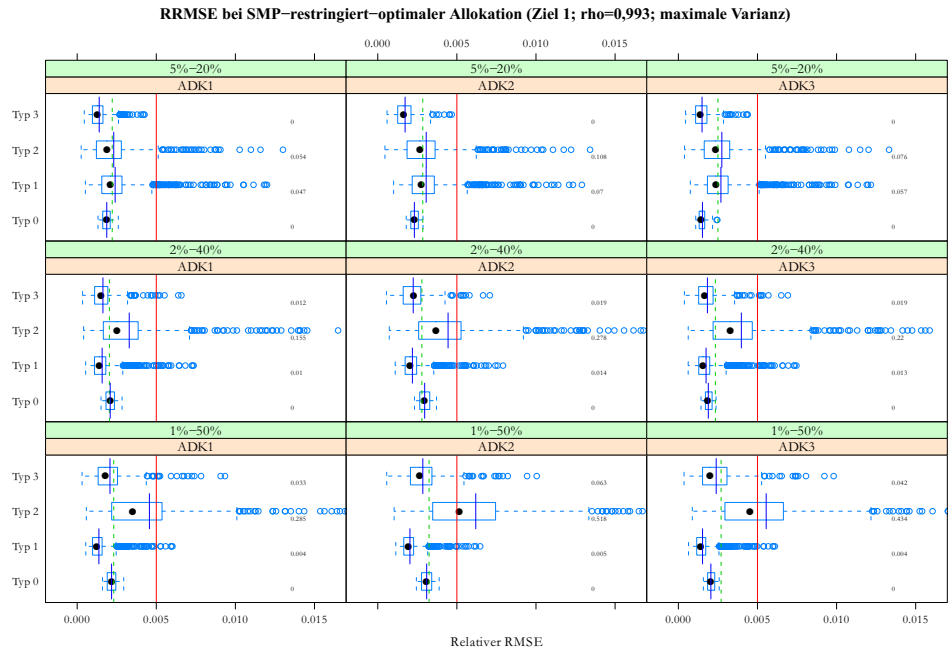


Abbildung 2.4: RRMSE für *Ziel 1* bei $\rho = 0{,}993$ bei drei Schichtungen und drei Entnahmeanteilsvariationen bei eingeschränkter SMP-optimaler Allokation

In Abbildung 2.4 wurde eine eingeschränkte SMP-optimale Allokation verwendet. Im Gegensatz zur Allokation zuvor wurden alle Schichten in den SMP-Typen 2 und 3, also den Resten in Verbandsgemeinden und Kreisen, auf den Wert des Mindestentnahmeanteils gesetzt. In Abbildung 2.5 wurde zusätzlich die etwas geringere Korrelation von $\rho = 0{,}987$ statt $\rho = 0{,}993$ zugrunde gelegt.

Man erkennt, dass sich zunächst die Qualität der Typen 2 und 3 verschlechtert, die der Stadtteile und Gemeinden jedoch nur unwesentlich verbessert. Der Effekt reduziert sich jedoch ganz erheblich, wenn ein höherer Mindestentnahmeanteil verwendet wird. Ebenso sieht man, dass die Schichtung ADK2 weniger geeignet ist. ADK1 und ADK3 lassen noch geringfügige Spielräume für verbesserte Stratifikationen.

Auffällig negativ wirkt sich die Modellannahme der geringeren Korrelation zwischen tatsächlich in den Anschriften vorhandenen Personen und den Anschriftengrößen aus. Die Ergebnisse von *Ziel 1* sind sehr sensitiv bezüglich der verwendeten Korrelation. Da außer dem Zensustest keinerlei valide Informationen hierzu vorliegen, bleibt die Wahl der Vorgaben und die damit einhergehenden Beurteilungen dem Auftraggeber vorbehalten.

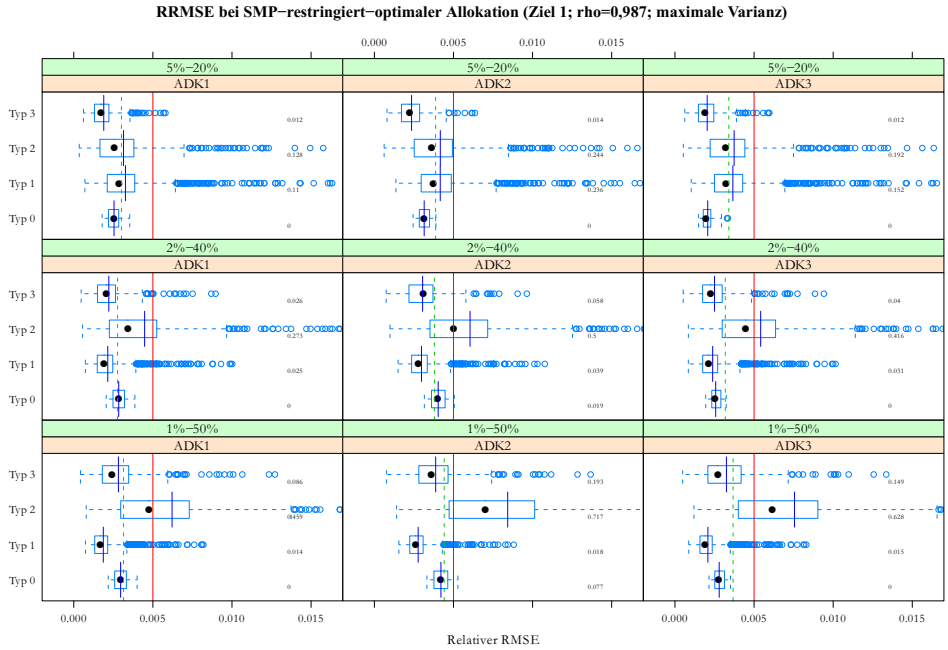


Abbildung 2.5: RRMSE für *Ziel 1* bei $\rho = 0{,}987$ bei drei Schichtungen und drei Entnahmeanteilsvariationen bei eingeschränkter SMP-optimaler Allokation

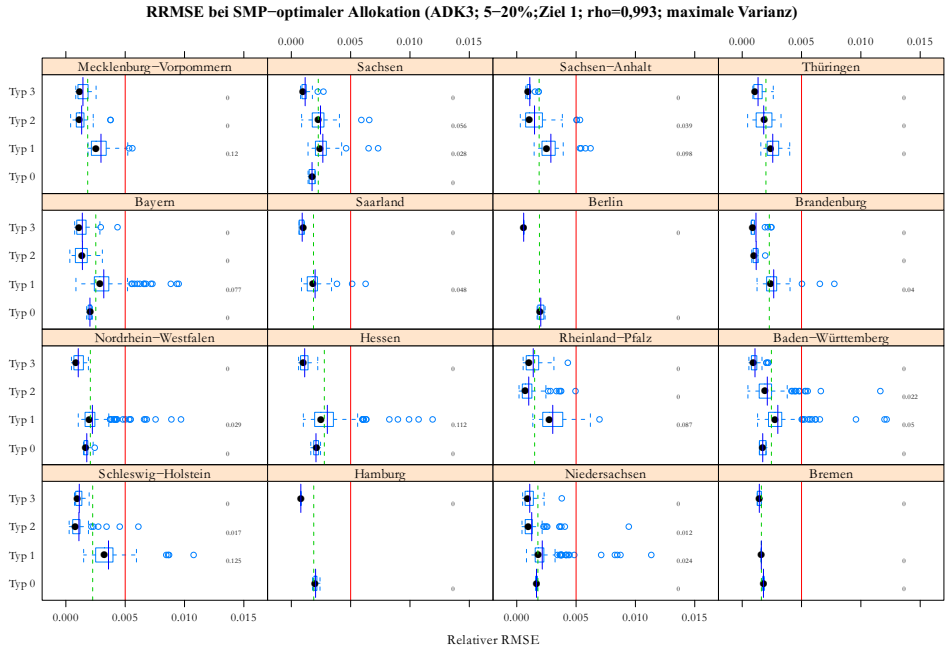


Abbildung 2.6: RRMSE für *Ziel 1* bei $\rho = 0{,}993$ bei ADK3-Schichtungen und Entnahmeanteil 5-20 % bei SMP-optimaler Allokation für die 16 Bundesländer

Schließlich sei noch die Aufteilung der Qualität auf die 16 Bundesländer dargestellt. Hierzu werden in Analogie zu Abbildung 2.3 die RRMSEs bei ADK3 und einem Entnahmeanteil von 5-20 % dargestellt. Man erkennt, dass sich prinzipiell ähnliche Verteilungen der RRMSEs ergeben, die *Ausreißer* jedoch eher vereinzelt vorliegen. Im Allgemeinen entstehen diese, wenn die maximalen Entnahmeanteile insbesondere in den oberen Schichten angenommen werden und in den zugehörigen SMPs keine weitere Reduktion des RRMSEs mehr stattfinden kann. Es liegen in solchen Fällen auffällig hohe Streuungen der Anschriftengrößen vor, die solche Probleme verursachen. Sofern im vorliegenden Datenmaterial noch einzelne Anstalten enthalten sind, darf noch von einer Reduktion solcher Ausreißer ausgegangen werden. Bei Verwendung von Minimax-Regeln (der Supremums-Norm) statt der 2-Norm würden in solchen Fällen ineffiziente Allokationen resultieren, da weitere mögliche Verbesserungen in relativ *besseren* Schichten nicht mehr durchgeführt werden könnten.

Letztendlich fiel die Entscheidung auf acht Schichten in personengleicher Allokation[11]. Die oberen und unteren Entnahmeanteile wurden schließlich auch in den vier Gemeindegrößenklassen (GemGK) unterschiedlich gewählt. Dabei wurde folgende Einteilung verwendet:

GemGK I 0 bis unter 10.000 Einwohner

GemGK II 10.000 bis unter 30.000 Einwohner

GemGK III 30.000 bis unter 100.000 Einwohner

GemGK IV ab 100.000 Einwohnern

Mit $p_h := \frac{m_h}{N_h} \leqslant \frac{n_h}{N_h} \leqslant \frac{M_h}{N_h} =: P_h$ gelten schließlich folgende Festlegungen:

Tabelle 2.3: Entnahmeanteile der Anschriften in den SMPs nach Gemeindegrößenklassen

	SMP-Typ									
	0		1		2 (RLP)		2 ($\overline{\text{RLP}}$)		3	
GemGK	p_h	P_h	p_h	P_h	p_h	P_h	p_h	P_h	p_h	P_h
I	—	—	—	—	—	—	—	—	0,05	0,05
II	—	—	0,05	0,50	0,05	0,50	0,05	0,05	0,05	0,05
III	—	—	0,04	0,40	0,04	0,40	0,05	0,05	0,05	0,05
IV	0,02	0,40	0,02	0,40	0,02	0,40	0,05	0,05	0,05	0,05

Insgesamt werden also Auswahlsätze zwischen 2 und 50 % verwendet, womit ein Gelman-Faktor von 25 resultiert, der in den einzelnen SMPs sogar 20 nicht überschreitet. In ländlichen Gegenden wird ein konstanter Auswahlsatz von 5 % verwendet.

In nachfolgender Abbildung 2.7 erkennt man, dass die so gewählten Rahmenbedingungen, wie erwartet, a priori eine Einhaltung der Präzisionsvorgaben erwarten lassen.

[11] Personengleiche Allokation bedeutet, dass jede Schicht gleich viele registrierte Personen umfasst. Eine Alternative wäre die gleiche Anzahl von Adressen in jeder Schicht gewesen, welche jedoch zu weniger effizienten Schätzungen führt.

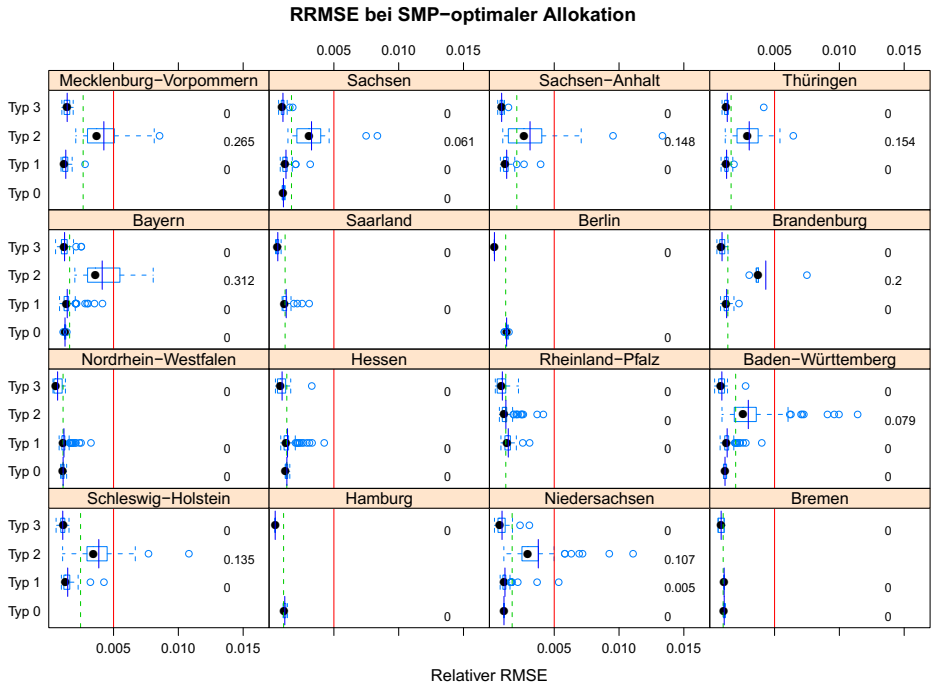


Abbildung 2.7: Theoretische relative RRMSEs in Bezug auf Bundesländer und SMP-Typen

Da später aber nicht notwendigerweise von einem homogenen $\rho = 0{,}993$ ausgegangen werden darf und die Regressionskoeffizienten aus der Stichprobe geschätzt werden müssen, sollte eine weitere Einschränkung der Variabilität der Auswahlsätze nicht vorgenommen werden, da sonst mit Effizienzverlusten zu rechnen ist, die sich negativ auf das Einhalten der Präzisionsvorgaben auswirken könnten.

2.3 Schätzmethoden im Zensus 2011

2.3.1 Vorbemerkungen zur Schätzung

Beim Zensus 2011 spielen Schätzmethoden sowohl bei *Ziel 1* als auch bei *Ziel 2* eine wesentliche Rolle. Zunächst soll die Zahl der Karteileichen und der Fehlbestände in allen Gemeinden ab 10.000 Einwohnern (SMP-Typ 1) beziehungsweise in Stadtteilen von Großstädten (SMP- Typ 0) ermittelt werden. Diese können jedoch nicht aus den Registern, sondern nur auf Basis von Schätzungen aus der Stichprobe gewonnen werden, also durch $\widehat{\tau}_K$ und $\widehat{\tau}_F$. Aus diesen Schätzwerten soll die amtliche Einwohnerzahl ermittelt werden. Sie wird im Folgenden als τ_Z bezeichnet und ergibt sich als

$$\widehat{\tau}_Z = \tau_R + \widehat{\tau}_F - \widehat{\tau}_K \quad . \tag{2.3.1}$$

Schätzwerte sind durch ein Dach gekennzeichnet. Die amtliche Einwohnerzahl $\widehat{\tau}_Z$ geht also aus dem Registertotalwert τ_R korrigiert um die geschätzte Zahl der Karteileichen $\widehat{\tau}_K$ und der Fehlbe-

stände $\widehat{\tau}_F$ hervor. Dies gilt für alle zu betrachtenden Areas $d = 1, \ldots, D$[12]. Für die Schätzung stehen sowohl Informationen aus der Stichprobe, als auch Informationen aus Registern zur Verfügung. Informationen über Karteileichen und Fehlbestände sind jedoch ausschließlich in der Stichprobe verfügbar. Die Registerinformationen umfassen Daten der Melderegister sowie der Bundesagentur für Arbeit. Diese Daten können als Hilfsinformation in den Schätzern verwendet werden. Die Anschriftengröße laut Register wurde bereits bei der Optimierung des Stichprobendesigns verwendet (siehe Abschnitt 2.2).

Im Rahmen der Untersuchungen wurden zunächst die EURAREA [13] Standard-Schätzmethoden als Grundlage verwendet. Diese Methoden gelten als State of the Art und als zuverlässig zur simultanen Schätzung auf Area-Ebene. Ideen und Grundlagen dieser Verfahren sowie einige für die Schätzung im Zensus 2011 relevante Erweiterungen werden nachfolgend beschrieben.

2.3.2 Design-basierte Schätzmethoden

Im Rahmen der klassischen Stichprobentheorie spielt insbesondere der Ziehungsprozess der Stichprobe eine wesentliche Rolle. In die Schätzung gehen die Wahrscheinlichkeiten ein, mit denen die Elemente der Gesamtheit bei gegebenem Auswahlverfahren in die Stichprobe gelangen. Man bezeichnet diese Wahrscheinlichkeit für Element i als *Inklusionswahrscheinlichkeit* π_i. Ein Schätzer ist Design-erwartungstreu, wenn sein Erwartungswert bezüglich des Auswahlverfahrens gleich dem Populationsparameter ist.

Die nachfolgenden Ausführungen können in der klassischen Stichprobenliteratur wie etwa Cochran (1977), Särndal et al. (1992) oder Lohr (1999) nachgelesen werden. Für die simultane Schätzung auf oft kleinräumigen Areas (Small Area-Estimation) sei auf Rao (2003) bzw. Lehtonen und Veijanen (2009) verwiesen.

2.3.2.1 Der Horvitz-Thompson-Schätzer

Der in der amtlichen Statistik am weitesten verbreitete Schätzer ist der Horvitz-Thompson-Schätzer (HT). Zur Schätzung von Totalwerten kann er als gewichtetes Mittel der Beobachtungswerte interpretiert werden, wobei die Gewichte die inversen Inklusionswahrscheinlichkeiten sind. Für die Schätzung des Totalwertes in Area d ist der HorvitzThompson-Schätzer definiert als

$$\widehat{\tau}_{Y,d}^{\mathrm{HT}} = \sum_{i \in S_d} w_i \cdot y_i \tag{2.3.2}$$

mit Designgewichten $w_i = 1/\pi_i$. Hierbei besteht S_d aus allen Stichprobenelementen (Anschriften) der Area d.

Dieser Schätzer ist Design-unverzerrt. Er verwendet nur Informationen, die in Area d beobachtet werden. Er ist daher ein sogenannter *direkter Schätzer*.

Für geschichtete Zufallsstichproben ist

$$\widehat{\tau}_{Y,d}^{\mathrm{HT}} = \sum_{h=1}^{H} N_{h,d} \cdot \overline{y}_{h,d} \quad . \tag{2.3.3}$$

12 Die Bezeichnungen Area und Domain werden im Folgenden nachfolgend synonym verwendet.

13 Links zu den Berichten des EURAREA-Projekts: www.ons.gov.uk/ons/guide-method/method-quality/general-methodology/spatial-analysis-and-modelling/eurarea/index.html bzw. www.dacseis.de, WP10

$N_{h,d}$ bezeichnet die Anzahl der Anschriften und $\overline{y}_{h,d}$ das arithmetische Mittel der Stichprobenwerte des Untersuchungsmerkmals in Schicht h der Area d. Die Gewichte w_i werden zur besseren Verständlichkeit auch $w_{h,i,d}$ geschrieben und es gilt im geschichteten Fall $w_{h,i,d} = {}^{N_{h,d}}/_{n_{h,d}}$ für alle Stichprobeneinheiten i in Schicht h und Domain d. Hilfsinformationen aus Registern werden nicht verwendet.

Der Varianzschätzer des HT-Schätzers im Falle geschichteter Zufallsstichproben ist

$$\widehat{\mathrm{V}}\left(\widehat{\tau}_{Y,d}^{\mathrm{HT}}\right) = \sum_{h=1}^{H} N_{h,d}^2 \cdot \frac{S_{h,d}^2}{n_{h,d}} \cdot \left(1 - \frac{n_{h,d}}{N_{h,d}}\right) \quad , \tag{2.3.4}$$

wobei $n_{h,d}$ der Stichprobenumfang und $S_{h,d}^2$ die Varianz der Stichprobenwerte des Untersuchungsmerkmals in Schicht h und Area d ist. Ein Design-erwartungstreuer Varianzschätzer ist durch

$$\widehat{\mathrm{V}}\left(\widehat{\tau}_{Y,d}^{\mathrm{HT}}\right) = \sum_{h=1}^{H} N_{h,d}^2 \cdot \frac{s_{h,d}^2}{n_{h,d}} \cdot \left(1 - \frac{n_{h,d}}{N_{h,d}}\right) \quad , \tag{2.3.5}$$

gegeben. In beiden Fällen wird die inferentiell-statistische Varianz verwendet. Bei großen Teilstichprobenumfängen $n_{h,d}$ liefert der Varianzschätzer sehr gute Ergebnisse. Bei kleinen Teilstichprobenumfängen hat der HT-Schätzer möglicherweise eine große Varianz und kann eine hohe Variabilität in den Varianzschätzungen aufweisen.

2.3.2.2 Der verallgemeinerte Regressionsschätzer (GREG)

Im Gegensatz zum HT-Schätzer wird beim verallgemeinerten Regressionsschätzer ein statistisches Modell verwendet. Geht man vom linearen Regressionsmodell

$$y = x' \cdot \beta + \varepsilon \tag{2.3.6}$$

mit den üblichen Annahmen für ε (siehe Särndal et al. 1992, S. 225) aus, erhält man als verallgemeinerten Regressionsschätzer für den Totalwert $\tau_{Y,d}$ bezüglich der Untersuchungsvariablen Y in Domain d

$$\widehat{\tau}_{Y,d}^{\mathrm{GREG}} = \sum_{i \in S_d} w_{i,d} y_{i,d} + \left(\sum_{i=1}^{N_d} x_{i,d} - \sum_{i \in S_d} w_{i,d} \cdot x_{i,d} \right)' \widehat{\beta} \tag{2.3.7}$$

mit

$$\widehat{\beta} = \left(\sum_{i \in S} w_i x_i x_i' \right)^{-1} \sum_{i \in S} w_i x_i y_i \quad . \tag{2.3.8}$$

Dabei sind x_i die Hilfsinformationen der i-ten Anschrift, die auch vektoriell vorliegen können. $\widehat{\beta}$ ist die Lösung der KQ-Schätzung des Regressionskoeffizienten im linearen Regressionsmodell (2.3.6) bezogen auf die gesamte Stichprobe S. Abweichend davon können die β auch für jede Domain separat oder basierend auf Gruppen g geschätzt werden, wobei sich diese Gruppen aus Areas und/oder Schichten zusammensetzen können – die Einteilung in Gruppen kann etwa die ursprüngliche Aufteilung in Areas (SMPs) vergröbern. Beim verallgemeinerten Regressionsschätzer handelt

es sich um ein *Modell-unterstütztes Verfahren*, da die Design-basierte Schätzung mit Hilfe eines Regressionsmodells korrigiert wird.

Im Rahmen der Small Area-Methodik wird der Schätzer (2.3.7) im Allgemeinen durch

$$\widehat{\tau}_{Y,d}^{\text{GREG}} = \sum_{i=1}^{N_d} x_{i,d}' \widehat{\beta}_g + \sum_{i \in S_d} w_{i,d} \cdot \underbrace{\left(y_{i,d} - x_{i,d}' \widehat{\beta}_g \right)}_{e_{i,d}} \tag{2.3.9}$$

beschrieben. Diese Schreibweise setzt sich aus einem synthetischen Teil und den gewichteten Residuen $e_{i,d}$ in der Stichprobe zusammen. Man unterscheidet folgende Klassen:

Kombinierte Schätzung: Es wird über alle Schichten hinweg ein einziges Regressionsmodell betrachtet;

Separate Schätzung: Es wird separat in jeder Schicht ein eigenes Regressionsmodell betrachtet.

Im Rahmen der klassischen Theorie werden diese Modelle eingehend in Münnich (1997) behandelt.

Bei Small Area-Modellen kommt eine weitere Unterscheidung in Form von Areas dazu. Dann kann es sinnvoll sein, gruppierte Schätzungen der β_g zu betrachten. Im Rahmen der Zensus-Untersuchungen wurden folgende Gruppierungen festgelegt:

- Voll kombinierte Schätzung über alle Areas und Schichten;
- SMP-separate Schätzung (kombiniert über die Schichten);
- Kreis-separate Schätzung (alle SMPs eines Kreises werden als Gruppe betrachtet, über die Schichten wird kombiniert geschätzt);
- Schicht-separate Schätzung;
- Voll separate Schätzung.

Eine detaillierte Darstellung gruppierter Schätzer kann in Särndal et al. (1992) bzw. Lehtonen und Veijanen (2009) nachgelesen werden.

Werden die β auf Basis von Gruppen geschätzt, die mehrere Areas umfassen, so liegt ein indirekter Schätzer vor, denn es werden auch Informationen *benachbarter* beziehungsweise ähnlicher Areas verwendet. Damit können Schätzungen stabilisiert werden, wenn in einzelnen Areas nur wenige Stichprobeninformationen vorliegen. Man spricht dann von *borrowing strength*, also dem Borgen von Stärke durch Verwendung von Informationen über die interessierende Area hinaus. Man spricht dann auch von indirekter Schätzung, da zur Schätzung auf Area d auch Informationen außerhalb dieser Area verwendet werden. Eine direkte Schätzung würde ausschließlich Informationen auf der zu betrachtenden Area d verwenden.

In Analogie zum Varianzschätzer des HT erhält man im geschichteten Fall für den GREG

$$\widehat{\mathrm{V}}\left(\widehat{\tau}_{Y,d}^{\text{GREG}}\right) = \sum_{h=1}^{H} N_{h,d}^2 \cdot \frac{s_{e,h,d}^2}{n_{h,d}} \cdot \left(1 - \frac{n_{h,d}}{N_{h,d}}\right) \quad , \tag{2.3.10}$$

wobei $s_{e,h,d}^2$ die Stichprobenvarianz der Residuen $e_{h,i,d}$ bezeichnet. Die Herleitung folgt unmittelbar aus der Darstellung (2.3.9), wobei die Variabilität von $\widehat{\beta}_g$ vernachlässigt wird. Wird β_g auf Gruppen

mit wenigen Beobachtungen geschätzt, dann tendiert der Varianzschätzer zu einer Unterschätzung der tatsächlichen Varianz.

Der GREG-Schätzer ist asymptotisch Design-unverzerrt. Die Verzerrung ist im Allgemeinen sehr gering, solange die Teilstichprobenumfänge nicht zu klein sind. Durch die Möglichkeit, diese Regressionsschätzer mit Hilfe von *g*-weights darzustellen (siehe Särndal et al. 1992), weist der (gruppierte) verallgemeinerte Regressionsschätzer eine wünschenswerte Eigenschaft auf: Er ist *vertikal kohärent* (siehe Särndal et al. 1992, S. 399-400). Das bedeutet, dass die Aggregation von Schätzungen auf Untergruppen stets der Schätzung auf der übergeordneten Ebene entspricht.

Der Präzisionsgewinn des GREG gegenüber dem HT kann bei Vorliegen nur einer Hilfsvariable bei kombinierter Hochrechnung einfach durch $1 - \rho^2$ angegeben werden, wobei ρ der Korrelationskoeffizient zwischen Untersuchungs- (Y) und Hilfsmerkmal (X) ist. Das Vorliegen der Registerdaten als Hilfsmerkmal bei *Ziel 1*-Fragestellungen legt die Verwendung des GREG nahe, weil von einer hohen Korrelation mit dem zu schätzenden Merkmal ausgegangen wird. Der GREG dient deshalb im Weiteren als Benchmark-Schätzer.

Der Umfang der zusätzlich zum HT benötigten Information ist beim GREG relativ gering: Es werden lediglich die Ausprägungen der Hilfsmerkmale für alle Elemente der Stichprobe benötigt. Zudem müssen die Ausprägungen der Hilfsmerkmale im Korrekturteil nur auf aggregiertem Niveau vorliegen. Damit können auch externe Register verwendet werden, auf die mit Hilfe einer Variablen in der Stichprobe verlinkt wird (siehe z. B. Wiegert und Münnich 2004). Diese Eigenschaft folgt aus der Linearität des synthetischen Teils in (2.3.9).

Zahlreiche Erweiterungen des Regressionsschätzers lassen sich einfach implementieren. Zum einen können nicht-parametrische Regressionen mögliche nicht-lineare Zusammenhänge geeignet abbilden, wie dies beispielsweise bei Splines der Fall ist (vgl. Münnich 1997). Aufgrund der erwarteten hohen Variabilität in der obersten Schicht wird auf eine eingehende Betrachtung dieser Methoden verzichtet. Verallgemeinerungen können durch erweiterte Schätzungen der β erreicht werden, etwa durch Verwendung von Multi-Level-Modellen oder der verallgemeinerten Momentenmethode. Eine eingehende Darstellung findet man in Lehtonen und Veijanen (2009).

2.3.3 Synthetische und kombinierte Schätzverfahren

Betrachtet man in Gleichung (2.3.9) nur den synthetischen Teil und ignoriert den auf der Stichprobe basierenden Residualterm, so erhält man eine rein synthetische Schätzung, bei der die Stichprobendaten ausschließlich zur Schätzung von β herangezogen werden. Prinzipiell können derartige synthetische Schätzungen auf sehr unterschiedlichen Modellen basieren. Ihre Motivation ergibt sich aus folgendem Sachverhalt:

Ausgangspunkt ist ein Schätzverfahren für jede Area $d = 1, \ldots, D$. Aufgrund von kleinen Area-Fallzahlen wird eine geforderte Genauigkeit möglicherweise nicht in allen Areas erfüllt. Jedoch sei der direkte Schätzer hinreichend genau für eine übergeordnete Gesamtheit, die aus mehreren oder allen Areas bestehen kann.

Deshalb wird ein indirekter Schätzer konstruiert, der Informationen des direkten Schätzers aus allen D Areas verwendet. Dies geschieht unter der Annahme, dass alle untergeordenten Areas dieselbe Struktur wie die übergeordnete Gesamtheit aufweisen. Es wird also die Information der übergeordneten Gesamtheit auf die einzelnen Areas *synthetisch übertragen*.

Die unterschiedlichen synthetischen Schätzer ergeben sich durch die Konkretisierung der *Struktur*, die im Allgemeinen auf Basis von Modellen erfolgt. Beispiele hierfür sind:

- Kennwerte der übergeordneten Gesamtheit werden unter Berücksichtigung der Größenproportionen auf die D Areas übertragen;
- Es werden identische Modelle unterstellt.

2.3.3.1 Synthetische Schätzung ohne Hilfsinformationen

Sind keinerlei Hilfsinformationen über die Areas verfügbar, dann ist der *naive synthetische* Schätzer für den Mittelwert in Area d gegeben durch

$$\widehat{\mu}_{Y,d}^{\text{Synth}} = \frac{\sum\limits_{i \in S} w_i \cdot y_i}{\sum\limits_{i \in S} w_i} = \widehat{\mu}_Y^{\text{Synth}} \qquad (d = 1, \ldots, D) \qquad .$$

Dieser naive Schätzer geht von der Annahme aus, dass die Mittelwerte für alle Areas gleich sind und wird daher auch als *National Sample Mean* bezeichnet. Dieser Schätzer kann für Mittelwerte oder Anteile berechnet werden. Bei der Schätzung von Totalwerten müssen die Umfänge der Areas N_d berücksichtigt werden. Ist dieser Umfang nicht bekannt, kann er mittels $\hat{N}_d = \sum_{i \in S_d} w_i = \sum_{i \in S_d} w_{i,d}$ geschätzt werden. $w_{i,d}$ hängt ausschließlich von der Einheit i ab. Der Index i, d bringt nur zum Ausdruck, dass die Einheit i in der Domain d liegt. Wir können für die Schätzung des Totalwerts $\tau_{Y,d}$ also

$$\widehat{\tau}_{Y,d}^{\text{Synth}} = \hat{N}_d \cdot \widehat{\mu}_{Y,d}^{\text{Synth}} = \sum_{i \in S_d} w_{i,d} \cdot \frac{\sum\limits_{i \in S} w_i \cdot y_i}{\sum\limits_{i \in S} w_i} \qquad (d = 1, \ldots, D)$$

verwenden. Der naive synthetische Schätzer liefert im Allgemeinen verzerrte Schätzungen.

2.3.3.2 Regressions-synthetische Schätzung

Sind Area-spezifische Totalwerte einer Hilfsvariablen X, bezeichnet mit $\tau_{X,d}$, verfügbar, dann ist

$$\widehat{\tau}_{Y,d}^{\text{SynthR}} = \tau'_{X,d}\widehat{\beta} \qquad \text{mit} \qquad \widehat{\beta} = \left(\sum_{i \in S} w_i \mathbf{x}_i \mathbf{x}_i' / c_i\right)^{-1} \left(\sum_{i \in S} w_i \mathbf{x}_i \mathbf{y}_i' / c_i\right) \tag{2.3.11}$$

der Regressions-synthetische Schätzer für $\tau_{Y,d}$. Dieser Schätzer entspricht dem synthetischen Teil des GREG, wenn β auf der übergeordneten Gesamtheit geschätzt wird, also dem voll kombinierten GREG. Die c_i sind näher zu spezifierende Konstanten, welche der Einfachheit halber zunächst auf $c_i \equiv 1$ festgelegt werden. Dabei wird angenommen, dass die Regressionsterme der Areas dem globalen Regressionsterm β entsprechen.

Der Regressions-synthetische Schätzer ist insbesondere dann sehr effizient, wenn Area-spezifische Effekte bezüglich der zugrunde liegenden Regression nicht (oder kaum) vorliegen. Der Schätzer kann in einfacher Weise auf Mittelwertschätzungen angepasst werden.

2.3.3.3 Synthetischer Schätzer, Modell A

Beim synthetischen Schätzer, Modell A, wird ein Multi-Level-Modell auf Personen-Ebene geschätzt:

$$y_{i,d} = x_d'\beta + u_d + e_{i,d} \qquad (2.3.12)$$
$$u_d \sim \text{iid } N(0;\sigma_u^2), \quad e_{i,d} \sim \text{iid } N(0;\sigma_e^2) \quad .$$

Genau genommen handelt es sich um ein Standard Unit-Level Random Intercept-Modell, auch als Random Effects-Modell bekannt. Die Hochrechnung wird anschließend über

$$\hat{\mu}_{Y,d}^{\text{SynthA}} = \mu_{X,d}'\hat{\beta} \qquad (2.3.13)$$

durchgeführt. Dieses Modell entspricht dem Regressions-synthetischen Schätzer bis auf den Unterschied, dass in das vorliegende Modell ein Random Effekt bezüglich der Areas aufgenommen wird. Die Schätzung erfolgt im Allgemeinen durch Maximum Likelihood (ML) oder Restricted Maximum Likelihood (REML) Methoden, wie sie in der Literatur vorgeschlagen werden (vgl. bspw. Goldstein 2003, S. 51 ff.).

Der MSE kann über

$$\widehat{\text{MSE}}\left(\hat{\mu}_{Y,d}^{\text{SynthA}}\right) = \hat{\sigma}_u^2 + \mu_{X,d}' \cdot \hat{V} \cdot \mu_{X,d} \qquad (2.3.14)$$

geschätzt werden (vgl. Rao 2003, S. 51 ff.), wobei $\hat{\sigma}_u^2$ die geschätzte Modellvarianz, $\mu_{X,d}$ die Mittelwerte der Hilfsvariablen und $\hat{V}$ den Design-basierten Varianzschätzer für $\hat{\mu}_{Y,d}^{\text{SynthA}}$ bezeichnen. Dieses und das folgende Modell sind elementare Bestandteile der klassischen Small Area-Schätzer, welche nachfolgend eingeführt werden (siehe auch die Forschungsberichte der Projekte EURAREA bzw. DACSEIS, WP10).

2.3.3.4 Synthetischer Schätzer, Modell B

Im Gegensatz zum synthetischen Schätzer auf Personen-Ebene werden beim synthetischen Schätzer auf Area-Ebene die Kovariaten nicht auf Beobachtungs-Ebene sondern auf Area-Ebene eingebracht. Man erhält

$$\bar{y}_d = \bar{x}_d'\beta + \xi_d \quad ; \quad \xi_d \sim iid\, N\left(0;\sigma_u^2 + \frac{\sigma_e^2}{n_d}\right) \qquad (2.3.15)$$
$$\text{mit} \quad \hat{\sigma}_e^2 = \frac{1}{n-D}\sum_{d=1}^{D}\sum_{i=1}^{N_d}(y_{i,d} - \bar{y}_d)^2 \quad .$$

Die Hochrechnung erfolgt wiederum über

$$\hat{\mu}_{Y,d}^{\text{SynthB}} = \mu_{X,d}'\hat{\beta} \quad . \qquad (2.3.16)$$

Auch die MSE-Schätzung wird in Analogie zum vorigen Modell durchgeführt.

Diese Modellierung hat den Vorteil, dass die verwendeten Hilfsinformationen lediglich auf Area-Niveau benötigt werden.

2.3.3.5 Zusammengesetzte Schätzer

Die synthetischen Schätzer haben den Vorteil, dass sie zumeist eine im Verhältnis zum GREG sehr kleine Varianz aufweisen. Allerdings können die Schätzwerte erhebliche Verzerrungen aufweisen, wenn das Modell die Strukturen nicht in allen Areas adäquat wiedergibt.

Vergleicht man synthetische mit Design-basierten Schätzungen, kann man schnell nachvollziehen, dass die Vor- und Nachteile sich gegenseitig ausgleichen. Tabelle 2.4 gibt einen Vergleich der Eigenschaften von Design-basierten und synthetischen Schätzmethoden (siehe Lehtonen und Veijanen 2009, S. 225). Die Güte der Schätzer der unten stehenden Kennwerte (Bias, Varianz und MSE) wird in beiden Schätzklassen stets Design-basiert betrachtet. Es sei darauf hingewiesen, dass nach Lehtonen und Veijanen (2009, S. 225) die Aussagen in der rechten Spalte auch für den EBLUP (siehe folgender Abschnitt auf Seite 46) gelten.

Tabelle 2.4: Design-basierte Eigenschaften von Modell-unterstützten und Modell-abhängigen Schätzern

	Design-basiert Modell-unterstützt	Modell-abhängig
	GREG	Synthetische Schätzer
Bias	Design-unverzerrt (appr.)	Design-verzerrt. Design-Verzerrung geht auch asymptotisch nicht notwendiger Weise mit steigenden Stichprobenumfängen gegen Null.
Varianz	Groß, wenn n klein	Kann selbst für kleine Domains klein sein.
	Varianz wird mit steigenden Stichprobenumfängen kleiner	
MSE	MSE $\approx$ Varianz	MSE = Varianz + Bias^2; kann groß sein, wenn Verzerrung wesentlich ist.
Konfidenzintervall	Valide Design-basierte Intervalle können konstruiert werden.	Valide Design-basierte Intervalle können nicht notwendigerweise erhalten werden.

Durch die hohe Komplementarität der beiden Ansätze lassen sich unmittelbar zusammengesetzte Schätzer konstruieren, welche die Vorteile beider Verfahren ausnutzen.

Es sei $\widehat{\tau}^{\text{dir}}_{Y,d}$ ein direkter Schätzer und $\widehat{\tau}^{\text{Synth}}_{Y,d}$ ein synthetischer Schätzer für $\tau_{Y,d}$. Dann ist der zusammengesetzte Schätzer $\widehat{\tau}^{\text{comp}}_{Y,d}$ definiert durch

$$\widehat{\tau}^{\text{comp}}_{Y,d} = \gamma_d \cdot \widehat{\tau}^{\text{dir}}_{Y,d} + (1 - \gamma_d) \cdot \widehat{\tau}^{\text{Synth}}_{Y,d} \tag{2.3.17}$$

mit $0 \leq \gamma_d \leq 1$ $(d = 1, \ldots, D)$.

Das wesentliche Problem besteht in einer geeigneten Wahl der Gewichte γ_d. Prinzipiell lassen sich zwei Methoden unterscheiden:

- Fallzahlabhängige Wahl von γ_d;
- Wahl eines optimierten γ_d.

Auf die Betrachtung der fallzahlabhängigen Methoden wird an dieser Stelle verzichtet. Einen Überblick gibt (Rao 2003, S. 60 ff.).

Der (Design-)MSE des zusammengesetzten Schätzers ist

$$\begin{aligned}\mathrm{MSE}(\widehat{\tau}_{Y,d}^{\mathrm{comp}}) = \gamma_d^2 \cdot \mathrm{MSE}(\widehat{\tau}_{Y,d}^{\mathrm{dir}}) + (1-\gamma_d)^2 \cdot \mathrm{MSE}(\widehat{\tau}_{Y,d}^{\mathrm{Synth}}) \\ + 2 \cdot \gamma_d \cdot (1-\gamma_d) \cdot \mathrm{E}\,(\widehat{\tau}_{Y,d}^{\mathrm{dir}} - \tau_{Y,d})(\widehat{\tau}_{Y,d}^{\mathrm{Synth}} - \tau_{Y,d}) \quad .\end{aligned} \tag{2.3.18}$$

Unter der Annahme, dass der Kovarianzterm relativ zum zweiten Term, d. h. zum MSE des synthetischen Schätzers, klein ausfällt (siehe Rao 2003, S. 58), erhält man nach Optimierung der quadratischen Funktion in γ_d schließlich

$$\gamma_d^{\mathrm{opt}} \doteq \frac{\mathrm{MSE}(\widehat{\tau}_{Y,d}^{\mathrm{Synth}})}{\mathrm{MSE}(\widehat{\tau}_{Y,d}^{\mathrm{Synth}}) + \mathrm{MSE}(\widehat{\tau}_{Y,d}^{\mathrm{dir}})} \quad . \tag{2.3.19}$$

Hierbei entstehen zwei Probleme. Zum Einen stellt sich die Frage, ob der Kovarianzterm in (2.3.18) tatsächlich vernachlässigbar ist. Zum Andern kann eine adäquate Schätzung der MSEs schwierig sein. Darüber hinaus ist zu überlegen, ob ein einheitliches γ_d in allen Areas sachgerecht ist. Wie nachfolgend zu sehen ist, sorgen variable Stichprobenumfänge für Variationen in den Gewichten γ_d.

Zwei Sonderfälle, die als zusammengesetzte Schätzer dargestellt werden können, bilden die beiden Standard-Schätzer der Small Area-Statistik, welche *Empirical Best Linear Unbiased Predictors* (EBLUP) sind. Auch ihnen liegen jeweils ein Unit-Level- sowie Area-Level-Modell zugrunde.

2.3.3.6 Empirical Best Linear Unbiased Predictors

Der *Best Linear Unbiased Predictor* (BLUP) ist eine gewichtete Kombination eines synthetischen Schätzers (z. B. SynthA oder SynthB) und eines direkten Schätzers (z. B. HT oder GREG). Das Gewicht γ_d ist durch die Modellvarianz σ_u^2 in Relation zur totalen Varianz $\sigma_u^2 + \frac{\sigma_e^2}{n_d}$ gegeben. Wenn die Modellvarianz im Verhältnis zur totalen Varianz relativ klein ist, wird mehr Gewicht auf die synthetische Komponente gelegt. Auf der anderen Seite erhält der direkte Schätzer mehr Gewicht, wenn der Area-Stichprobenumfang n_d wächst. Zu beachten ist aber, dass in beide Komponenten des zusammengesetzten Schätzers nur ein Modell eingeht, damit ein eindeutiges β geschätzt wird. Bei den vorliegenden beiden Varianten, dem Unit- und dem Area-Level-Modell, wird ein Multi-Level-Modell zugrundegelegt.

Werden die Varianzkomponenten des BLUP durch die Stichprobendaten geschätzt, spricht man vom Empirical BLUP oder EBLUP.

Eine weitere Unterscheidung innerhalb der Klasse der EBLUP ergibt sich durch die Verwendung unterschiedlicher synthetischer Schätzer. Wird als synthetischer Schätzer der SynthA verwendet, spricht man vom EBLUPA. Der EBLUPA geht auf das Modell von Battese, Harter und Fuller (vgl. Battese et al. 1988) zurück. Der EBLUPA-Schätzer für den Domain-Mittelwert $\mu_{Y,d}$ ist gegeben durch

$$\begin{aligned}\widehat{\mu}_{Y,d}^{\mathrm{EBLUPA}} &= \widehat{\gamma}_d^{\mathrm{A}} \cdot \widehat{\mu}_{Y,d}^{\mathrm{GREG}} + (1-\widehat{\gamma}_d^{\mathrm{A}}) \cdot \widehat{\mu}_{Y,d}^{\mathrm{SynthA}} \\ &= \widehat{\gamma}_d^{\mathrm{A}} \cdot (\widehat{\mu}_{Y,d}^{\mathrm{GREG}} - \widehat{\mu}_{X,d}'\widehat{\beta}) + \mu_{X,d}'\widehat{\beta} \quad ,\end{aligned} \tag{2.3.20}$$

mit

$$\hat{\gamma}_d^{\mathrm{A}} = \frac{\widehat{\sigma}_{u,A}^2}{\widehat{\sigma}_{u,A}^2 + \widehat{\sigma}_{e,A}^2/n_d} \quad . \tag{2.3.21}$$

Das Subskript A zeigt an, dass die Varianzkomponenten σ_u^2 und σ_e^2 in Abhängigkeit von Modell A (siehe Abschnitt 2.3.3.3) geschätzt werden.

Daneben ergibt sich der EBLUPB analog (vgl. Fay und Herriot 1979). Der EBLUPB ist eine gewichtete Kombination des synthetischen Schätzers B und des HT-Schätzers. Das Gewicht γ_d ist, in Analogie zum EBLUPA, der Quotient aus der Modellvarianz der Area Komponenten σ_u^2 und der totalen Varianz $\sigma_u^2 + \frac{\sigma_e^2}{n_d}$. Die Schätzung der γ_d erfolgt analog zu oben mit dem Unterschied, dass die Varianzkomponenten σ_u^2 und σ_e^2 nun in Abhängigkeit von Modell B geschätzt werden, also gilt:

$$\hat{\gamma}_d^{\mathrm{B}} = \frac{\widehat{\sigma}_{u,B}^2}{\widehat{\sigma}_{u,B}^2 + \widehat{\sigma}_{e,B}^2/n_d} \quad . \tag{2.3.22}$$

Man beachte, dass die Schätzung individueller Varianzen bei Area-Informationen nicht ohne weitere Annahmen möglich ist.

Eine geeignete Modellierung ist bei der Verwendung dieser Klasse von Schätzern unabdingbar. Nähere Informationen zu den Schätzern sind den Berichten der Forschungsprojekte EURAREA und DACSEIS zu entnehmen. Beide Schätzer sind die Standard-Schätzer im Unit- bzw. Area-Level-Modell.

Zuletzt sei noch darauf hingewiesen, dass wie zuvor die Registerinformationen nur als Aggregate auf Ebene der Areas, gegebenenfalls auch auf den Schichten, benötigt werden. Mikroinformation wird nicht benötigt, sofern die Information in der Stichprobe als vorhanden angenommen werden darf.

2.3.4 Erweiterte Small Area-Schätzer

2.3.4.1 Der Pseudo-EBLUP von You und Rao

Im Standard Unit-Level-Modell (EBLUPA) werden Informationen zum Stichprobendesign nicht berücksichtigt. Design-unverzerrte Ergebnisse sind daher nur bei SRS zu erwarten. You und Rao (2002) verwenden im Gegensatz zum Standard Unit-Level-Modell die Designgewichte. Weiterhin schätzen sie die Parameter des Modells unter Nebenbedingungen, so dass automatisch die Benchmark-Eigenschaft im Sinne des nationalen Gesamtwertes erfüllt wird. Hierzu werden die Designgewichte reskaliert. Seien $w_{i,d}$ die inversen Inklusionswahrscheinlichkeiten in Area d, dann sind die korrespondierenden reskalierten Gewichte $\tilde{w}_{i,d}$:

$$\tilde{w}_{i,d} = \frac{w_{i,d}}{\sum\limits_{\ell \in S_d} w_{\ell,d}} \quad , \qquad d = 1, \ldots, D \tag{2.3.23}$$

Als direkter Schätzer für $\mu_{Y,d}$ bietet sich

$$\widehat{\mu}_{Y,d} = \overline{y}_{d\tilde{w}} = \sum_{i \in S_d} \tilde{w}_{i,d} y_{i,d}$$

an. Man erhält dann folgenden Schätzer für die Area-Mittelwerte:

$$\begin{aligned}\widehat{\mu}_{Y,d}^{\text{YOURAO}} &= \widehat{\gamma}_{d\tilde{w}}\overline{y}_{d\tilde{w}} + (\mu_{X,d} - \widehat{\gamma}_{d\tilde{w}}\overline{x}_{d\tilde{w}})^{'}\widehat{\beta}_{\tilde{w}} \\ &= \mu_{X,d}^{'}\widehat{\beta}_{\tilde{w}} + \widehat{u}_{d\tilde{w}}\end{aligned} \tag{2.3.24}$$

mit

$$\begin{aligned}\widehat{u}_{d\tilde{w}} &= \widehat{\gamma}_{d\tilde{w}}(\overline{y}_{d\tilde{w}} - \overline{x}_{d\tilde{w}}^{'}\widehat{\beta}_{\tilde{w}}), \\ \widehat{\gamma}_{d\tilde{w}} &= \frac{\widehat{\sigma}_u^2}{\widehat{\sigma}_u^2 + \widehat{\sigma}_e^2\delta_d^2}, \\ \delta_d^2 &= \sum_{i\in S}\tilde{w}_{i,d}^2\end{aligned}$$

und

$$\widehat{\beta}_{\tilde{w}} = \left(\sum_{d=1}^{D}\gamma_{d\tilde{w}}\overline{x}_{d\tilde{w}}\overline{x}_{d\tilde{w}}^{\prime}\right)^{-1}\left(\sum_{d=1}^{D}\gamma_{d\tilde{w}}\overline{x}_{d\tilde{w}}\overline{y}_{d\tilde{w}}\right) \tag{2.3.25}$$

Für die Herleitung und die Eigenschaften des Schätzers verweisen wir auf You und Rao (2002).

2.3.4.2 Binomial-synthetischer Schätzer

Neben den *herkömmlichen* synthetischen Schätzern existieren noch weitere, speziellere Schätzer. Hierzu zählen beispielsweise Schätzer, die nicht auf Modellen mit zugrundeliegender Normalverteilung, sondern auf zugrundeliegender Binomial- oder Poissonverteilung basieren, und durch deren Verwendung Effizienzgewinne erzielt werden können. Durch die geeignetere Struktur dieser Modelle für kategoriale Daten sind bessere Schätzungen für sehr kleine Areas zu erwarten. Weiterhin können bei diesen Modellen, im Gegensatz zu den auf Normalverteilung basierenden linearen Modellen, bei der Schätzung von Anteilen keine Schätzwerte unter 0 oder über 1 resultieren. Allerdings erfordern diese Binomial- oder Poissonmodelle sowohl in der Schätzung, aufgrund von notwendigerweise iterativen Verfahren, wie auch bei der Vorhersage, aufgrund der Nicht-Linearität der Hochrechnung, einen erheblich höheren Rechen- und Speicheraufwand.

Derzeit existiert noch kein überzeugendes Konzept, wie eine Varianzschätzung auf einem Datensatz der im Zensus 2011 auftretenden Größe durchgeführt werden kann. Erste allgemeine Überlegungen zur Varianzschätzung für diese Modelle sind bei González-Manteiga et al. (2007) zu finden.

Die auf Binomial- und Poissonmodellen basierenden Schätzer werden in Anlehnung an den Best Prediction (BP) Ansatz von Jiang und Lahiri (2001) mit einem ähnlichen Setup wie in Münnich et al. (2009) und González-Manteiga et al. (2007) berechnet. Die Schätzung erfolgte über das Paket *lme4*

in *R*. Für die Poisson-Modelle wurde der Loglink $g(x) := log(x)$ und für die Binomial-Modelle der Logitlink $g(x) := logit(x)$ verwendet. Die Modelle lauten wie folgt:

$$
\begin{aligned}
y_{i,d} &\sim \text{Binomial}(n_{i,d}; \theta_{i,d}) \\
&\text{bzw.} \\
y_{i,d} &\sim \text{Poisson}(\theta_{i,d}) \\
&\text{und} \\
g(\theta_{i,d}) = \eta_{i,d} &= x_{i,d}^{'}\beta + u_d, \quad i \in S, \\
u_d &\sim N(0, \sigma_u^2) \qquad .
\end{aligned}
$$

Die Vorhersage für die synthetischen Schätzer erfolgt über

$$\widehat{\eta}_{i,d}^{\text{Synth}} = x_{i,d}^{'}\widehat{\beta} \qquad .$$

Die Vorhersage für den best prediction Schätzer erfolgt über

$$\widehat{\eta}_{i,d}^{BP} = x_{i,d}^{'}\widehat{\beta} + \widehat{u}_d \qquad .$$

Die Rücktransformation und die Aggregation ergibt sich aus

$$
\begin{aligned}
&\widehat{\theta}_{i,d}^{\text{Synth}} = g^{-1}(\widehat{\eta}_{i,d}^{\text{Synth}}) \quad \text{bzw.} \quad \widehat{\theta}_{i,d}^{\text{BP}} = g^{-1}(\widehat{\eta}_{i,d}^{\text{BP}}) \\
&\widehat{\tau}_d^{\text{Synth}} = \sum_{i=1}^{N_d} \widehat{\theta}_{i,d}^{\text{Synth}} \quad \text{bzw.} \quad \widehat{\tau}_d^{\text{BP}} = \sum_{i=1}^{N_d} \widehat{\theta}_{i,d}^{\text{BP}}, \; d = 1, \ldots, D \qquad .
\end{aligned}
$$

Aus dieser allgemeinen Form wurden folgende Schätzer untersucht:

Tabelle 2.5: Bezeichnung der generalisierten Multi-Level-Modelle

	Binomial		Poisson	
	BP	SYNTH	BP	SYNTH
Ungewichtete	BINBP	BINSYNTH	POIBP	POISYNTH
Direkte Designgewichte	BINBPW	BINSYNTHW	POIBPW	POISYNTHW
Reskalierte Designgewichte	BINBPW2	BINSYNTHW2	POIBPW2	POISYNTHW2

2.3.5 MSE-Schätzung

Zur Beurteilung der Genauigkeit von Schätzern wird im Allgemeinen die Varianz des Schätzers angegeben. Prinzipiell interessiert die Verteilung des Schätzers, die nach dem zentralen Grenzwertsatz von Lindeberg-Lévy in der Praxis vereinfacht als normal angenommen wird. Unter dieser Annahme reicht die Varianz des Schätzers, um die Schätzverteilung anzugeben. Daraus lassen sich dann auch unmittelbar Konfidenzintervalle berechnen. Diese Vorgehensweise setzt allerdings die Unverzerrtheit eines Schätzers voraus.

Der HT und GREG (asymptotisch) erfüllen diese Eigenschaft, nicht jedoch synthetische Schätzverfahren. Beim EBLUP muss von einer Design-Verzerrung ausgegangen werden, so dass statt der Varianz des Schätzverfahrens der MSE betrachtet werden muss.

Die Varianzschätzer für den HT und den GREG wurden bereits in (2.3.5) bzw. (2.3.10) angegeben. Beide Varianzschätzer erweisen sich in der Praxis als sehr geeignet. Lediglich in Sonderfällen, bei denen der synthetische Teil des GREG-Schätzers nicht mehr vernachlässigt werden kann, tritt eine Unterschätzung der tatsächlichen Schätzvarianz auf. In diesen Fällen sollten statt Design-basierter Verfahren Small Area-Methoden zur Anwendung kommen. Lediglich bei sehr kleinen Anteilen der Beobachtungsvariablen sind beide Verfahren nicht mehr geeignet, was aus einer nicht zufriedenstellenden Asymptotik resultiert (vgl. Münnich 2008). Diese kann jedoch nur mit Hilfe von Modellannahmen *repariert* werden, welche auch nur dann funktioniert, wenn die Modellannahmen in geeigneter Weise zutreffen.

Im Falle des Modells von Battese, Harter und Fuller bzw. EBLUPA gestaltet sich die MSE-Schätzung wesentlich komplizierter, weil neben der klassischen Randomisierung aufgrund der Stichprobenziehung die Modellschätzung integriert werden muss. Als besonders geeignet hat sich der MSE-Schätzer von Prasad und Rao (siehe Prasad und Rao 1990) erwiesen.

Der MSE des EBLUPA wird in drei Komponenten aufgeteilt (siehe Prasad und Rao 1990 und Datta und Lahiri 2000):

$$\mathrm{MSE}\left(\widehat{\mu}_{d,\mathrm{EBLUPA}}(\widehat{\psi})\right) \approx g_{1d}(\widehat{\psi}) + g_{2d}(\widehat{\psi}) + 2g_{3d}(\widehat{\psi}) \quad . \tag{2.3.26}$$

Dabei ist $\widehat{\psi} = (\widehat{\sigma}_u^2, \widehat{\sigma}_e^2)'$. Die drei Komponenten ergeben sich mit

$$\gamma_d = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2/n_d}$$

zunächst aus

$$g_{1d}(\psi) = (1-\gamma_d)\,\sigma_u^2 = \frac{\sigma_u^2\sigma_e^2}{n_d\sigma_u^2 + \sigma_e^2} \tag{2.3.27}$$

$$g_{2d}(\psi) = (\mu_{X,d} - \gamma_d\bar{x}_d)^{'}\,(X^{'}V^{-1}(\psi)X)^{-1}\,(\mu_{X,d} - \gamma_d\bar{x}_d) \tag{2.3.28}$$

und

$$g_{3d}(\psi) = n_d^{-2}(\sigma_u^2 + n_d^{-1}\sigma_e^2)^{-3}\left[\sigma_e^4 I^{uu} + \sigma_u^4 I^{ee} - 2\sigma_u^2\sigma_e^2 I^{ue}\right] \quad , \tag{2.3.29}$$

wobei

$$\begin{aligned}
I^{uu} &= 2a^{-1}\textstyle\sum_{d=1}^{D}[(n_d-1)\sigma_e^{-4} + w_d^{-2}] \\
I^{ee} &= 2a^{-1}\textstyle\sum_{d=1}^{D} n_d^2 w_d^{-2} \\
I^{ue} &= -2a^{-1}\textstyle\sum_{d=1}^{D} n_d w_d^{-2} \\
a &= \left[\textstyle\sum_{d=1}^{D} n_d^2 w_d^{-2}\right]\left[\textstyle\sum_{d=1}^{D}\{(n_d-1)\sigma_e^{-4} + w_d^{-2}\}\right] - \left(\textstyle\sum_{d=1}^{D} n_d w_d^{-2}\right)^2 \\
w_d &= \sigma_e^2 + n_d\sigma_u^2 \quad .
\end{aligned}$$

Diese drei Komponenten sind nicht berechenbar, da sie vom unbekannten Parameter $\psi = (\sigma_u^2, \sigma_e^2)$ abhängen. Deshalb werden $g_{2d}(\hat{\psi})$ und $g_{3d}(\hat{\psi})$ als Schätzer für g_{2d} and g_{3d} verwandt. Der Schätzer $g_{1d}(\hat{\psi})$ für $g_{1d}(\psi)$ ist verzerrt, weshalb Prasad und Rao (1990, S. 166) einen um den Bias korrigierten Schätzer $g_{1d}(\hat{\psi}) + g_{3d}(\hat{\psi})$ verwenden. Hieraus resultiert der MSE-Schätzer (2.3.26) von Prasad und Rao. Prasad und Rao (1990, S. 166) zeigen, dass die einzelnen Schätzer nur einen Bias von kleiner Ordnung haben und somit akkurat sind.

Alternativ können auch Resampling-Methoden verwendet werden. Die MSE-Schätzung lässt sich prinzipiell auch mit delete one cluster-Jackknife beziehungsweise mit einem *parametrischen Bootstrap* durchführen. Beide Methoden sind in Jiang und Lahiri (2001) beschrieben. Aufgrund des sehr hohen Aufwandes bei deren Berechnung wird auf eine nähere Betrachtung dieser Methoden an dieser Stelle verzichtet.

Wie später ersichtlich wird, kann der MSE-Schätzer (2.3.26) von Prasad und Rao auch beim Schätzer von You und Rao verwendet werden, auch wenn er das Stichprobendesign im Sinne von You und Rao nicht korrekt behandelt. Ein modifizierter MSE-Schätzer für den Schätzer von You und Rao wurde unlängst durch Torabi und Rao (2010) entwickelt. Dieser MSE-Schätzer verwendet jedoch aufwändige Matrizengleichungen, die sich zurzeit noch nicht auf die Daten des Zensus anwenden lassen.

2.4 Chi-Quadrat Verfahren

2.4.1 Einführung in das Dual-System Modell

Wir betrachten eine Gesamtheit U von Personen mit unbekanntem Umfang N_P und nehmen an, dass sie geschlossen ist, also keine Geburt, Todesfall oder Migration stattfindet. Man geht davon aus, dass der Datenerzeugungsprozess durch ein multinomiales Modell beschrieben werden kann, wobei eine Einheit $i \in U$ in eine Liste C oder in eine zweite Liste S, in beide oder in keine eingetragen wird. Die Parameter der Multinomialverteilung seien mit p_K, p_F, p_{00}, p_{11} bezeichnet, d. h. wir unterstellen konstante Wahrscheinlichkeiten für jede Person i, einer der vier Zellen anzugehören. p_A bezeichnet die Wahrscheinlichkeit, für

A = K nur zur Liste C

A = F nur zur Liste S

A = 00 zu beiden Listen C und S

A = 11 zu keiner der beiden Listen C und S

zu gehören.

Nach N_P unabhängigen Wiederholungen gemäß diesem Modell ergeben sich die Zellhäufigkeiten in der 2×2 Kontingenztabelle 2.6.

Tabelle 2.6: Die 2×2 Kontingenztabelle

		Liste S: drin	Liste S: draußen	
Liste C	drin	N_{00}	N_K	N_C
	draußen	N_F	N_{11}	
		N_S		N_P

Beide so erzeugten Listen enthalten Überdeckungsfehler (coverage error). Um diesen coverage error zu schätzen, kann eine Methode Verwendung finden, die auf der *capture-recapture* Methode

basiert und bei der man den Umfang der geschlossenen Population U schätzt. Der bekannteste Schätzer ist der Petersen-Schätzer (siehe Petersen 1896)

$$\hat{N}_P = N_{\mathtt{C}} \frac{N_{\mathtt{S}}}{N_{00}} \quad .$$

Hinter diesem Ansatz steckt die Annahme der capture-recapture Methode, dass die Anteile in $\mathtt{C}$ und $\mathtt{S}$ gleich sind.

Da es aus Kostengründen zu teuer ist, beide Listen vollständig zu erheben, geht man davon aus, dass nur die erste Liste $\mathtt{C}$ aus einem Zensus kommt, aus der zweiten Liste $\mathtt{S}$ dagegen nur eine Stichprobe gezogen wird, mittels derer man $N_{\mathtt{S}}$ und, in Verbindung mit $\mathtt{C}$, auch N_{00} schätzt. N_P kann dann durch

$$\hat{N}_P^{\mathrm{DSE}} = N_{\mathtt{C}} \frac{\hat{N}_{\mathtt{S}}}{\hat{N}_{00}} \quad ,$$

geschätzt werden. Dieser Schätzer wird Dual System-Schätzer (DSE) genannt.

Das hier beschriebene Vorgehen kann auf mehr als zwei Listen erweitert werden und ist zum Beispiel in Zaslavsky (1989) beschrieben.

DSE ist berechenbar, da die Schätzung nicht von N_{11} abhängt, d. h. vollständig die Personen ignoriert, die weder in der ersten Liste noch in der Stichprobe erscheinen. Allerdings verlangt DSE, dass alle Einheiten in U dieselbe „capture“ Wahrscheinlichkeit $p_{00} + p_K$ haben, in die erste Liste aufgenommen zu werden, und dass ebenfalls alle Einheiten in U dieselbe „recapture“ Wahrscheinlichkeit $p_{00} + p_F$ haben, in die Stichprobe zu kommen. Diese Annahme ist nicht realistisch. Eine demografische Analyse zeigt, dass Personen in verschiedenen Alters- und Geschlechtsgruppen unterschiedliche gruppenspezifische Wahrscheinlichkeiten haben (Robinson et al. 1993). Die Abweichung von der Homogenität zwischen den einzelnen Gruppen findet dadurch Berücksichtigung, dass die Population zunächst in homogene Klassen zerlegt wird, in denen mittels DSE einzeln geschätzt wird und am Schluss diese Schätzungen geeignet zu einer Gesamtschätzung zusammengefasst werden.

2.4.1.1 Dual-System Schätzung in den USA und der Schweiz

Im Jahr 2000 wurde in den USA eine spezielle Erhebung durchgeführt, die Accuracy and Coverage Evaluation (A. C. E.), zusätzlich zum Zensus 2000. Dies hatte zum Ziel, eine Überdeckungs-Evaluation vorzunehmen, um N_S/N_{00} mittels DSE zu schätzen. Diese spezielle Erhebung A. C. E. entspricht in der Terminologie von Kapitel 2.4.1 der Liste S – und ist in der US-Literatur besser unter dem Namen „P-sample“ bekannt.

Der DSE hatte in den USA im Jahr 2000 die folgende Gestalt

$$\hat{N}_P^{\mathrm{USA,2000}} = \hat{N}_{\mathtt{C}} \frac{\hat{N}_{\mathtt{S}}}{\hat{N}_{00}}.$$

Eine tiefergehende Modifikation dieses Verfahrens ist bei Hogan (2003) zu finden.

Diese in den USA entwickelte Schätzmethode war die Grundlage für eine Überdeckungsevaluation des Schweizer Census 2000. Die Schätzung der Fehlbestände (undercount) baute auf einer P-Stichprobe auf, die einige Monate nach dem Zensus gezogen wurde. Die Karteileichen (overcount) wurden mittels einer Stichprobe ermittelt, die aus der Zensus-Datenmenge gezogen wurde. Beide Stichproben waren Teil des Swiss Coverage Survey (SCS), der am 5. Dezember 2000 erhoben wurde.

Mittels dieser in Renaud (2004) beschriebenen Methode wurde eine fehlerhafte Zählung von 0,4 % der Bevölkerung ermittelt (die geschätzte Karteileichenrate) und andererseits konstatiert, dass der Zensus 1,6 % der Population nicht enthielt (die geschätzte Fehlbestandsrate). Beide Raten wurden im DSE kombiniert und führten zu einer geschätzten Netto-Fehlbestandsrate von 1,4 % für den Schweizer Zensus 2000. Alle numerischen Ergebnisse, auch die für Untergruppen, finden sich in Renaud (2004; 2007). Die Stichprobenauswahl findet man im Detail in Renaud (2001).

2.4.1.2 Dual-System Schätzung in Deutschland

Beim Register-gestützten Zensus 2011 in Deutschland hat man die Register aus den Einwohnermeldeämtern der Gemeinden zur Verfügung. Wir bezeichnen mit τ_R die bekannte Anzahl der registrierten Einwohner einer Gemeinde. Das Register R enthält unter anderem eine unbekannte Anzahl τ_K von Personen (Karteileichen), die nicht oder nicht mehr in der Gemeinde leben. Auf der anderen Seite gibt es eine unbekannte Zahl τ_F von Personen (Fehlbestände), die in der Gemeinde leben, aber nicht im Einwohnermeldeamt registriert sind.

Zusätzlich zu den Melderegistern ist auch ein vollständiges Anschriften- und Gebäuderegister (AGR) verfügbar. Es sei N die bekannte und vollständige Anzahl aller Anschriften. N_P bezeichnet die Summe aller Personen in U, die an diesen Anschriften wohnen oder die in den Melderegistern verzeichnet sind. Dann gilt $N_P = \tau_K + \tau_{00} + \tau_F = \tau_Z + \tau_K$, mit τ_{00} als der Zahl der Personen, die sowohl registriert sind als auch tatsächlich in der Gemeinde wohnen.

Ein wichtiges Ziel des deutschen Zensus 2011 ist die Schätzung der wahren Bevölkerungszahl $\tau_Z = \tau_{00} + \tau_F$. Wir bezeichnen die Gesamtheit dieser Population als Z. $\tau_R = \tau_{00} + \tau_K$ ist, wie bereits erwähnt, bekannt.

Wir modifizieren das bisherige Vorgehen und wenden die Theorie des DSE auf den Register-gestützten Zensus an, um die wahre Zahl der Bevölkerung τ_Z in einer Gemeinde in Deutschland zu schätzen.

Das Modell

Wie in Abschnitt 2.4.1 bereits ausgeführt wurde, geht man davon aus, dass in der geschlossenen Population U ein Zufallsexperiment mit Wahrscheinlichkeit p_{iF}, p_{i00} und p_{iK}, bestimmt, ob eine Einheit $i \in U$ zu einer der drei disjunkten Kategorien $F, 00$ oder K der Population U gehört. Dabei bezeichne $p_{iF} = P(i \notin R, i \in Z)$, $p_{i00} = P(i \in R, i \in Z)$ und $p_{iK} = P(i \in R, i \notin Z)$. Für die Wahrscheinlichkeiten $p_{iZ} = P(i \in Z)$ und $p_{iR} = P(i \in R)$ erhalten wir

$$p_{iZ} = p_{i00} + p_{iF}, \qquad p_{iR} = p_{i00} + p_{iK}. \tag{2.4.1}$$

Tabelle 2.7 drückt die Beziehung zwischen den Kategorien K, 00, F und ihren Häufigkeiten τ_K, τ_{00}, τ_F nach N_P unabhängigen Wiederholungen in der Population U aus. Man beachte, dass das Register R die Rolle des Zensus beim klassischen DSE in den USA entspricht.

Tabelle 2.7: Die drei Kategorien mit ihren Häufigkeiten

Register R				
	drin	K	τ_K	$\}\tau_R$
		00	τ_{00}	$\}\tau_Z$
	draußen	F	τ_F	
			N_P	

Unter anderem waren die Reduktion des Befragungsaufwandes für den Zensus aber auch eine Reduktion der Belastung für die Bürger Gründe für einen Register-gestützten Zensus. Nur für einen Teil der Bürger wird eine direkte Befragung notwendig. Die dazu gezogene Stichprobe S kommt aus dem Anschriftenregister einer Gemeinde. Alle an einer ausgewählten Anschrift wohnenden Personen werden in die Befragung einbezogen. Bezüglich s, mit n_P Personen, die in n Anschriften tatsächlich wohnen oder dort registriert sind, liefert das Zufallsexperiment sechs mögliche Ergebnisse mit den Wahrscheinlichkeiten $p_{S,iK}$, $p_{S,i00}$, $p_{S,iF}$, $p_{\bar{S},iK}$, $p_{\bar{S},i00}$ und $p_{\bar{S},iF}$, $\forall i \in U$. Dabei ist $\bar{S}$ die komplementäre Menge von S in U. Wir unterstellen wieder $\forall i \in U$ homogene Wahrscheinlichkeiten $p_{iK} = p_K$, $p_{i00} = p_{00}$, $p_{iF} = p_F$. Tabelle 2.8 zeigt die Häufigkeiten dieses Zufallsexperiments, wobei $i \in \bar{S}$ bedeutet, dass Person i nicht in einer Anschrift der Stichprobe S wohnt. Man beachte, dass die letzte Spalte einer Zeile nicht die Summe der beiden vorherigen Spalten ist sondern Schätzungen für die Populationsumfänge für die drei Kategorien K, 00, und F.

Tabelle 2.8: 3×2 Häufigkeitstabelle

		Stichprobe s		
Register R	drin	$\tau_{K,S}$	$\tau_{K,\bar{S}}$	$\hat{\tau}_K$
		$\tau_{00,S}$	$\tau_{00,\bar{S}}$	$\hat{\tau}_{00}$
	draußen	$\tau_{F,S}$	$\tau_{F,\bar{S}}$	$\hat{\tau}_F$
		n_P	$N_P - n_P$	$\hat{N}_P$

Die Stichprobe S mit n_P Personen aus U wird im Durchschnitt aus $\frac{n_P}{N_P} \times 100\% = \theta \times 100\%$ der registrierten Personen bestehen, da alle Personen in den ausgewählten Anschriften in die Untersuchung einbezogen werden. Wir nehmen an, dass alle Personen in der Stichprobe S in eine der drei Kategorien $\{K, 00, F\}$ fehlerfrei eingeordnet werden können. Der Maximum Likelihood-Schätzer von p_j für dieses Modell ist dann gegeben durch $\hat{p}_j = \frac{\tau_{j,S}}{n_P} = \frac{\tau_j}{N_P}, j \in \{K, 00, F\}$.

Wie bereits erwähnt, ist die Zahl τ_R der registrierten Personen bekannt und der wahre Populationsumfang τ_Z ist zu schätzen. Wegen (2.4.1) schätzt DSE τ_Z in einer Gemeinde in Deutschland durch

$$\hat{\tau}_Z^{\text{DSE-GER}} = \tau_R \frac{\tau_{S,00} + \tau_{S,F}}{\tau_{S,00} + \tau_{S,K}} = \tau_R \frac{\tau_{S,Z}}{\tau_{S,R}} \quad , \tag{2.4.2}$$

d. h. $\hat{\tau}_Z^{\text{DSE-GER}}$ ist ein Verhältnisschätzer.

Da die Personen in den Anschriften geklumpt sind, haben wir es auf Personenebene mit einer Klumpenstichprobe zu tun. Sind die Anschriften geschichtet und/oder wählen wir Anschriften und damit die dazu gehörenden Personen mit unterschiedlichen Wahrscheinlichkeiten aus, wird eine Anschrift i in der Stichprobe S mit einem Designgewicht w_i versehen, das der Inversen der Inklusionswahrscheinlichkeit der ausgewählten Anschrift entspricht. Der Dual System-Schätzer ist dann

$$\hat{\tau}_Z^{\text{DSE-GER}} = \tau_R \frac{\sum_{i \in S} w_i \tau_{i,Z}}{\sum_{i \in S} w_i \tau_{i,R}} = \tau_R \frac{\hat{\tau}_Z}{\hat{\tau}_R}. \tag{2.4.3}$$

Man beachte, dass der DSE in Deutschland nur drei Kategorien hat im Gegensatz zu vier beim klassischen DSE-Ansatz. Wir gehen also davon aus, dass es keine Personen gibt, die nicht im Register R verzeichnet sind und auch nicht in der Gemeinde wohnen.

Man beachte weiter, dass man im Register-gestützten Zensus in Deutschland nicht an der Schätzung von τ_{00} sondern von τ_Z interessiert ist. Diese Kennzahl wird mittels der Stichprobe S geschätzt, die dem P-sample beim klassischen DSE in den USA entspricht. Allerdings wird kein E-Sample wie in den USA gezogen, d. h. es wird keine zusätzliche Stichprobe aus dem Register R in Deutschland gezogen.

Der Chapman-Schätzer

Um den Homogenitätsannahmen des Modells innerhalb von Klassen zu genügen, werden homogene *Klassen* $k \in \{1, \ldots, K\}$ gebildet und der DSE darin angewendet. Es kommt allerdings vor, dass es in der Stichprobe S keine Person der Klasse k im Register R gibt, d. h. $\tau_{S,R,k} = 0$ ist, oder niemand aus der Klasse k an der ausgewählten Anschrift lebt, d. h. $\tau_{S,Z,k} = 0$ ist. Im ersten Fall können wir keine Schätzung in (2.4.3) berechnen, da im Nenner eine 0 steht. Im zweiten Fall ist nicht klar, ob eine Stichprobennull oder eine strukturelle Null vorliegt. Unter einer strukturellen Null versteht man den Wert in einer Zelle, die in der Population keine Beobachtung enthält und der daher der Wert 0 zugeordnet wird. Der Wert einer Zelle wird als Stichprobennull bezeichnet, wenn er keine strukturelle Null ist, jedoch die Zelle in der Stichprobe keine Beobachtung enthält. Auch einer solchen Zelle wird der Wert 0 zugeordnet.

Wegen dieser Beschränkung des DSE betrachtete Chapman (1951) einen modifizierten Schätzer. Diese Modifikation besteht darin, eine 1 zu $\tau_{S,00}$, $\tau^*_{S,00} = \tau_{S,00} + 1$ zu addieren, was auch eine Änderung der Randhäufigkeiten $\tau^*_{00} = \tau_{00} + 1$, $n^*_P = n_P + 1$, $N^*_P = N_P + 1$ zur Folge hat und daher $\theta^* = \dfrac{n_P + 1}{N_P + 1} = \dfrac{n^*_P}{N^*_P}$ ist.

Der Chapman-Schätzer für τ_Z in einer Klasse k einer Gemeinde in Deutschland ist gegeben durch

$$\hat{\tau}^{\text{CHAP-GER}}_{Z,k} = (\tau_{R,k} + 1)\,\frac{\hat{\tau}_{Z,k} + 1}{\hat{\tau}_{R,k} + 1} - 1 \tag{2.4.4}$$

und wird für alle $k \in \{1, \ldots, K\}$ berechnet. Weder Zähler noch Nenner können 0 sein. Damit vermeidet man zweifelhafte Ergebnisse in den Situationen, die zuvor erwähnt wurden.

2.4.2 Strukturerhaltende Schätzer

Neben den Klassenumfängen beim DSE, ist man häufig an der Schätzung von Anzahlen an Personen in sogenannten *Domains* interessiert, zum Beispiel an 18-24 Jahre alten Frauen. Wir bezeichnen solche Teilgesamtheiten als Domains d, die einen leeren oder nicht-leeren Durchschnitt mit der Klasse h haben können. Man kann einen strukturerhaltenden Schätzer (Structure PREserving Estimator, SPREE) (Purcell und Kish 1980, Rao 2003, S. 53) verwenden, um den Umfang der Domain d zu schätzen. Dadurch modifiert man die Schätzungen für die Domains in Übereinstimmung mit anderen bekannten Anzahlen für diese Domains, so dass diese neuen Anzahlen aufsummiert die Anzahl in der Klasse h schätzen. In der Simulationsstudie werden die Registerzahlen als bekannte Werte verwendet.

2.4.2.1 SPREE

Eine Möglichkeit, einen strukturerhaltenden Schätzer für $\tau_{Z,d}$ in einer Gemeinde zu definieren, ist gegeben durch

$$\hat{\tau}_{Z,d}^{\text{SPREE}} = \sum_{k=1}^{K} \frac{\tau_{R,dk}}{\tau_{R,k}} \hat{\tau}_{Z,k}, \tag{2.4.5}$$

wobei $\tau_{R,dk}$ der Durchschnitt der Register-Datenmenge R in Domain d mit der Register-Datenmenge R in Klasse k ist. $\hat{\tau}_{Z,k}$ bezeichnet den DSE- oder Chapman-Schätzer in einer Klasse k. Durch den SPREE-Schätzer wird also die Struktur eines Registers auf die Stichprobenwerte der Variablen Z übertragen.

2.4.2.2 Verallgemeinerter strukturerhaltender Schätzer, GSPREE

SPREE könnte nicht nur auf eine Gemeinde sondern auf das gesamte Bundesland angewendet werden. Es ist aber auch möglich, mittels eines loglinearen Modells die Interaktionen zwischen den Gemeinden zu berücksichtigen und die Ergebnisse aus der Zensus-Stichprobe an die Struktur der Registerwerte anzupassen. Zhang und Chambers (2004) entwickelten den verallgemeinerten strukturerhaltenden Schätzer, GSPREE (Generalized Structure PREserving Estimator). Beim GSPREE geht man von einem saturierten Modell der logarithmierten Registerwerte aus, $\log \tau_R = X_1\beta_1 + X_2\beta_2$, um durch diese Reparametrisierung Werte für β_2 zu erhalten. Dabei ergibt sich die Länge des Vektors $\log \tau_R$ als Produkt der Zahl der betrachteten Domains, der Zahl der betrachteten Klassen und der Zahl der betrachteten Gemeinden. Der Vektor β_2 gehört zu dem Teil X_2 der Designmatrix X, der sich auf die Interaktionseffekte bezieht. β_1 ist der Vektor der Haupteffekte und X_1 der entsprechende Teil der Designmatrix X, der sich auf die Haupteffekte bezieht. Da wir ein saturiertes Modell haben, sind β_1 und β_2 bekannt.

Nun werden sogenannte Pseudoschätzungen durch ein verknüpftes loglineares Modell modelliert, bei dem $X_2\beta_2$ aus dem saturierten Modell übernommen wird. In der Simulationsstudie wurde für die Pseudoschätzungen der Chapman-Schätzer $\hat{\tau}_Z^{\text{CHAP}}$ verwendet und mit dem folgenden zweiten loglinearen Modell: $\log \hat{\tau}_Z^{\text{CHAP}} = X_1\beta_3 + \alpha X_2\beta_2$ verknüpft. Der Parameter α modelliert etwaige Zeitveränderungen, die wegen der verschiedenen Erhebungszeitpunkte zwischen dem Register R und dem Zensus Z nicht auszuschließen sind.

Verwendet man $\hat{\beta}_3$ und $\hat{\alpha}$ als Schätzungen aus dem zweiten Modell und β_2 aus dem ersten Modell, lassen sich die Anzahlen in jeder betrachteten Domain, jeder betrachteten Klasse und jeder betrachteten Gemeinde durch GSPREE mittels

$$\log \hat{\tau}_Z^{\text{GSPREE}} = X_1\hat{\beta}_3 + \hat{\alpha}X_2\beta_2 \tag{2.4.6}$$

schätzen. Ist $\alpha = 1$ wird GSPREE auch einfach SPREE genannt. Allerdings wird SPREE in (2.4.5) nur in einer Gemeinde angewendet.

2.4.2.3 Der Chi-Quadrat-Schätzer als Alternative zu GSPREE

Die in den vorangegangenen Abschnitten vorgestellten Ansätze bieten sich an, um *Ziel 1* Variablen strukturerhaltend zu schätzen. Die multidimensionale Kreuzkombination dieser Variablen wird als *Hypercube* bezeichnet. Sollen jedoch unter anderem *Ziel 2* Variablen in den Hypercube mit aufgenommen werden, so eignen sich diese Ansätze nicht mehr, da sie unter anderem auf der Kenntnis von Zellbesetzungen in der Population beruhen. Bei *Ziel 2* Variablen muss davon ausgegangen werden, dass durch die geringe Fallzahl innerhalb der Zellen des Hypercubes präzise Schätzungen

auf NUTS2-Ebene nicht mehr möglich sind. Univariate Schätzungen der Ränder eines Hypercubes sind jedoch mit hinreichender Präzision möglich, etwa durch den GREG-Schätzer. Der Hypercube-Schätzer macht sich diese Eigenschaft zunutze und erfüllt damit eine Kohärenz-Eigenschaft (siehe Statistisches Bundesamt 2006, S. 15 ff.). Hierbei bedient er sich bestimmter Strukturen höherer Ebene (Bundesland) und bildet diese auf NUTS2-Ebene ab.

Um die Entwicklung des (zunächst für den zweidimensionalen Fall definierten) χ^2-*Schätzers* zu verdeutlichen, gehen wir von einer Häufigkeitstabelle M aus und suchen eine andere Häufigkeitstabelle N, die in gewissem Sinne der ursprünglichen Häufigkeitstabelle M möglichst nahe ist. Auf diese Weise transformieren wir die Struktur von M in N. Die Nähe von N zu M wird über die übliche χ^2-Distanz gemessen, die auch in Rao (2003, S. 54) erwähnt wird. Es ist zwar nicht mehr gesichert, dass die odds-ratios gleich bleiben (vgl. Rao 2003, S. 55), dies wird aber mehr als aufgewogen durch die Tatsache, dass exakte analytische Lösungen mit relativ einfachen Varianzschätzungen (siehe Abschnitt Varianzschätzung auf Seite 60) möglich sind.

Gegeben sei also eine $I \times J$ Matrix $M = \left(m_{ij}\right)$ mit $m_{ij} > 0$. Wir suchen eine $I \times J$ Matrix $N^{\chi^2} = \left(n_{ij}\right)$ mit vorgegebenen Rändern, die der ursprünglichen Matrix M am nächsten ist, d. h.

$$\sum_{i,j} \frac{\left(\frac{n_{ij}}{n} - \frac{m_{ij}}{m}\right)^2}{m_{ij}}$$

als Funktion der n_{ij} minimal ist mit $m = \sum_{i,j} m_{ij}$ und $n = \sum_{i,j} n_{ij}$ unter den Nebenbedingungen, dass $n^*_{i\cdot} = \sum_{j=1}^{J} n_{ij}$ und $n^*_{\cdot j} = \sum_{i=1}^{I} n_{ij}$ bekannt und gegeben sind.

Die Lösung der obigen Problemstellung erhält man über den folgenden Lagrange Ansatz:

$$\tilde{f}(\, n_{11}, \ldots, n_{IJ}) = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(\frac{n_{ij}}{n} - \frac{m_{ij}}{m}\right)^2}{m_{ij}} - \sum_{i=1}^{I} \tilde{\lambda}_i \left(\sum_{j=1}^{J} n_{ij} - n^*_{i\cdot}\right) - \sum_{j=1}^{J} \tilde{\mu}_j \left(\sum_{i=1}^{I} n_{ij} - n^*_{\cdot j}\right) \quad .$$

Dies ist äquivalent zu

$$f(n_{11}, \ldots, n_{IJ}) = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{n_{ij}^2}{m_{ij}} - 2\sum_{i=1}^{I} \lambda_i \left(\sum_{j=1}^{J} n_{ij} - n^*_{i\cdot}\right) - 2\sum_{j=1}^{J} \mu_j \left(\sum_{i=1}^{I} n_{ij} - n^*_{\cdot j}\right) \quad .$$

Ableiten und Nullsetzen ergibt

$$n_{ij}^{\chi^2} = \left(\lambda_i + \mu_j\right) m_{ij} \quad .$$

Die Ränder lassen sich daher schreiben als

$$n^*_{i\cdot} = \lambda_i m_{i\cdot} + \sum_{j=1}^{J} \mu_j m_{ij}$$
$$n^*_{\cdot j} = \sum_{i=1}^{I} m_{ij}\lambda_i + \mu_j m_{\cdot j} \quad .$$

In Matrixschreibweise dargestellt ergibt sich

$$A \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_I \\ \mu_1 \\ \vdots \\ \mu_J \end{pmatrix} = \begin{pmatrix} n^*_{1\cdot} \\ \vdots \\ n^*_{I\cdot} \\ n^*_{\cdot 1} \\ \vdots \\ n^*_{\cdot J} \end{pmatrix} \tag{2.4.7}$$

wobei

$$A = \begin{pmatrix} m_{1\cdot} & \cdots & 0 & m_{11} & \cdots & m_{1J} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & m_{I\cdot} & m_{I1} & \cdots & m_{IJ} \\ m_{11} & \cdots & m_{I1} & m_{\cdot 1} & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ m_{1J} & \cdots & m_{IJ} & 0 & \cdots & m_{\cdot J} \end{pmatrix} = \begin{pmatrix} D_{Me} & M \\ M' & D_{e'M} \end{pmatrix} \tag{2.4.8}$$

mit D_{Me} als Diagonalmatrix der Zeilensummen beziehungsweise $D_{e'M}$ als Diagonalmatrix der Spaltensummen von M. Man erhält als Lösung für $(\lambda_1, \ldots, \lambda_I, \mu_1, \ldots, \mu_J)'$

$$\begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_I \\ \mu_1 \\ \vdots \\ \mu_J \end{pmatrix} = A^+ \begin{pmatrix} n^*_{1\cdot} \\ \vdots \\ n^*_{I\cdot} \\ n^*_{\cdot 1} \\ \vdots \\ n^*_{\cdot J} \end{pmatrix} + \left(Id_{I+J} - A^+A\right) \cdot z \tag{2.4.9}$$

wobei A^+ die Moore-Penrose Inverse von A bedeutet und Id_{I+J} die $I+J$ Einheitsmatrix ist.

Als Lösung bestimmen wir die Elemente von N^{χ^2} als

$$n^{\chi^2}_{ij} = \left(\lambda_i + \mu_j\right) m_{ij} \quad .$$

Man beachte, dass z in (2.4.9) ein beliebiger Vektor ist. Wenn alle $m_{i\cdot}$ und $m_{\cdot j}$ positiv sind, ist A vom Rang $I + J - 1$ mit $A \begin{pmatrix} e_I \\ -e_J \end{pmatrix} = 0$ und $Id_{I+J} - A^+A = \frac{1}{I+J} \begin{pmatrix} E_{II} & -E_{IJ} \\ -E_{JI} & E_{JJ} \end{pmatrix}$(siehe Gabler 1990).

e_K ist ein Vektor von Einsen der Länge K mit $K = I$ oder $K = J$ und $E_{ab} = e_a e_b'$. Nach Gabler (1990) kann A^+ ausgedrückt werden durch

$$A^+ = \begin{pmatrix} D_{Me} + \frac{E_{II}}{I+J} & M - \frac{E_{IJ}}{I+J} \\ M' - \frac{E_{JI}}{I+J} & D_{e'M} + \frac{E_{JJ}}{I+J} \end{pmatrix}^{-1} - \frac{1}{I+J} \begin{pmatrix} E_{II} & -E_{IJ} \\ -E_{JI} & E_{JJ} \end{pmatrix} \quad . \quad (2.4.10)$$

Wegen

$$\begin{pmatrix} E_{II} & -E_{IJ} \\ -E_{JI} & E_{JJ} \end{pmatrix} \begin{pmatrix} n_{1\cdot} \\ \vdots \\ n_{I\cdot} \\ n_{\cdot 1} \\ \vdots \\ n_{\cdot J} \end{pmatrix} = 0 \quad \text{und} \quad \begin{pmatrix} E_{II} & -E_{IJ} \\ -E_{JI} & E_{JJ} \end{pmatrix} z = \begin{pmatrix} e_I \\ -e_J \end{pmatrix} \zeta$$

mit $\zeta = \left(e_I', -e_J'\right) \cdot z$, kann der zweite Term von (2.4.10) in $\lambda_i + \mu_j$ unberücksichtigt bleiben.

Verallgemeinerung auf *k* Variablen

Der vorgestellte Ansatz ist sehr flexibel und kann leicht auf den Fall von k Variablen verallgemeinert werden. Dann ist folgender Ausdruck zu minimieren

$$\sum_{i,j,\cdots,k} \frac{\left(\frac{n_{ij\cdots k}}{n} - \frac{m_{ij\cdots k}}{m}\right)^2}{m_{ij\cdots k}}$$

und zwar unter den Nebenbedingungen

$$n^*_{i\cdots} = \sum_{\text{nicht } i} n_{j\cdots k}, \quad n^*_{\cdots j\cdots} = \sum_{\text{nicht } j} n_{i\cdots k}, \quad \cdots, \quad n^*_{\cdots k} = \sum_{\text{nicht } k} n_{ij\cdots} \quad .$$

In Analogie zu oben ergibt der Lagrange Ansatz

$$n_{ij\cdots k} = m_{ij\cdots k} \left(\lambda_i + \mu_j + \cdots + \nu_k\right)$$

mit

$$\begin{pmatrix} D_1 & M_{12} & \cdots & M_{1k} \\ M_{21} & D_2 & \cdots & M_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ M_{k1} & M_{k2} & \cdots & D_k \end{pmatrix} \begin{pmatrix} \underline{\lambda} \\ \underline{\mu} \\ \vdots \\ \underline{\nu} \end{pmatrix} = \begin{pmatrix} n^*_{i\cdots} \\ n^*_{j\cdot\cdot} \\ \vdots \\ n^*_{\cdots k} \end{pmatrix} \quad , \quad (2.4.11)$$

wobei M_{ab} die zweidimensionale Häufigkeitstabelle der a-ten und b-ten Variablen basierend auf den m-Werten bezeichnet und $\underline{\lambda} = (\lambda_1, \ldots, \lambda_i, \ldots, \lambda_I)'$, $\underline{\mu} = (\mu_1, \ldots, \mu_j, \ldots, \mu_J)'$, $\cdots$,

$\underline{\nu} = (\nu_1, \ldots, \nu_k, \ldots, \nu_K)'$ definiert ist. D_ℓ bezeichnet die univariaten Häufigkeiten der m-Werte der ℓ-ten Variablen als Diagonalmatrix.

Varianzschätzung

Es seien $\left(\hat{T}_1, \ldots, \hat{T}_I\right)$ unabhängige Punktschätzer für $(T_1, \ldots, T_I)$ mit Varianzen $(V_1, \ldots, V_I)$ und Varianzschätzern $\left(\hat{V}_1, \ldots, \hat{V}_I\right)$. Teilt man die I Punktschätzer über SPREE auf die J Domains auf, erhält man

$$\begin{array}{cccc|c} p_{11}\hat{T}_1 & p_{12}\hat{T}_1 & \cdots & p_{1J}\hat{T}_1 & \hat{T}_1 \\ p_{21}\hat{T}_2 & p_{22}\hat{T}_2 & \cdots & p_{2J}\hat{T}_2 & \hat{T}_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ p_{I1}\hat{T}_I & p_{I2}\hat{T}_I & \cdots & p_{IJ}\hat{T}_I & \hat{T}_I \end{array}$$

Die Koeffizienten p_{ij} werden als bekannt vorausgesetzt $\left(\text{z.B. } p_{ij} = {}^{m_{ij}}/_{m_{i\cdot}}, \text{ siehe Gleichung (2.4.5)}\right)$. Dieser Fall liegt bei *Ziel 1* Fragestellungen vor (siehe Abbildung 3.28) und wird in Abschnitt 3.5.2 genauer dargestellt. Daher können die Spaltensummen berechnet werden als

$$\begin{array}{ccccc|c} & & & & & \hat{T}_1 \\ & & & & & \hat{T}_2 \\ & & & & & \vdots \\ & & & & & \hat{T}_I \\ \hline \sum_{i=1}^{I} p_{i1}\hat{T}_i & \sum_{i=1}^{I} p_{i2}\hat{T}_i & \cdots & \sum_{i=1}^{I} p_{iJ}\hat{T}_i & & \end{array}$$

Wir wenden den χ^2-Ansatz an. λ_i und μ_j sind lineare Funktionen der Ränder und daher gegeben durch

$$\begin{aligned} \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_I \\ \mu_1 \\ \vdots \\ \mu_J \end{pmatrix} &= \begin{pmatrix} D_{Me} + \frac{E_{II}}{I+J} & M - \frac{E_{IJ}}{I+J} \\ M' - \frac{E_{JI}}{I+J} & D_{e'M} + \frac{E_{JJ}}{I+J} \end{pmatrix}^{-1} \begin{pmatrix} \hat{T}_1 \\ \vdots \\ \hat{T}_I \\ \sum_{i=1}^{I} p_{i1}\hat{T}_i \\ \vdots \\ \sum_{i=1}^{I} p_{iJ}\hat{T}_i \end{pmatrix} \\ &= \begin{pmatrix} D_{Me} + \frac{E_{II}}{I+J} & M - \frac{E_{IJ}}{I+J} \\ M' - \frac{E_{JI}}{I+J} & D_{e'M} + \frac{E_{JJ}}{I+J} \end{pmatrix}^{-1} \begin{pmatrix} Id_I \\ P' \end{pmatrix} \begin{pmatrix} \hat{T}_1 \\ \vdots \\ \hat{T}_I \end{pmatrix} \\ &= G \begin{pmatrix} \hat{T}_1 \\ \vdots \\ \hat{T}_I \end{pmatrix} \end{aligned} \tag{2.4.12}$$

mit der $I \times J$ stochastischen Matrix $P = \left(p_{ij}\right)$ und $Pe_J = e_I$.

In Analogie zu oben erhalten wir

$$n_{ij}^{\chi^2} = \left(\lambda_i + \mu_j\right) \cdot m_{ij} \qquad .$$

Wegen der Unabhängigkeit der Punktschätzer $\hat{T}_1, \ldots, \hat{T}_I$ ist

$$var \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_I \\ \mu_1 \\ \vdots \\ \mu_J \end{pmatrix} = var \left[G \begin{pmatrix} \hat{T}_1 \\ \vdots \\ \hat{T}_I \end{pmatrix} \right] = G \cdot var \begin{pmatrix} \hat{T}_1 \\ \vdots \\ \hat{T}_I \end{pmatrix} \cdot G' = G \cdot D_V \cdot G' \qquad , \quad (2.4.13)$$

und daher

$$\begin{aligned} var\left(n_{ij}^{\chi^2}\right) &= m_{ij}^2 var\left(\lambda_i + \mu_j\right) \\ &= m_{ij}^2 \left(var\left(\lambda_i\right) + var\left(\mu_j\right) + 2 \cdot cov\left(\lambda_i, \mu_j\right)\right) \qquad . \end{aligned}$$

Ein Varianzschätzer ist daher gegeben durch

$$\begin{aligned} \widehat{var}\left(n_{ij}^{\chi^2}\right) &= m_{ij}^2 \widehat{var}\left(\lambda_i + \mu_j\right) \\ &= m_{ij}^2 \left(\widehat{var}\left(\lambda_i\right) + \widehat{var}\left(\mu_j\right) + 2 \cdot \widehat{cov}\left(\lambda_i, \mu_j\right)\right) \qquad . \end{aligned} \quad (2.4.14)$$

Bemerkung 1. Sind die Schätzer $\hat{T}_1, \ldots, \hat{T}_I$ nicht unabhängig, kann die Formel leicht durch den Einschluss von Kovarianzen in den Außerdiagonalelementen von $D_{\hat{V}}$ verallgemeinert werden. Schätzungen für die Kovarianzen sind etwa mittels Jackknife möglich.
Bemerkung 2. Wegen der Nichtnegativität der Matrix in (2.4.13), ist auch der Schätzer in (2.4.14) nichtnegativ, da er aus einer Teilmatrix von (2.4.13) um die Hauptdiagonale gebildet wird.
Bemerkung 3. Im Kapitel 5 wird gezeigt, dass $n_{ij}^{\chi^2} = p_{ij}\hat{T}_i$ falls $p_{ij} = \dfrac{m_{ij}}{m_{i\cdot}}$, d. h. die χ^2-Lösung ist mit der DSE/SPREE Lösung identisch. Die Anpassung des DSE/SPREE auf M ändert den Schätzer nicht.

Allgemeiner haben wir in (2.4.11)

$$\begin{pmatrix} D_1 & M_{12} & \cdots & M_{1k} \\ M_{21} & D_2 & \cdots & M_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ M_{k1} & M_{k2} & \cdots & D_k \end{pmatrix} \begin{pmatrix} \underline{\lambda} \\ \underline{\mu} \\ \vdots \\ \underline{\nu} \end{pmatrix} = \begin{pmatrix} P_1' \\ P_2' \\ \vdots \\ P_k' \end{pmatrix} \begin{pmatrix} \hat{T}_1 \\ \vdots \\ \hat{T}_I \end{pmatrix} \qquad ,$$

wobei M_{ab} die zweidimensionale Häufigkeitstabelle der a-ten und b-ten Variablen basierend auf den m-Werten ist. D_ℓ bezeichnet wieder den Rand der m-Werte der ℓ-ten Variablen als Diagonalmatrix. P_i sind $I \times J_i$ stochastische Matrizen, die die Allokation von $\hat{T}_1, \ldots, \hat{T}_I$ auf die J_i Werte der i-ten Variablen ($I = 2, \ldots, k$) beschreiben. Speziell gilt $P_1' = Id_I$.

Als Beispiel sei der Vektor

$$\begin{pmatrix} P_1' \\ P_2' \\ \vdots \\ P_k' \end{pmatrix}$$

gegeben durch

$$\begin{pmatrix} P_1' \\ P_2' \\ \vdots \\ P_k' \end{pmatrix} = \begin{pmatrix} D_1 \\ M_{21} \\ \vdots \\ M_{k1} \end{pmatrix} \cdot D_1^{-1} \quad .$$

MSE Schätzung

Wir nehmen zunächst das folgende (SPREE) Modell an:

$$n_{ij} = p_{ij} T_i \tag{2.4.15}$$

wobei $p_{ij} = \frac{m_{ij}}{m_{i\cdot}}$ mit $m_{i\cdot} = \sum_{j=1}^{J} m_{ij}$. In dem betrachteten Fall sei $m_{ij} = \tau_{R,ij}$, also die Anzahl der registrierten Personen in der Hypercube-Zelle ij. Der χ^2-Schätzer für n_{ij} ist

$$n_{ij}^{\chi^2} = p_{ij} \hat{T}_i \quad ,$$

wobei $\hat{T}_i = \frac{\hat{\tau}_{Z,i}}{\hat{\tau}_{R,i}} \tau_{R,i}$ ist, also der Verhältnisschätzer mit der Anzahl der registrierten Personen als Hilfsvariablen.

Die Frage ist, ob sich der mittlere quadratische Fehler von $n_{ij}^{\chi^2}$ schätzen lässt. Im Folgenden wird gezeigt, dass sich zumindest eine Abschätzung angeben lässt, die umso besser ist, je mehr das Modell, d. h. die Assoziationsstruktur, der Realität entspricht.

Wir betrachten den Schätzer

$$\widehat{var}\left(n_{ij}^{\chi^2}\right) + \left(n_{ij}^{\chi^2} - \widehat{n}_{ij}\right)^2 \quad , \tag{2.4.16}$$

wobei $\widehat{n}_{ij}$ ein (asymptotisch) erwartungstreuer Schätzer für n_{ij} ist, zum Beispiel der HT- oder der GREG-Schätzer. Allerdings haben beide Schätzer für kleine Stichprobenumfänge eine hohe Varianz, weshalb sie nicht direkt als Schätzer für die Zellen des Hypercubes Verwendung finden, sondern nur in die Abschätzung des MSE des χ^2-Schätzers eingehen.

Genauer gilt, falls $\widehat{n}_{ij} \geq 1, \forall i, j$

$$\begin{aligned} E\left[\widehat{var}\left(n_{ij}^{\chi^2}\right) + \left(n_{ij}^{\chi^2} - \widehat{n}_{ij}\right)^2\right] &= var\left(n_{ij}^{\chi^2}\right) + E\left(n_{ij}^{\chi^2} - \widehat{n}_{ij}\right)^2 \\ &\geq var\left(n_{ij}^{\chi^2}\right) + \left(E\left(n_{ij}^{\chi^2}\right) - n_{ij}\right)^2 \\ &= E\left(n_{ij}^{\chi^2} - n_{ij}\right)^2 \\ &= \text{MSE}\left(n_{ij}^{\chi^2}\right) \quad , \end{aligned} \tag{2.4.17}$$

d. h. wir haben mit (2.4.16) einen konservativen Schätzer, der im Erwartungswert den MSE überschätzt. Im Falle $n_{ij}^{\chi^2} = p_{ij}\hat{T}_i$ mit $p_{ij} = \frac{m_{ij}}{m_{i\cdot}}$ ist der MSE umso näher an der Varianz, je weniger sich $p_{ij} = \frac{m_{ij}}{m_{i\cdot}}$ von $\frac{n_{ij}}{n_{i\cdot}}$ unterscheidet, d. h. die angenommene Assoziationsstruktur stimmt.

Es bleibt anzumerken, dass der χ^2-Schätzer Ähnlichkeiten zum Kalibirierungsansatz aufweist. Diese herauszuarbeiten wird Gegenstand weiterer Forschungen sein.

2.4.2.4 Der GREG-kombinierte Chi-Quadrat-Schätzer

Beim χ^2-Schätzer wird eine Struktur einer eventuell höheren Ebene auf eine tiefere Ebene transferiert, z. B. vom Bundesland auf die NUTS2-Ebene. Dies erscheint in den Fällen sinnvoll zu sein, in denen direkte Zell-Schätzungen aufgrund geringer Fallzahlen in der Stichprobe zu ungenau sind. Geringer bedeutet nach Drew et al. (1982) und Ghosh und Rao (1994) bei uneingeschränkter Zufallsauswahl, dass die Anzahl der registrierten Personen $n_{R,i}$ in der Stichprobe der Hypercube-Zelle $i (i = 1, \ldots, I)$ kleiner als die erwartete Anzahl der registrierten Personen ist, also

$$n_{R,i} < n_R \frac{N_{R,i}}{N_R} \quad ,$$

wobei $N_{R,i}$ die Anzahl der registrierten Personen in der Gesamtheit in Hypercube-Zelle i bezeichnet. $n_R = \sum_{i=1}^{I} n_{R,i}$ und $N_R = \sum_{i=1}^{I} N_{R,i}$ sind die Anzahl der registrierten Personen in der Stichprobe und in der Gesamtheit der Hypercube-Zelle i.

Im Weiteren wird ein kombinierter Schätzer vorgestellt, der sich aus einer Konvexkombination des GREG- und des χ^2-Schätzers ergibt. Die Konvexkombination ergibt sich in Anlehnung an Rao (2003, 60) wie folgt zu

$$\hat{\tau}_Z = \phi \cdot \hat{\tau}^{\text{GREG}} + (1 - \phi) \cdot \hat{\tau}^{\chi^2}$$

wobei

$$\phi = \begin{cases} \left(\frac{\hat{\tau}_R^{\text{HT}}}{\tau_R}\right)^2 & \text{falls } \frac{\hat{\tau}_R^{\text{HT}}}{\tau_R} < 1 \\ 1 & \text{sonst} \end{cases} .$$

Natürlich lassen sich auch andere Konvexkombinationen aufstellen, die gegebenenfalls vom absoluten Stichprobenumfang in den Zellen abhängen.

Schätzungen für den MSE des χ^2- und des kombinierten Schätzers liegen bisher nur auf Basis von Simulationen vor und sind Gegenstand zukünftiger Forschungsarbeiten. Auch die weiteren DSE-Schätzer sowie die strukturerhaltenden Schätzer (SPREE und GSPREE) werden im Folgenden nicht weiter betrachtet, da sie hinsichtlich ihrer Implementation für die im Zensus relevanten Fragestellungen noch nicht hinreichend erforscht sind. Erste Ergebnisse und Simulations-Resultate sind in Dostal (2012) zu finden.

3 Aufbau und Ergebnisse der Simulationsstudie

3.1 Aufbau der Simulationsstudie

Bei der Stichprobe für den deutschen Zensus handelt es sich um eine Erhebung, bei der die Eigenschaften großer Stichproben im Allgemeinen gültig sind. Auf der anderen Seite sollen aber auch Ergebnisse für Small Areas und/oder Small Domains ausgewiesen werden. In diesem Fall basieren die Schätzungen dann auf einer sehr kleinen Stichprobenbasis.

Um herauszufinden, welches Stichprobendesign und welche Schätzer für den deutschen Zensus geeignet sind, müssen also zum Teil gegenläufige Theorie-Ansätze miteinander verglichen werden. Dieser Vergleich kann im Allgemeinen nicht analytisch durchgeführt werden. Deshalb sind Simulationen das Mittel der Wahl, um zu zeigen, welche Kombination von Stichprobendesign und Schätzern am geeignetsten ist (vgl. Münnich 2008). Um eine solche Simulation zu ermöglichen, wird eine Simulationsgrundgesamtheit benötigt. Diese erlaubt es, ein breites Spektrum an verschiedenen Stichprobendesigns und Schätzmethoden zu entwickeln und zu testen. Im Laufe des gesamten Zensus-Stichprobenforschungsprojekts wurde eine Vielzahl an unterschiedlichen Vorgehensweisen zur Datengenerierung angewendet, die im Folgenden vorgestellt werden (siehe auch Kolb 2012). Teilweise wurden diese Verfahren bereits im Rahmen anderer Projekte, wie dem DACSEIS-Projekt (Münnich und Wiegert 2001, Devroye 1986) oder dem AMELI-Projekt (Alfons, Filzmoser, Hulliger, Kolb, Kraft, Münnich und Templ 2011, Alfons, Burgard, Filzmoser, Hulliger, Kolb, Kraft, Münnich, Schoch und Templ 2011), erfolgreich implementiert.

Zur Erzeugung der Simulationsgesamtheit standen die anonymisierten Melderegister-Daten für Gesamtdeutschland (im Folgenden MR-Daten) zur Verfügung. Da die MR-Daten als deterministischer Block angesehen werden können, dienen die 85.790.381 Einträge als Rahmen der Simulation. Es werden also keine neuen Personen hinzu generiert, sondern lediglich weitere Variablen für diesen schon vorhandenen Block synthetisch generiert.

Neben den Grundvariablen in den MR-Daten (siehe hierzu Tabelle 3.1) und den darin enthaltenen sogenannten Strukturvariablen, welche die Zugehörigkeit zu einer administrativen Einheit beziehungsweise zu einem Postleitzahlenbereich oder einer Anschrift[14] anzeigen (siehe Tabelle 3.2), werden in einem ersten Schritt Informationen über mögliche Karteileichen und Fehlbestände hinzu simuliert. Dazu werden einzelne Personen, verschiedenen Modellen folgend, als nicht im Melderegister erfasste (Fehlbestände) oder fälschlicherweise im Melderegister erfasste Personen (Karteileichen) ausgewiesen (vgl. Kapitel 3.2.3 Münnich et al. 2009 und Burgard und Münnich 2010). Die verwendeten Karteileichen- und Fehlbestandsmodelle basieren auf dem Zensustest 2001 (Burgard und Münnich 2010).[15]

Mit diesem Datensatz ist bereits die Analyse der *Ziel 1*-Fragestellungen möglich. Für *Ziel 2*-Fragestellungen sind jedoch weitere Variablen notwendig. Diese müssen synthetisch zur Simulationsgesamtheit hinzu generiert werden, da sie nicht in den Registern vorhanden sind. Es handelt sich dabei zum Beispiel um Variablen zu Ausbildungs- und Erwerbsprofilen oder zum Bildungsniveau. Für die Generierung dieser synthetischen Daten wird der Mikrozensus als Grundlage herangezogen, um möglichst genaue Informationen über die Verteilung der einzelnen Variablen in der Grundgesamtheit zu erhalten. In Abschnitt 3.2 wird auf das Vorgehen bei der Generierung der verschiedenen synthetischen Variablen eingegangen.

---

[14] Nicht bekannt in einer Anschrift ist die Zugehörigkeit der Personen zu Haushalten.

[15] Für nähere Informationen zum Zensustest 2001 siehe Schäfer (2004).

Erläuterungen zu den untersuchten Stichprobendesigns erfolgten bereits in Abschnitt 2.2. Für jedes Design wurden 1.000 voneinander unabhängige Stichproben gezogen und gespeichert. Die verschiedenen Schätzungen wurden dann für jede einzelne Stichprobe berechnet. Die resultierenden 1.000 Schätzergebnisse eines Schätzers für ein Stichprobendesign sind eine Monte-Carlo-Approximation der tatsächlichen Verteilung des Schätzers für eben dieses Stichprobendesign auf Basis der vorliegenden Daten. Aus dieser Monte-Carlo-Approximation können verschiedene Maße wie der *relative root mean squared error* (RRMSE), der *relative bias* (RBias) oder die *relative dispersion* (RDisp) berechnet werden. Diese Maße können zur Evaluation der Güte des Schätzers bei gegebenem Stichprobendesign herangezogen werden. Die Definition der Maße und die Vorgehensweise bei ihrer Berechnung finden sich in Kapitel 6.

Zusätzlich zu diesem Simulationsgrundgerüst können noch verschiedene Szenarien in die Simulation eingebaut werden. So wurden zum Beispiel die Auswirkungen der verschiedenen Karteileichen- und Fehlbestandsmodelle auf die Schätzungen untersucht. Hierzu wurden die Schätzungen für jedes Karteileichen- und Fehlbestandsmodell separat wiederholt. Auf ähnliche Weise wurden auch andere Szenarien betrachtet. Diese Szenarien sind hauptsächlich in Abschnitt 3.6 dargestellt.

3.2 Die Zensus Simulationsgesamtheit

3.2.1 Vorgehen zur Erzeugung weiterer synthetischer Variablen

Eine wichtige Datenquelle zur Erzeugung der synthetischen Variablen ist der anonymisierte Mikrozensus aus dem Jahr 2006 (MZ06). Er hat 785.681 Einträge und ist damit viel zu klein, um ihn direkt für die Simulation zu verwenden. An den hochgerechneten Mikrozensus-Daten lassen sich aber wichtige Informationen über die Struktur der Daten bis hinunter auf Kreisebene ablesen.

Ziel der synthetischen Datengenerierung war es, eine gewisse Heterogenität in den Daten zu erzeugen. Das heißt, die Lage- und Streuungsmaße sollten sich für die administrativen Einheiten stärker unterscheiden.

Zu den Variablen aus dem Melderegister (Strukturvariablen und demographische Merkmale) und der Information, ob es sich um Karteileichen oder Fehlbestände handelt, werden die synthetischen Variablen blockweise auf Basis von Kreuztabellen hinzu generiert. Es wurde also schrittweise vorgegangen und die bereits erzeugten synthetischen Variablen zur Erzeugung weiterer Variablen verwendet. Dieses Vorgehen wurde gewählt, um die Konsistenz zwischen den erzeugten synthetischen Variablen zu gewährleisten. Die Kreuztabellen basieren auf den mit den Standardhochrechnungsfaktoren (MZ06 EF951) hochgerechneten Kreisergebnissen aus dem Mikrozensus. Damit ist auch die Konsistenz zwischen den hochgerechneten Stichprobenergebnissen aus dem Mikrozensus und den auf Kreisen aggregierten synthetischen Populationswerten gegeben.

Wichtig ist hierbei, dass die Variablen, die aus den MR-Daten gegeben sind, sich auch in der Kreuztabelle wiederfinden lassen. Die Kreuztabelle, aus der Ausprägungen gezogen werden, ist also eine m-dimensionale Kreuztabelle. Die Anzahl der Dimensionen setzt sich aus der Summe der Zahl der MR-Variablen und der Zahl der synthetischen Variablen zusammen. Zu den synthetischen Variablen zählen zum einen die bereits erzeugten Zusatzvariablen, in deren Abhängigkeit eine neue Variable erzeugt wird, und zum anderen die zu erzeugende Variable selber. Es muss also eine multivariate Verteilung $F(x_1, x_2, \ldots, x_k)$ bestimmt werden. Aus den Melderegistern sind Variablen $X_1, \ldots, X_k$ gegeben. Die Verteilungen weiterer Variablen lassen sich durch

$$F(x_{k+1}, \ldots, x_n | x_1, \ldots, x_k) = \frac{F(x_1, \ldots, x_n)}{F(x_1, \ldots, x_k)} \tag{3.2.1}$$

berechnen. Allgemein gilt:

$$F(x_1, \ldots, x_n) = F(x_1) \cdot F(x_2|x_1) \cdot \ldots \cdot F(x_n|x_1, \ldots, x_{n-1}) \tag{3.2.2}$$

Die einzelnen (Blöcke von) Variablen lassen sich demnach rekursiv berechnen. Die Reihenfolge der Simulation der Variablen sollte keine Bedeutung haben, da aus der gleichen gemeinsamen Verteilung gezogen wird.

Eine besondere Rolle spielen die sogenannten Nullzellen. Zu unterscheiden ist zwischen Stichprobennullen, also Merkmalskombinationen, die in der Realität aber nicht in der Stichprobe auftauchen, und strukturellen Nullen, also solchen Merkmalskombinationen, die auch in der Population nicht auftreten können. Um die Basis der Stichprobe zu erhöhen, wurde für solche Merkmalskombinationen, die im Mikrozensus für den jeweiligen Kreis nicht vorkamen, die Kreuztabelle für Regierungsbezirke gebildet und zur Ziehung verwendet. Falls die Merkmalskombination auch hier nicht existierte, wurde wieder eine Ebene höher gezogen, also auf der Ebene der Bundesländer. Wo die Merkmalskombination selbst auf Bundesebene nicht vorhanden war, wurde die Anzahl der Variablen verringert. Bei der Berücksichtigung von immer mehr Variablen wird bei der Erzeugung neuer synthetischer Variablen irgendwann ein Punkt erreicht, an dem eine auffällig hohe Zahl an strukturellen Nullen in den hochdimensionalen Kreuztabellen auftritt. Es werden an dieser Stelle immer die Kreuztabellen aus darüber liegenden administrativen Einheiten herangezogen. Durch dieses Vorgehen werden zur Erzeugung von synthetischen Variablen in verschiedenen Kreisen die gleiche Basisverteilungen verwendet. Die Population ist somit insgesamt homogener, als sie es für den Fall gewesen wäre, in dem unterschiedliche Basisverteilungen verwendet werden. Es gilt demnach einen Trade-off zwischen der Erhaltung von Zusammenhängen und der Einhaltung der Randwerte auf Kreisniveau respektive einer heterogenen Population zu gewährleisten. Dies kann durch die Bildung latenter Klassen (vgl. Linzer und Lewis. 2007) oder durch die Auslassung einzelner Variablen bei der Bildung der Kreuztabellen gewährleistet werden. Bei latenten Klassen werden diese zur Bildung der Kreuztabellen anstatt der einzelnen Variablen zugrunde gelegt. Beide Vorgehensweisen machen aber ein logisches Editing notwendig.

3.2.2 Gelieferte Daten und Struktur der Simulationsgesamtheit

Ausgangspunkt der erzeugten Simulationsgesamtheit waren anonymisierte Melderegisterdaten, die folgende für die Erzeugung der Simulationsgesamtheit wichtigen Variablen enthalten (siehe Tabelle 3.1).

Tabelle 3.1: Melderegistervariablen

Variable	MZ06	Name	Code	Beschreibung
SEX	EF32	Geschlecht	1	männlich
			2	weiblich
AGE	EF44	Alter	metrische Altersangabe - über 95-Jährige eine Klasse	
NAT	-	Staatsangehörigkeit	1	deutsch
			2	nicht deutsch
			8	doppelte Staatsbürgerschaft
FST	EF49	Familienstand	1	ledig
			2	verheiratet
			3	verwitwet
			4	geschieden
			5	eingetragene Lebenspartnerschaft
			6	Familienstatus unbekannt
WST	-	Wohnstatus	0	alleinige Wohnung
			1	Hauptwohnsitz
			2	Nebenwohnsitz
			9	Wohnungsstatus ungeklärt

Die Melderegistervariablen aus Tabelle 3.1 sind zusammen mit den Strukturvariablen in Tabelle 3.2 und der Information über Karteileichen und Fehlbestände die wichtigsten Variablen für die Simulation der *Ziel 1*-Fragestellungen.

Tabelle 3.2: Strukturvariablen

Variable	Name	Ausprägungen
BLA	Bundesland	1:16
ADR	Anschrift	eindeutige Codierung aller Anschriften von 1:18.251.008
ADG	Anschriftengröße	Anzahl der Personen in einer Anschrift von 1:1.770
GEM	Gemeinde	eindeutige Codierung aller Gemeinden von 1:12.243
SMP	Sampling Point	eindeutige Codierung aller Stichprobenbasiseinheiten 1:2.391
SDT	Stadtteile	eindeutige Codierung aller Stadtteile 00:12
KRS	Kreis	eindeutige Codierung aller Kreise von 1:429
AGS	-	Amtlicher Gemeindeschlüssel

Der Simulationsrahmen von 85.790.381 Einträgen umfasst einzelne Personen auch doppelt. Es handelt sich hierbei um Personen, die neben ihrem Hauptwohnsitz noch eine weitere Wohnung angemeldet haben. Informationen über die absoluten und relativen Häufigkeiten sind der Tabelle 3.3 zu entnehmen. Lediglich zur Bestimmung der Schichtungen, die der Stichprobenziehung zugrunde liegen, aber nicht für die folgenden Simulationen, wurden zusätzlich noch die Bewohner von Sonderanschriften berücksichtigt.

Tabelle 3.3: Status der Wohnung

Ausprägungen	1 = alleinige Wohnung	2 = Hauptwohnung	3 = Nebenwohnung
Abs. Häufigkeiten	77.221.140	4.450.282	4.118.959
Rel. Häufigkeiten	90 %	5,2 %	4,8 %

Die wichtigste Strukturvariable ist die Variable SMP (Stichprobenbasiseinheit bzw. Sampling Point), welche die Zugehörigkeit zu einem speziellen Sampling Point anzeigt. Da die Genauigkeitsanforderungen unter Zuhilfenahme der Sampling Points dargestellt wurden (siehe Abschnitt 2.1.2 bzw. 2.1.3), wird später auf diese Variable vermehrt zurückgegriffen, um die Ergebnisse der Simulation zu visualisieren. In Abbildung 3.1 ist der Anteil der Personen mit einer bestimmten Ausprägung an der Gesamtbevölkerung des jeweiligen Sampling Points dargestellt. Die Verteilung dieser Anteile in allen Sampling Points ist besonders gut anhand sogenannter Violinplots sichtbar (siehe Kapitel 6). Die Violinplots geben genauso wie die Boxplots einen Überblick über die Verteilung der Daten. Dabei wird die Dichte der Daten in vertikaler Richtung abgetragen. Gegenüber den Boxplots lassen sich anhand der Violinplots auch Aussagen treffen, ob es sich beispielsweise um eine bimodale Verteilung der Daten handelt.

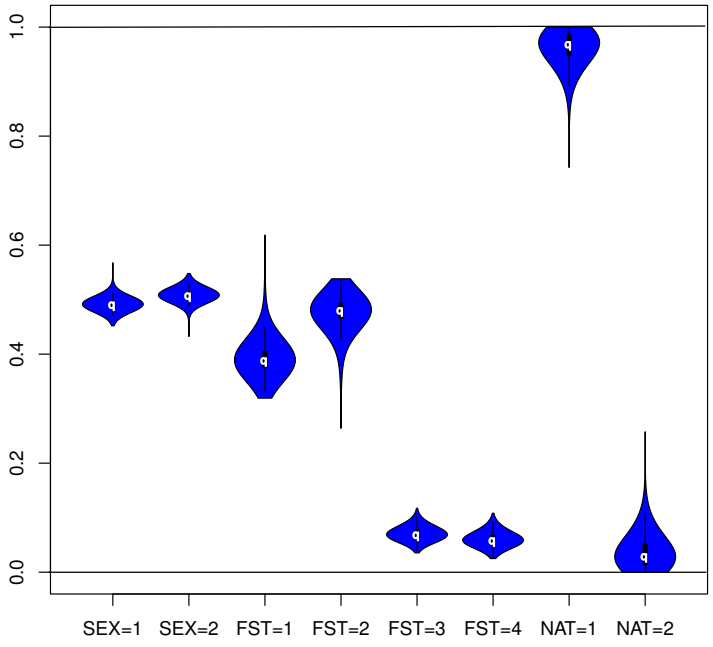


Abbildung 3.1: Verteilung der Anteile von Ausprägungen der Melderegistervariablen über die Sampling Points

Es ist zu sehen, dass der Anteil der männlichen Personen (SEX=1) in allen Sampling Points ungefähr bei 50 % liegt. Demgegenüber ist der Anteil der Personen mit Familienstand geschieden (FST=4) relativ gering (6 %) und der Anteil der Personen deutscher Staatsangehörigkeit (NAT=1) mit durchschnittlich 96 % relativ hoch. Die erste Figur ganz links zeigt die Anteile der männlichen Personen innerhalb von den Sampling Points. Der oberste Punkt liegt knapp oberhalb von 0,5 und der unterste Punkt der Figur liegt knapp unterhalb von 0,5. Das bedeutet, dass die Anteile zwischen den Sampling Points nicht so stark variieren. Anders ist es bei den ersten beiden Ausprägungen des Familienstands (FST=1 und FST=2) und den Ausprägungen deutsche (NAT=1) beziehungsweise nicht deutsche Staatsangehörigkeit (NAT=2). Hier ist die Variation deutlich höher. Später wird anhand von Violinplots für die erzeugten synthetischen Variablen gezeigt, dass diese Variation noch größer sein kann. Bezüglich der Ausprägung Personen mit einer anderen als der deutschen Staatsangehörigkeit (NAT=2) muss festgestellt werden, dass in vier Sampling Points keine Angehörigen dieser Personengruppe wohnhaft sind. In dem Sampling Point mit dem höchsten Anteil dieser Personengruppe liegt der Anteil der Personen mit nicht deutscher Staatsangehörigkeit bei 25,8 %.

An dieser Stelle wird der Simulationsdatenbestand um synthetische Variablen erweitert. Basis für die Erzeugung der synthetischen Variablen sind die hochgerechneten Werte aus dem Mikrozensus 2006. Um die MR-Daten mit den Daten des Mikrozensus 2006 zu verbinden, wurde die Zahl der Ausprägungen angeglichen. Beispielsweise wurden die Ausprägungen der Variable an die Ausprägungen der Variable EF49 im Mikrozensus angepasst.

Wichtig ist es, Anforderungen an die zu erzeugende synthetische Gesamtheit zu stellen, die nach Erzeugung überprüft werden können. Die wichtigste Anforderung ist, dass die synthetische Gesamtheit möglichst realitätsnah sein sollte. Zur Beurteilung steht hier wiederum der Mikrozensus zur Verfügung. In früheren Simulationen hat sich zudem gezeigt, dass eventuell vorhandene Heterogenitäten von besonderem Interesse für die Schätzung sind. Dies ist ein weiteres Anliegen bei der Erzeugung der synthetischen Variablen.

3.2.3 Karteileichen und Fehlbestände

Ein wichtiges Ziel des Zensus-Stichprobenforschungsprojekts ist es, Schätzmethoden zur Schätzung der tatsächlichen Bevölkerungszahl vorzuschlagen. Um dies zu überprüfen, wird der synthetischen Gesamtheit die Information beigefügt, ob es sich bei der Person um eine Karteileiche oder einen Fehlbestand handelt. Diese Information basiert auf den Ergebnissen von Logit-Modellen. Der erste Ansatz bei der Modellierung von Karteileichen (KAL) und Fehlbeständen (FEB) war es, eine Vorhersage auf klassischen Logit-Modellierungen zu treffen. Dabei wird jeweils ein Vektor für Karteileichen und Fehlbestände erzeugt. Im KAL-Vektor werden Karteileichen mit einer 1 gekennzeichnet, alle anderen Elemente haben den Eintrag 0. Analog gilt dies auch für den Fehlbestandsvektor. Die Korrelation zwischen der Zahl der registrierten Personen und der Zahl der Personen, die tatsächlich in der Anschrift wohnen, wurde bei diesen Modellen zunächst mit 0,993 veranschlagt.[16]

Allerdings liefert dieser Ansatz sehr homogene Ergebnisse, die in dieser Form nicht der Realität entsprechen. Demgegenüber resultieren aus einer Multi-Level Logit-Modellierungen realistischere KAL/FEB-Vektoren sowohl in Bezug auf unterschiedliche Anteile von Karteileichen und Fehlbeständen, als auch mit Hinblick auf die Existenz von Strukturen, die mehr Heterogenität aufweisen.[17] Dieser Ansatz führt allerdings dazu, dass die Korrelation zwischen Zensusbevölkerung und Registerbevölkerung pro Anschrift nicht mehr durchgängig bei 0,993 angenommen werden kann. In

[16] Die Zahl 0,993 beruht auf einer Vorgabe des Auftraggebers.

[17] Bei der Erstellung der Karteileichen- und Fehlbestandsmodelle wurde die jeweilige Wohnung als Hauptsitzwohnung angenommen, sofern der Wohnstatus ungeklärt war (also $WST = 9$).

Abbildung 3.2 sind drei verschiedene Arten von Korrelationen abgebildet. Die ursprüngliche homogene Korrelation ist im Panel ganz rechts dargestellt (Korrelation_993). Sie wurde mit einem deterministischen Algorithmus erzeugt. Die Korrelationen im linken Panel (Korrelation_I3) wurden auf Basis eines Multi-Level-Modells erzeugt und die Korrelationen im mittleren Panel (Korrelation_Syn) basieren auf einer Vorhersage ohne Modell, welche zum Ziel hatte, die Korrelation von 0,993 zu erreichen (zu sehen sind Ergebnisse für Rheinland-Pfalz). Eine genauere Beschreibung der Erzeugung der Karteileichen- und Fehlbestandsmodelle ist auf S. 73 gegeben.

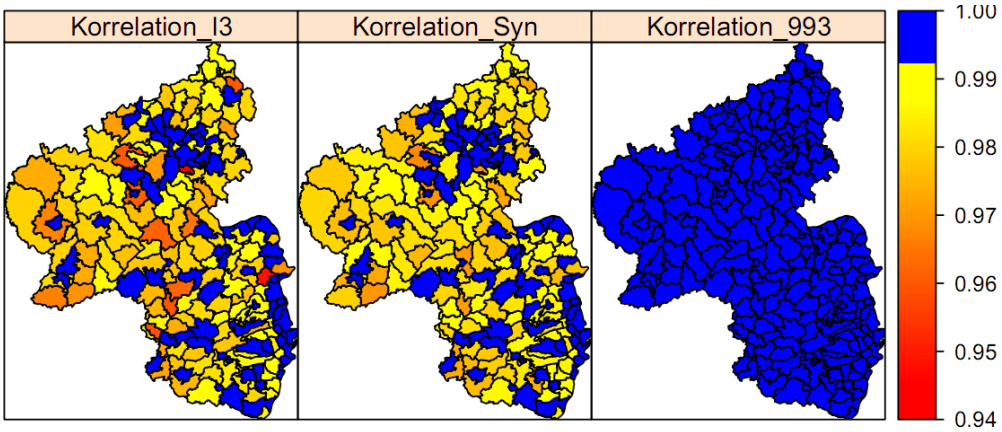


Abbildung 3.2: Karte zur Korrelation zwischen Register- und Zensusbevölkerung in Rheinland-Pfalz

Bei der synthetischen Erzeugung spielen die bereits angesprochenen Probleme mit zu großen Designgewichten im Zensustest eine Rolle. Dieses Problem thematisiert vor allem Gelman in seinem diskutierten Papier (Gelman 2007) und Meng in seinem Kommentar (Meng et al. 2009). Es geht hier insbesondere darum, dass das Verhältnis von dem größten zum kleinsten Designgewicht 10 nicht überschreiten sollte und es inakzeptabel ist, wenn dieses Verhältnis größer als 100 ist. Letztere Situation ist aber für den Zensustest gegeben, dem die Modellierung von Karteileichen und Fehlbeständen zugrunde liegt. Zusätzlich ergibt sich das Problem, dass stark differierende Populationen (Schätzung vs. Prediction) vorhanden sind. Die synthetische Population ist um ein Vielfaches größer als der Zensustest, auf dem die Schätzung beruht.

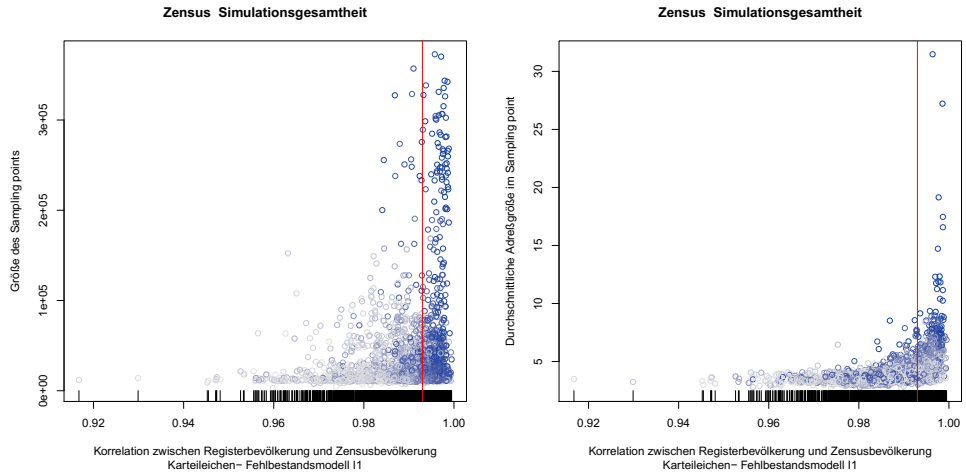


Abbildung 3.3: Korrelation zwischen Register- und Zensusbevölkerung

Liegt die Korrelation zwischen der Registerbevölkerung und der Zensuspopulation tatsächlich bei 0,993 und höher, stehen mit den Registerinformationen extrem starke Hilfsinformationen bei der Schätzung der Bevölkerung zur Verfügung. Allerdings zeigt die Abbildung 3.3, dass die Korrelation zwischen der Zahl der registrierten Personen und der Zahl der Personen, die tatsächlich in der Anschrift wohnen, in einigen Fällen deutlich unter 0,993 liegt.

In Abbildung 3.3 ist der Zusammenhang zwischen der Korrelation zweier Bevölkerungsgrößen und der Größe des Sampling Points (linke Grafik) beziehungsweise der durchschnittlichen Anschriftengröße (rechte Grafik) dargestellt. Je dunkler die Einfärbung der Punkte in der linken Grafik ist, desto größer ist die durchschnittliche Anschriftengröße pro Sampling Point. Wenn ein Sampling Point in einem sehr hellen Grauton eingefärbt ist, dann bedeutet dies, dass die mittlere Anschriftengröße sehr gering ist. Die kleinste mittlere Anschriftengröße pro Sampling Point liegt in der Zensus-Simulationsgesamtheit bei 2,7 Personen pro Anschrift. Die größte durchschnittliche Personenanzahl pro Anschrift liegt bei 31,4. Diese hohen Durchschnittszahlen tauchen hauptsächlich für die Sampling Points vom Typ 0, also Stadtteilen auf. Auf der x-Achse sind die Korrelationen und auf der y-Achse die Größe der Sampling Points abgetragen. Prinzipiell lässt sich ein Zusammenhang zwischen der Größe des Sampling Points und der Korrelation feststellen. Dieser Zusammenhang kann darauf zurückgeführt werden, dass in größeren Sampling Points auch durchschnittlich größere Anschriften auftauchen. In kleinen Sampling Points, in denen zudem die durchschnittliche Anschriftengröße sehr klein ist, kann die Korrelation also sehr stark einbrechen.

Für die finalen Simulationen wurden drei verschiedene Karteileichen- und Fehlbestandsmodellierungen verwendet:

I1 Das Karteileichen- und Fehlbestandsmodell **I1** wurde anhand eines Multi-Level Logit-Modells erzeugt. Dabei wurden verschiedene Kovariaten verwendet. Bei der Generierung wurde die Karteileichen- und Fehlbestandsstruktur aus dem Zensustest berücksichtigt. Das Modell enthält geklumptes Auftreten von Karteileichen und Fehlbeständen sowie zufällige Niveauunterschiede zwischen Gemeinden und Bundesländer. Es ist unter Berücksichtigung der Informationen aus dem Zensustest als eine Art Worst-Case-Szenario anzusehen.

Syn Ist ein Algorithmus zur Erzeugung von Karteileichen und Fehlbeständen, der eine rein zufällige Auswahl der Karteileichen und Fehlbestände durchführt. Dieser Algorithmus basiert nicht auf einem Modell. Er ist also insbesondere nicht von irgendwelchen Registervariablen abhängig. Dieses Modell eignet sich insbesondere zur Überprüfung des Einflusses des Karteileichen- und Fehlbestandsmodells auf die Qualität der *Ziel 2* Schätzungen, weil sehr viele unterschiedliche Korrelationen zwischen der Zensusbevölkerung und der Registerbevölkerung pro Anschrift enthalten sind.

T993 Um den Vorgaben des Statistischen Bundesamtes der Korrelation der Registeranzahl und der Zensus Anzahlen von 0,993 je SMP zu entsprechen, wurde eigens für Rheinland-Pfalz ein Karteileichen- und Fehlbestandsmodell erstellt. Dies weist die Korrelation von annähernd 0,993 pro SMP auf. Dieses Modell ist geeignet zu überprüfen, wie sich die Korrelation der Zensusbevölkerung zur Registerbevölkerung auf die Qualität der Schätzungen auswirkt.

Nähere Informationen zur Modellierung der Karteileichen- und Fehlbestandsmodelle sind bei Burgard und Münnich (2010) zu finden.

3.2.4 Editing

Die synthetischen Variablen sind das Ergebnis eines stochastischen Zufallsprozesses. Es ist möglich, dass bei diesem Zufallsprozess Kombinationen von Variablen erzeugt werden, die in der Realität nicht vorkommen. Der Arbeitsschritt, bei dem Variablen überprüft und gegebenenfalls korrigiert werden, ist dem Aufarbeiten der Umfragedaten ähnlich und wird Editing genannt.

Insbesondere aus der Kombination der Variable Alter und den Bildungsvariablen sowie den Erwerbsvariablen ergeben sich strukturelle Nullen, die bei einem nach der Erzeugung durchzuführenden Editing zu berücksichtigen sind. Als strukturelle Nullen werden Variablenkombinationen aus zwei Mikrozensusvariablen definiert, die im gesamten Mikrozensus nicht vorkommen.

Neben dem Editing für die synthetisch generierten Variablen muss auch für die MR-Daten ein Editing durchgeführt werden. Dies liegt unter anderem darin begründet, dass durch das Anonymisieren der Daten Inkonsistenzen aufgetreten sind. Einige Variablen müssen zudem so umkodiert werden, dass sie zu den Ausprägungen der Variablen im Mikrozensus 2006 passten. Auch durch diesen Umkodierungsprozess können logische Inkonsistenzen auftreten. Beispielsweise war in den MR-Daten nur das Geburtsdatum enthalten, benötigt wurde aber das Alter in Jahren, dieses Alter darf minimal 0 Jahre und maximal 110 Jahre betragen. Nach oben muss eine Altersgrenze gewählt werden. Wo diese Grenze genau liegt, dürfte nur von theoretischer Bedeutung sein.

Es muss weiter überprüft werden, ob für alle Personen innerhalb einer Anschrift der gleiche amtliche Gemeindeschlüssel auftaucht. Dies ist eine wichtige Anforderung, die beispielsweise für den Grundbestand der MR-Daten kontrolliert werden muss. Zudem benötigt man eine Anschriftenlaufnummer, die über ganz Deutschland hinweg eindeutig ist. Diese Anforderung ist von enormer Bedeutung für den Stichprobenziehungsprozess.

Für die synthetischen Variablen gibt es Editing Regeln, die wie oben bereits beschrieben, teilweise automatisch erfüllt werden, deren Einhaltung aber teilweise nochmals geprüft werden muss. Eine exemplarische Auflistung ist in Tabelle 3.4 gegeben.

Tabelle 3.4: Beispiele für Editing Regeln

Regel	Betrifft Variablen
Weibliche Wehrdienst- und Zivildienstleistende sollte es nicht geben.	SEX und EF117
Wehrdienstleistende und Zivildienstleistende müssen Deutsche sein.	NAT und EF117
Personen mit einem Universitätsabschluss müssen mindestens 21 Jahre sein.	AGE und ISCED
Personen die erwerbstätig sind, müssen mindestens 15 Jahre sein.	ILO und AGE

3.2.5 Synthetisch generierte Daten

Im Folgenden werden nun die erzeugten Variablen mit ihren Ausprägungen und Verteilungen vorgestellt. Grundsätzlich orientiert sich das Programm der synthetisch generierten Daten an den Informationsanforderungen, die von verschiedenen Seiten an den Zensus gestellt werden. Im Kernprogramm der EU/ECE-Empfehlungen für den Zensus 2011 werden verschiedene Merkmale zur Bearbeitung vorgeschlagen. Dazu zählen unter anderem erwerbs- und bildungsstatistische Merkmale. Um diese Merkmale in die Simulation einfließen zu lassen, wurden die nachfolgenden Variablen synthetisch generiert.

3.2.5.1 Die Variable ISCED Stufen

Die Bildungsvariable (ISCED[18] Stufen) ist eine interessierende Variable des Kernprogramms und baut auf der Mikrozensus Variable Höchster beruflicher oder allgemeiner Abschluss (MZ06 EF540) auf. Dabei wird auf die international gebräuchlichen ISCED97 Stufen (Stand: 12.02.2003) zurückgegriffen. Die ISCED Stufen sind ein von der OECD herausgegebenes Instrument zum Vergleich verschiedener Ausbildungsniveaus. Tabelle 3.5 gibt einen Überblick über diese Ausbildungsniveaus.

[18] ISCED steht hier für International Standard Classification of Education.

Tabelle 3.5: Ausprägungen der Variable Höchster beruflicher oder allgemeiner Abschluss (ISCED) EF540 im Mikrozensus 2006

Code	Beschreibung
11	ISCED 1 (Ohne allgemeinen und ohne beruflichen Abschluss)
21	ISCED 2 (Haupt-/Realschulabschluss ohne berufl. Abschluss; Haupt-/Realschulabschluss mit Anlernausbildung, berufl. Praktikum oder Berufsvorbereitungsjahr; ohne allgemeinen Abschluss, aber mit Anlernausbildung, berufl. Praktikum oder Berufsvorbereitungsjahr)
31	ISCED 3 (Hoch-/Fachhochschulreife; Lehrausbildung; berufsqualifizierender Abschluss an einer Berufsfachschule/Kollegschule, 1-jährige Schule des Gesundheitswesens Abschluss einer Lehrausbildung, Vorbereitungsdienst für den mittleren Dienst in der öffentlichen Verwaltung)
41	ISCED 4a, b (Hoch-/Fachhochschulreife und Lehrausbildung; berufsqualifizierender Abschluss an einer Berufsfachschule/Kollegschule, 1-jährige Schule des Gesundheitswesens)
51	ISCED 5a (Fachhochschule, Hochschule)
52	ISCED 5b (Meister-/Techniker- oder gleichwertiger Fachschulabschluss, Abschluss einer 2- oder 3- jährigen Schule des Gesundheitswesens, Abschluss einer Fach- oder einer Berufsakademie; Abschluss der Fachschule der DDR; Abschluss einer Verwaltungsfachhochschule Meister-/Technikerausbildung oder gleichwertiger Fachschulabschluss, Abschluss einer 2- oder 3- jährigen Schule des Gesundheitswesens, Abschluss einer Fachakademie oder einer Berufsakademie)
61	ISCED 6 (Promotion)
90	Keine Angabe
99	Entfällt (Kind unter 15 Jahren)

Einige Ausprägungen der Mikrozensus-Variablen wurden für die entsprechende Variable der synthetischen Population zusammengefasst. Des Weiteren wurden aus dem Mikrozensus 2006 neben der Variable EF540 noch weitere Informationen hinzugezogen. Zur Zuordnung der Mikrozensus-Bevölkerung zu den synthetischen ISCED-Stufen wurde die Variable Höchster beruflicher Ausbildungs- oder Hochschul-/Fachhochschulabschluss (MZ06 EF312) und die Variable Höchster weiterer beruflicher Abschluss (MZ06 EF316) verwendet.

Es ergeben sich daraus die folgenden ISCED Ausprägungen:

1. Grundbildung
   Besuch der Grundschule als höchstes Ausbildungsniveau.
2. Sekundarbildung Unterstufe
   Besuch der Sekundarstufe 1. In etwa bis zum Ende der Schulpflicht.
3. Sekundarbildung Oberstufe
   Besuch der Sekundarstufe 2 und der dualen Berufsbildung.
4. Postsekundäre Bildung
   Erreichen der Hochschulreife.
5. Tertiäre Bildung, erste Stufe
   Hochschulausbildung. z. B. Diplom, Bachelor, Master ...

6. Tertiäre Bildung, Forschungsqualifikation
   Postgraduale Ausbildung, Promotion oder Habilitation als höchsten Bildungsabschluss.

In der Simulationsgesamtheit treten die Codes (Ausprägungen) mit folgenden Häufigkeiten auf:

Tabelle 3.6: Ausprägungen und Häufigkeiten der Variable ISCED

Code	Häufigkeit	Beschreibung
1	11.119.688	Grundbildung
2	7.284.351	Sekundarbildung Unterstufe
3	8.542.488	Sekundarbildung Oberstufe
4	14.266.715	Postsekundäre Bildung
5	28.808.211	Tertiäre Bildung, erste Stufe
6	10.739.200	Tertiäre Bildung, Forschungsqualifikation
9	5.027.177	nicht spezifiziert

3.2.5.2 Erwerbsvariable ILO

Eine weitere wichtige Information, die ebenfalls Teil des Kernprogramms der EU/ECE-Empfehlungen ist, betrifft die Erwerbstätigkeit der Personen in der Simulationsgesamtheit. Nach dem ILO Konzept werden Personen als erwerbslos definiert, sofern sie ohne jegliche Beschäftigung sind, innerhalb der letzten vier Wochen vor der Befragung aktiv auf der Suche nach einer Erwerbstätigkeit sind und sofort in der Lage wären, eine Beschäftigung anzutreten. Die Variable ILO, die für die synthetische Zensusgesamtheit (ZSD) erzeugt wurde, basiert auf den Variablen Erwerbstyp (EF29) und überwiegender Lebensunterhalt (EF401) im Mikrozensus 2006. Des Weiteren wurden zur Zusammenstellung der Variable ILO für den Mikrozensus folgende Variablen verwendet: Situation der Arbeitssuche (EF279), Schulbesuch in den letzten 12 Monaten (EF288) und der Variable Alter (für die Personen, die noch nicht das nationale Mindestalter für die Erwerbstätigkeit erreicht haben). Diese Variable hat dann die folgenden Ausprägungen und relativen Häufigkeiten[19] (siehe Tabelle 3.7):

Tabelle 3.7: Ausprägungen der Variable ILO

Code	Beschreibung	Relative Häufigkeit (MZ06)	Relative Häufigkeit (ZSD)
1	Erwerbstätig	0,45	0,45
2	Erwerbslos	0,05	0,05
3	Sonstige Nichterwerbsperson	0,02	0,02
4	Nichterwerbsperson und Rente, Pension, Eigenes Vermögen, Pflegeversicherung	0,22	0,23
5	Nichterwerbsperson und Unterhalt durch Eltern/Lebenspartner etc.	0,25	0,23
6	Nichterwerbsperson und sonstige Unterstützungen	0,01	0,01
9	Entfällt	0	0

[19] Die relativen Häufigkeiten für den Mikrozensus 2006 sind mit der Variable EF951 hochgerechnet. Bei dieser Variable handelt es sich um den Hochrechnungsfaktor für das Quartal.

Die Variable Alter spielt eine bedeutende Rolle bei der Erzeugung der synthetischen Variablen bezüglich des Erwerbslebens. Bei einigen Variablen gibt es Altersstufen, aus denen sich direkt eine Editing-Regel ableiten lässt. So darf eine erwerbstätige Person beispielsweise nicht jünger als 15 Jahre sein. Eine Aufteilung in Altersklassen ist notwendig, da die Anzahl der möglichen Merkmalsausprägungen sonst zu hoch wäre.

3.2.5.3 Berufsgruppen (EF117)

Die Variable EF117 des Mikrozensus 2006 gibt Informationen über die Stellung im Beruf. Diese Variable ist zum einen für die Hypercubes von Interesse. Zum anderen ist die Schätzung dieser Variable interessant, da die Ausprägungen dieser Variable regional sehr unterschiedlich verteilt sind.

Tabelle 3.8: Ausprägungen der Variable Stellung im Beruf (EF117)

Code	Beschreibung
01	Selbstständiger ohne Beschäftigte
02	Selbstständiger mit Beschäftigten
03	Mithelfender Familienangehöriger
04	Beamter, Richter
05	Angestellter
06	Arbeiter, Heimarbeiter
07	kaufm./techn. Auszubildender
08	gewerbl. Auszubildender
09	Zeit-/Berufssoldat (einschl. Bundespolizei und Bereitschaftspolizei)
10	Grundwehrdienstleistender
11	Zivildienstleistender
12	Entfällt (Nichterwerbstätige)
13	Keine Angabe

Bei der Variable Stellung im Beruf ist es besonders wichtig, den Zusammenhang zu weiteren Variablen aus dem Mikrozensus aufzunehmen und in der synthetischen Population richtig wiederzugeben. Der Zusammenhang zwischen kategorialen Variablen kann besonders gut mit Mosaikplots dargestellt werden (siehe Kapitel 6). Mit dieser Art von Diagrammen lassen sich bedingte Häufigkeiten erkennen. Die Größe der Flächen zeigt jeweils an, wie häufig eine Merkmalskombination vorkommt.

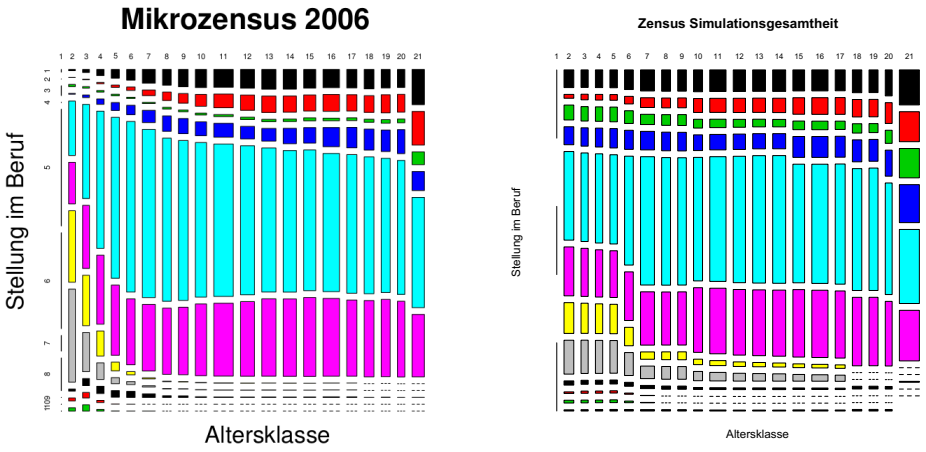


Abbildung 3.4: Zusammenhang zwischen Altersklassen und Stellung im Beruf (Variable EF117)

Auf der linken Seite in Abbildung 3.4 befindet sich ein Mosaikplot zu den Häufigkeiten der Merkmalskombinationen der Variablen Altersklasse und Stellung im Beruf für den Mikrozensus 2006. In der rechten Grafik ist der entsprechende Mosaikplot noch einmal für die Zensus-Simulationsgesamtheit zu sehen. Die Variable Altersklasse resultiert aus der Unterteilung des Alters in 21 Klassen. In der ersten Altersklasse befinden sich alle Personen die jünger als 15 Jahre alt sind. Diese Grenze für die erste Klasse wurde gewählt, da Personen, die jünger als 15 Jahre alt sind, keine Stellung im Beruf haben dürfen. In der zweiten Altersklasse befinden sich Personen, die mindestens 15 Jahre aber höchstens 19 Jahre alt sind. Jede weitere Altersklasse umfasst die nachfolgenden fünf Jahre. Bei den Ausprägungen für Stellung im Beruf wurden aus Gründen der übersichtlicheren Darstellung die letzten beiden Ausprägungen weggelassen. Schaut man nur auf die linke Grafik, so ist zu sehen, dass die Häufigkeiten der Ausprägungen für Stellung im Beruf, verglichen zwischen den verschiedenen Altersklassen, deutlich unterschiedlich sind. Diese Struktur würde zu der Annahme führen, dass die Variable Stellung im Beruf von der Variable Altersklasse abhängig ist. Vergleicht man die linke mit der rechten Grafik, so sieht man, dass die Abhängigkeitsstruktur, die für den Mikrozensus gegeben ist, in der synthetischen Population nicht exakt wiedergegeben werden konnte, aber wichtige Elemente in beiden Grafen wiederzuerkennen sind.

Da die Variable EF117 (Stellung im Beruf) sehr viele Ausprägungen hat, wurde die Anzahl in der später vorgestellten Schätzung auf drei Berufsgruppen begrenzt, die besonders wichtig sind (siehe Tabelle 3.9). Es handelt sich dabei um die Gruppe der Angestellten, der Beamten und der Selbstständigen.

Tabelle 3.9: Die zu schätzenden Berufsgruppen

Code	Ausprägung	Bedeutung
A	EF117=5	Angestellte
B	EF117=4	Beamte
S	EF117=1 oder 2	Selbstständige

In Abbildung 3.5 ist für alle Sampling Points der Anteil dieser drei Gruppen an der Gesamtbevölkerung eines Sampling Points für das Bundesland Nordrhein-Westfalen (NRW) zu sehen. Es ist zu erkennen, dass der Anteil der angestellten Personen (EF117A) gerade in und um die großen Städte sehr hoch ist, während die Anteile der Beamten und Selbstständigen, die allgemein unter dem Wert für die Angestellten liegen, über die Sampling Points in NRW relativ homogen verteilt sind. Ähnliche Ergebnisse sind auch für die anderen Bundesländer zu beobachten, wie dies in Abbildung 3.6 erkennbar ist.

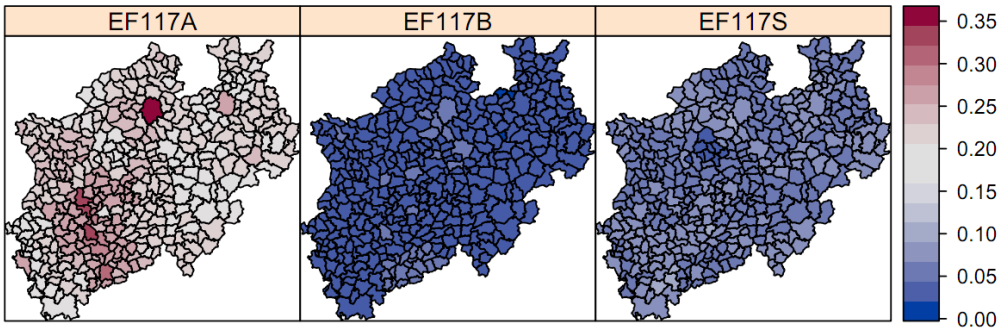


Abbildung 3.5: Räumliche Verteilung der Ausprägungen der Variable EF117 - Stellung im Beruf - in Nordrhein-Westfalen

Abbildung 3.6 zeigt die Anteile der Angestellten für die vier Bundesländer Nordrhein-Westfalen (NRW), Rheinland-Pfalz (RLP), Berlin (BER) und Mecklenburg-Vorpommern (MVP). Offensichtlich nimmt Berlin in Bezug auf die Berufsgruppen in der synthetischen Simulationsgesamtheit eine Sonderstellung ein. In Berlin ist der Anteil der Angestellten an der Gesamtbevölkerung in allen Sampling Points (es sind in Berlin nur Sampling Points vom Typ 0 vorhanden) geringer als die Anteile der Angestellten in den Sampling Points der anderen Länder.

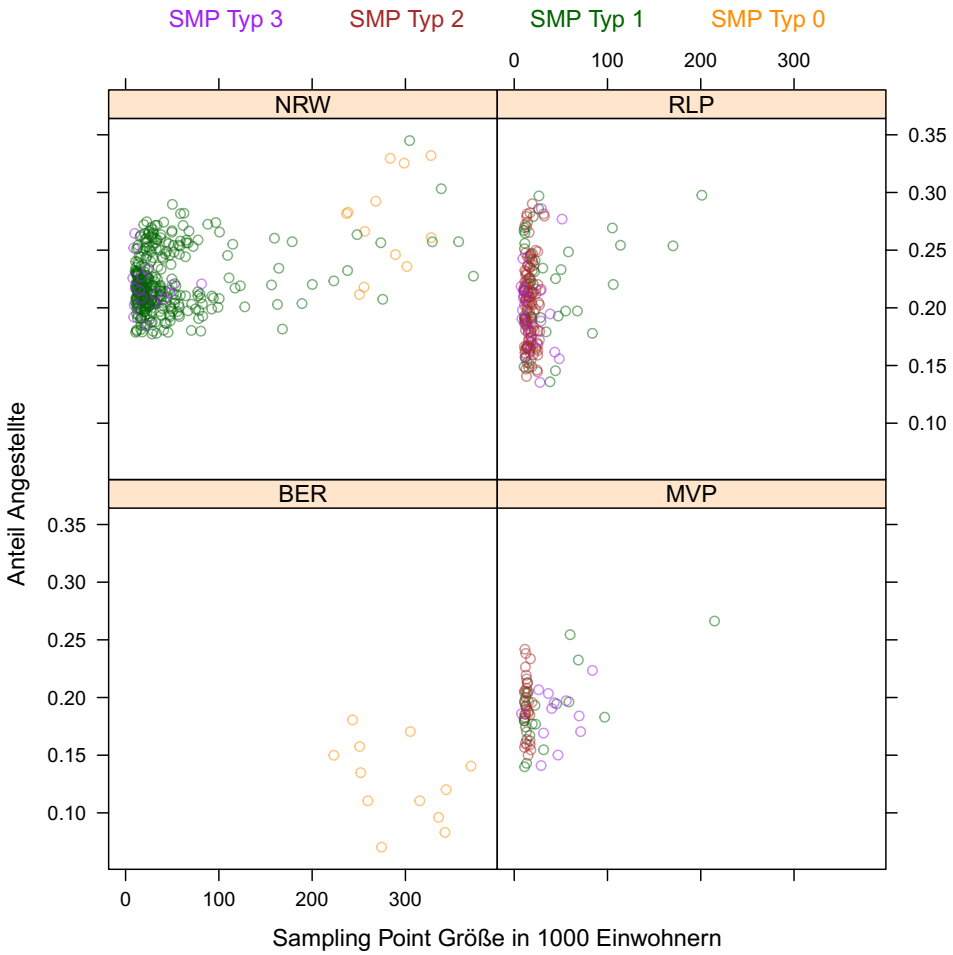


Abbildung 3.6: Anteil Angestellte (Variable EF117)

In Abschnitt 3.2.2 wurde bereits erwähnt, dass die MR-Daten relativ homogen sind. Die Anteile der Personen, die pro Sampling Point eine bestimmte Merkmalsausprägung aufweisen, streuen nicht sehr stark zwischen den Sampling Points. Demgegenüber ist in Abbildung 3.7 zu sehen, dass diese Anteile für die synthetisch generierten Variablen teilweise sehr viel mehr streuen. Es ist also eine größere Heterogenität gegeben.

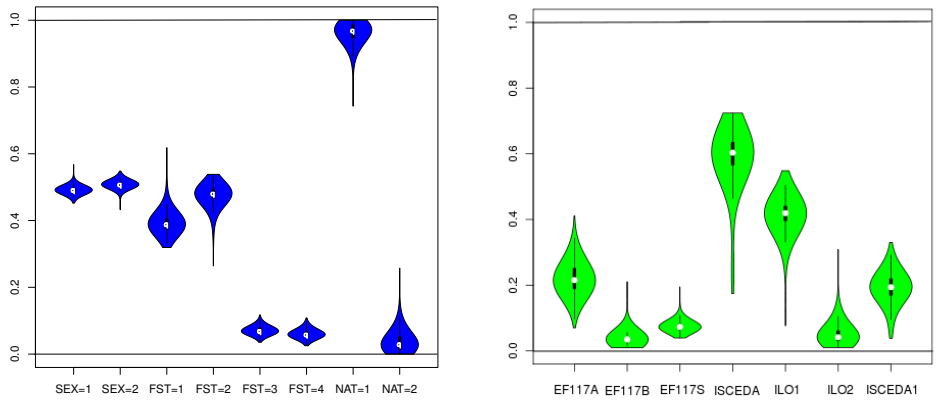


Abbildung 3.7: Vergleich der Variation von Melderegistervariablen und synthetischen Variablen

3.2.5.4 Die Variablen für den Hypercube

Der Hypercube, der mit dem Auftraggeber vereinbart wurde, umfasst die Variablen SEX Geschlecht, EF310 Höchster allgemeiner Schulabschluss und EF401 Überwiegender Lebensunterhalt. Die Variable SEX ist bereits im anonymisierten Melderegisterdatenmaterial enthalten. Bei der Variable EF401 handelt es sich um eine Variable, die Auskunft über das Erwerbsprofil einer Person gibt. Die Ausprägungen der Variable sind aus Tabelle 3.10 abzulesen.

Tabelle 3.10: Überwiegender Lebensunterhalt (Variable EF401)

Code	Variable
1	Erwerbstätigkeit/Berufstätigkeit
2	Arbeitslosengeld I, II
3	Rente, Pension
4	Unterhalt durch Eltern, Ehepartner/Ehepartnerin, Lebenspartner/Lebenspartnerin oder andere Angehörige
5	Eigenes Vermögen, Ersparnisse, Zinsen, Vermietung, Verpachtung, Altenteil
6	Sozialhilfe, - geld, Grundsicherung, Asylbewerberleistungen
7	Leistungen aus einer Pflegeversicherung
8	Sonstige Unterstützungen (z. B. BAföG, Vorruhestandsgeld, Stipendium) Bezug von Renten/Pensionen

Die Variable EF310 ist neben ISCED eine weitere Variable, die Informationen über das Ausbildungsprofil der jeweiligen Person gibt. In Tabelle 3.11 sind die Ausprägungen der Variable im synthetischen Simulationsdatenbestand zu sehen. Berücksichtigt werden nur Schulabschlüsse, die bereits beendet wurden, deshalb entfällt die Variable für Personen, die jünger als 15 Jahre alt sind.

Tabelle 3.11: Ausprägungen der Variable Höchster allgemeiner Schulabschluss (EF310)

Code	Beschreibung
1	Haupt-(Volks-)schulabschluss
2	Abschluss der allgemeinbildenden Polytechnischen Oberschule der ehemaligen DDR
3	Realschulabschluss (Mittlere Reife) oder gleichwertiger Abschluss
4	Fachhochschulreife
5	Allgemeine oder fachgebundene Hochschulreife (Abitur)
6	Ohne Angabe
7	Entfällt (Kinder unter 15 Jahren, Schüler an allgemeinbildenden Schulen; Personen ohne allg. Schulabschluss)

Für die Schätzung des Hypercubes ist von besonderem Interesse, mit welcher Häufigkeit die Merkmalskombinationen der beiden Variablen EF310 und EF401 auftreten. In der linken Grafik der Abbildung 3.8 wird die relative Häufigkeit des Auftretens der 56 möglichen Ausprägungskombinationen gezeigt. Die Merkmalskombination EF401=1 und EF310=5 tritt dabei mit 16 % am häufigsten auf. Es ist auch zu sehen, dass es sich bei einigen Merkmalskombinationen (auch Zellen genannt) um strukturelle Nullen handelt. Diese Zellen sind dann dunkel eingefärbt.

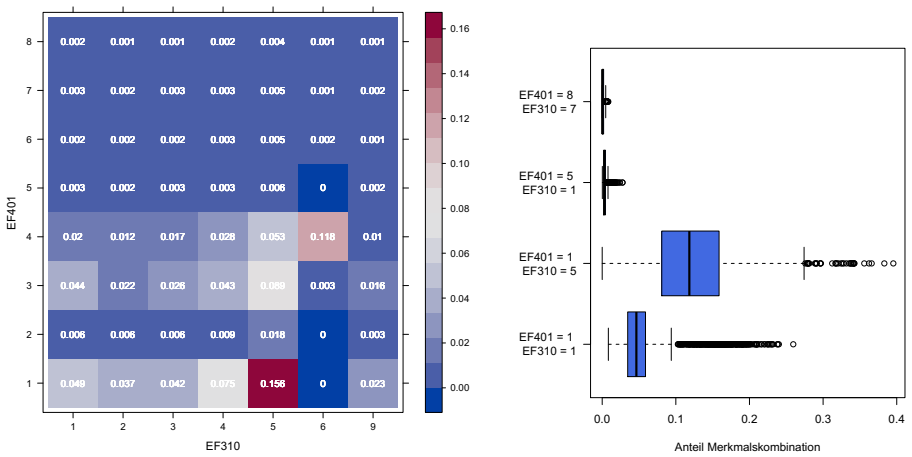


Abbildung 3.8: Relative Häufigkeiten bzgl. der *Ziel 2* Hypercube-Variablen EF310 und EF401 sowie Boxplots

Auf der rechten Seite der Abbildung 3.8 ist der Anteil des Auftretens der Merkmalsausprägung je Sampling Point in einem Boxplot abgetragen. In einem Sampling Point trifft die Merkmalskombination EF401=1 und EF310=5 auf 40 % der Personen im Sampling Point zu. Allerdings gibt es auch Sampling Points, in denen diese Merkmalskombination weitaus seltener vorkommt. Während diese Merkmalsausprägung in allen Sampling Points mindestens einmal vorhanden ist, kommt die Merkmalskombination EF401=8 und EF310=9 nur in 54 % aller Sampling Points mindestens einmal vor. Es ist zu berücksichtigen, dass im dreidimensionalen Hypercube zu den beiden hier in den Merkmalskombinationen analysierten Variablen noch die Variable SEX hinzukommt. Der Hypercube besteht also aus 112 Zellen.

3.2.5.5 Die Daten der Bundesagentur für Arbeit (BA-Daten)

Einer der Paradigmenwechsel, die im Rahmen des Zensus 2011 durchgeführt wurden, bezieht sich auf die Verwendung von Registern zur Unterstützung der Schätzungen. Ein prominentes Beispiel ist die Verwendung des Melderegisters zu diesem Zweck. Ein weiteres Beispiel sind Daten der Bundesagentur für Arbeit. Dabei ist allerdings zu beachten, dass die Ausgangssituation hier eine etwas andere ist. Dies liegt vor allem daran, dass die Qualität dieser Register nicht vergleichbar mit der Qualität der Melderegister ist. Zudem können vermutlich nur Aussagen über die Zahl der Erwerbstätigen pro Anschrift gemacht werden.

Aufgrund des Datenschutzes war es nicht möglich, Mikrodaten der Bundesagentur für Arbeit in die Simulation einzubauen. Um dennoch Aussagen über mögliche Effizienzgewinne machen zu können, wurde eine Variable (EWT) künstlich erzeugt, die das Vorhandensein von Informationen aus den Registern der Bundesagentur für Arbeit simulieren soll. Es wurden zwei verschiedene Versionen dieser Variable EWT erzeugt, zum einen eine Variable, bei der die Korrelation zwischen Zahl der Personen in einer Anschrift, die erwerbstätig sind (ILO=1), und der Zahl der Personen, die laut den synthetischen Daten der BA erwerbstätig sind, 60% beträgt (EWT60) und zum anderen ein weiteres Szenario, in dem die gleiche Korrelation 90% beträgt. Es ist hier anzumerken, dass das 90% Szenario bei weitem realistischer erscheint.

Die erzeugte Variable EWT basiert ebenfalls auf Informationen des Auftraggebers über die Daten der Bundesagentur für Arbeit. Um die Anonymität zu gewährleisten, wurden hier pseudo SMPs[20] gebildet.

Es wurden zur Erstellung der Variable folgende in Tabelle 3.12 dargestellten Informationen verwendet.

Tabelle 3.12: Bereitgestellte Informationen zur Erstellung der Variable EWT

Information	Beschreibung
an_{grkl}	Anschriftengrößenklasse gemessen an der registrierten Bevölkerung (HW+NW): 1/2/3/4-5/6-9/10+ , Einsteller codiert mit 1,...,6.
m_{svp}	Mittelwert der Zahl der svp Beschäftigten[a]
m_{al}	Mittelwert der Zahl der Arbeitslosen
s_{svp}	Standardabweichung der Zahl der svp Beschäftigten
s_{al}	Standardabweichung der Zahl der Arbeitslosen
n	Zahl der Anschriften, die der Mittelwertbildung und der Standardabweichung zugrundeliegen

[a] svp: sozialversicherungspflichtig

Diese Informationen dienen hauptsächlich dem Zweck, ein Bild über die Struktur der Sozialversicherungsdaten zu gewinnen. Es wurden weiterhin für die Erzeugung der Variable EWT die MR-Variablen AGE (Alter), BLA (Bundesland), ADR (Anschrift) und SEX (Geschlecht) sowie die bereits erstellte synthetische Variable ILO (Erwerbsstatus) verwendet. Analog zum Vorgehen bei der Erzeugung der Karteileichen- und Fehlbestandvektoren werden auch für die EWT-Variable Logit-Modelle genommen. Mit Hilfe der Parameter des Binomialmodells wird dann eine Vorhersage gemacht.

[20] Dieser Pseudo SMP hat 12 Stellen, bei Gemeinden ab 10.000 Einwohnern handelt es sich um den AGS12, bei Kreisresten (Kreis ohne große Gemeinden) sind es die ersten fünf Stellen des AGS12, die restlichen 7 Stellen sind mit Nullen aufgefüllt.

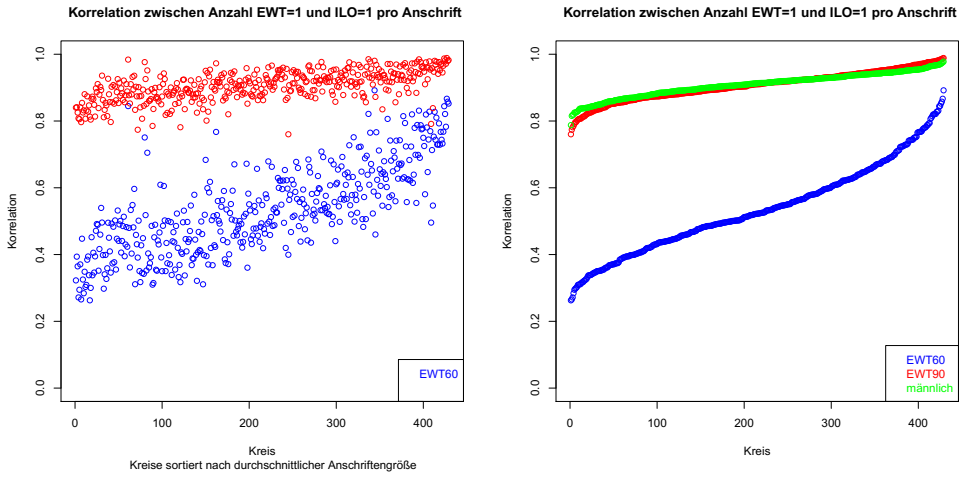


Abbildung 3.9: Vergleich der Korrelation der Variable EWT

Ziel der Generierung der Variable EWT ist es zu zeigen, welchen Einfluss die Qualität der Daten der Bundesagentur für Arbeit haben können. In der linken Grafik der Abbildung 3.9 werden die beiden Szenarien der Variable gezeigt. Die roten Punkte stellen die Korrelation der 90 % Variable dar, während die blauen Punkte die Korrelationen der 60% Variable kennzeichnen. An der Abbildung 3.9 ist zu erkennen, dass die Bezeichnungen nicht wörtlich aufzufassen sind. Jeder Punkt steht für die Korrelation zwischen der Zahl der Ausprägungen ILO=1 und der Zahl der Ausprägungen EWT=1 pro Anschrift in einem Kreis. Die maximale Korrelation für einen Kreis bei der EWT90 Variable liegt demzufolge bei 0,99, während die minimale Korrelation bei 0,76 liegt. Für die EWT60 Variable liegt der Minimalwert bei 0,26 und der Maximalwert bei 0,89. Mit diesen beiden Versionen der Variable EWT sind also zwei ganz unterschiedliche Qualitäten für eine Hilfsvariable abgebildet. Welche Auswirkungen diese unterschiedlichen Qualitäten auf die Schätzergebnisse haben, wird später gezeigt. In der rechten Grafik der Abbildung 3.9 ist zu sehen, dass der Vorteil des EWT60 Szenarios darin liegt, dass die Ausprägungen deutlich heterogener über die Kreise verteilt sind. In dieser Abbildung sind die Kreise nach der Höhe des Anteils der jeweiligen Ausprägung sortiert. Die grünen Punkte zeigen dabei die Korrelation zwischen Zahl der Männer in einer Anschrift und der Anschriftengröße. Die Verteilung der Korrelationen in den Kreisen ist vergleichbar mit der Korrelation der EWT90 Variable mit den Anschriftengrößen. Die Korrelationen der EWT60 Variable sind deutlich anders strukturiert.

3.2.5.6 Die Variable Zuzugsjahr

Die synthetische Variable Zuzugsjahr orientiert sich an der Variable EF367 im Mikrozensus 2006. Der Erzeugung der Variable Zuzugsjahr liegt ein zweistufiges Verfahren zugrunde. Die Variable wird nur für Personen angegeben, die eine andere als die deutsche Staatsangehörigkeit haben. Die Werte, die aus dem Mikrozensus stammen, werden als Randverteilung verwendet. Eine wichtige Hilfsvariable ist die Herkunft der Personen.

Es wurde eine synthetische Variable Zuzugsjahr für den Melderegisterdatenbestand generiert. Leider weisen einige Kreise im Mikrozensus 2006 keine Personen auf, die zu einer der fünf folgenden Ausprägungen gehören (dies ist in Abbildung 3.10 zu sehen):

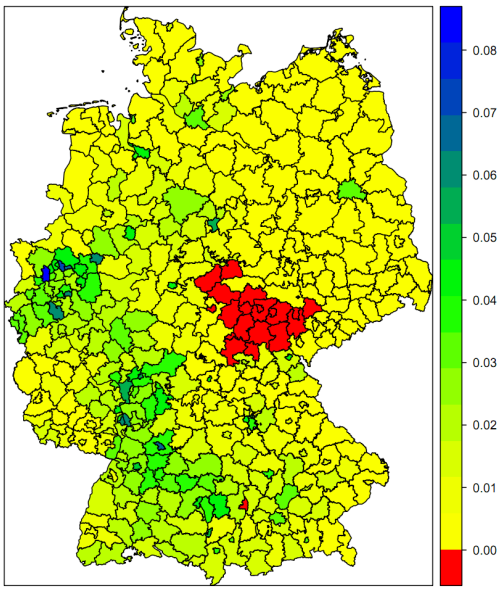


Abbildung 3.10: Anteil der Personen mit türkischer Staatsangehörigkeit an der Gesamtpopulation im Kreis

1. Türken
2. Griechen
3. Italiener
4. Personen aus dem Gebiet ehemaliges Jugoslawien
5. Osteuropäer
6. Andere

Deswegen mussten wir für diese Kreise Informationen aus anderen Kreisen zu Hilfe nehmen. Bei einer reinen Ziehung aus der Häufigkeitsverteilung aus dem gesamten Mikrozensus wäre es jedoch zu einer äußerst homogenen Variable gekommen. Um dies zu verhindern, wurde ein Modell zur Schätzung der Populationsparameter verwendet. Das Modell bildet dann die Basis für eine Vorhersage. Der resultierende Vektor hat die Größe des Simulationsdatenbestands. Zur Erstellung des Modells war ein zweistufiges Verfahren nötig, da ein großer Teil der Personen mit ausländischer Staatsbürgerschaft in Deutschland geboren ist. So wurde zunächst analog zur Modellierung der Karteileichen und Fehlbestände ein Multi-Level Logit-Modell verwendet, um die in Deutschland geborenen Ausländer zu identifizieren. Auf die Ausländer, die nach diesem Modell nicht in Deutschland geboren wurden, wurde ein weiteres Multi-Level-Modell angewendet, um das Jahr des Zuzugs vorherzusagen. Hierbei wurde die Variable Alter im Nachhinein als Editing-Argument herangezogen.

In Abbildung 3.11 wurden die relativen Häufigkeiten des Zuzugs von Personen abgebildet. Für die vierte Gruppe gab es beispielsweise zwei große Wellen der Zuwanderung um die Jahre 1970 und 1990 herum, in denen jeweils über 6 % der Personen dieser Staatsangehörigkeiten zugewandert sind. Wie an dieser Abbildung zu erkennen ist, konnte die Verteilung der Zuzugsjahre der Ausländer

in der synthetischen Population nicht identisch zur Verteilung der Zuzugsjahre im Mikrozensus modelliert werden. Jedoch wurden die wichtigsten Eigenschaften, die bei der Simulation Probleme bereiten könnten, modelliert:

- Zweigipflige Verteilungen
- relativ breite Verteilungen
- einzelne extreme Piks

Es ist festzuhalten, dass beispielsweise die Schätzergebnisse, die aus der Simulation resultieren, nicht den tatsächlichen Gegebenheiten entsprechen. Allerdings ist die synthetische Population ausreichend, um die Eigenschaften der Schätzer in Bezug auf die vorgegebenen Fragestellungen im Allgemeinen zu untersuchen.

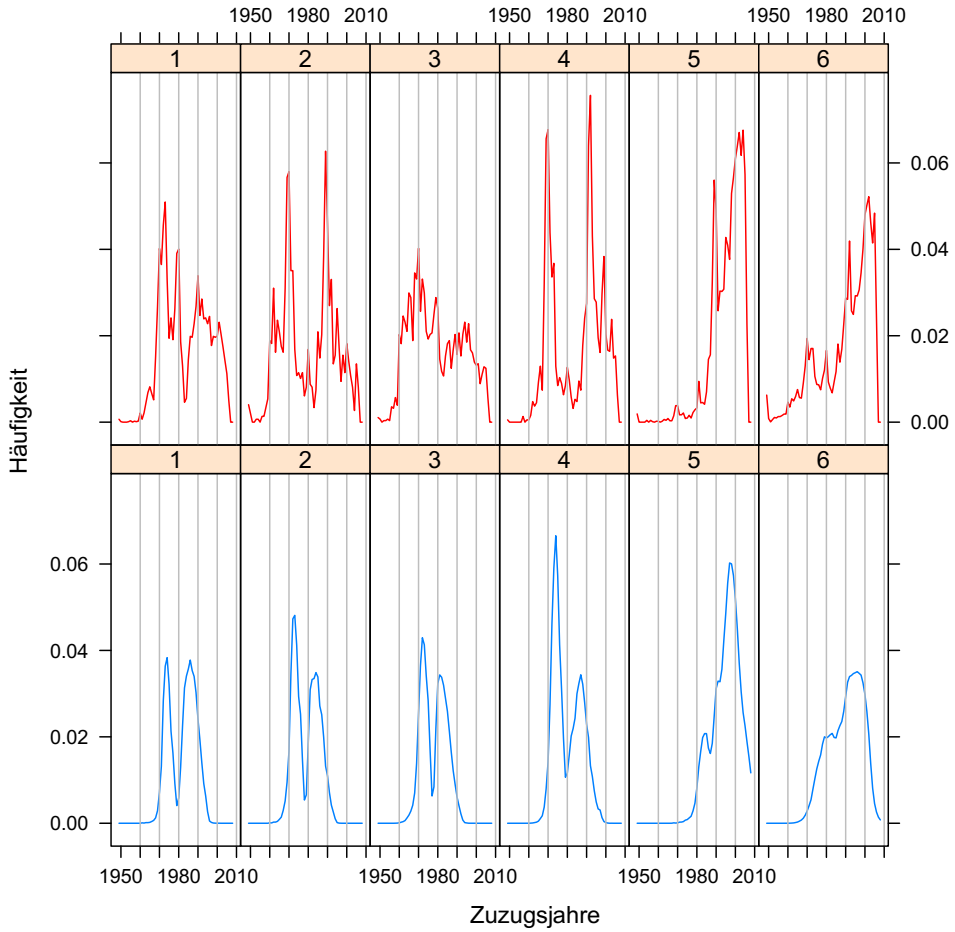


Abbildung 3.11: Verteilung des Zuzugsjahr im Mikrozensus (rot) und in der Simulationspopulation (blau) für die Personen mit Staatsangehörigkeit 1,. . . ,6

3.3 *Ziel 1*: Fehlbestände, Karteileichen und die amtliche Einwohnerzahl

3.3.1 Schätzung der amtlichen Einwohnerzahl

Die zentrale Fragestellung des Zensus ist die Ermittlung der amtlichen Einwohnerzahl (im Folgenden auch Zensusbevölkerung genannt). Voruntersuchungen ergaben, dass in Bezug auf die Zensusbevölkerung der GREG-Schätzer am geeignetsten erscheint. Dies ist in der Tatsache begründet, dass Small Area-Verfahren auf den SMP-Typen 0 und 1 kaum Verbesserungen ermöglichen, die GREG-Schätzer die Genauigkeitsanforderungen zu erfüllen scheinen und bei der Varianzschätzung Design-basierte Verfahren klar im Vorteil sind.

Aus diesem Grund werden an dieser Stelle zur Schätzung der Zensusbevölkerung nur Ergebnisse des GREG-Schätzers dargestellt. Es sei darauf hingewiesen, dass die Ergebnisse nicht ohne Weiteres auf demografische Untergruppen übertragen werden können.

Von den in Abschnitt 2.3.2.2 aufgeführten Varianten des GREG wurden voll kombinierte und die gruppierte Versionen untersucht. Aufgrund der Schichtung der Anschriften nach der Anschriftengröße ist die Schätzung durch einen voll separaten GREG nicht möglich, da in einzelnen Schichten und SMPs nur eine Ausprägung der Anschriftenregistergröße auftritt. In diesem Fall ist keine lineare Regression schätzbar, da die für die Parameterschätzung benötigte Inverse der Varianz-Kovarianzmatrix nicht existiert. Eine Übersicht über die Gruppierungen befindet sich in Tabelle 3.13. Eine Zusammenfassung von Schichten, die zur GREG-Schätzung herangezogen werden kann, ist in Tabelle 3.14 aufgeführt.

Tabelle 3.13: Bezeichnung der verschiedenen verwendeten GREG-Schätzer in *Ziel 1*

Name des Schätzers	Gruppierungsvariable
COM	voll kombinierte Schätzung über das gesamte Bundesland
D-SEP	separate Schätzung nach SMP und kombiniert über die Schichten innerhalb der SMP
KRS-SEP	separate Schätzung nach Kreisen und kombiniert über die Schichten und SMP der Kreise
SMP-Typ SEP	separat über die SMP-Typen und kombiniert über SMP und Schichten
S-SEP	separat über die Schichten und kombiniert über alle SMP
GS-SEP	separat über gruppierte Schichten und kombiniert über alle SMP
BLA-SEP	separat über die Bundesländer

Tabelle 3.14: Zuordnung der Schichten zu Gruppen für den GREG GS-SEP-Schätzer

Schicht	1	2	3	4	5	6	7	8
GS	1	1	1	2	2	3	3	4

Die nachfolgenden drei Abbildungen enthalten für die in Tabelle 3.13 aufgeführten Gruppierungen die RRMSEs, die relativen Verzerrungen und die relativen Dispersionen der GREG-Schätzungen für die amtlichen Einwohnerzahlen in allen Bundesländer 01,....,16. Zur Interpretation des Abbildungtyps siehe Kapitel 6. Die Codierung der Bundesländer ist in Abschnitt 7 zu finden.

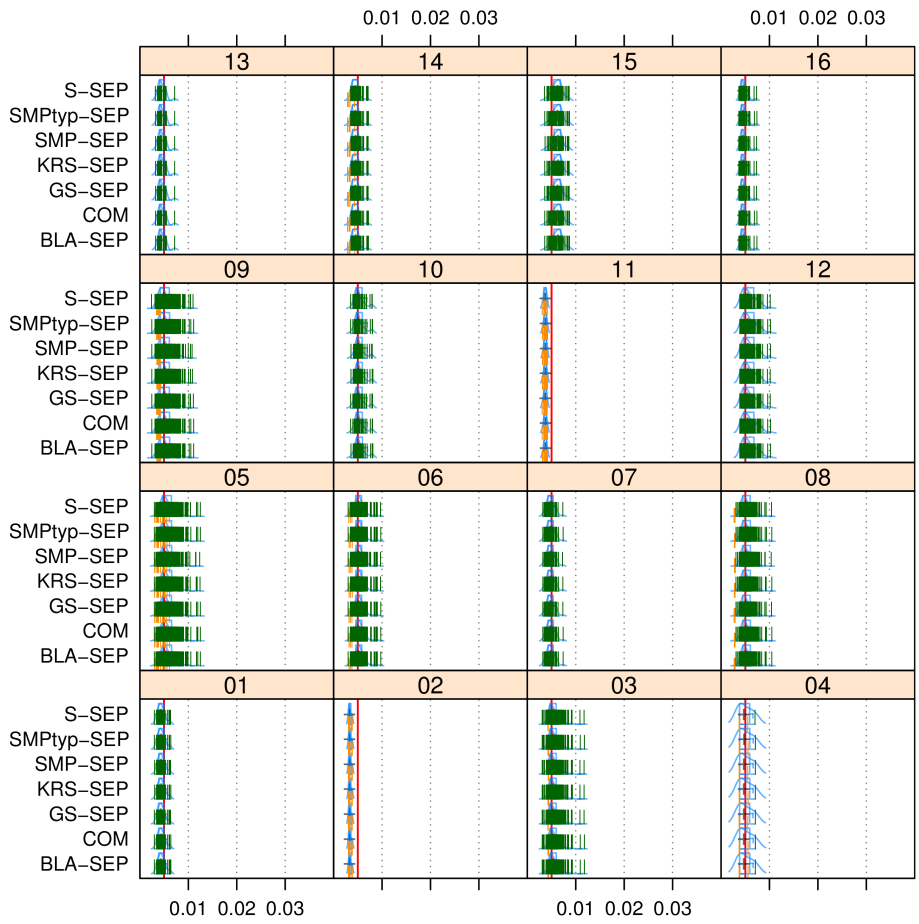


Abbildung 3.12: RRMSE der Schätzung der amtlichen Einwohnerzahl in allen 16 Bundesländern nach Modell I1

Zunächst wird nochmals ausdrücklich erwähnt, dass im Rahmen der *Ziel 1*-Untersuchungen insbesondere die SMP-Typen 0 und 1 sowie in Rheinland-Pfalz auch SMP-Typ 2 im Vordergrund stehen. Alle weiteren Fälle werden nicht den Präzisionsanforderungen ausgesetzt. Insgesamt zeigen sich kaum besondere Auffälligkeiten. Geringfügige Sensitivitäten können in einzelnen SMPs auftreten, zeigen sich aber insgesamt bei der Betrachtung der RRMSEs als kaum auffällig.

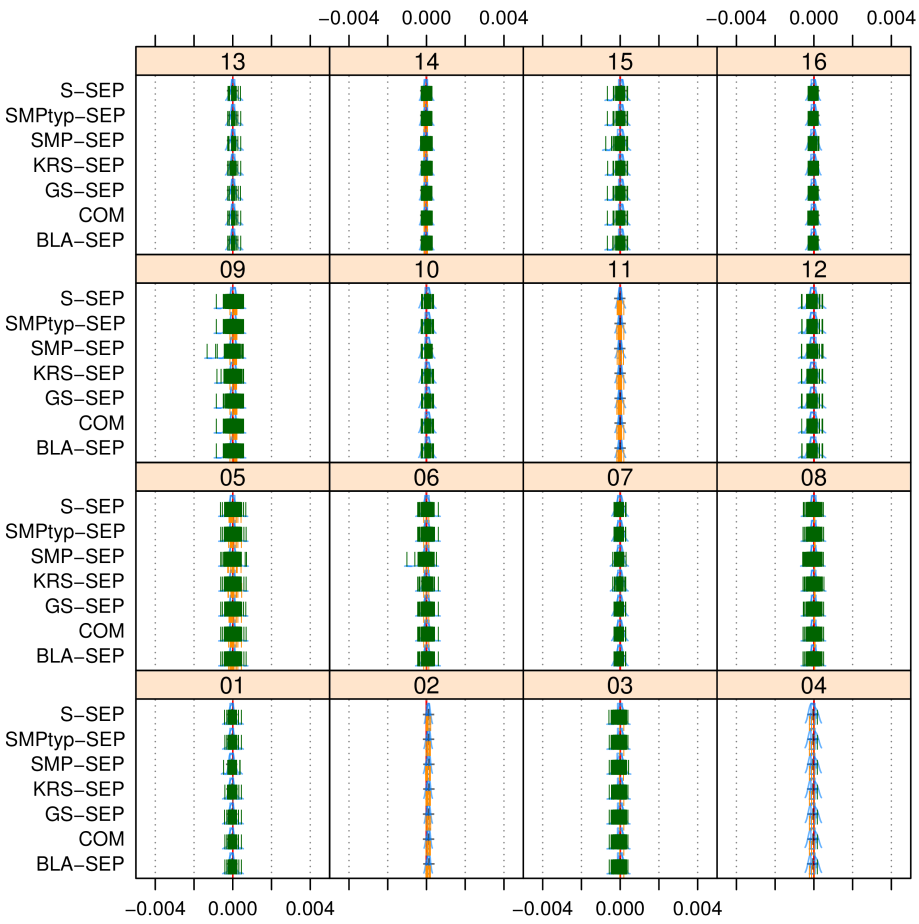


Abbildung 3.13: Relativer Bias der Schätzung der amtlichen Einwohnerzahl in allen 16 Bundesländern nach Modell I1

Insgesamt zeigt sich, dass kaum nennenswerte Unterschiede zwischen den verschiedenen Varianten auftreten. Detailuntersuchungen zur Empfindlichkeit der β-Schätzungen (siehe Abschnitt 2.3.2.2) legen jedoch nahe, keine Schicht-separaten Schätzungen zu verwenden. Unter den bestehenden Annahmen kann demnach auch der SMP-separate Regressionsschätzer verwendet werden.

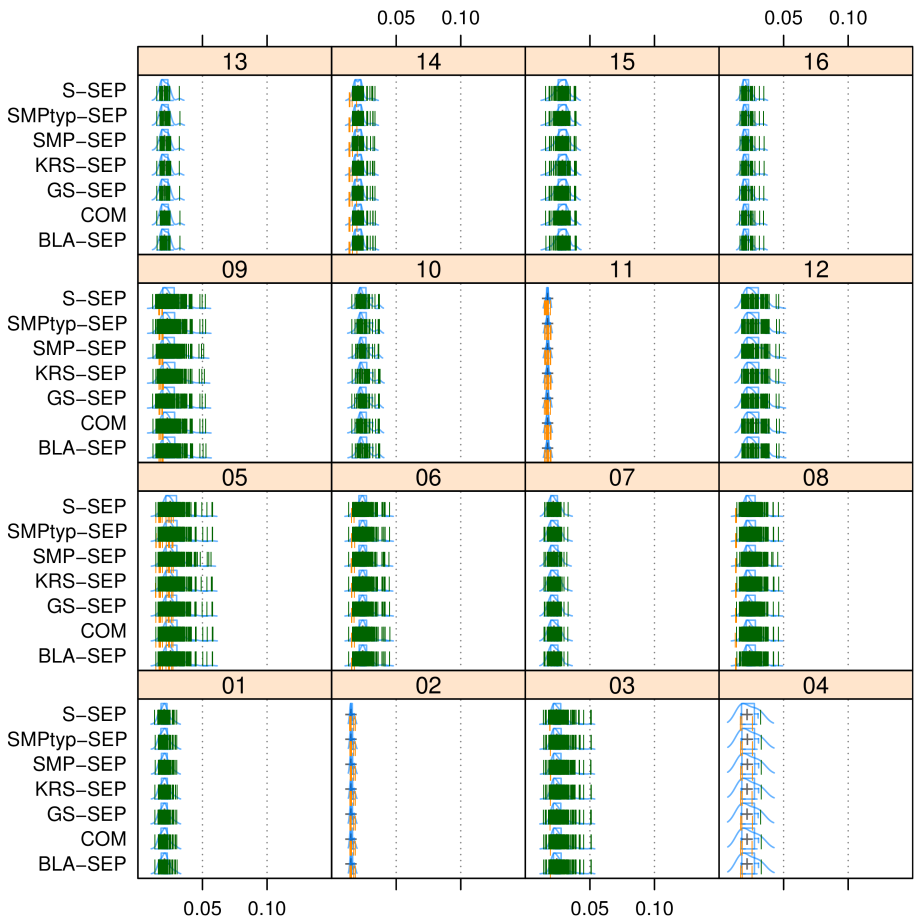


Abbildung 3.14: Relative Dispersion der Schätzung der amtlichen Einwohnerzahl in allen 16 Bundesländern nach Modell I1

Leichte Abweichungen zur theoretischen Betrachtung der Erfüllung der Präzisionsanforderungen von 0,5% ergeben sich aus dem Karteileichen- und Fehlbestandsmodell. Wie zu erwarten war, resultieren beim I1-Modell zum Teil hohe RMSEs, so dass bei einigen SMPs die Anforderungen nicht erfüllt werden. Das I1-Modell liefert von der Homogenitätsstruktur her zwar realistische Bestände, die aber insgesamt zu niedrigeren Korrelationen führen als sie von der amtlichen Statistik vorgegeben waren. Wesentlich besser würde sich ein theoretisches Modell erweisen, das die 0,993-Korrelationen strikt erfüllt. Solche Modelle in der Simulation zu implementieren, erweist sich indes als sehr problematisch, da die Korrelationen von Stichprobe zu Stichprobe stark variieren können. Alternative Modelle ergaben keine Verbesserungen. Lediglich sehr homogene Karteileichen- und Fehlbestandsmodelle, die für die große Simulation nicht flächendeckend vorlagen, zeigten Verbesserungen gegenüber dem I1-Modell.

Die Varianzschätzung liefert Ergebnisse, die man erwarten würde. Die Residualvarianzschätzungen leisten in allen Fällen akzeptable Ergebnisse, welche sich mit den Empfehlungen der Punktschätzungen decken.

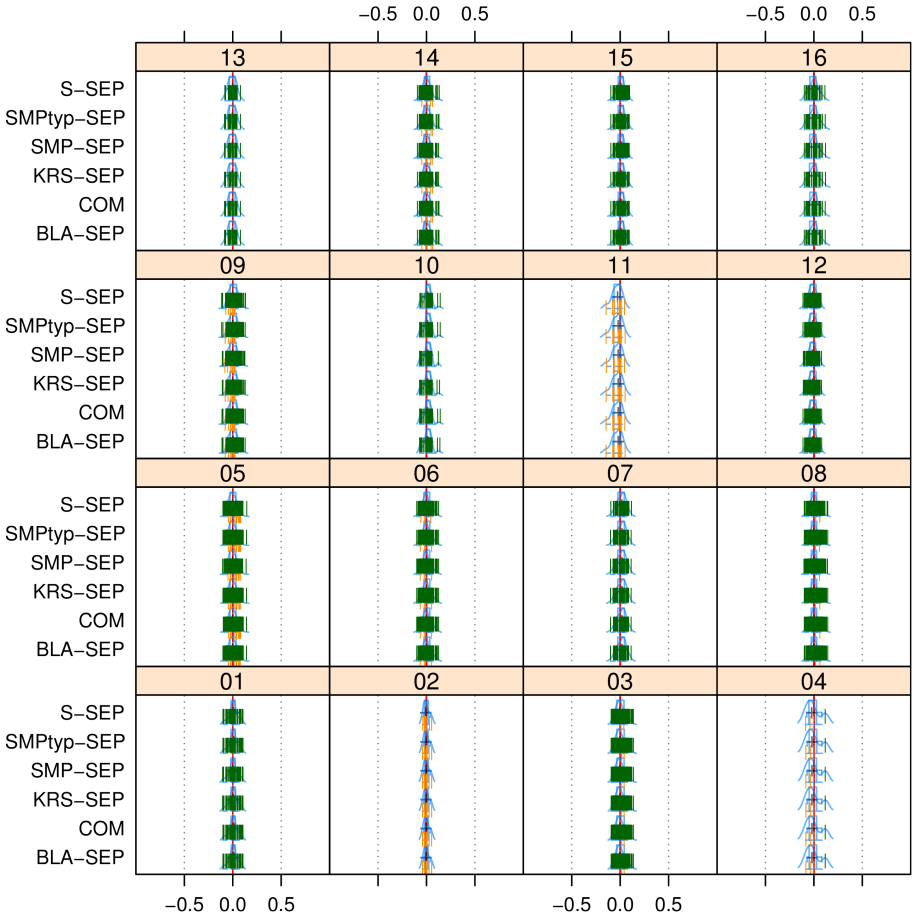


Abbildung 3.15: Relative Verzerrung der Varianzschätzungen zur Schätzung der amtlichen Einwohnerzahl in allen 16 Bundesländern bei Modell I1

Bei näherer Betrachtung sollten noch die Konfidenzintervallüberdeckungsraten herangezogen werden. In Abbildung 3.16 sind diese gegen die relativierten Konfidenzintervalllängen dargestellt. Ideal sind möglichst kurze Konfidenzintervalle, welche die nominale Überdeckungsrate exakt erfüllen. Wiederum zeigen sich hier nur geringfügige Unterschiede. Marginale Nachteile können hier beim SMP-separaten GREG beobachtet werden.

Zusammenfassend wird angemerkt, dass die Verwendung des konservativen GREG-Schätzers als besonders geeignet bewertet werden muss. Zwar ließen sich mit speziellen Small Area-Modellierungen geringfügig bessere Schätzergebnisse erzielen, die jedoch wegen der Design-Verzerrtheit

und der weniger präzisen Varianzschätzung in der Anwendung und Interpretation gewisse Probleme bereiten, die man bei *Ziel 1* besser vermeidet.

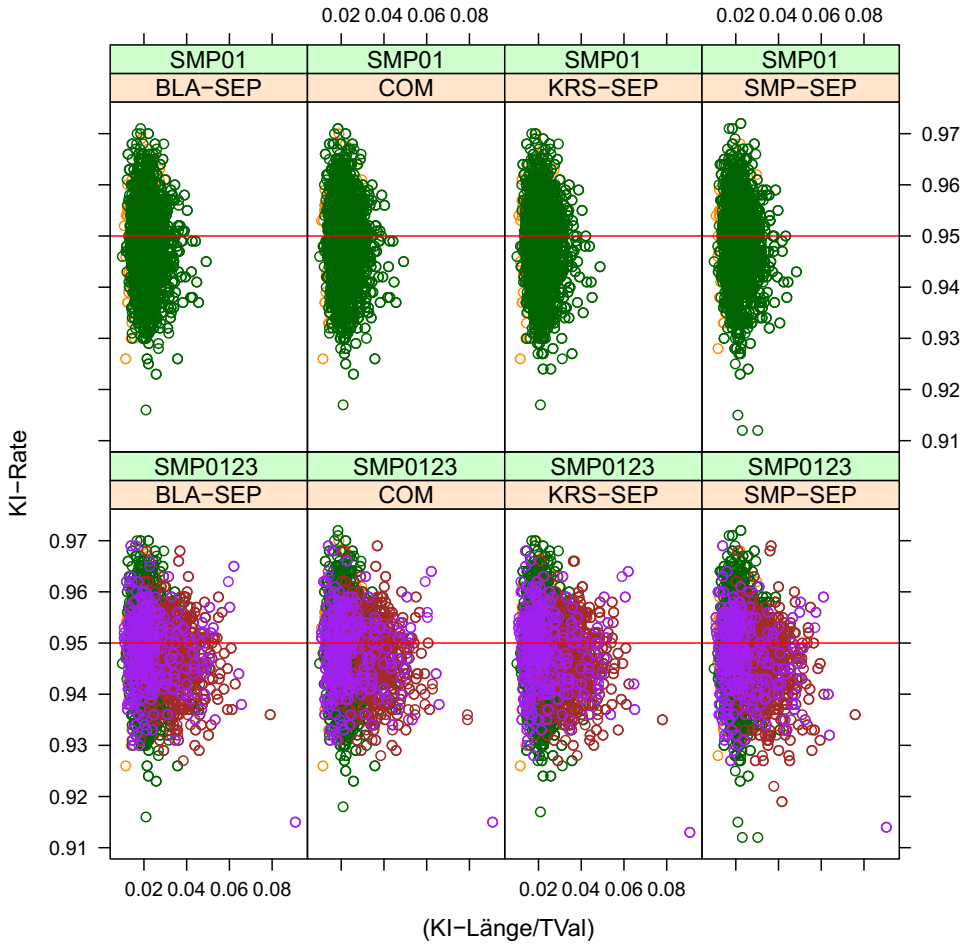


Abbildung 3.16: KI-Überdeckungsrate versus relative KI-Länge bei der Schätzung der amtlichen Einwohnerzahl beim Modell I1 ($1 - \alpha = 0{,}95$)

3.3.2 Schätzung von Karteileichen und Fehlbeständen

Die Schätzung von Karteileichen- und Fehlbeständen stellt sich als schwierig dar, da zu erwarten ist, dass deren Anteil in der Grundgesamtheit sehr gering ist. Weiterhin ist im Rahmen des Stichprobenforschungsprojekts keine Qualitätsanforderung in Bezug auf die Karteileichen- und Fehlbestandsschätzungen gestellt worden. Um die Kohärenz der Karteileichen- und Fehlbestandsschätzungen mit den Gesamtbevölkerungsschätzungen und Registerabzügen zu garantieren, wird eine dritte Kennzahlschätzung herangezogen. Die dritte zu schätzende Größe, die als eine Art Proxy zur Ermittlung der Karteileichen- und Fehlbestandswerte verwendet wird, ist die Zahl der paarigen Fälle. Paarige Fälle sind als Personen definiert, die sowohl im Register stehen als auch in der

Zensusbevölkerung vertreten sind. Mit anderen Worten: sie sind weder Karteileichen noch Fehlbestände.

In der Zensusbevölkerung sind sowohl die paarigen Fälle als auch die Fehlbestände enthalten, nicht aber die Karteileichen. Somit lassen sich die geschätzten Fehlbestände als Differenz der geschätzten Zensusbevölkerung und der geschätzten paarigen Fälle berechnen:

$$\widehat{\tau}_F^D = \widehat{\tau}_Z^S - \widehat{\tau}_P^S \tag{3.3.1}$$

In der Registerbevölkerung sind die Karteileichen, aber logischerweise nicht die Fehlbestände enthalten. Somit lassen sich die geschätzten Karteileichen aus der Differenz der Registerbevölkerung zur Zahl der geschätzten paarigen Fälle berechnen.

$$\widehat{\tau}_K^D = \tau_R - \widehat{\tau}_P^S \tag{3.3.2}$$

An die Schätzwertegenauigkeit bezüglich der Zahl der Karteileichen und Fehlbestände werden keine besonderen Anforderungen gestellt. Deshalb werden sie hier nur kurz behandelt.

In Abbildung 3.17 werden die RRMSEs der Karteileichen- und Fehlbestände je SMP (vertikaler Strich) in den 16 Bundesländern dargestellt. Wie zu sehen ist, sind die RRMSEs der Karteileichen über die verschieden gruppierten GREG relativ stabil zwischen 10 und 30%. Wenn die Qualitätsanforderung der *Ziel 2*-Variablen angesetzt würde (Anteil der Karteileichen ungefähr 3,5%), dann wären die SMPs mit einem RRMSE von unter 28,57% in einem akzeptablen Bereich. Hieran ist zu erkennen, dass das obige Vorgehen unter Berücksichtigung der kleinen Fallzahlen zu guten Ergebnissen für die Karteileichen- und Fehlbestandsschätzungen führt. Zu den Fehlbeständen sei hier angemerkt, dass sie zwar absolut gesehen höhere RRMSEs aufweisen als die Karteileichenschätzungen, jedoch in Relation zu ihrem Anteil unter Berücksichtigung der *Ziel 2*-Anforderungen noch günstiger abschneiden als die Karteileichenschätzungen. Die anderen Karteileichen- und Fehlbestandsmodelle liefern ähnliche Resultate.

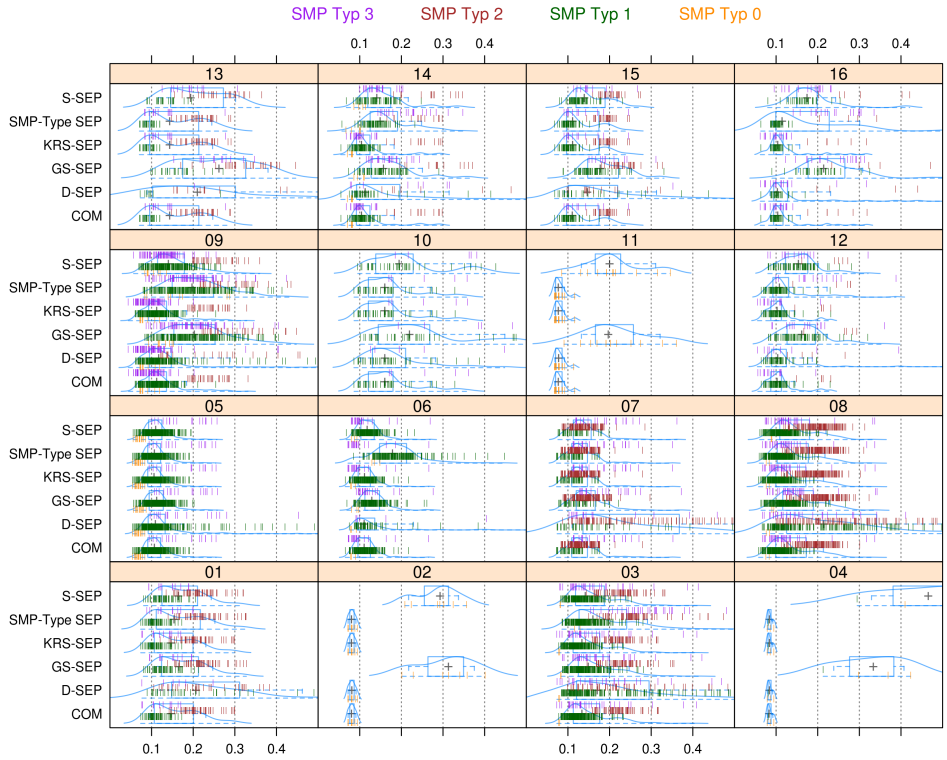


Abbildung 3.17: RRMSE der Karteileichenschätzung in allen 16 Bundesländern nach Modell I1

3.4 *Ziel 2*: Schätzung von Zusatzvariablen

Im folgenden Abschnitt werden nun die Ergebnisse der Simulation von *Ziel 2* Variablen vorgestellt. Die Simulationen wurden auf fünf Bundesländer begrenzt. Bei diesen Bundesländern handelt es sich um Nordrhein-Westfalen (NRW, Code 05), Rheinland-Pfalz (RLP, Code 07), Baden-Württemberg (BAW, Code 08), Berlin (BER, Code 11) und Mecklenburg-Vorpommern (MPV, Code 13). Mit dieser Auswahl wurde versucht, eine ausgewogene Mischung zwischen eher ländlich und eher städtisch geprägten Regionen sowie zwischen Ost- und Westdeutschland und zwischen Flächenbundesländern und Stadtstaaten zu gewährleisten. Schätzungen für Gesamtdeutschland hätten einen deutlich höheren Rechenaufwand vorausgesetzt. Das Grundproblem bei den *Ziel 2*-Fragestellungen ist, dass die Modelle bei einigen Fragestellungen zu wenig Erklärungsgehalt haben.

3.4.1 Schätzungen von Bildungsniveau gemäß ISCED

Für die Variable ISCED wurden sieben Fragestellungen untersucht, die in Tabelle 3.15 aufgeführt sind. Bei jeder Fragestellung handelt es sich um die Frage, wie viele Personen innerhalb einer Anschrift der vorgestellten Personengruppe angehören.

Tabelle 3.15: Fragestellungen bezüglich der Variable ISCED

Variable	Beschreibung
ISCEDA	ISCED - Stufen 1 und 2
ISCEDA1	ISCED - Stufe 1
ISCEDA2	ISCED - Stufe 2
ISCEDB	ISCED - Stufen 3 und 4
ISCEDB1	ISCED - Stufe 3
ISCEDB2	ISCED - Stufe 4
ISCEDC	ISCED - Stufen 5 und 6

Folgendes Kovariablen-Modell lag den Schätzungen zugrunde:

$$y \sim ADG + AGE1 + AGE2 + AGE3 + SEX \tag{3.4.1}$$

Dabei ist y eine Null-Eins-Variable. Entweder gehört die Person der in Tabelle 3.15 vorgestellten Personengruppe an oder nicht. Als Hilfsvariablen wurden die in Tabelle 3.16 aufgeführten Variablen verwendet. Dabei handelt es sich um die Anschriftengröße, drei verschiedene Altersklassen und um das Geschlecht SEX.

Tabelle 3.16: Hilfsvariablen zur Schätzung der ISCED Fragestellungen

Variable	Inhalt
ADG	Anschriftengröße
AGE1	Alter der Person ≤ 19
AGE2	$20 \leq$ Alter der Person < 40
AGE3	$40 \leq$ Alter der Person < 60
SEX	Geschlecht

In Abbildung 3.18 sind die Ergebnisse (RRMSE) der Schätzung für die drei Hauptfragestellungen bei der Variable ISCED (siehe Tabelle 3.15) zu sehen. In der oberen Zeile wird die Zugehörigkeit zur Gruppe C geschätzt. In der zweiten Zeile sind entsprechend die Ergebnisse für die Schätzung der Gruppe B und in der dritten Zeile diejenigen für die Gruppe A zu sehen. Die Genauigkeitsanforderung orientiert sich an dem Anteil der jeweiligen Personengruppe im Sampling Point.

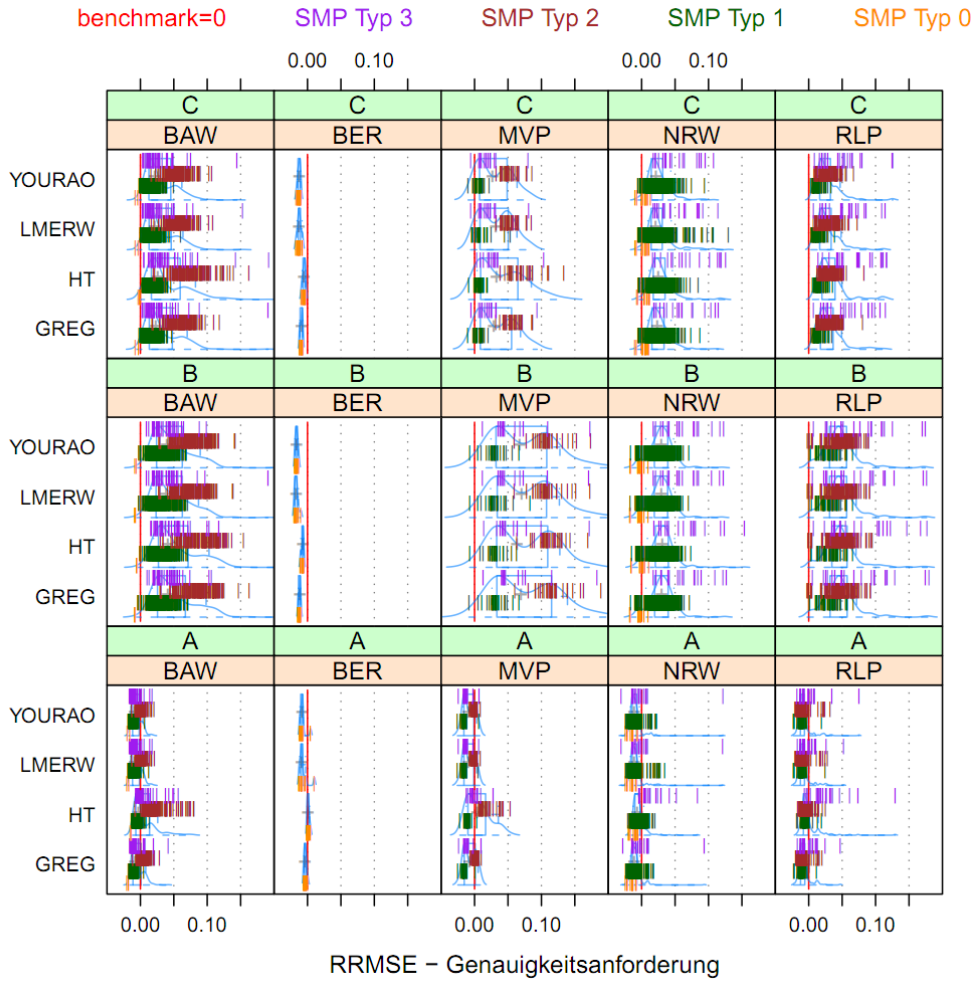


Abbildung 3.18: RRMSE - Schätzung ISCED

Die Nulllinie ist durch eine rote Linie gekennzeichnet. Es ist zu sehen, dass die durchschnittlichen Ergebnisse für die Schätzung der drei ISCED-Stufen in Berlin bis auf eine Ausnahme links der roten Linie sind. Sie erfüllen also die Genauigkeitsanforderungen. In den anderen Bundesländern ist der RRMSE zumeist höher als erlaubt. Besonders problematisch sind die Ergebnisse in den Kreisresten (SMP Typ 3). Die Ergebnisse für den SMP Typ 0 sind im Falle von Berlin besser als in Nordrhein-Westfalen. Das könnte darauf hindeuten, dass die Qualität der Ergebnisse mit der Größe der Sampling Points zunimmt.

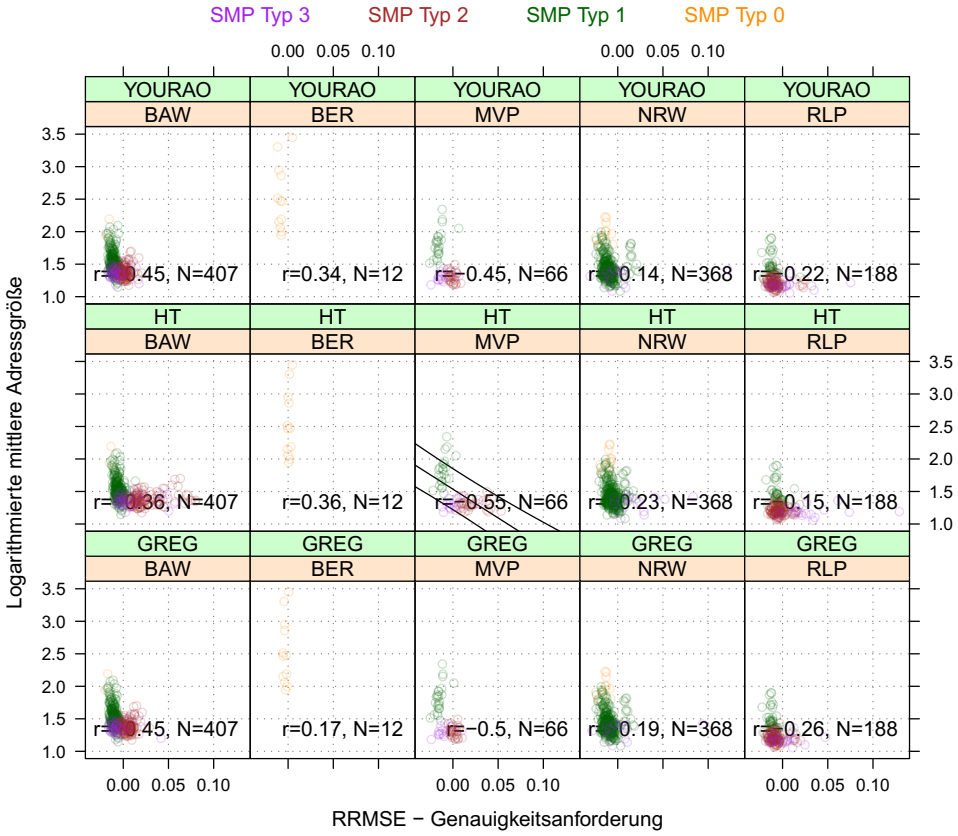


Abbildung 3.19: RRMSE nach SMP Typ und mittlerer Anschriftengröße - Schätzung ISCED

In Abbildung 3.19 ist die logarithmierte mittlere Anschriftengröße gegen den RRMSE abgebildet. Jeder Punkt steht für diese Kombination innerhalb eines SMPs, die SMPs sind farblich gekennzeichnet. Sofern ein signifikanter Zusammenhang zwischen den beiden Größen für ein Bundesland und einen Schätzer besteht, sind Regressionslinien eingezeichnet. Dies ist allerdings nur für den HT-Schätzer in Mecklenburg-Vorpommern der Fall. Der LMERW verhält sich hierbei genauso wie der YOURAO, sodass es genügt, letzteren darzustellen.

Ein Faktor, der einen größeren Einfluss auf den RRMSE hat, ist die Zahl der Anschriften in einem SMP. In Abbildung 3.20 ist die Korrelation zwischen der Zahl der Anschriften und dem RRMSE abgetragen. Die Regressionslinie ist nur für den Fall eingezeichnet, dass die Korrelation zwischen den beiden Merkmalen größer als 0,5 ist. Neben der Korrelation r ist noch die Zahl der Beobachtungen N abgetragen. Der Zusammenhang ist vor allem bei den Sampling Points des Typs 3 vorhanden (dritte Spalte). Für den Sampling Point vom Typ 1 findet man bei keinem Schätzer eine Korrelation, die größer als 0,5 ist.

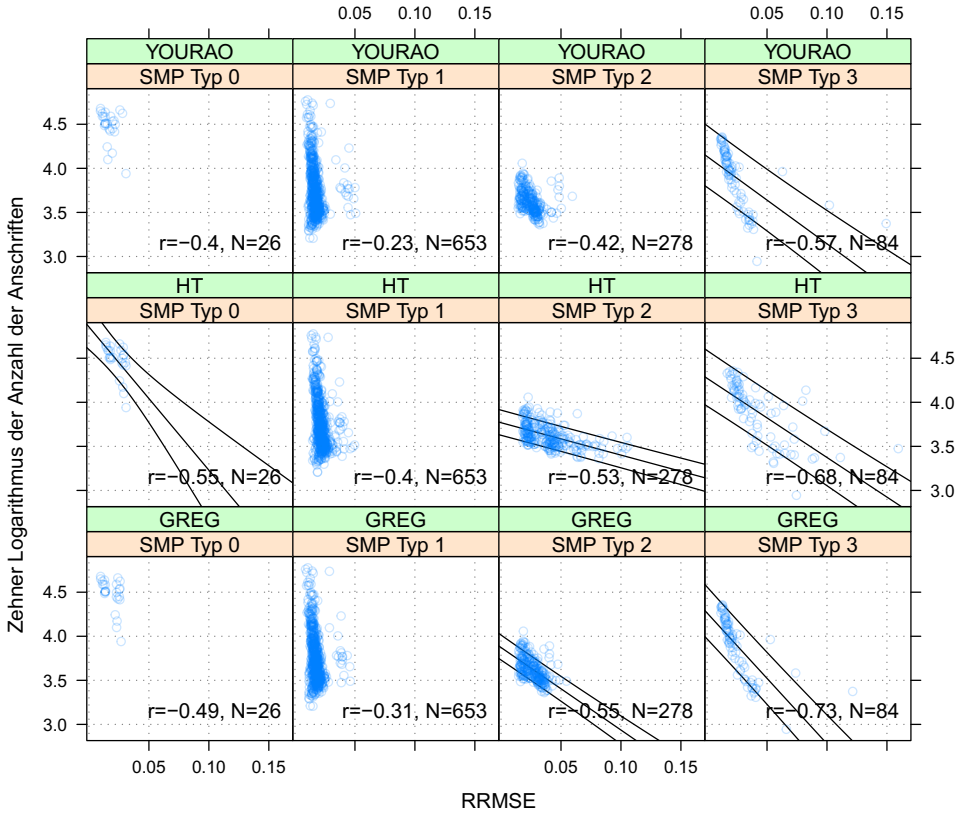


Abbildung 3.20: RRMSE nach SMP Typ und Zahl der Anschriften - Schätzung ISCED

In Abbildung 3.21 ist zu sehen, dass der RRMSE bei den synthetischen Schätzern hauptsächlich durch den Bias getrieben ist. Der relative Bias ist bei jedem Typ der ISCED Schätzung und in jedem Bundesland größer als der relative Bias beim Horvitz-Thompson- und beim GREG-Schätzer. Die Ergebnisse sind für die Fragestellung A am besten, die schlechtesten Resultate bezüglich des relativen Bias werden bei der Fragestellung C erzielt.

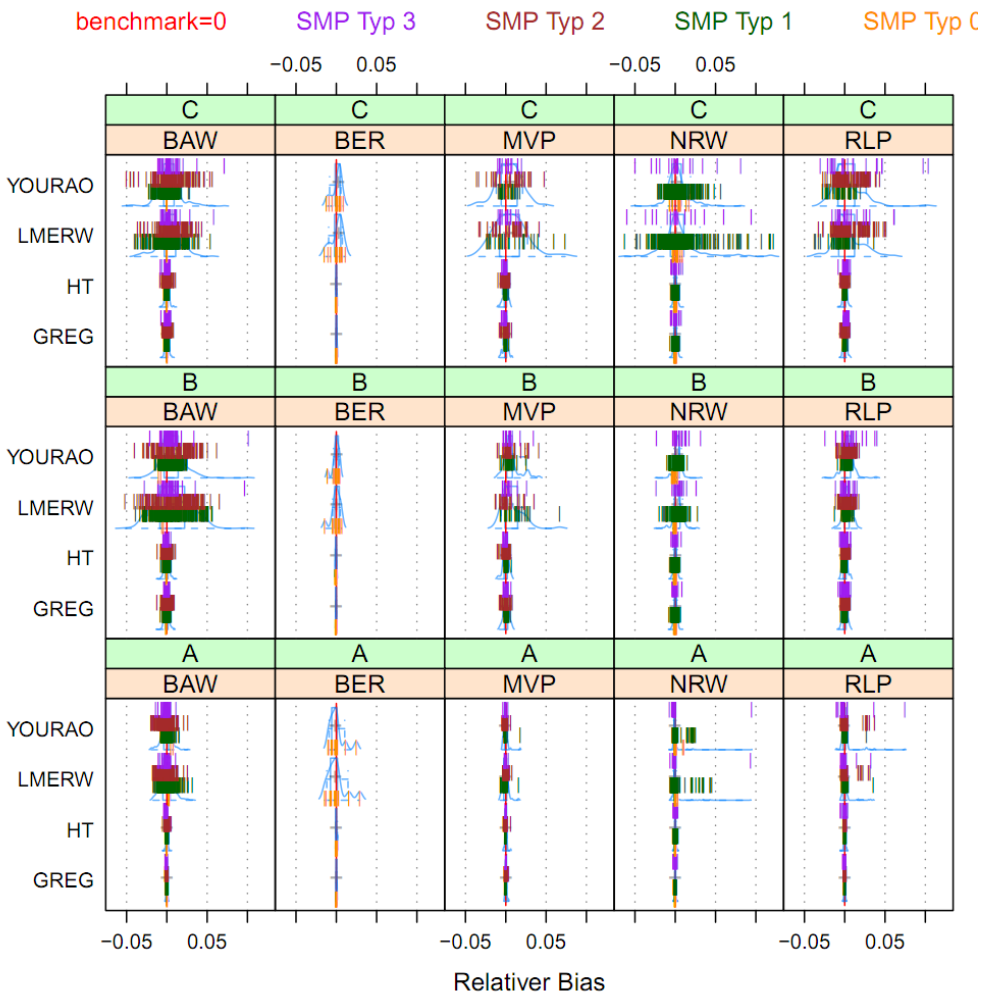


Abbildung 3.21: Relativer Bias - Schätzung von ISCED

3.4.2 Schätzungen von Erwerbstätigkeit gemäß ILO

Der zweite Typ von Fragestellungen im Bereich der Zusatzvariablen betrifft die Schätzung von Variablen, die mit dem Erwerbsprofil der Person zu tun haben. Im vorliegenden Fall handelt es sich um die Schätzung der Erwerbstätigkeit nach ILO Standard. Bei dieser Fragestellung ist zwischen sechs verschieden Fragen zu unterscheiden. Diese sind von ILO1 bis ILO6 durchnummeriert. Dabei bezeichnet die Zahl jeweils die Ausprägung, nach der ausgezählt wird. In der Fragestellung ILO1 wird also die Zahl der Erwerbstätigen geschätzt, bei der zweiten wird die Zahl der erwerbslosen Personen geschätzt, bei der dritten wird die Zahl der sonstigen Nichterwerbspersonen geschätzt und so fort.

Zur Bearbeitung dieser sechs Fragestellungen wird jeweils das folgende Modell verwendet:

$$y \sim ADG + AGE1 + AGE2 + AGE3 + SEX \tag{3.4.2}$$

Die zu schätzende Variable y ist für die erste Fragestellung eine Dummy-Variable, die den Wert 1s annimmt, wenn die Person erwerbstätig ist. Als Hilfsvariablen werden die in Tabelle 3.17 aufgeführten Variablen verwendet.

Tabelle 3.17: Hilfsvariablen zur Schätzung der ILO Fragestellungen

Variable	Inhalt
ADG	Anschriftengröße
AGE1	Alter der Person ≥ 15
AGE2	$18 \leq$ Alter der Person < 25
AGE3	Alter der Person > 65
SEX	Geschlecht

Bis auf die Anschriftengröße handelt es sich bei den verwendeten Hilfsvariablen um Dummy-Variablen.

In Abbildung 3.22 ist der RRMSE der Schätzungen für die Fragestellung ILO1 zu sehen. In einer Reihe sind jeweils die Ergebnisse der drei Schätzer für die fünf untersuchten Bundesländer abgetragen. Jeder Punkt entspricht dem Ergebnis in einem Sampling Point. Die Sampling Points wurden je nach Typ farblich gekennzeichnet. Auf der y-Achse ist die Zahl der Anschriften im Sampling Point abgetragen, während auf der x-Achse die Differenz zwischen dem zu erreichenden Benchmark und dem RRMSE der jeweiligen Schätzung eingezeichnet ist. Wenn der Punkt links des Ursprungs liegt, ist die Zielvorgabe erfüllt. Vor allem bei den Ergebnissen des Horvitz-Thompson-Schätzers in Baden-Württemberg (08) ist zu erkennen, dass die Sampling Points mit einer kleinen Zahl an Anschriften Probleme bereiten. Im Allgemeinen ist aber nicht von einem Einfluss der Zahl der Anschriften auf die Ergebnisse der Schätzung auszugehen.

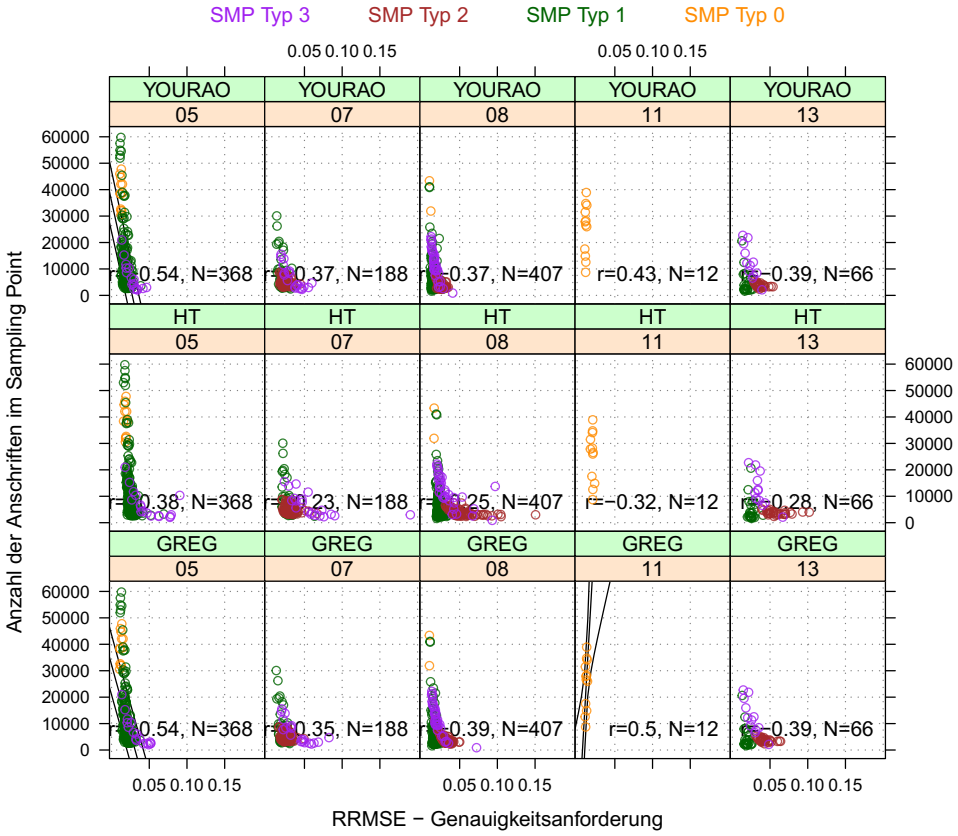


Abbildung 3.22: RRMSE für die Schätzung der Fragestellung ILO1

3.4.3 Schätzungen von Berufsgruppen

Bezüglich der Berufsgruppen wurde folgendes Modell geschätzt:

$$y \sim ADG + AGE1 + AGE2 + AGE3 + SEX \tag{3.4.3}$$

Dabei bezeichnet ADG wieder die Anschriftengröße. AGE1 ist eine Dummy-Variable, die den Wert 1 annimmt, wenn die betreffende Person älter als 15 Jahre ist und 0, wenn sie jünger als 15 Jahre oder genau 15 Jahre alt ist (siehe auch Tabelle 3.17). Die anderen beiden Altersvariablen sind ebenfalls Dummy-Variablen. AGE2 nimmt den Wert 1 an, wenn die Person mindestens 18 Jahre aber jünger als 25 Jahre ist. Die Variable AGE3 nimmt den Wert 1 an, wenn die betreffende Person über 65 Jahre alt ist. Hinzu kommt wieder die Variable SEX, das Geschlecht.

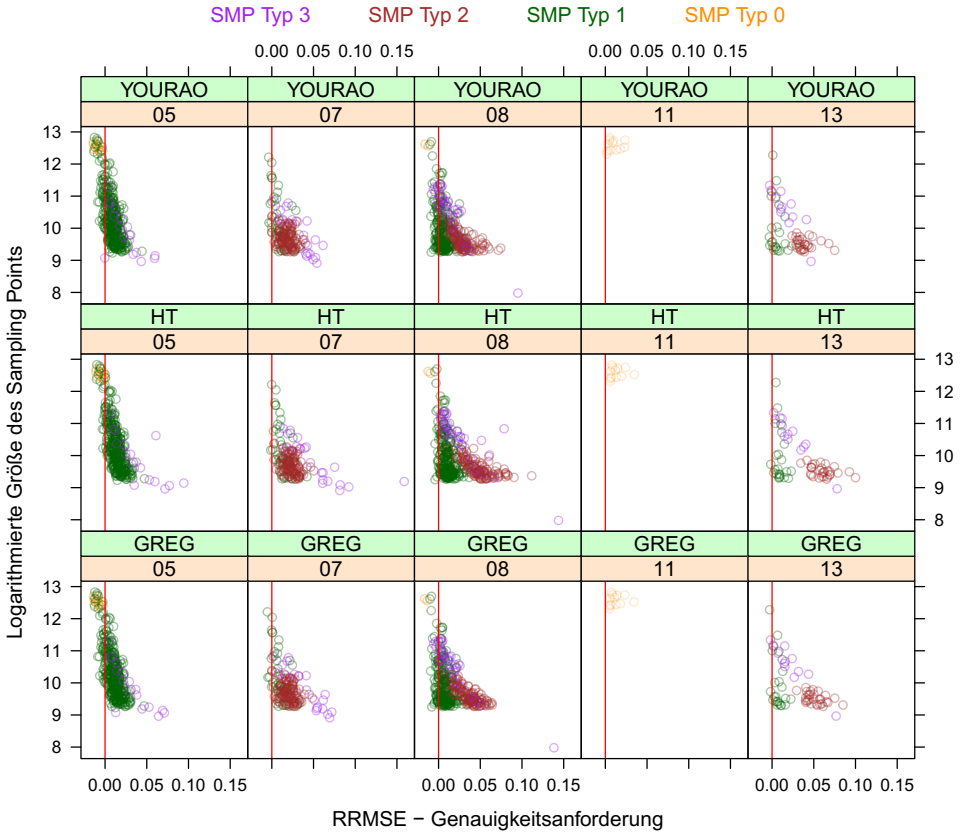


Abbildung 3.23: RRMSE - Schätzung von EF117A

In Abbildung 3.23 sind die RRMSEs für die Schätzung der Zahl der Angestellten in den fünf Bundesländern abgetragen. Es sind die Ergebnisse für die Schätzer YOURAO, HT und GREG zu sehen. Auf der y-Achse in jedem Panel ist die logarithmierte Zahl der Personen im Sampling Point abgetragen, während auf der x-Achse das durchschnittliche Ergebnis des RRMSE über 1.000 Simulationen abzüglich des Benchmarks abzulesen ist. Werte über Null bedeuten, dass der Schätzer die Genauigkeitsanforderung in der Area nicht erfüllt.

Zu sehen ist, dass in Nordrhein-Westfalen (05) vor allem die Schätzung in den SMPs vom Typ 3 problematisch ist.

3.4.4 Schätzungen von Zuzugsjahren

Bei der Schätzung von Zuzugsjahren soll das Jahr geschätzt werden, in dem Personen mit ausländischer Staatsbürgerschaft nach Deutschland zugezogen sind. Da die Schätzung für alle in Deutschland lebenden Nationalitäten sehr umfangreich wäre und bei einer Auszählung extrem viele Sampling Points mit strukturellen Nullen resultieren würden, wurde die Zahl der Personengruppen, für die Schätzungen zu liefern sind, auf fünf Gruppen reduziert (siehe Tabelle 3.18).

Tabelle 3.18: Fünf Teilgesamtheiten

Code	Nationaliät
T	Türken
G	Griechen
I	Italiener
E	Personen aus dem Gebiet ehemaliges Jugoslawien
O	Osteuropäer

Bei dieser Simulation war es vor allem wichtig herauszufinden, wie schnell die Schätzungen für solch kleine und regional stark unterschiedlich auftretende Häufigkeiten an ihre Grenzen stoßen. Bei der Schätzung ist zu beachten, dass es sich, wenn der Zuzug für Einzeljahre geschätzt werden soll, um ein seltenes Ereignis handelt.

$$y \sim ADG + AGE\_STA + SEX + GEB \tag{3.4.4}$$

ADG bezeichnet wieder die Anschriftengröße. Bei der Variable AGE_STA handelt es sich um eine Dummy-Variable, die den Wert 1 annimmt, wenn es sich um eine Person mit ausländischer Staatsangehörigkeit handelt, die älter als 15 Jahre ist. Hinzu kommt die Variable SEX (Geschlecht) und eine weitere Variable GEB. Diese setzt sich aus verschiedenen Dummy-Variablen zusammen, die den Wert 1 annehmen, wenn die betreffende Person der Teilgesamtheiten aus Tabelle 3.18 angehört und deren Alter größer als 2007 minus das zu schätzende Jahr ist.

Die Schätzung einzelner Zuzugsjahre in manchen Teilgesamtheiten hat sich hierbei als schwierig herausgestellt, da zum Teil sehr geringe Fallzahlen pro Jahr und Land sowie SMP auftreten. Das Problem bei diesen seltenen Ereignissen ist die Asymptotik. Zudem fällt eine sehr große Zahl an Schätzungen an, wenn jedes Jahr einzeln geschätzt wird. Aus diesem Grund wurde die Zahl der zugezogenen Personen je oben genannter Teilgesamtheit innerhalb eines 10 Jahres-Intervalls geschätzt.

Für einige Zeitintervalle und Gruppen von Nationalitäten resultieren trotz der oben beschriebenen Zusammenfassungen keine Ergebnisse. In Abbildung 3.24 wird der RRMSE für die Schätzung der türkischen Bevölkerungsgruppe gezeigt. Dabei haben die Überschriften der Panels die folgende Bedeutung:

Tabelle 3.19: Bedeutung der Überschriften in Abbildung 3.24

Überschrift	Bedeutung
	Zahl der Personen, die die türkische Staatsangehörigkeit haben und
T_1_1960	zwischen 1950 und 1960 nach Deutschland gezogen sind.
T_1_1970	zwischen 1960 und 1970 nach Deutschland gezogen sind.
T_1_1980	zwischen 1970 und 1980 nach Deutschland gezogen sind.
T_1_1990	zwischen 1980 und 1990 nach Deutschland gezogen sind.

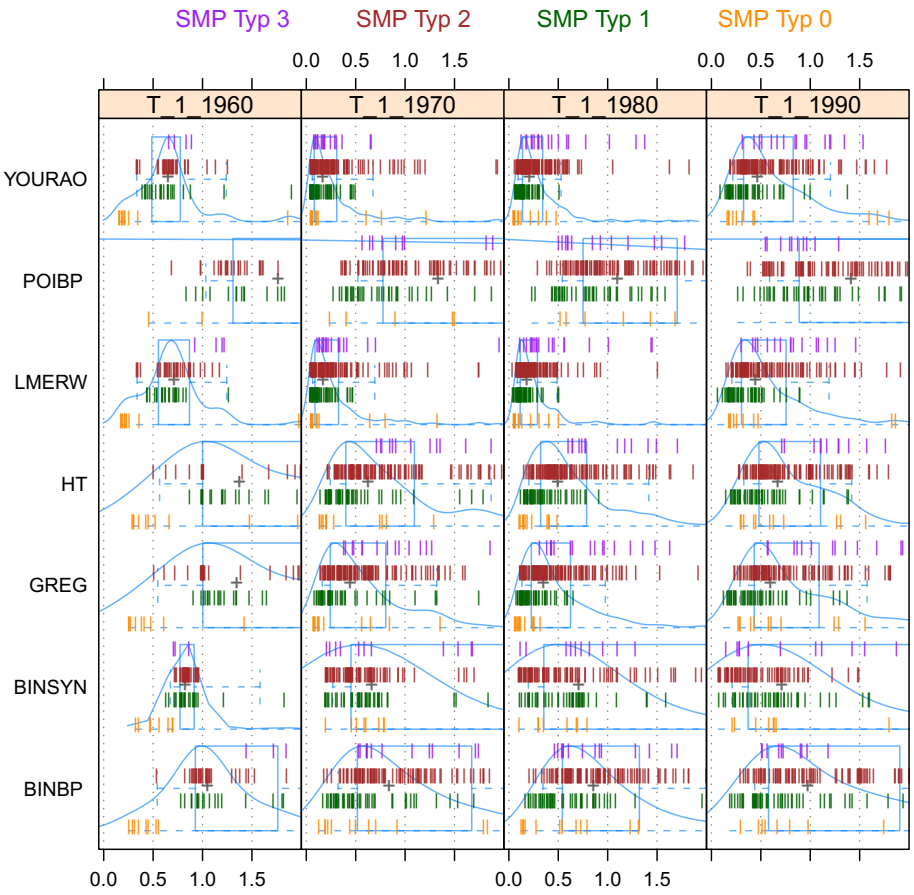


Abbildung 3.24: RRMSE für die Schätzungen von ausgewählten Zuzugsjahren

In Abbildung 3.24 ist zu sehen, dass die Schätzer YOURAO und LMERW für die Schätzung dieser Fragestellung am besten abschneiden.

Eine Übersicht über die Rechenzeiten, die für die Schätzungen der Fragestellung Zuzugsjahr benötigt werden, gibt die Abbildung 3.25. Es handelt sich dabei um die Zeit in Sekunden, die für einen Schätzer, für eine Stichprobe und für alle Domains benötigt wird. Dabei steht jeder Boxplot für einen Schätzer. Dargestellt sind die Rechenzeiten für das Bundesland Nordrhein-Westfalen.

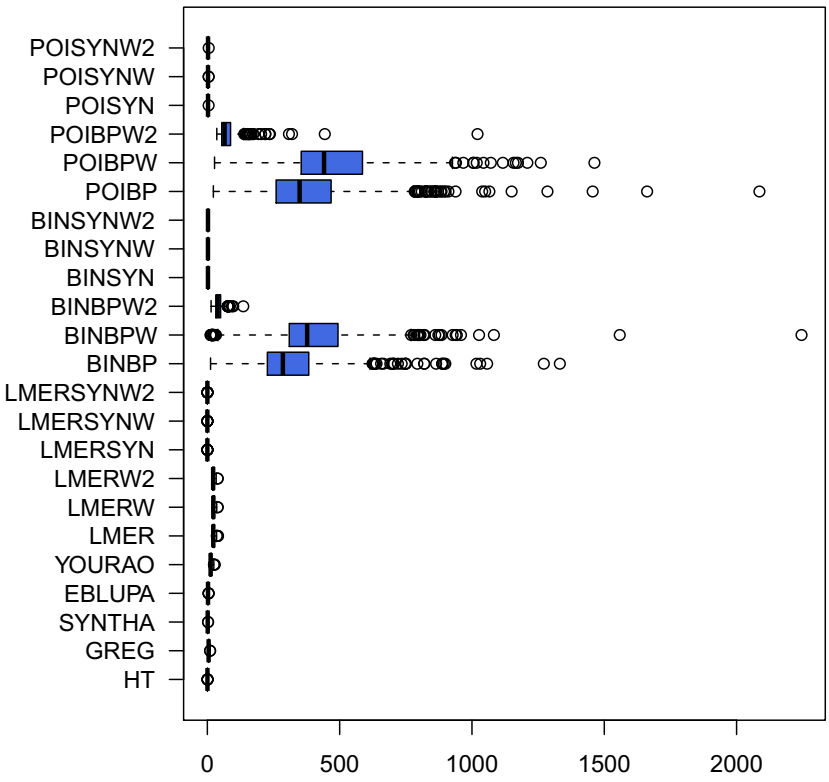


Abbildung 3.25: Rechenzeiten der Schätzungen für das Zuzugsjahr in den Simulationen in Nordrhein-Westfalen in Sekunden

Es wird deutlich, dass die Zeit für den Horvitz-Thompson-Schätzer am geringsten ist. Es ist weiterhin zu sehen, dass die Zeit, die für die gewichteten Schätzer (beispielsweise POIBPW oder BINBPW) benötigt wird, sehr stark variiert. Im Mittel (Median) braucht der POIBPW und der gewichtete Schätzer BINBPW am längsten.

3.5 Hypercube

3.5.1 Einleitung in die Schätzung von Hypercubes

Im Register-gestützten deutschen Zensus 2011 sollen neben einer Totalwertschätzung auf SMP-Ebene auch Schätzungen der Zellbesetzungen sogenannter *Hypercubes*, das sind mehrdimensio-

nale Häufigkeitstabellen, vorgenommen werden. Abbildung 3.26 verdeutlicht den Aufbau eines solchen Hypercubes am Beispiel von drei Variablen[21].

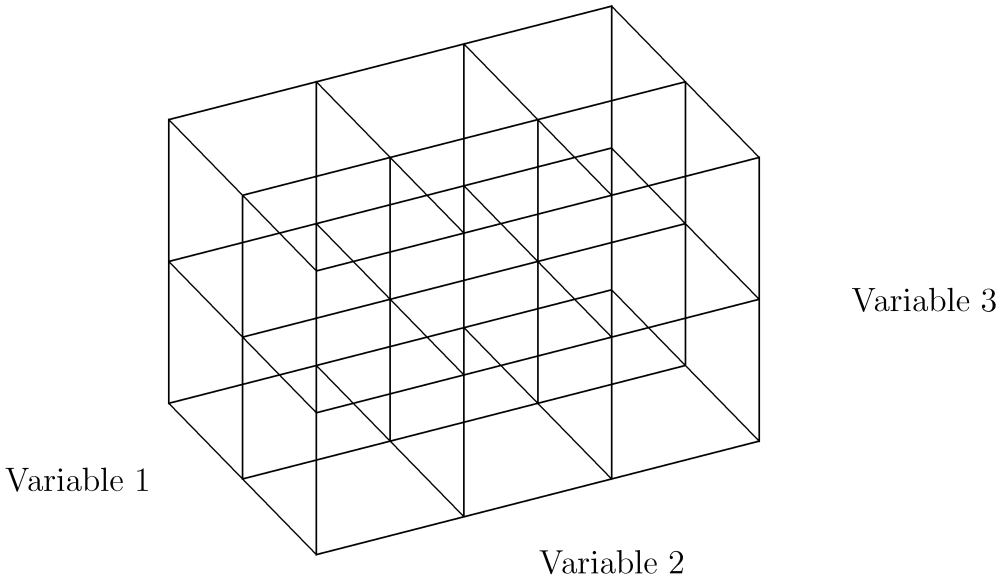


Abbildung 3.26: Schematischer Aufbau eines dreidimensionalen Hypercubes

Je nach Zusammensetzung der Variablen, die den Hypercube aufspannen, kann es dabei vorkommen, dass einzelne Zellen leer sind. Diese strukturellen Nullzellen sollte ein geeignetes Schätzverfahren berücksichtigen. Darüber hinaus ist eine weitere Anforderung an einen geeigneten Schätzer, dass er kohärente Ergebnisse hervorbringt. Das bedeutet etwa, dass die Summe aller Zell-Schätzungen bei Festhalten einer Dimension der unbedingten Randwertschätzung entspricht. In der folgenden Abbildung ergibt dementsprechend die Summe der Zellschätzungen der rot markierten Zellen den Totalwert der Randschätzung. Wenn die Variable 1 etwa dem Geschlecht entspricht, dann sollte ein geeigneter Hypercube-Schätzer so beschaffen sein, dass die Summe aller rot markierten Zellschätzungen gerade die Anzahl der Frauen ergibt.

Bei der Entscheidung für einen Hypercube-Schätzer gilt es weiterhin zu unterscheiden, ob *Ziel 1*- oder *Ziel 2*-Variablen den Hypercube aufspannen. Falls es sich bei den Hypercube-Variablen ausschließlich um *Ziel 1*-Variablen handelt, so kann der verallgemeinerte strukturerhaltende Schätzer (GSPREE) oder der χ^2-Schätzer (siehe Abschnitt 2.4.2) verwendet werden, wobei die Anpassung an die Struktur der Registerwerte in den Zellen erfolgt. Gehen jedoch zu *Ziel 1* auch *Ziel 2*-Variablen in den Hypercube ein, so wird in den Zellen entweder an die Struktur einer höheren Aggregat-Ebene angepasst und strukturerhaltend geschätzt oder eine Kombination von direktem Schätzer und strukturerhaltendem Schätzer verwendet. Drei mögliche Schätzer werden in den folgenden Abschnitten näher beschrieben und ihre Güte anhand von Simulationsstudien im darauf folgenden Abschnitt bewertet.

[21] Aus Gründen der Darstellung werden hier nur drei Variablen verwendet, im Beispiel sind das später Geschlecht, höchster allgemeiner Schulabschluss (EF310) und überwiegender Lebensunterhalt (EF401). Prinzipiell kann ein Hypercube ein hochdimensionales Gebilde sein.

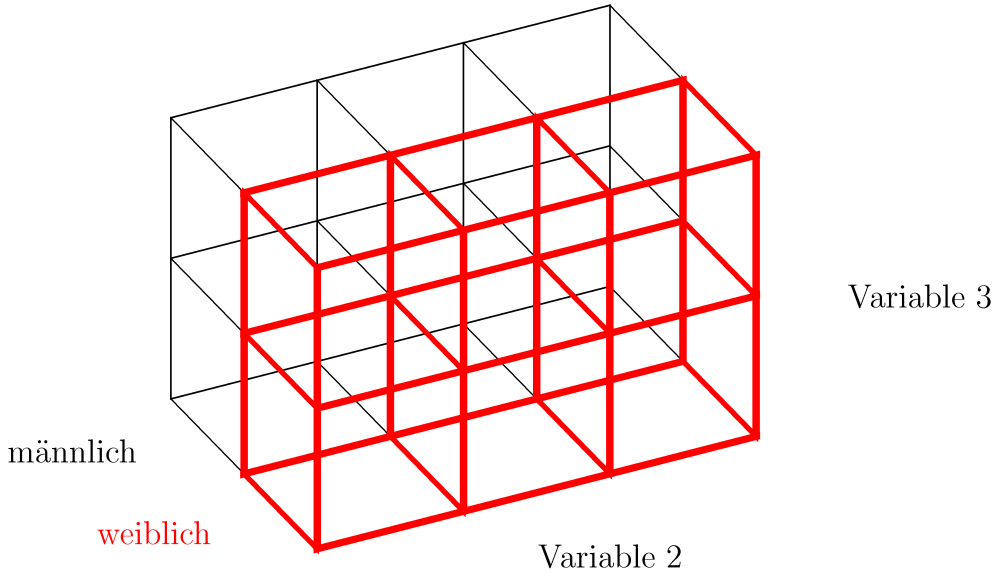


Abbildung 3.27: Ausschnitt (rot) aus dem dreidimensionalen Hypercube

Die Entscheidung für einen der drei oben genannten Schätzer hängt von der Ebene ab, auf der verlässliche Schätzungen der Zusammenhangstruktur (engl. „association structure“) zur Verfügung stehen. Im Fall des deutschen Zensus kann zwischen zwei Möglichkeiten unterschieden werden. Liegen für die einzelnen Zellen verlässliche Schätzungen auf NUTS2-Ebene direkt vor, etwa bei *Ziel 1*-Variablen, dann lässt sich das Vorgehen bei der Bestimmung der Zusammenhangsstruktur wie in Abbildung 3.28 dargestellt beschreiben.

Liegen dagegen keine verlässlichen Schätzungen auf NUTS2-Ebene in den Zellen sondern nur auf den Rändern vor, so muss die Zusammenhangsstruktur in den Zellen auf der nächst höheren Ebene, also dem Bundesland, geschätzt werden. Abbildung 3.29 beschreibt das Vorgehen grafisch.

3.5.2 Ergebnisse *Ziel 1*

Die Punkte in den folgenden Abbildungen geben für jede Hypercube-Zelle die RRMSEs des χ^2-, des GREG und des YOURAO-Schätzers der amtlichen Einwohnerzahl (τ_Z), Karteileichen (τ_K) und Fehlbestände (τ_F) für einen von den Variablen Geschlecht, Nationalität, Familienstatus und Altersklasse aufgespannten vierdimensionalen Hypercube in den Bundesländern Rheinland-Pfalz und Berlin wieder. Die Zusammenhangsstruktur entspricht der Matrix M aus den Erläuterungen in Abschnitt 2.4.2.3. Als Randschätzer auf SMP-Ziel-Ebene wird der GREG-Schätzer verwendet. Die den Hypercube aufspannenden Variablen haben die in den Tabellen 3.20 und 3.21 wiedergegebenen Ausprägungen. Damit ergibt sich in jeder SMP eine maximale Anzahl von $4 \times 4 \times 7 = 112$ Hypercube-Zellen. Diese Zahl kann jedoch kleiner ausfallen, wenn in einer SMP nicht alle Hypercube-Zellen besetzt sind. Auf der x-Achse sind die aus der Grundgesamtheit der Simulationsstudie bekannten Hypercube-Zellumfänge abgetragen – *kleine Zellen* befinden sich also links auf der x-Achse, *große Zellen* dagegen rechts. In Rheinland-Pfalz kann der Hypercube so aus maximal $112 \times 188 = 21\,056$ Zellen bestehen.

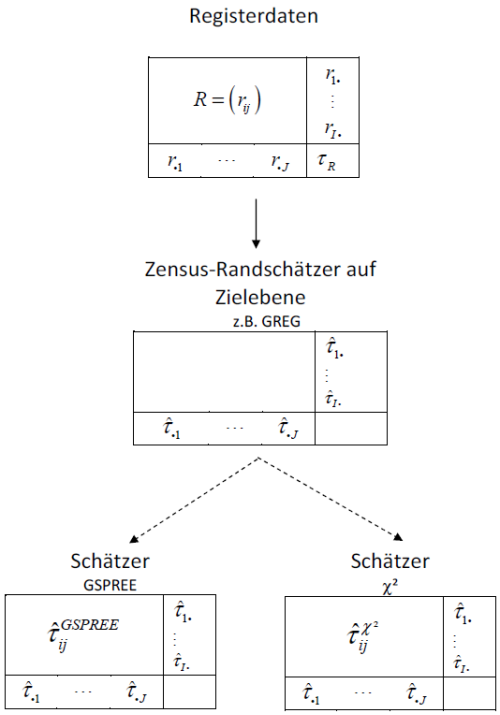


Abbildung 3.28: Schematischer Ablaufplan bei Register als Zusammenhangsstruktur

Bei der Schätzung von τ_Z in den Zellen des Hypercubes zeigt sich in Abbildung 3.30, dass der GREG bei Zellen mit großem τ_Z einen sehr kleinen RRMSE aufweist. Sind die Zellumfänge klein, weist der χ^2-Schätzer im Verhältnis dazu für einige kleine Zellen einen sehr kleinen RRMSE auf (zu erkennen an der Klumpung der Punkte nahe der x-Achse im Bereich von $\tau_Z = 1000$ bis ca. $\tau_Z = 2500$). Auch bei der Schätzung von Karteileichen zeigt sich in Abbildung 3.31, dass der χ^2-Schätzer wiederum vor allem bei einigen kleinen Zellen einen sehr kleinen RRMSE aufweist, in einigen Zellen jedoch auch deutlich größere RRMSEs hat als der GREG. Der YOURAO-Schätzer liegt zwischen dem GREG- und dem χ^2-Schätzer. Bei der Schätzung von Fehlbeständen kann man aus Abbildung in 3.32 erkennen, dass sich im Großen und Ganzen ein sehr ähnliches Bild ergibt wie bei der Schätzung von Karteileichen. Allerdings weisen alle Schätzer ein höheres Niveau des RRMSEs auf, was an der kleineren Schätzgröße liegt.

Insgesamt lässt sich festhalten, dass mit einer Zunahme der Schätzgröße von τ_F über τ_K nach τ_Z eine Verschiebung der Punktwolke in Richtung der x-Achse auftritt, die RRMSEs also kleiner werden.

Die oben gemachten Aussagen zeigen sich auch in Berlin, wie den Abbildungen 3.33 bis 3.35 zu entnehmen ist.

Zusammenfassend lässt sich sagen, dass keiner der Schätzer alle anderen Schätzer dominiert. In großen SMPs bzw. Zellen sollte der GREG-Schätzer verwendet werden. Geht es um eine Minimierung des maximal erwarteten RRMSEs in sehr kleinen Zellen, so kann der YOURAO-Schätzer empfohlen werden. Im folgenden Abschnitt wird zudem ein kombinierter Schätzer aus GREG- und χ^2-

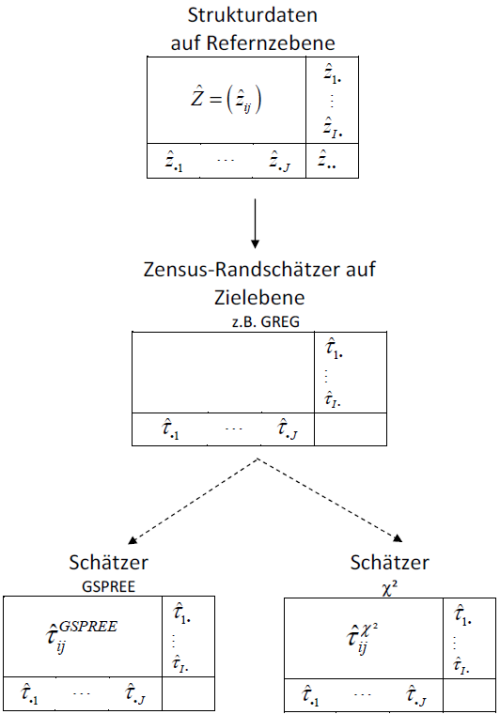


Abbildung 3.29: Schematischer Ablaufplan bei Schätzung auf Bundesland-Ebene als Zusammenhangsstruktur

Schätzer betrachtet, der die Ergebnisse weiter stabilisiert. Der GREG-Schätzer sollte ein umso höheres Gewicht erhalten, je größer die Zellpopulation ist.

3.5.3 Ergebnisse *Ziel 2*

Die folgenden Abbildungen geben die RRMSEs des χ^2-, des GREG-, des GREG-kombinierten χ^2- und des YOURAO-Schätzers für einen dreidimensionalen Hypercube aufgespannt von den Variablen Geschlecht, höchster allgemeiner Schulabschluss EF310 und überwiegender Lebensunterhalt EF401 in den Bundesländern Nordrhein-Westfalen, Rheinland-Pfalz und Baden-Württemberg wieder. Die beiden letztgenannten Variablen sind *Ziel 2*-Variablen, die auf der NUTS2-Ebene nicht verlässlich genug geschätzt werden können. Daher wird die Zusammenhangsstruktur auf Bundesland-Ebene geschätzt (s. Abbildung 3.29). Die Zusammenhangsstruktur entspricht der Matrix M aus den Erläuterungen in Abschnitt 2.4.2.3. Als Randschätzer auf NUTS2-Ebene wird der GREG-Schätzer verwendet. Die den Hypercube aufspannenden Variablen haben die in Tabelle 3.22 gegebenen Ausprägungen.

In den einzelnen Panels der Abbildungen 3.36–3.38 sind jeweils die $2\times7\times8 = 112$ RRMSEs des χ^2-, des GREG- und des kombinierten Schätzers in Abhängigkeit von den amtlichen Einwohnerzahlen τ_Z Werten dargestellt. Darüber hinaus ist die Toleranzgrenze für die geforderte Präzision als grüne Linie eingezeichnet. Werte auf der x-Achse im Bereich der grünen Linie entsprechen Anteilswerten unter $1/15 = 6,67\,\%$, an die keine expliziten Präzisionsanforderungen gestellt werden. Dies trifft in den folgenden Abbildungen auf fast alle Zellen in den genannten Bundesländern zu.

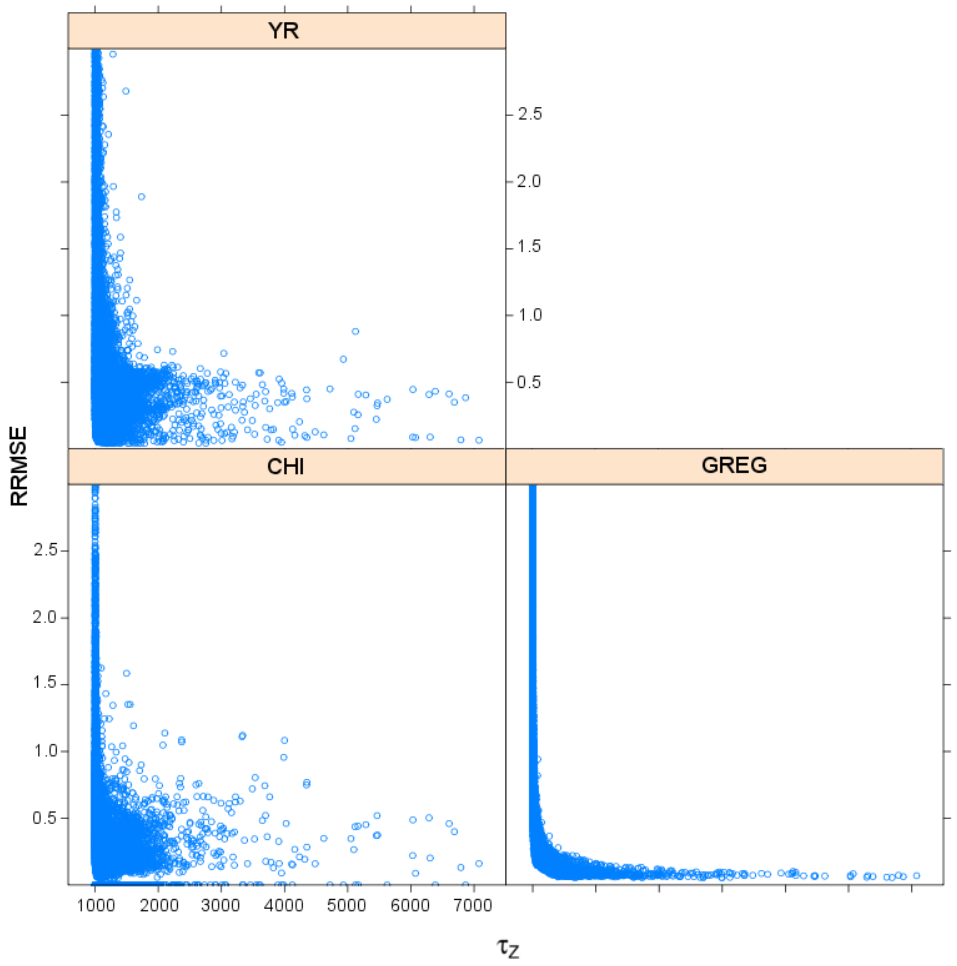


Abbildung 3.30: RRMSEs der drei Schätzer für die amtliche Einwohnerzahl τ_Z in den SMPs von Rheinland-Pfalz

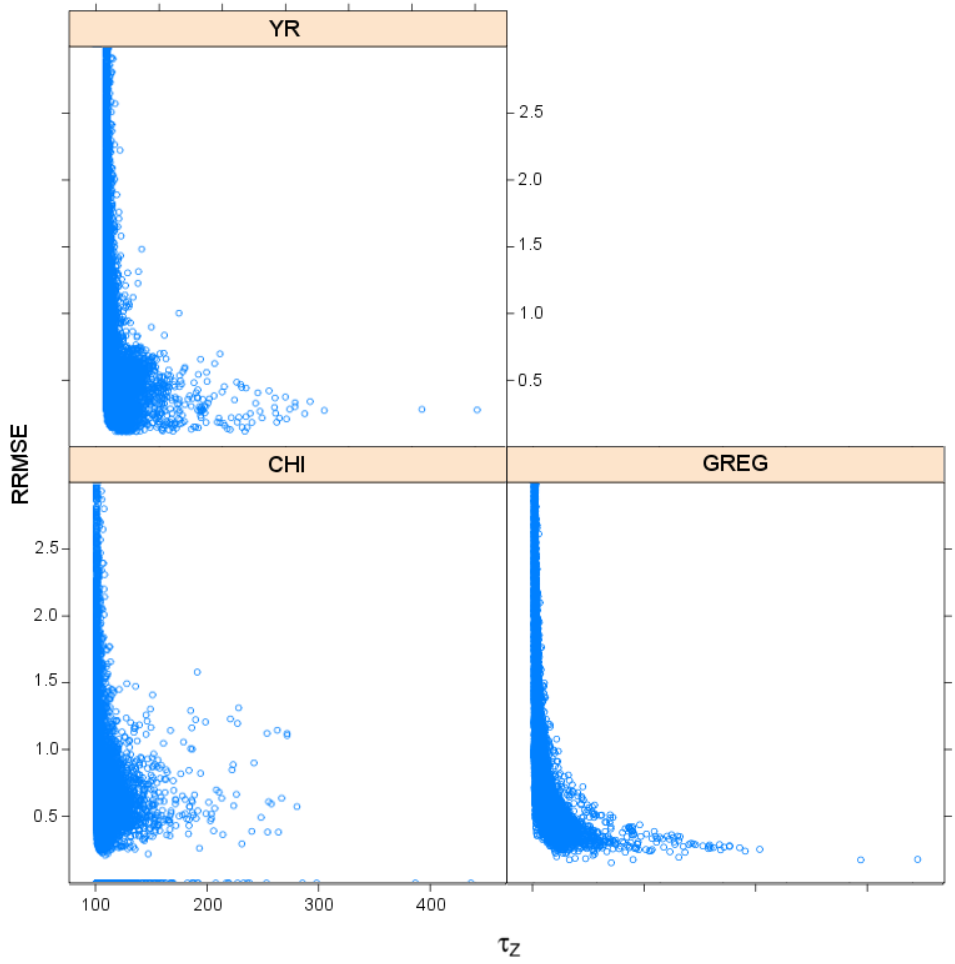


Abbildung 3.31: RRMSEs der drei Schätzer für Karteileichen τ_K in den SMPs von Rheinland-Pfalz

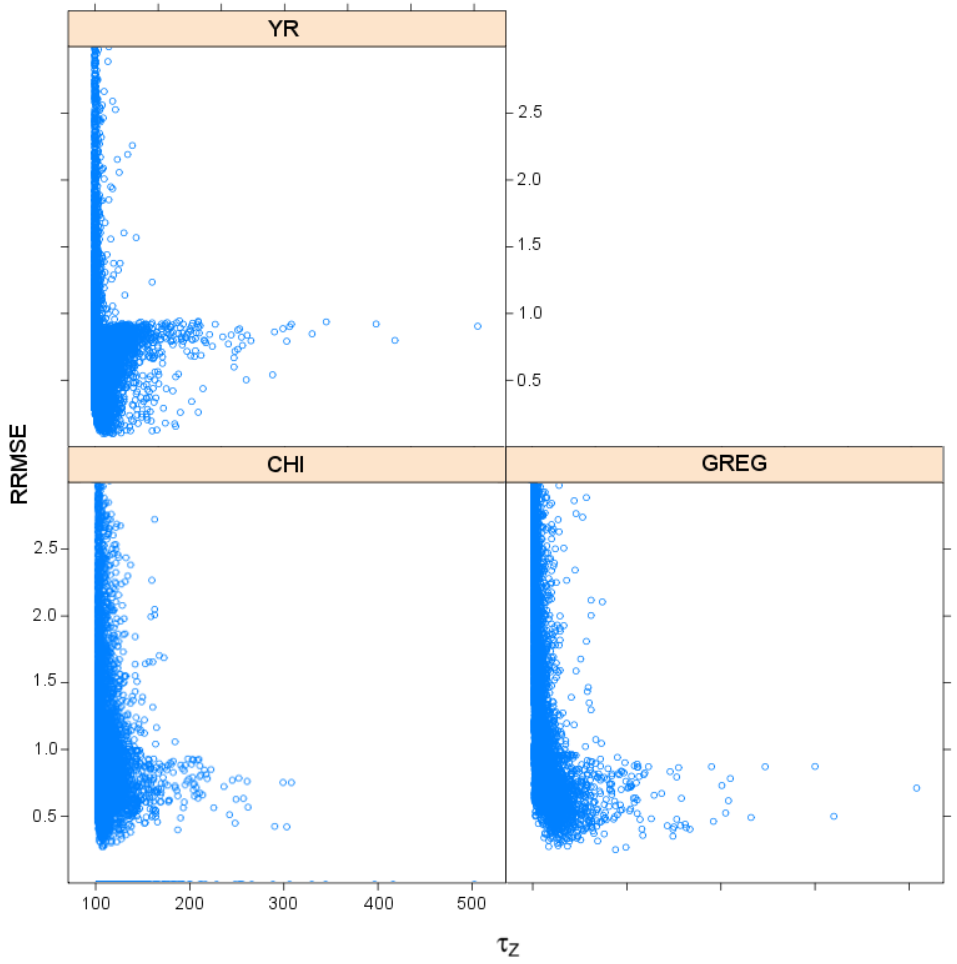


Abbildung 3.32: RRMSEs der drei Schätzer für Fehlbestände τ_F in den SMPs von Rheinland-Pfalz

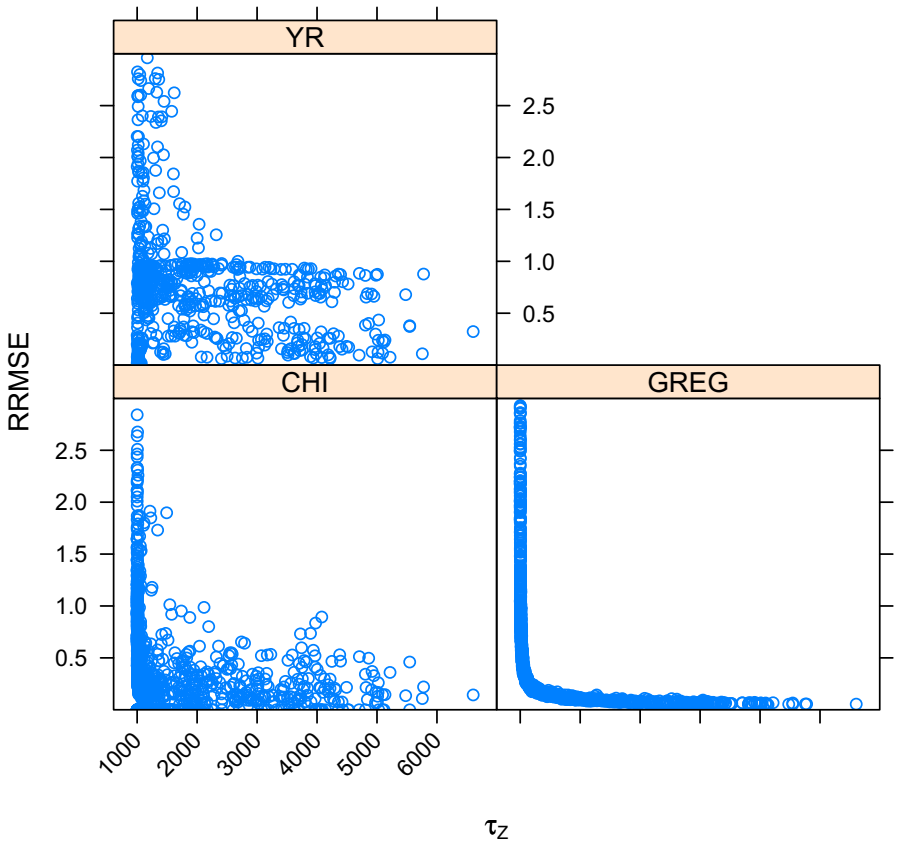


Abbildung 3.33: RRMSEs der drei Schätzer für die amtliche Einwohnerzahl τ_Z in den SMPs von Berlin

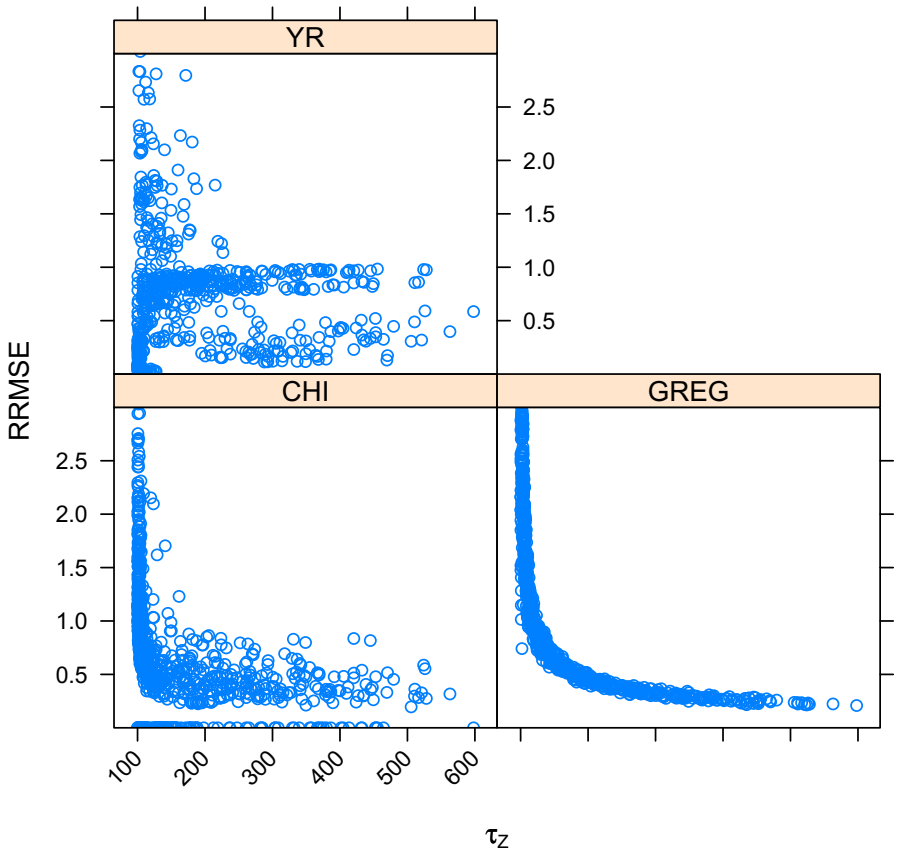


Abbildung 3.34: RRMSEs der drei Schätzer für Karteileichen τ_K in den SMPs von Berlin

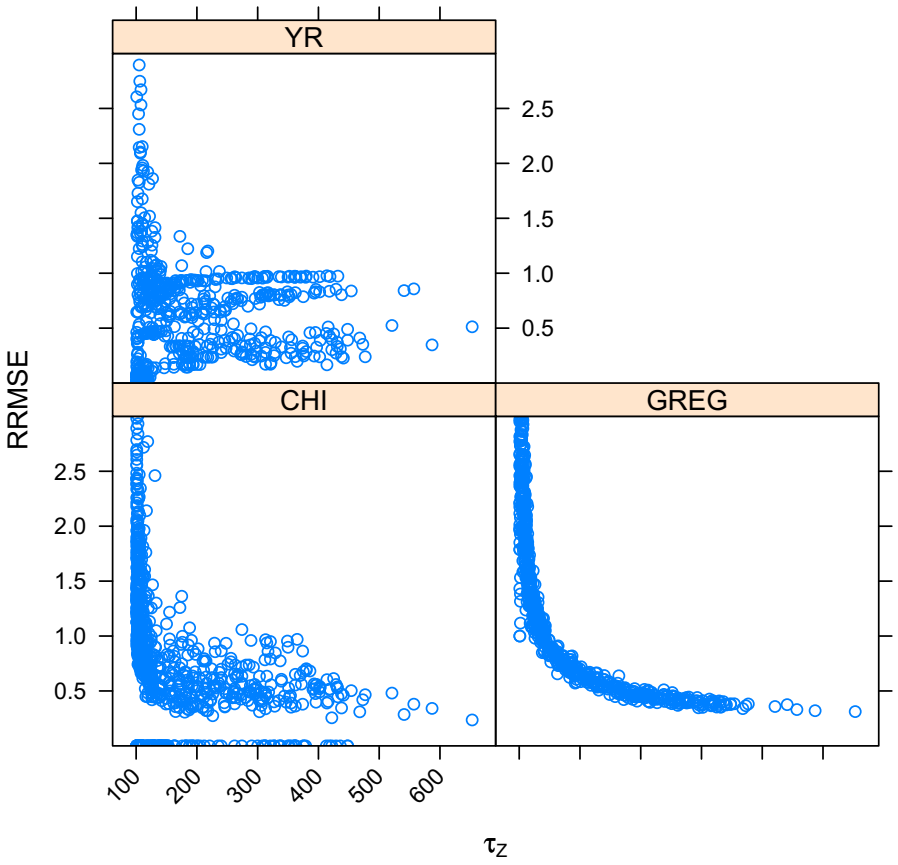


Abbildung 3.35: RRMSEs der drei Schätzer für Fehlbestände τ_F in den SMPs von Berlin

Tabelle 3.20: Übersicht über die Ausprägungen der den Hypercube aufspannenden Variablen für *Ziel 1* zur Schätzung von Karteileichen

SEX / NAT	FAM	AKL
männlich/deutsch	ledig	unter 18
männlich/nicht deutsch	verheiratet	18 bis unter 25
weiblich/deutsch	verwitwet	25 bis unter 30
weiblich/nicht deutsch	geschieden	30 bis unter 40
		40 bis unter 50
		50 bis unter 65
		65 oder älter

Tabelle 3.21: Übersicht über die Ausprägungen der den Hypercube aufspannenden Variablen für *Ziel 1* zur Schätzung von Fehlbeständen

SEX / NAT	FAM	AKL
männlich/deutsch	ledig	unter 18
männlich/nicht deutsch	verheiratet	18 bis unter 30
weiblich/deutsch	verwitwet	30 bis unter 50
weiblich/nicht deutsch	geschieden	50 oder älter

In Abbildung 3.36 können wir sehen, dass die RRMSEs des GREG-Schätzers in allen NUTS2-Gebieten wie bei *Ziel 1* stark von der Schätzgröße abhängt. Der χ^2-Schätzer ist im Vergleich zu allen anderen Schätzern dann am kleinsten, wenn die Zellen eher klein sind, weist jedoch hier auch unter Umständen einzelne hohe Werte auf, z. B. in NUTS2-Gebieten 1, 3, 7 und 9. Der kombinierte Schätzer (COMB) liegt über alle Zellen zwischen dem GREG- und dem χ^2-Schätzer.

Ein ähnliches Bild wie in Nordrhein-Westfalen ergibt sich auch in Rheinland-Pfalz. Hier fällt die einzige vorkommende Verletzung der Präzisionsanforderungen in der Zelle 111 (männlich/Haupt-(Volks-)schulabschluss/Erwerbs- (Berufs-)tätig) auf, von der alle Schätzer betroffen sind.

Auch in Baden-Württemberg ist zu erkennen, dass in allen NUTS2-Gebieten alle Schätzer die Präzisionsanforderungen deutlich übererfüllen. In NUTS2-Gebiet 2 und 4 fällt auf, dass der χ^2- und der kombinierte Schätzer kleine Ausreißer aufweisen, die jedoch in Zellen mit Anteilswerten unter $1/15 = 6,67$ % vorkommen, an die keine expliziten Präzisionsanforderungen gestellt werden. Erfreulicherweise erfüllen in fast allen Zellen alle Schätzer die Präzisionsanforderungen, selbst da, wo sie explizit nicht gefordert werden, weil der Anteil in der Gesamtheit zu gering ist. Dies ist auch nicht weiter verwunderlich, da es sich bei der Untergliederung bei *Ziel 2* um NUTS2-Gebiete handelt, die eine größere Menge an Anschriften umfassen als die SMPs bei *Ziel 1*.

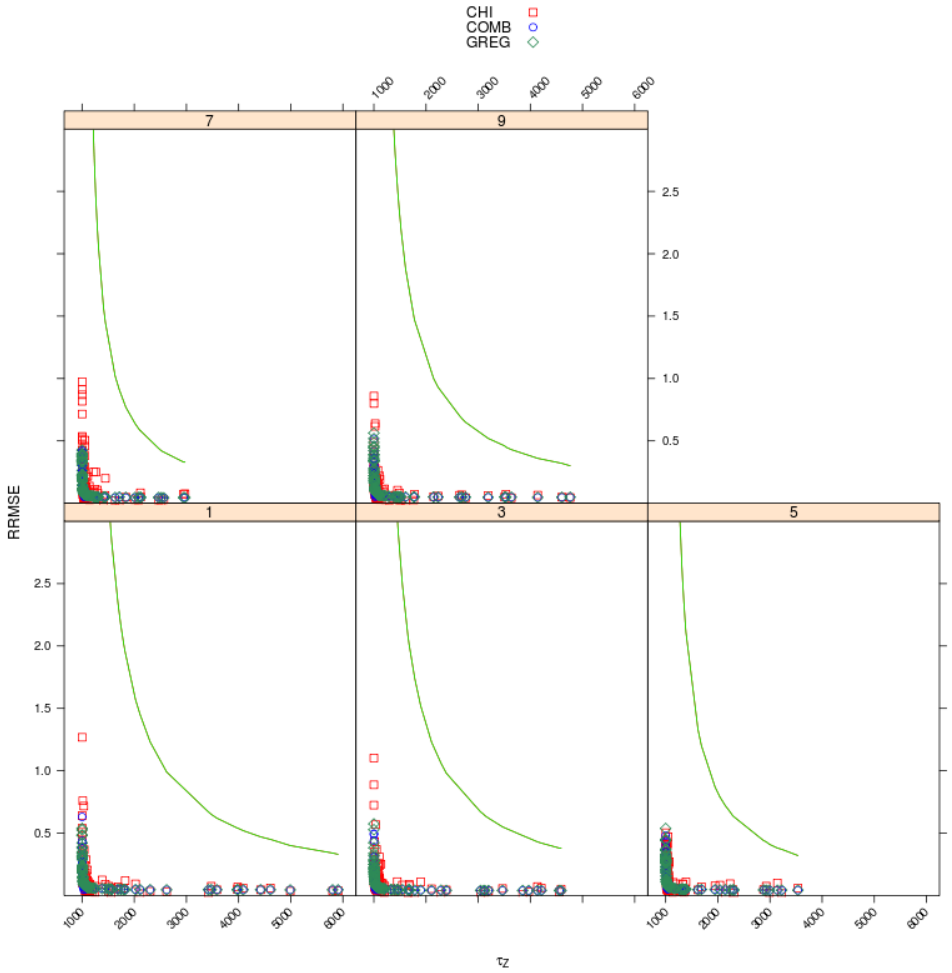


Abbildung 3.36: RRMSEs des χ^2-, GREG- und kombinierten Schätzers in den fünf NUTS2-Gebieten in Nordrhein-Westfalen

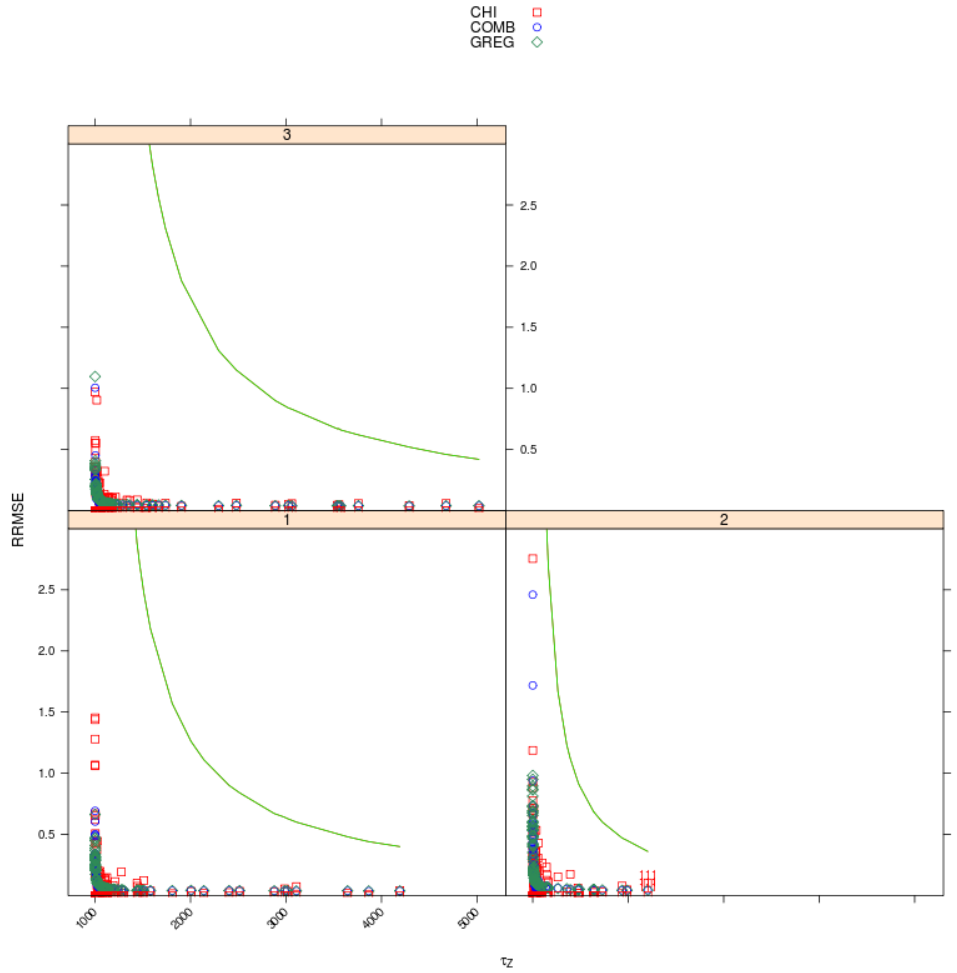


Abbildung 3.37: RRMSEs des χ^2-, GREG- und kombinierten Schätzers in den drei NUTS2-Gebieten in Rheinland-Pfalz

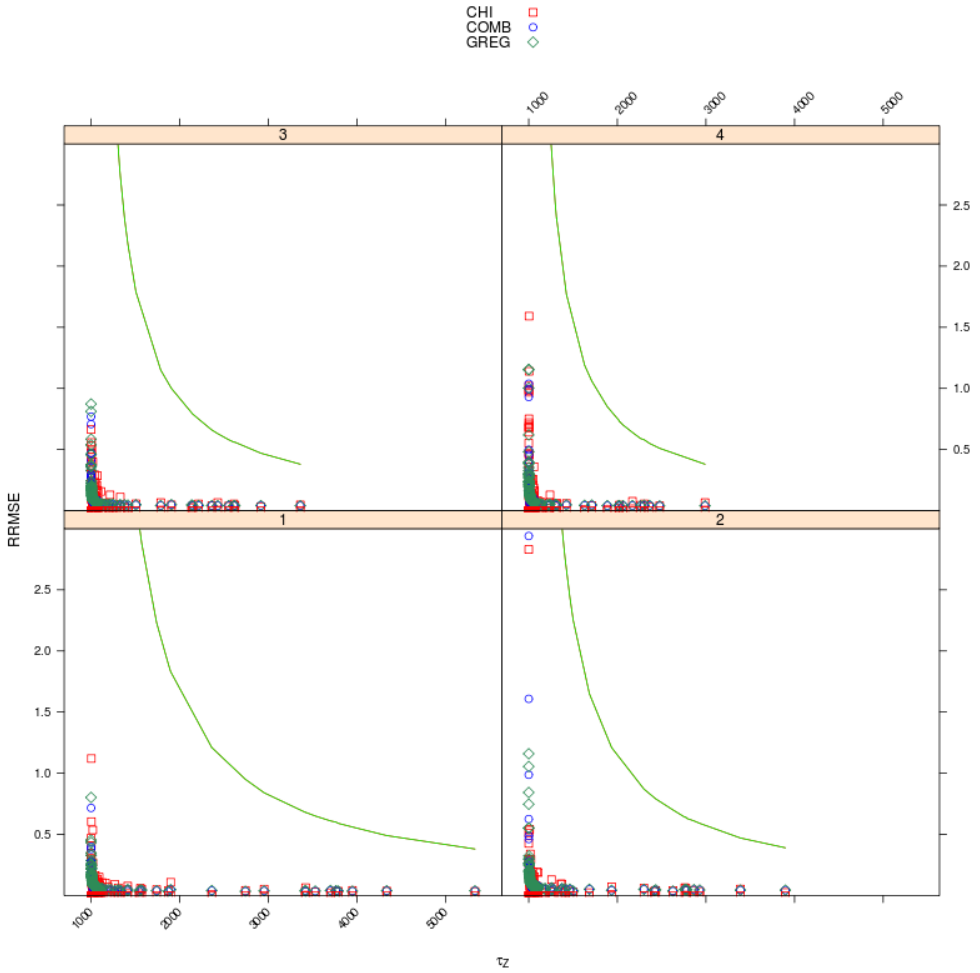


Abbildung 3.38: RRMSEs des χ^2-, GREG- und kombinierten Schätzers in den vier NUTS2-Gebieten in Baden-Württemberg

Tabelle 3.22: Übersicht über die Ausprägungen der den Hypercube aufspanndenen Variablen

SEX	EF310	EF401
männlich	Haupt- (Volks-)schulabschluss	Erwerbstätig/Berufstätig
weiblich	Abschluss der allgemeinbildenden Polytechnischen Oberschule der DDR	Arbeitslosengeld I, II
	Realschulabschluss (Mittlere Reife) oder gleichwertiger Abschluss	Rente, Pension
	Fachhochschulreife	Unterhalt durch Eltern, Ehepartner / Ehepartnerin, Lebenspartner / Lebenspartnerin oder andere Angehörige
	Allgemeine oder fachgebundene Hochschulreife (Abitur)	Eigenes Vermögen, Ersparnisse, Zinsen, Vermietung, Verpachtung, Altenteil
	Ohne Angabe	Sozialhilfe, - geld, Grundsicherung, Asylbewerberleistungen
	Entfällt (Kinder unter 15 Jahren, Schüler an allgemeinbildenden Schulen; Personen ohne allgemeinen Schulabschluss)	Leistungen aus einer Pflegeversicherung
		Sonstige Unterstützungen (z. B. BAföG, Vorruhestandsgeld, Stipendium)

3.6 Sonderprobleme

Bei der Untersuchung der *Ziel 1*- und *Ziel 2*- Fragestellungen emergierten verschiedene Sonderproblematiken, welche gesondert untersucht wurden. Hierbei handelt es sich um Fragestellungen, die sowohl direkt auf die Produktion der Schätzergebnisse Einfluss nehmen, als auch um Fragestellungen die zusätzliches Verbesserungspotential erörtern sollten. Eine Verbesserungsmöglichkeit bietet zum Beispiel die Verwendung weiterer Register.

3.6.1 Fallstudie zur Verwendung weiterer Register

In den vorhergehenden Abschnitten wird die Verwendung von Variablen des Melderegisters zur Verbesserungen vor allem der *Ziel 1*-Fragestellungen beschrieben. Neben den Daten aus dem Melderegister könnten nun auch weitere Register zur Verfügung stehen. Im Folgenden Abschnitt werden deshalb Szenarien simuliert, in denen Daten aus anderen Registern zusätzlich zur Verfügung stehen. Dabei handelt es sich in diesem Abschnitt zunächst um Daten der Bundesagentur für Arbeit, im nächsten Abschnitt wird dann das Vorhandensein weiterer Informationen simuliert.

Im vorliegenden Beispiel wird geschätzt, ob eine Person erwerbstätig/berufstätig ist, also die Variable *überwiegender Lebensunterhalt* (UBS) den Wert 1 annimmt. Als Hilfsvariablen für die Schätzung werden neben der Anzahl der Personen in der Anschrift (ADN), den Altersklassen der Personen (15–24, 25–39, 40–64), dem Geschlecht der Bewohner auch die Registervariable Erwerbstätigkeit (EWT; Daten der BA) verwendet. Dabei sind zwei Versionen dieser Variable, wie auf Seite 83 beschrieben, berücksichtigt. Es wird hier also mit einer Szenarioanalyse gearbeitet, bei denen die folgenden drei Szenarios unterschieden werden:

00A Es werden keine Informationen aus dem Register der Bundesagentur für Arbeit verwendet.

60A Die Variable EWT wird bei der Schätzung verwendet. Die Anzahl der Personen, für die EWT=1 zutrifft, hat einen mittlerer Zusammenhang zur Anzahl der Personen, für die EWT=1 zutrifft.

90A Die Variable EWT90 wird verwendet. Die Anzahl der Personen, für die EWT=1 zutrifft, hat einen hohen Zusammenhang zur Anzahl der Personen, für die EWT=1 zutrifft.

Folgende Modelle werden also geschätzt:

00A UBS: y ~ ADN + AGE1 + AGE2 + AGE3 + SEX
60A UBSEWT: y ~ ADN + AGE1 + AGE2 + AGE3 + SEX + EWT
90A UBSEWT90: y ~ ADN + AGE1 + AGE2 + AGE3 + SEX + EWT90

3.6.1.1 GREG

Der GREG ist in der untersten Zeile der Abbildung 3.39 auf der nächsten Seite dargestellt. Wie in allen fünf dargestellten Bundesländern zu sehen ist, verbessert sich die Schätzung durch die Hinzunahme der EWT90 Variable beträchtlich. So erfüllen fast alle SMP-Typ 1-Schätzungen durch die Verwendung von EWT90 die Qualitätsanforderungen. Ohne diese Variable sind es zwar deutlich mehr als die Hälfte, die sie erfüllen, aber eben wesentlich weniger. Auch bei SMP-Typ 2 und 3 zeigt sich eine eindeutige Verbesserung durch die Verwendung von EWT90, die dazu führt, dass selbst für diese SMP-Typen, für die keine Qualitätsanforderung gesetzlich festgeschrieben ist, viele den selben Benchmark-Kriterien wie für SMP-Typ 0 und 1 standhalten.

3.6.1.2 YOURAO und LMERW

Sowohl der YOURAO- wie auch der LMERW-Schätzer verhalten sich in Bezug auf das Hinzufügen der Kovariate EWT90 genauso wie der GREG. Hierbei erzielt der YOURAO-Schätzer in diesem Setting leicht präzisere Ergebnisse als der LMERW.

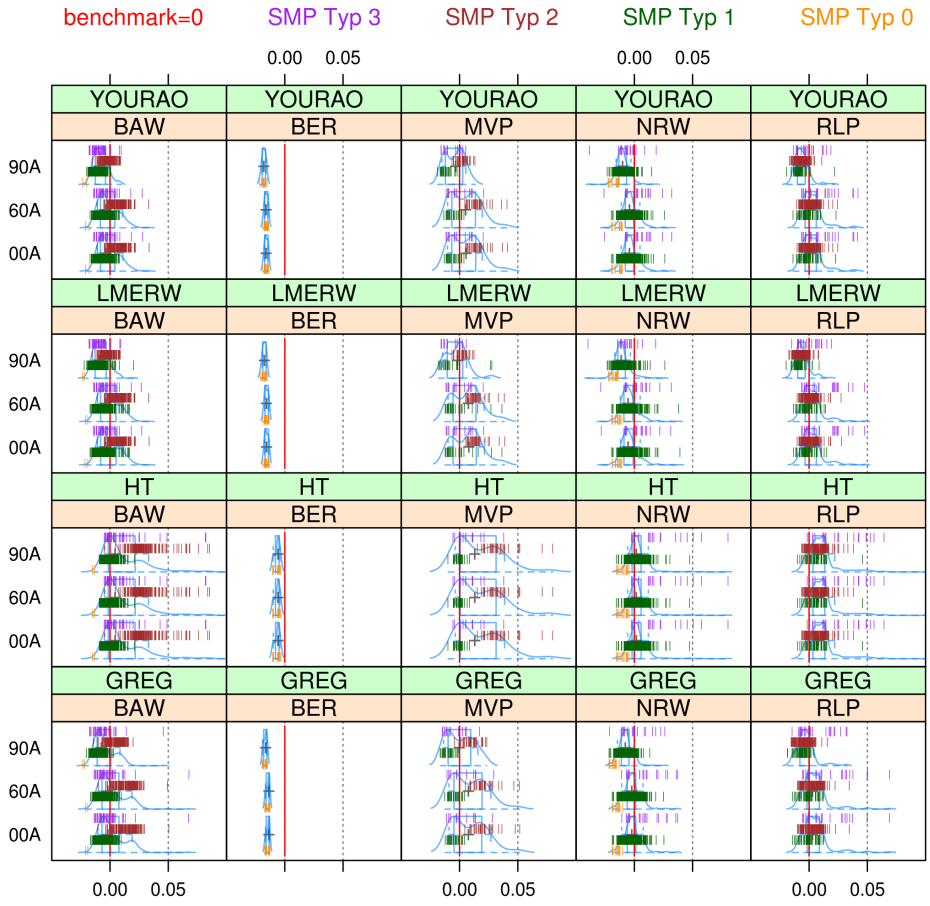


Abbildung 3.39: Schätzung von UBS=1 mit verschieden starken Kovariaten

3.6.2 Fallstudie zur Modellbildung mit unkorrelierten Variablen

Von Seiten des Auftragsgebers bestand der Wunsch zu erfahren, ob das Hinzufügen einzelner Variablen, die nicht in einem Zusammenhang mit der Untersuchungsvariablen stehen, zu Problemen bei der Schätzung führen würde. Hierzu wurden drei Szenarios (**A**, **B**, **C**) berechnet. Das Modell **A** stellt das schon zuvor besprochene Modell aus Registervariablen dar, **B** und **C** hingegen sind um unkorrelierte Kovariablen erweiterte Modelle. Weiterhin wurde mit Hinzunahme der hoch korrelierten Variablen EWT90 (bezeichnet mit: **90A**, **90B**, **90C**) und ohne dieselbe (Bezeichnet mit: **00A**, **00B**, **00C**) die Fragestellung UBS=1 geschätzt.

Dies wird in der Grafik 3.40 veranschaulicht. Es wird der RRMSE abzüglich der durch den Anteilwert der Beobachtung je SMP gegebenen Qualitätsanforderung für fünf Bundesländer (NRW, RLP, BAW, BER, MVP), vier Schätzer (HT, GREG, LMERW, YOURAO) und drei unterschiedliche Kovariaten dargestellt. Ein Strich unter Null bedeutet, dass ein SMP die Genauigkeitsanforderung erfüllt. Das Erreichen der Genauigkeitsanforderungen für die Stadtteile stellt sich für die Schätzung dieser Fragestellung als unproblematisch heraus. Der HT profitiert von einer Verbesserung des Modells nicht, da er kein Modell verwendet.

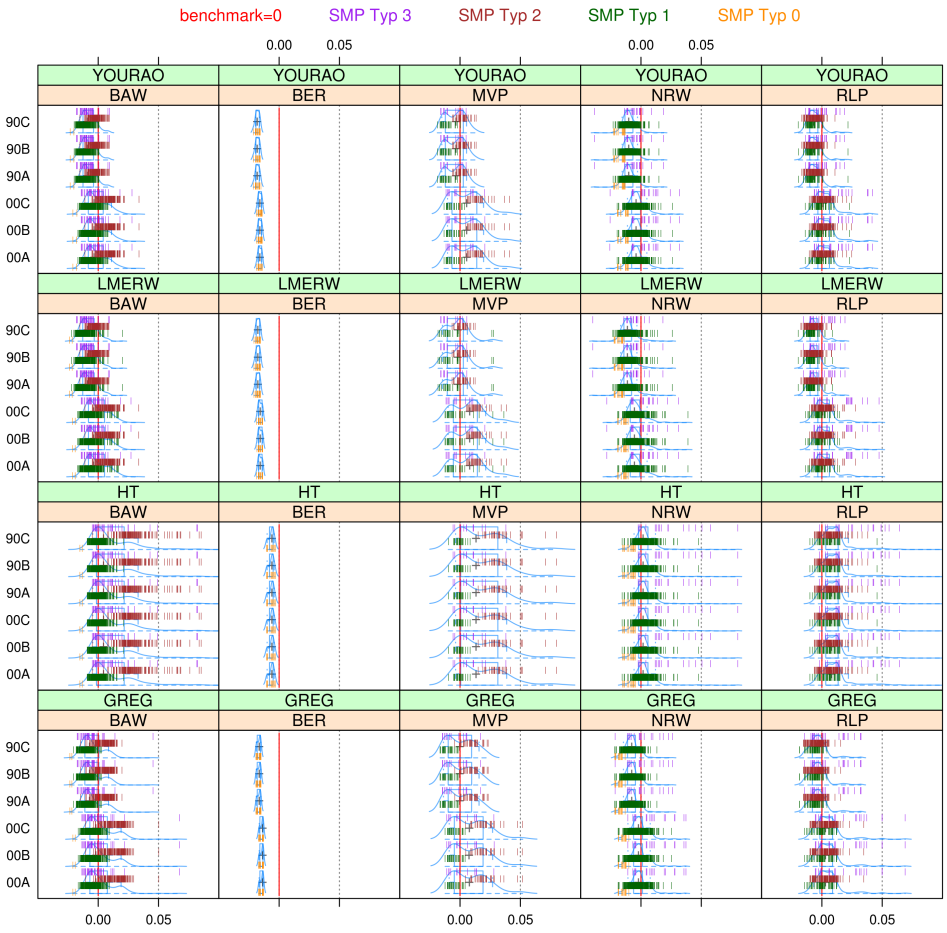


Abbildung 3.40: Schätzung von UBS=1 mit zum Teil zufälligen Kovariaten

Unterschiede in der Schätzqualität sind durch diese Modellierungen nur durch die Hinzunahme der hoch korrelierten Kovariaten EWT90 zu beobachten. Weder das erweiterte Modell **B** noch die Erweiterung **C** wirken sich negativ auf die RRMSEs der Schätzungen des GREG-, YOURAO- und des LMERW-Schätzers aus. Da der HT-Schätzer kein Modell verwendet bleibt er unberührt von der Modellierung.

3.6.3 Einfluss des Karteileichen- und Fehlbestandsmodells auf *Ziel 2*-Ergebnisse

Der Einfluss des Karteileichen- und Fehlbestandsmodells auf die Qualität der Schätzung der Gesamtpopulation wurde zuvor bereits besprochen. Hier stellt sich die Frage, welchen Einfluss die unterschiedliche Modellierung der Karteileichen- und Fehlbestände auf die Qualität der Schätzung von *Ziel 2*-Variablen hat. Zu diesem Zweck wurde die Schätzungen mit den Karteileichen- und Fehlbestandsmodellen *I1* und *Syn993* exemplarisch für die Fragestellung UBS durchgeführt.

Wie an Grafik 3.41 exemplarisch am Beispiel Rheinland-Pfalz zu sehen ist, ist der Einfluss des Karteileichen- und Fehlbestandsmodells bei dieser *Ziel 2*-Fragestellung vernachlässigbar. Die Schätzergebnisse für SMP-Typ 1 verändern sich nur in einem geringen Ausmaß. Bei SMP-Typ 2 und 3 sind hingegen vereinzelte Veränderungen zu erkennen, welche jedoch weder auffällig noch systematisch sind. sollten die Karteileichen- und Fehlbestandsmodelle Interaktionen mit der zu untersuchenden Fragestellung aufweisen, dann sollte dies sicher näher untersucht werden. In den hier durchgeführten Simulationen konnte das jedoch nicht beobachtet werden.

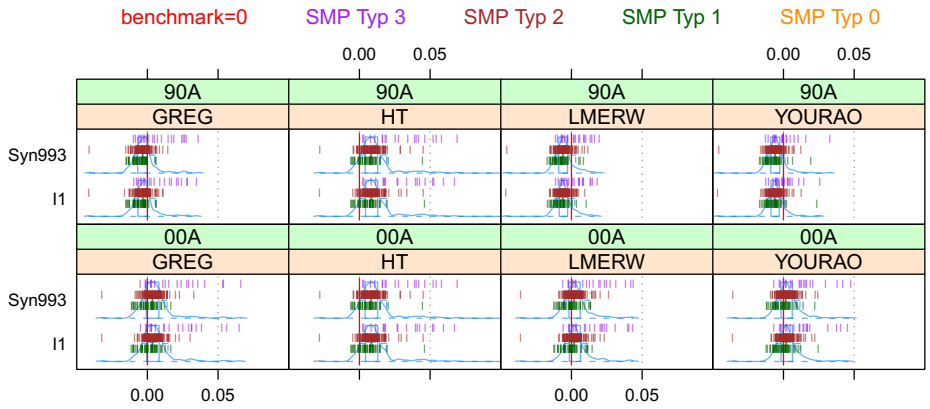


Abbildung 3.41: Schätzung von UBS=1 mit zwei verschiedenen Karteileichen- und Fehlbestandsmodellierungen

3.6.4 Vertikale Kohärenz der Schätzungen

Unter *vertikaler Kohärenz* versteht man die Aggregierbarkeit von Schätzwerten zu übergeordneten Schätzwerten im Sinne von Hierarchieebenen. Beispielsweise sollten die Summen der Schätzwerte für die Stichprobenbasiseinheiten die zusammen einen Kreis ergeben den Kreisschätzwerten entsprechen. Es ist wichtig zu untersuchen, inwieweit sich die Schätzer hier unterschiedlich verhalten. Dabei muss auch untersucht werden, ob unterschiedliche Modelle für Karteileichen und Fehlbestände die Beurteilung beeinflussen.

Bei der Schätzung der ISCED Fragestellungen ergaben sich bezüglich der Kohärenz die folgenden Ergebnisse. In Abbildung 3.42 ist die Abweichung der auf Kreisen aggregierten SMP-Ergebnisse mit den Kreisergebnissen für vier Schätzer dargestellt. Es ist zu sehen, dass die Ergebnisse des GREG für SMP's und Kreise am wenigsten voneinander abweichen, da die Boxplots alle auf der Nullinie liegen. Am schlechtesten schneidet bei dieser Betrachtung der EBLUPA ab, hier weichen die beiden Ergebnistypen teilweise um bis zu 10 % ab. Während in der ersten Grafik die Ergebnisse für das Karteileichen- und Fehlbestandsmodell Syn993 dargestellt sind, sind in Abbildung 3.43 die Ergebnisse für das KAL/FEB Modell I1 zu sehen. Insgesamt scheint das Karteileichenmodell keinen großen Einfluss auf die Kohärenz der Schätzungen zu haben.

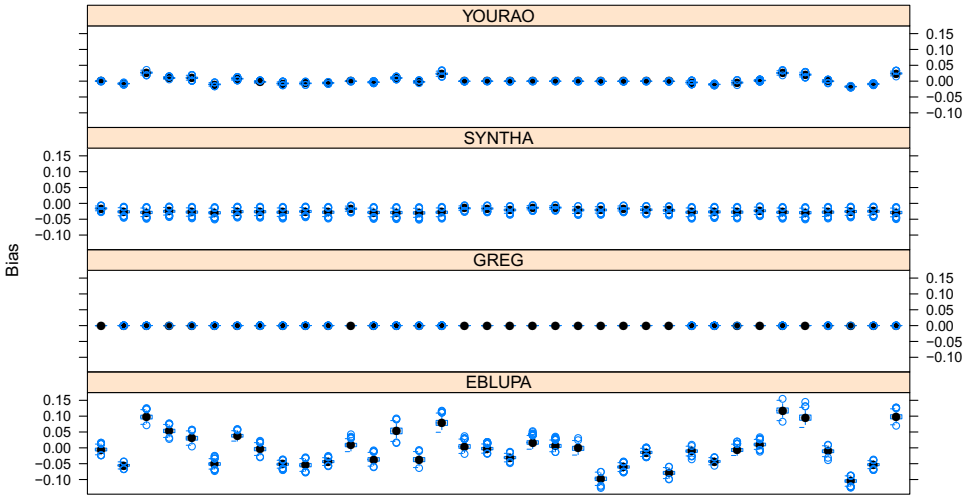
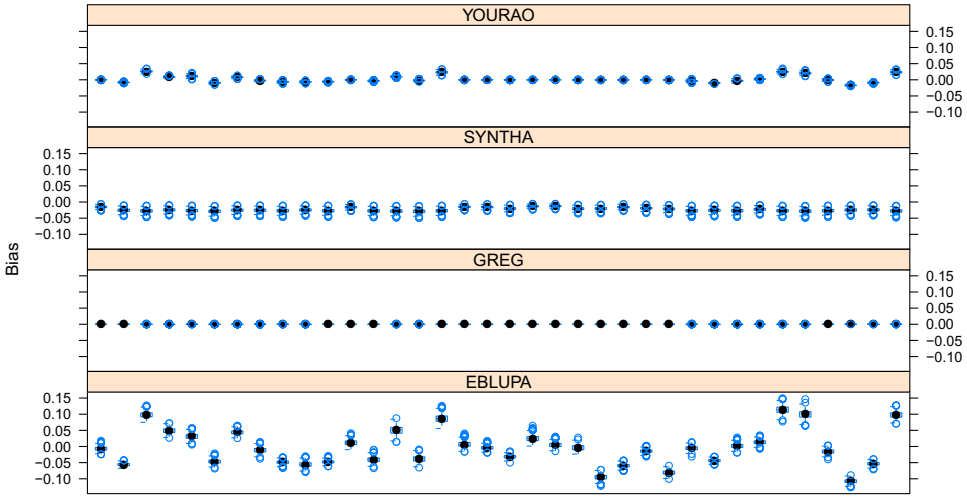


Abbildung 3.42: Kohärenz von ISCED in KRS (Syn993)



-

Abbildung 3.43: Kohärenz von ISCED in KRS (I1)

3.6.5 Disparitätsbetrachtung der Schätzungen

Beim Vergleich von Schätzwerten über die verschiedenen Gebiete (Areas: SMP oder KRS) interessieren die relativen Abstände und damit die Verteilung des Gesamtmerkmalsbetrages auf die einzelnen Gebiete. Dies kann mit der Lorenzkurve dargestellt werden. Es soll in jeder einzelnen Schätzung die

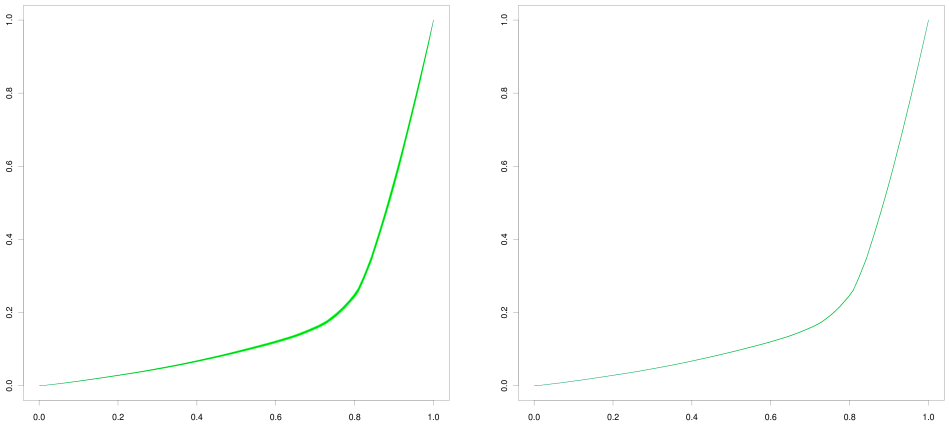


Abbildung 3.44: Lorenzinferenzkurven für SRS (HT (links) und GREG (rechts))

Disparität der wahren Werte erzielt werden. Zur Überprüfung werden die Werte der Grundgesamtheit benötigt. Da diese nur in Simulationen zur Verfügung stehen, ist die Überprüfung auch nur innerhalb einer Simulationsumgebung möglich.

Insbesondere für den Länderfinanzausgleich ist es von enormer Bedeutung die Disparität der Gemeinden richtig darzustellen. Dabei sollte jede Gemeinde an der richtigen Stelle auftauchen. Die Disparität wird im Folgenden mit Hilfe von Lorenzkurven dargestellt. Ziel der Lorenzinferenzkurven ist es, die Disparitäten der geschätzten Größen mit den tatsächlichen Größen zu vergleichen. In der Abbildung 3.44 ist dies exemplarisch für die Schätzung der Gesamtbevölkerung dargestellt.

Die kaum sichtbare blaue Linie kennzeichnet die Lorenzkurve der tatsächlichen Bevölkerungsgrößen zwischen den Sampling Points. Zusätzlich sind für die R=1.000 Läufe der Simulation die korrespondierenden Lorenzkurven jedes Laufs für die geschätzten Größen aufgezeigt. Weichen die Proportionen der geschätzten von den tatsächlichen Bevölkerungsgrößen ab, erkennt man abweichende Lorenzkurven. Idealerweise sollten alle Lorenzkurven übereinander liegen, denn dann würde in jeder Stichprobe durch die Schätzung die tatsächliche Disparität wiedergegeben. Aus Abbildung 3.44 wird erkenntlich, dass die Schätzungen überwiegend die tatsächliche Disparität wiedergeben. Allerdings ist ebenso sichtbar, dass die Präzision beim GREG spürbar besser ist. Dies gilt im Prinzip für alle durchgeführten Simulationen, wobei der Effekt bei weniger ausgeprägten KAL-/FEB-Modellen und geschichteter Auswahl geringer ist. Insofern kann erfreulicherweise von einer adäquaten Proportionalität der Schätzergebnisse in *Ziel 1* ausgegangen werden.

Eine Detailbetrachtung erlaubt die Abbildung 3.45. Dargestellt sind in den Boxplots jeweils die Spearmanschen Rangkorrelationen der 1.000 geschätzten Populationsverteilungen zur wahren Populationsverteilung über alle SMPs. Streng genommen könnte eine Lorenzkurve identisch aussehen, wenn alle Werte einfach permutiert werden. Betrachtet man zusätzlich die Spearman-Boxplots, so erkennt man, dass kaum Änderungen der Ränge der geschätzten SMP-Größen auftreten. Hiermit zeigt sich, dass im Allgemeinen davon auszugehen ist, dass die einzelnen Gemeinden in der Schätzung *proportional gut* behandelt werden.

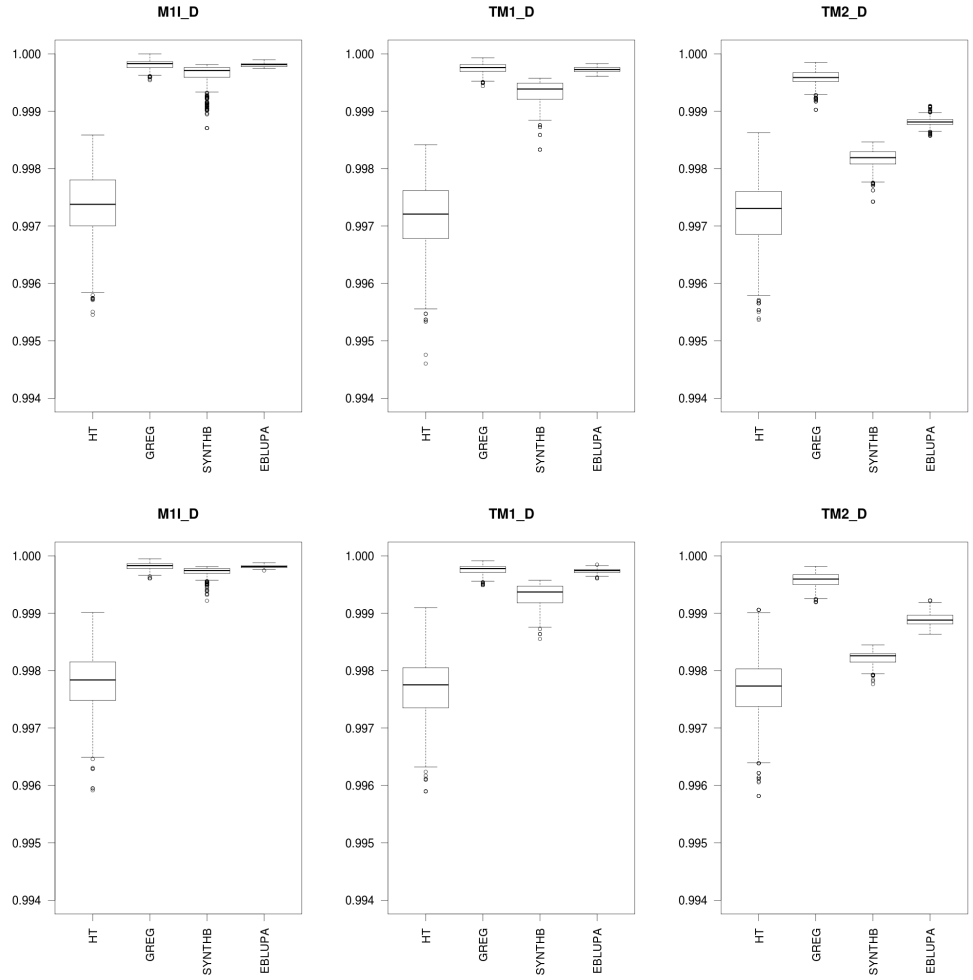


Abbildung 3.45: Boxplot der Spearman-Korrelationen zur tatsächlichen Bevölkerungsverteilung

In den Abbildungen 3.46 und 3.47 sind Ergebnisse der Lorenzinferenzen bezüglich einer *Ziel 2*-Fragestellung zu sehen.

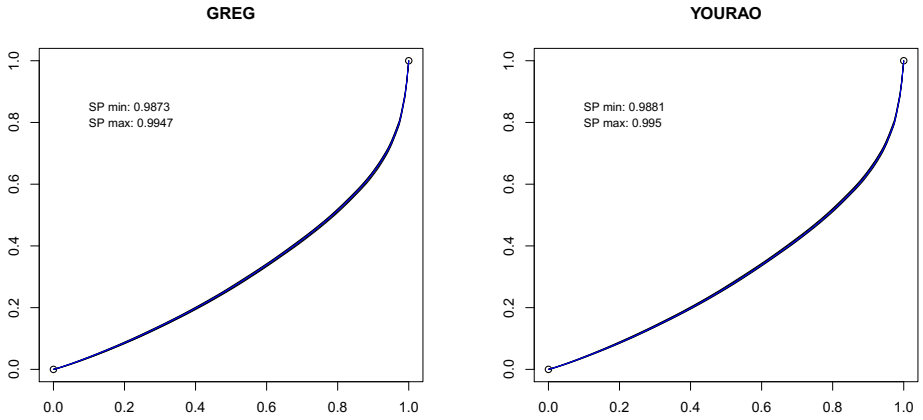


Abbildung 3.46: Disparität: EF117A (I)

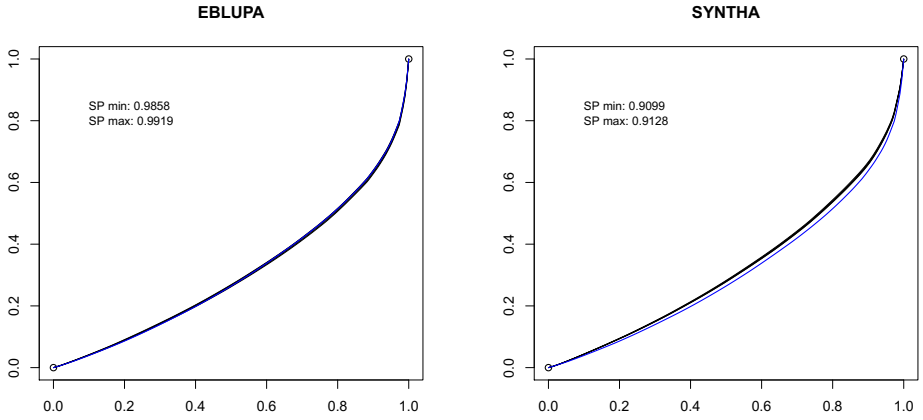


Abbildung 3.47: Disparität: EF117A (II)

3.6.6 Gewichtung bei Small Area-Methoden

Allgemeine Gewichtungen in Stichprobenverfahren folgen dem Prinzip

$$\widehat{\tau}_{d,\mathrm{W}} = \sum_{i \in S_d} \frac{g_i}{\pi_i} \cdot y_i \qquad . \tag{3.6.1}$$

Die Darstellung erfolgt in Analogie zum HT-Schätzer, wobei zusätzliche Gewichte g_i angegeben sind. Im Rahmen der GREG-Schätzer lassen sich eindeutige Gewichtungsvektoren ermitteln, welche g-weights genannt werden. Diese Gewichtungsfaktoren entstammen einer Kalibrierung mit quadratischer Zielfunktion (vgl. Deville und Sarndal 1992). Die Schätzungen gemäß (3.6.1) sind im

Allgemeinen dann sehr gut, wenn die Stichprobenumfänge nicht zu gering sind und die Anzahl der Hilfsvariablen nicht zu groß.

Im Kontext der Small Area-Methoden entstehen besondere Probleme bei der Kalibrierung. Bei einer großen Zahl von Kalibrierungsvariablen bei gleichzeitig geringem Gesamtbestand derselben können teils sehr hohe oder gar negative Gewichte auftreten, welche zu wenig hilfreichen Schätzungen führen können. Darüber hinaus spielt das Problem der Kohärenz der Schätzungen eine zusätzliche wesentliche Rolle. Während kaum effiziente Registervariablen für die Kalibrierung zur Verfügung stehen, liefern Design-basierte und Modell-basierte Schätzverfahren vielfach recht unterschiedliche Ergebnisse. Es ist aber davon auszugehen, dass auf Kreisen, insbesondere bei univariaten Fragestellungen, und Gemeinden unterschiedliche Schätzverfahren zum Zuge kommen. Eine Gewichtung müsste diese nicht kohärenten Schätzungen zusammenfassen können, was zu nicht lösbaren Zielräumen führt.

Daraus folgt, dass kein allgemeiner Gewichtungsvektor angegeben werden kann, welcher alle Nebenbedingungen erfüllt. Theoretisch wäre es denkbar, dass man im Sinne eines Schicht- und Area-separaten Regressionsschätzers Kalibrierungsgewichte ermittelt. Diese können aber kaum geeignet sein, da sie erstens nur auf den Registervariablen aufbauen und zweitens unter der hohen Variabilität der separaten Regressionsschätzung leiden. Ganz sicher sind sie nicht mehr kohärent zu möglichen Small Area-Schätzungen.

Münnich, Sachs und Wagner entwickeln derzeit einen neuen Ansatz, der diese verschiedenen Nebenbedingungen in Form geeigneter Zielkompromissfunktionen behandelt. Damit lassen sich später Gewichtungsvektoren ermitteln, welche zu kohärenten Schätzungen führen und gleichzeitig das Ausmaß der Inkohärenz zwischen den Schätzungen auf den verschiedenen Ebenen angibt. Aufgrund dann zahlreich vorhandener Nebenbedingungen, die berücksichtigt werden müssen, ist aber davon auszugehen, dass die Genauigkeit auf sehr kleinräumigen Schätzungen unzureichend ist, und dass der klassische Residualvarianzschätzer kaum noch zu vernünftigen Ergebnissen führt.

3.6.7 Varianzschätzung

Am Beispiel der Fragestellung ISCEDA wird das Verhalten der Varianzschätzer bezüglich *Ziel 2*-Variablen erklärt. Die Bezeichnungen der Schätzer sind in Einklang mit den zuvor besprochenen.

Dargestellt werden in Abbildung 3.48 die Konfidenzintervallüberdeckungsraten der in der Simulation errechneten Konfidenzintervalle. Bei einem Konfidenzintervall mit $\alpha = 0{,}05$ sollten somit 95 % der Konfidenzintervalle den wahren Wert überdecken (die Überdeckungsrate der Area sollte also auf der roten Linie liegen).

Wie anhand der Abbildung 3.48 zu sehen ist, überdecken die 95 % Konfidenzintervalle die aus den Varianzschätzern der GREG-Schätzer **COM** und **KRS-SEP** berechnet werden in 95% der Fälle den wahren Wert. Lediglich wenige SMPs liegen leicht darunter. Der GREG-Schätzer **D-SEP** hingegen, der nach SMP separat schätzt, überschätzt wie auch unterschätzt die Varianz des Punktschätzers zum Teil erheblich. Überschätzung bedeutet hierbei, dass die Überdeckungsraten über 95 % liegen. Solange die Konfidenintervalllängen nicht zu groß werden, ist dies jedoch noch akzeptabel. Es bedeutet nur, dass die Konfidenzintervalle breiter sind als eigentlich notwendig. Überdeckungsraten, die Unterhalb der 95 % liegen, sind dabei jedoch problematischer, da der Varianzschätzer eine nicht erreichte Genauigkeit ausgibt. Varianzschätzungen, die über 95 % Konfidenzintervallüberdeckungsrate ausweisen, sind somit *konservative Varianzschätzer*.

Der Varianzschätzer von You und Rao (2002) ist ebenso ein konservativer Varianzschätzer. Die Konfidenzintervalle überdecken stets den wahren Wert. Dies ist sicherlich auch dem Umstand

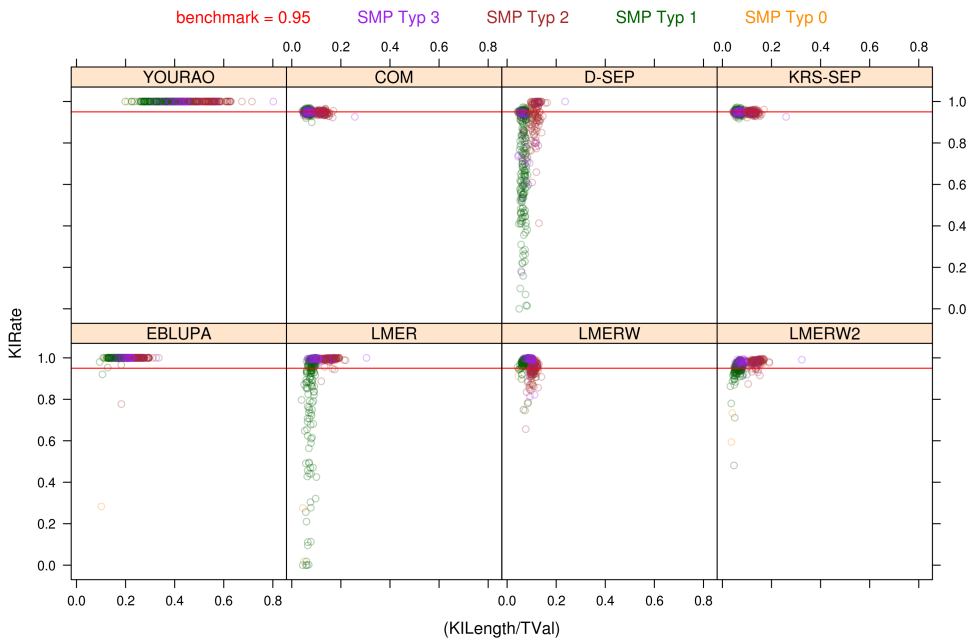


Abbildung 3.48: Konfidenzintervallüberdeckungsrate zur Begutachtung der Varianzschätzung für die Fragestellung ISCEDA in BAW

geschuldet, dass die Konfidenintervalllängen im Vergleich zu den anderen Schätzern relativ groß sind. Wenn es darum geht, kurze Intervalllängen zu bekommen, scheint auf den ersten Blick der EPLUPA-Varianzschätzer gute Dienste zu leisten. Jedoch ist eine der wenigen Areas, deren Konfidenzintervallüberdeckungsrate nicht bei 95 % liegt, gerade ein SMP vom Typ 0. Ein Varianzschätzer, der gerade in großen Areas seine Probleme hat, erscheint im Rahmen von Anwendungen im deutschen Zensus eher ungeeignet.

Erfreuliche Ergebnisse liefert die Anwendung des Prasad-Rao Varianzschätzers bei den Unit-Level Small Area-Schätzern LMER, LMERW und LMWERW2. Beim LMER, der ungewichteten Variante, hat der Prasad-Rao-Varianzschätzer noch einige Problem, ein überdeckendes Konfidenzintervall zu erzeugen. Dies ist sicherlich auch eine Folge der leicht verzerrten Schätzungen. Wie in Grafik 3.49 zu erkennen ist, weist der LMER-Schätzer in vielen Areas eine vergleichsweise hohe Verzerrung über die Simulationsdurchgänge auf. Wesentlich häufigere und stärkere Verzerrungen als zum Beispiel die gewichteten Schätzer LMERW und LMERW2.

Beide Informationen kombiniert betrachtet ergeben ein eindeutiges Bild vom Zusammenhang der Verzerrung und der Konfidenzintervallüberdeckungsrate. Dies ist in Abbildung 3.50 dargestellt. Auf der X-Achse ist der RBias aufgetragen und auf der Y-Achse die Konfidenzintervallüberdeckungsrate. Die rote Linie gibt wiederum die zu erreichende Konfidenzintervallüberdeckungsrate an, die blaue Linie symbolisiert Unverzerrtheit. Es ist ein eindeutiger Zusammenhang von verzerrter Schätzung und nicht einhalten der Konfidenzintervallüberdeckungsrate zu erkennen. Je verzerrter eine Schätzung, desto geringer ist dessen Konfidenzintervallüberdeckungsrate. An dieser Abbildung ist weiterhin zu sehen, dass der LMERW im Vergleich dieser drei Schätzer am besten abschneidet. Einerseits erreichen nur wenige Areas die Konfidenzintervallüberdeckungsrate von 95 % nicht, andererseits

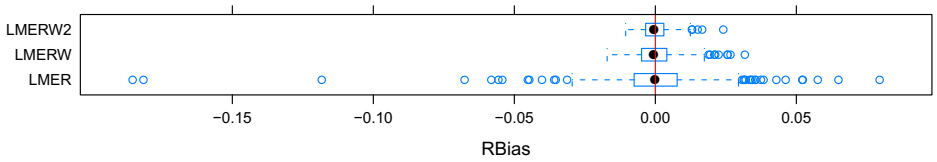


Abbildung 3.49: Relativer Bias für die Schätzer LMER, LMERW und LMERW2 für die Fragestellung ISCEDA in BAW

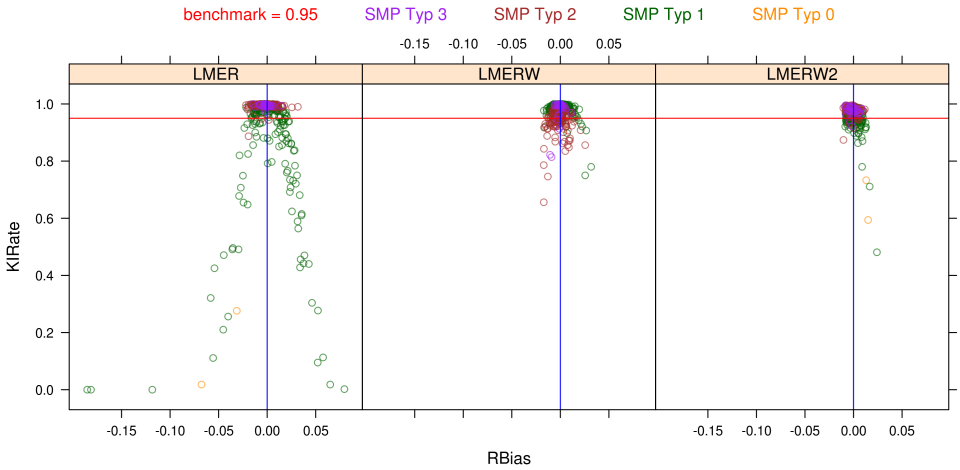


Abbildung 3.50: Konfidenzintervallüberdeckungsrate zur Begutachtung der Varianzschätzung für die Fragestellung ISCEDA in BAW

sind die Schätzergebnisse auch kaum verzerrt. Der LMERW2 weist zwar eine geringere Verzerrung auf und noch mehr Areas als beim LMERW erreichen die 95 % Konfidenzintervallüberdeckungsrate, allerdings schneiden bei diesem auch SMP-Typ 0 Areas schlecht ab. Dies ist wie zuvor erwähnt im Rahmen eines Zensus wenig opportun.

Abschließend lässt sich bemerken, dass im Hinblick auf die Varianzschätzung der GREG sehr gut abschneidet. Hierbei reussieren vor allem der voll-kombinierte GREG (COM) und der Kreis-separate GREG (KRS-SEP). Der SMP-separate GREG (D-SEP) hat wiederum erhebliche Probleme, die Varianz richtig zu schätzen.

Bei den Small Area-Schätzern fällt der Varianzschätzer von YOURAO sehr konservativ aus. Die Konfidenzintervalllängen sind zum Teil erheblich länger als die von den anderen Schätzern. Man kann sich aber relativ sicher sein, dass diese auch den wahren Wert umschließen. Hier würde sicherlich der Varianzschätzer von Torabi und Rao (2010) zu erheblichen Verbesserungen führen. Dieser ist zur Zeit aufgrund seiner Komplexität nur auf sehr kleinen Datensätzen anwendbar. Er sollte aber in Zukunft beachtet werden, falls verbesserte Implementationen gefunden werden. In Bezug auf die Varianzschätzung sind weiterhin der EBLUPA und die Schätzer LMER und LMERW2 nicht zu empfehlen, da diese auch größere Areas zum Teil sehr schlecht erfassen. Der LMERW

hingegen hat sowohl kurze Konfidenzintervalllängen, wie auch wenige Areas, die die Anforderung der 95 %-Konfidenzintervallüberdeckungsrate nicht erreichen. Auch die Varianzen in den großen Areas werden gut erfasst.

3.6.8 Modelldiagnostik

Bei der Verwendung statistischer Modelle spielt im Allgemeinen die Wahl der Variablen eine wesentliche Rolle. Darüber hinaus entsteht die Frage inwieweit die theoretischen Annahmen, die in einem Modell getroffen werden, auch tatsächlich empirisch gerechtfertigt sind. Im Gegensatz zur klassisch statistischen Modellierung, bei der Erklärungsgehalt im Vordergrund steht, werden bei den Zensus-Anwendungen Modelle *nur* zur Verbesserung der eigentlichen Survey-Schätzung benötigt (vgl. auch Knobelspies und Münnich (2008)).

Im Folgenden wird eine kurze Übersicht über Verfahren der Modelldiagnostik gegeben. Dies beinhaltet einerseits die Wahl des *besten* Modells und andererseits die Auswahl eines geeigneten Schätzers.

3.6.8.1 Modelldiagnostik für einen gegebenen Schätzer

Rao (2003), S. 75, schreibt:

> *Model diagnostics can be used to find suitable model(s) that fit the data well. Such model diagnostics include residual analysis to detect departures from the assumed model, selection of auxiliary variables for the model, and case-deletion diagnostics to detect influential observations.*

Alle untersuchten Schätzer mit Ausnahme des HT-Schätzer verwenden Modelle zur Verbesserung der (Survey-) Schätzung. Dabei unterscheidet sich der GREG als Design-basierter Schätzer hinsichtlich der Bedeutung des Modells wesentlich von den anderen Small Area-Schätzern. Im GREG wird das Modell lediglich als Unterstützung der Schätzung verwendet, wohingegen bei (synthetischen) Small Area-Schätzern das Modell die Grundlage der Schätzung bildet.

Modelldiagnostik beim GREG

Wie in Kapitel 2.3.2 beschrieben ist der GREG asymptotisch designunverzerrt. Diese Aussage ist unabhängig vom Modell. Mit anderen Worten die Wahl des Modells beeinflusst nicht die grundsätzliche Asymptotik. Je besser das Modell die Varianz der abhängigen Variable erklärt, desto genauer kann das Modell die direkte Schätzung unterstützen. Im Rahmen der Modelldiagnostik stellt sich hierbei die Frage, woran ein besseres Modell zu erkennen ist. Da es sich beim hier betrachteten GREG um eine lineare Regression handelt, kommen die klassischen Modelldiagnostikwerkzeuge zur Anwendung. Insbesondere ist der Korrelationskoeffizient R^2 von großer Bedeutung, da die Höhe desselben einen direkten positiven Einfluss auf die geschätzte Varianz der Punkt-Schätzung hat. Weiterhin interessiert die Normalverteiltheit der Residuen und deren Homoskedastizität. Wenn diese beiden Annahmen in der linearen Regression verletzt sind, dann ist die lineare Regression nicht mehr unbedingt der effizienteste Schätzer. Die Relevanz dieser theoretischen Überlegungen zur Modelldiagnostik relativieren sich in der praktischen Anwendung im Rahmen der deutschen Zensus-Schätzungen, da gar nicht genug Kovariaten vorliegen, um ein solches assistierendes Modell zu optimieren. Das Herausnehmen und Hinzufügen verschiedener Variablen um das geeignetste Modell zu finden beschränkt sich auf ein sehr überschaubares Set an Variablen. Bei Berücksichtigung des Ziels einer Modelldiagnostik im prädiktiven Kontext der Schätzungen eines Zensus stellen sich somit vor allem zwei Argumentationsstränge dar.

Hohes R^2 Wie bereits erwähnt ist es von Interesse, dass das R^2 möglichst hoch ist, da dies zu einer verbesserten Schätzung des Punktschätzers und somit auch zu einer geringeren geschätzten Varianz des Punktschätzer führt. Es kann gezeigt werden, dass bei Hinzunahme einer weiteren Kovariate sich das R^2 erhöht. Führt die Hinzunahme dieser weiteren Kovariate allerdings zu Multikollinearität, so gibt es keine eindeutige Lösung für die Parameter des Regressionsmodells. Somit beschränkt sich in Bezug auf den Korrelationskoeffizient R^2 die Modelldiagnostik auf das Aufspüren von Multikollinearität und deren Vermeidung. Da eine Multikollinearität an sich schon zu Problemen bei der Inversion der Varianz-Kovarianz-Matrix des β-Schätzers der linearen Regression führt, wird diese Problematik im Allgemeinen automatisch durch die Software diagnostiziert.

Adäquates Modell Bei kleinen Stichprobenumfängen kann unter Umständen das Hinzufügen vieler Variablen zu volatilen Parameter-Schätzungen führen. Deswegen bestrafen viele Modellwahl-Kriterien wie AIC und BIC oder das adjustierte R^2 die Verwendung vieler Variablen in Relation zur Stichprobe. Die im deutschen Zensus gezogenen Stichprobenumfänge sind in Relation zu den möglichen vorhandenen Registerkovariaten extrem hoch, so dass hier nicht zu erwarten ist, das die Hinzunahme einer Kovariate problematisch für die Schätzung werden könnte. Weiterhin ist es im prädiktiven Ansatz beim GREG auch nicht von Wichtigkeit, die β's optimal zu schätzen, da das Modell lediglich unterstützend wirkt.

Modelldiagnostik bei Small Area-Modellen

Im Gegensatz zum GREG wird das Modell bei Small Area-Schätzern nicht lediglich als Unterstützung einer direkten Schätzung verwendet, sondern ist die Basis des Schätzers. Daraus resultiert auch, dass diese Schätzer Modell-unverzerrt sind, aber eben nicht Design-unverzerrt. Modell-unverzerrt bedeutet aber insbesondere, dass vorausgesetzt wird, dass das Modell wahr ist. Ein wahres Modell ist außerhalb der Naturwissenschaften kaum zu finden. Nichts desto trotz können Modelle hilfreich sein, und zumindest Teile der Zusammnehänge erklären.

Den Unit-Level Small Area-Schätzern liegt ein Random Intercept Multi-Level-Modell zu Grunde. Bei bekannten Varianzkomponenten $\sigma_\varepsilon, \sigma_u$ kann der β-Schätzer des Multi-Level-Modells auch als gewichtete Regression geschrieben werden. D. h. auch hier können die klassischen Modelldiagnostikwerkzeuge angewandt werden. Zusätzlich kann noch die Annahme der Normalverteiltheit der Area-Effekte u_d überprüft werden.

Im Fall eines Binomial-Schätzers gibt es weiterhin die Möglichkeit die Devianzresiduen im Sinne eines Binned-Plots darzustellen, um die Annahme der Linearität des Modell in Bezug auf die transformierte abhängige Variable zu überprüfen.

Beim Schätzer von YOURAO ist die Wahl des Modells fast wie beim GREG zu handhaben, da durch die *self-weighting property* eine Design-Unverzerrtheit der Aggregation der Areas automatisch gegeben ist. Somit ist, relativ zu den anderen Small Area-Schätzer, eine hohe Sicherheit bezüglich der Punktschätzer gegeben. Dies gilt selbst wenn die Varianzkomponenten aufgrund des Modells extrem schlecht geschätzt wurden.

Modelldiagnostik zur Wahl zwischen verschiedenen Schätzern

Die klassischen Methoden der Modelldiagnostik erlauben zumeist nur den Vergleich der Modelle innerhalb einer Schätzmethodik. Oft werden die Log-Likelihoods von verschiedenen Schätzmethodiken verglichen. Dies ist im Allgemeinen nur zulässig wenn die Likelihood des einen Modells

die des anderen Modells umhüllt. Mit anderen Worten, der Parameter Raum des einen Modells muss ein Überraum des Paramter Raums des anderen Modells sein. Ein Vergleich zwischen der Log-Likelihood eines Binomial-Schätzers mit der eines linearen Schätzers hat im Allgemeinen somit keinen Erklärungsgehalt.

Eine weitere Möglichkeit der Überprüfung verschiedener Schätzer ist der Vergleich der geschätzten MSEs. Ist im Prinzip von konsistenten Punktschätzungen auszugehen, ist es eventuell ratsam, denjenigen Schätzer auszuwählen, der die geringsten geschätzten Varianzen aufweist. Bei Verwendung synthetischer Schätzungen muss jedoch stets die Gültigkeit des Modells beachtet werden.

3.6.8.2 Zusammenfassende Bemerkungen

Soll die Modelldiagnostik bei der Auswahl von Hilfsvariablen verwendet werden, ergibt sich die Schwierigkeit, dass bei prädiktiven Ansätzen auf Grund der sehr geringen Anzahl an möglichen Hilfsvariablen kaum eine Wahl besteht.

Im Rahmen der vorgeschlagenen Methoden sind Verfahren der Modelldiagnostik vor allem bei tiefer gehenden Small Area-Modellierungen einsetzbar. Diese kann man indes nicht verallgmeinert anwenden. Jede (bedeutsame) Fragestellung müsste separat untersucht werden, was den Aufwand bei der Implementation erheblich erhöht.

3.6.9 Rundung

Als einfaches Argument für die Notwendigkeit des Rundens erscheint die Tatsache, dass man nicht beliebig viele Nachkommastellen in einer Tabelle darstellen kann und also ein wie auch immer geartetes Runden notwendig ist. In der Regel heißt die Frage Auf- oder Abrunden. Das konventionelle ist ein deterministisches Runden, das bei Nachkommawerten kleiner 5 abrundet und ansonsten aufrundet.

Tabelle 3.23: Deterministische Rundung

Originalwert	gerundeter Wert
13,4	13
10,9	11
15,8	16
12,9	13

Die Schwachstelle dieses Rundens zeigt sich erst bei der Summation in den einzelnen Spalten.

Tabelle 3.24: Rundungstabelle

	Originalwert	gerundeter Wert
	13,4	13
	10,6	11
	15,5	16
	12,5	13
Summe	52,0	53

Es kann offensichtlich der gerundete Wert der Summe der Originalwerte von der Summe der gerundeten Werte deutlich abweichen. In der amtlichen Statistik liest man häufig die Fußnote „Abweichungen in den Summen ergeben sich durch das Runden der Zahlen.“

Um diese Schwachstelle der Nichtadditivität beim deterministischen Runden zu beseitigen, werden kontrollierte stochastische Rundungsmethoden angewendet. Derartige Rundungen finden auch im Bereich der Tabellengeheimhaltung Anwendung (siehe Giessing (2008)).

Ein weiteres Anwendungsfeld des Rundens sind Tabellen, bei denen die Randsummen vorgegeben sind. Ausgehend von diesen werden die Zellen mittels proportionaler Aufteilung oder über Algorithmen wie der IPF (Iterative Proportionale Algorithmus) ausfüllt, die aber nicht ganzzahlig zu sein brauchen und daher gerundet werden müssen. Allgemeiner besteht das Problem darin, dass man eine $I \times J$ Ausgangstabelle $A = (a_{ij})$ mit Zeilenhäufigkeiten $a_{i.} = \sum_{j=1}^{J} a_{ij}$ bzw. Spaltenhäufigkeiten $a_{.j} = \sum_{i=1}^{I} a_{ij}$ hat und eine naheliegende Kontingenztabelle $B = (b_{ij} = round(a_{ij}))$ mit ganzen Zahlen b_{ij} und der Eigenschaft sucht, dass sich nicht nur die einzelnen Zellhäufigkeiten b_{ij} um weniger als 1 von a_{ij} unterscheiden, sondern auch alle Ränder $|b_{i.} - a_{i.}| < 1$ und $|b_{.j} - a_{.j}| < 1$ erfüllen. In sozialwissenschaftlichen Bevölkerungsumfragen wie etwa dem ALLBUS steht man bei der Auswahl von Gemeinden schnell vor einem solchen Problem, wenn man etwa versucht, einen Stichprobenumfang proportional zu den Bevölkerungszahlen aus der Gesamtheit auf die Zellen aufzuteilen, die durch Kreise und eine Typisierung der Gemeinden definiert sind. Im Bereich des Zensus steht man beispielsweise bei den Hypercubes (s. Kapitel 3.5) vor dieser Aufgabe, da die Ausgangswerte auf Schätzungen beruhen, die in der Regel keine ganzzahligen Ergebnisse liefern.

Eine Übersicht über verschiedene Möglichkeiten des Rundens ist in Salazar-González (2002) gegeben.

Das Cox-Verfahren Cox (1987) ist ein häufig auf zweidimensionale Tabellen angewendetes Rundungsverfahren, das folgende Eigenschaften hat:

1. **Benachbarte Werte.** Jeder Wert a_{ij} der Ausgangstabelle wird zu einer benachbarten ganzen Zahl b_{ij} gerundet, d.h. $b_{ij} = round(a_{ij}) = [a_{ij}]$ oder $b_{ij} = round(a_{ij}) = [a_{ij}] + 1$, wobei $[x]$ als größte ganze Zahl kleiner gleich x definiert ist.
2. **Nullbeschränktheit.** Man fordert $|b_{ij} - a_{ij}| < 1$, d.h. $b_{ij} = a_{ij}$, wenn a_{ij} eine ganze Zahl ist.
3. **Additivität.** Die gerundeten Werte b_{ij} lassen sich zu einem gerundeten Wert der Zeilen- oder Spaltensummen aufaddieren, formal:
   - $\sum_{j=1}^{J} b_{ij} = round(a_{i.}) \forall i$
   - $\sum_{i=1}^{I} b_{ij} = round(a_{.j}) \forall j$
   - $\sum_{i=1}^{I} round(a_{i.}) = \sum_{j=1}^{J} round(a_{.j}) = round(\sum_{i=1}^{I} \sum_{j=1}^{J} (a_{ij}))$
4. **Unverzerrtheit.** Die Rundung sollte im Erwartungswert dem ursprünglichen Ausgangswert entsprechen, d.h. $E(b_{ij} - a_{ij}|a_{ij}) = 0$.

Beim Cox-Verfahren wird von der Ausgangsmatrix A zunächst der ganzzahlige Teil abgezogen, d.h. $a_{ij} = [a_{ij}] + d_{ij}$ mit $0 \leq d_{ij} < 1$ und das Verfahren auf $D = (d_{ij})$ angewendet. Wenn D die Nullmatrix ist, ist A bereits die gesuchte Lösung. Sonst wählt man einen Zyklus, also einen geschlossenen Pfad $(i_1, j_1), (i_1, j_2), (i_2, j_2), ..., (i_k, j_k + 1) = (i_k, j_1)$ wobei die entsprechenden d_{ij} in diesem Pfad $0 < d_{ij} < 1$ erfüllen müssen. Dass ein solcher Pfad existiert, solange mindestens ein d_{ij} nicht 0 oder 1 ist, lässt sich beweisen. Weiter sei

$$d_{-} = min_{1 \leq q \leq k}[d_{i_q j_q}, 1 - d_{i_q i_{q+1}}]$$

$$d_+ = min_{1 \leq q \leq k}[1 - d_{i_q j_q}, d_{i_q j_{q+1}}]$$

mit $0 < d_- < 1, 0 < d_+ < 1$ definiert. Nun wählt man d_- mit Wahrscheinlichkeit $P_- = \frac{d_+}{d_- + d_+}$ und d_+ mit Wahrscheinlichkeit $P_+ = \frac{d_-}{d_- + d_+}$. Diese Auswahlwahrscheinlichkeiten garantieren die zuvor geforderte Unverzerrtheit. Falls d_- ausgewählt wurde, wird Zelle (i_1, j_1) um d_- Einheiten verringert und die nächste Zelle des Zyklus entsprechend um d_- Einheiten erhöht. So fährt man auf diesem Zyklus abwechselnd bis zum Schluss fort. Falls die so veränderte Matrix noch Brüche enthält, wiederholt man die ganze Prozedur. Da bei jedem Iterationsschritt mindestens ein Bruch in eine ganze Zahl verwandelt wird und alle ganzen Zahlen unverändert bleiben, bricht der Prozess nach weniger Iterationen ab als es Brüche in $D = (d_{ij})$ gibt.

Wir demonstrieren das Verfahren an einem Beispiel. In diesem Beispiel sind die Randsummen der Spalten ganzzahlig, die Randsummen der Zeilen bis auf kleine Rundungsfehler ebenfalls. Beim Cox-Verfahren allgemein bräuchte nur die Gesamtsumme aller Matrixwerte ganzzahlig sein. Ausgangsmatrix im Beispiel ist

$$A = \begin{pmatrix} 30,4700235 & 10,0365190 & 39,114091 & 10,379367 \\ 0,5941915 & 0,5042813 & 3,278510 & 3,623017 \\ 4,9357850 & 5,4591997 & 5,607399 & 5,997616 \end{pmatrix} \quad .$$

mit Gesamtsumme 120. Der ganzzahlige Teil von A ist

$$[A] = \begin{pmatrix} 30 & 10 & 39 & 10 \\ 0 & 0 & 3 & 3 \\ 4 & 5 & 5 & 5 \end{pmatrix}$$

so dass das eigentliche Verfahren auf die Matrix

$$D = \begin{pmatrix} 0,4700235 & 0,0365190 & 0,114091 & 0,379367 \\ 0,5941915 & 0,5042813 & 0,278510 & 0,623017 \\ 0,9357850 & 0,4591997 & 0,607399 & 0,997616 \end{pmatrix}$$

angewendet werden muss. Nach 9 Iterationen erhält man als Lösung

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 \end{pmatrix}$$

und daher zusammen mit dem ganzzahligen Teil

$$\left(\begin{array}{cccc|c} 31 & 10 & 39 & 10 & 90 \\ 1 & 0 & 3 & 4 & 8 \\ 4 & 6 & 6 & 6 & 22 \\ \hline 36 & 16 & 48 & 20 & 120 \end{array}\right) .$$

Leider ist eine einfache Verallgemeinerung des Cox-Verfahrens auf höhere Dimensionen schwierig. Einen anderen Ansatz findet man dafür in Doerr et al. (2006), einen entsprechenden Algorithmus für Excel in Salazar-González und Schoch (2004).

3.6.10 Rechenzeiten und Speicherbedarf

Im Rahmen der Simulation konnten gute Proxies gefunden werden, um die Rechenzeit einzelner Schätzer und deren Speicherbedarf zu analysieren. Allerdings beziehen sich diese Angaben ausschließlich auf beobachtete Simulationszeiten innerhalb des Programms R. Dabei sind die Rechenschritte hauptsächlich in Fortran, C und C++ umgesetzt. Andere Statistikprogramme werden sich daher in ähnlichen Zeit- und Speicherbereichen bewegen. Eine leichte Beschleunigung ist durch die vollständige Umsetzung der Schätzer in vorkompiliertem Programmcode möglich.

3.6.10.1 Rechenzeiten

In Abbildung 3.51 sind die Bundesländer, für die bezüglich der Fragestellung ISCEDA-Simulationen durchgeführt wurden, nach der Anzahl der Sampling Points sortiert. Die gleiche Reihenfolge würde sich ergeben, wenn nach der Anzahl der Adressen sortiert werden würde. Das Bundesland Nordrhein-Westfalen hat also sowohl die meisten Anschriften als auch die meisten Sampling Points, während Berlin sowohl gemessen an der Anzahl der Sampling Points als auch gemessen an der Zahl der Anschriften das kleinste evaluierte Bundesland ist. Dargestellt ist die mittlere Rechenzeit einer Simulation (über die 1.000 Simulationsdurchläufe hinweg). Grundsätzlich steigt die mittlere Rechenzeit also mit der Anzahl der SMPs im Bundesland.

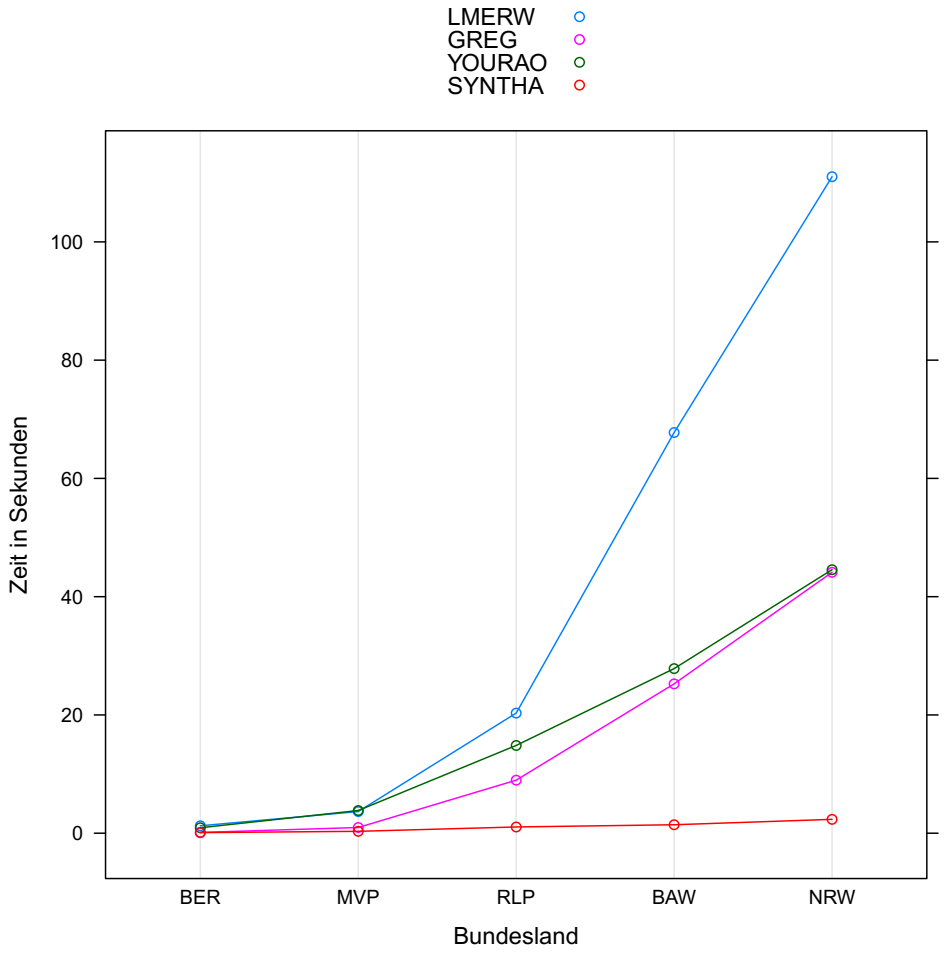


Abbildung 3.51: Mittlere benötigt Zeit zur Simulation der ISCEDA-Fragestellung

Ein weiterer wichtiger Unterschied ist zu erkennen, wenn man die Zeiten miteinander vergleicht, die benötigt wurden, um unterschiedliche Fragestellungen zu simulieren. Die längste Simulationsdurchlaufszeit für die ISCEDA-Fragestellung dauerte 111,02 Sekunden (LMERW-Schätzer in NRW). Bei der Simulation der Fragestellung Zuzugsjahr resultierten Schätzungen des Binomial- und des Poisson-Schätzers von bis zu 3.985,5 Sekunden. Im Mittel dauert ein Simulationsdurchlauf für den LMERW-Schätzer hier 10,502 Sekunden mit einem Maximalwert bei 31,92 Sekunden.

In Abbildung 3.52 wird die benötigte Zeit der Simulationsdurchläufe für die Fragestellung Zuzug von türkischen Staatsbürgern zwischen 1970 und 1980 dargestellt. Hier ist zu sehen, dass die Simulationen in Rheinland-Pfalz deutlich länger brauchen als die Simulationen in Berlin.

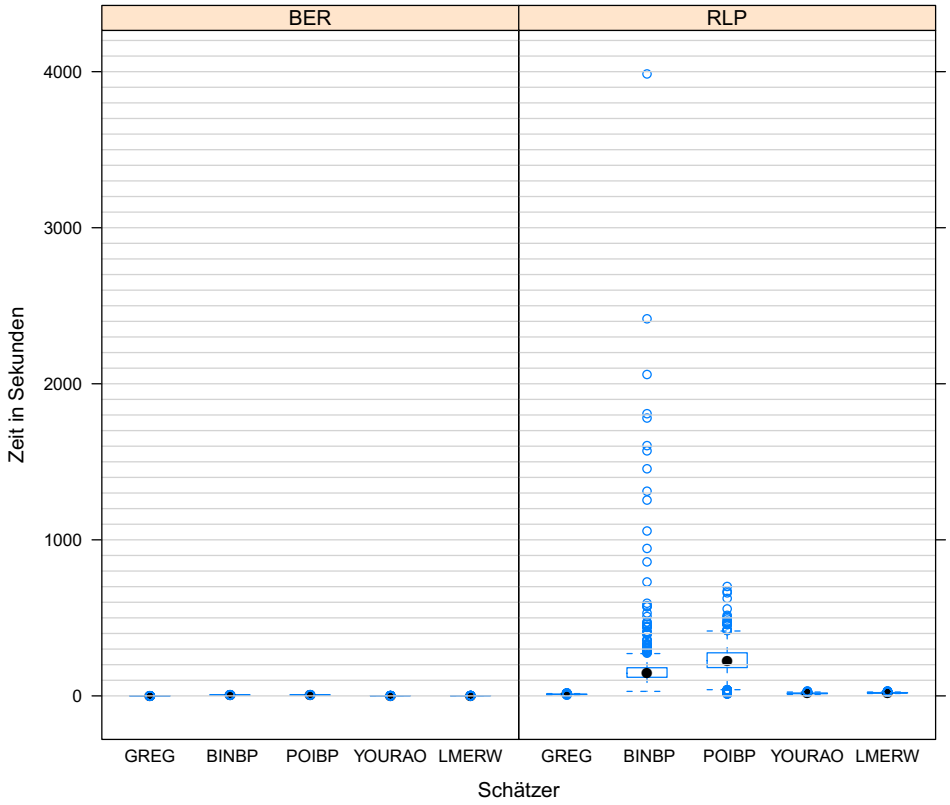


Abbildung 3.52: Benötigte Zeit zur Simulation der Zuzugsjahr Fragestellung

Die χ^2-Methode erwies sich als weniger aufwändig, als GREG- bzw. YOURAO-Schätzungen. Bei *Ziel 1* zeigten sich die höchsten Werte beim GREG mit 388,12 Sekunden in Rheinland-Pfalz und 214,25 Sekunden in Berlin. Die χ^2-Methode lieferte hier Rechenzeiten von 212,13 bzw. 167,02 Sekunden.

3.6.10.2 Speicherbedarf

Eine theoretische Berechnung der maximalen Matrixgrößen ist möglich. Das ist jedoch nur ein Proxy für den Gesamtbedarf an den Speicher. Viele Variablen müssen bereitgehalten werden, um Auswertungen zu machen, verschiedene Modelle auszuprobieren und weitere Berechnungen durchzuführen. Deswegen scheint es sinnvoller zu sein, den größten beobachteten RAM-Verbrauch anzugehen. Dieser resultiert bei Binomialschätzern, da deren Vorhersage nur über eine vollbesetzte Grundgesamtheitsmatrix möglich ist (aufgrund der Nicht-Linearität). Bei Binomialschätzungen wurden für die Variable Zuzugsjahr maximal 20 GB RAM verbraucht. Im Allgemeinen wurden aber bei linearen Schätzmethoden die 16 GB RAM nicht überschritten. Also wäre eine konservative Empfehlung 16 GB RAM pro Prozess vorzuhalten. Aufgrund der relativ zu den sonstigen Kosten geringen Anschaffungskosten von RAM, ist eine konservative Ausstattung durchaus empfehlens-

wert, da ansonsten die Gefahr droht, dass Prozesse inmitten von Berechnungen vom Betriebssystem als systemkritisch eingestuft werden und möglicherweise ohne Warnung beendet werden.

4 Empfehlungen für den Zensus 2011

Ziel des Zensus Stichprobenforschungsprojekts war die Erforschung von Stichprobendesigns und Schätzmethoden für die Haushaltsstichprobe des Register-gestützten Zensus 2011. Dabei spielten zwei Ziele eine zentrale Rolle: Einerseits soll die amtliche Einwohnerzahl ermittelt werden. Dazu ist notwendig, die Karteileichen und Fehlbestände flächendeckend zu schätzen. Andererseits sind zahlreiche Häufigkeitsverteilungen von Variablen, die nicht in den Registerdaten vorhanden sind, wie etwa Erwerbs- oder Ausbildungsvariablen, zu berechnen. Die zu ermittelnden Werte müssen hierbei nicht nur auf einem hohen Aggregationsniveau, wie zum Beispiel auf der Ebene von Bundesländern, sondern vor allem auf kleinräumigen Nachweiseinheiten geschätzt werden.

Da die Gemeindestrukturen in Deutschland zwischen den Bundesländern erhebliche Unterschiede aufweisen, galt es zunächst, Stichprobenbasiseinheiten geeignet zu definieren, auf denen die vorgegebenen Genauigkeitsanforderungen für *Ziel 1* und *Ziel 2* eingehalten werden können. Diese Stichprobenbasiseinheiten erlauben die Implementierung einer unter Box-Constraints optimalen, geschichteten Zufallsstichprobe von Anschriften.

Um die Effizienz des Stichprobendesigns und die Genauigkeit von Schätzverfahren zu analysieren, wurde eine realitätsnahe Simulationsgesamtheit erstellt, welche über 85 Millionen Einträge umfasste. Das wesentliche Problem bei der Erzeugung dieser Simulationsgesamtheit bestand darin, dass die Struktur der zu erzeugenden Variablen nicht für ganz Deutschland flächendeckend bekannt war. Insbesondere können versteckte Heterogenitäten Probleme bereiten.

Die zentrale Aufgabe des Projekts war es, Schätzer zu finden, die hinreichend belastbare Ergebnisse bezüglich beider Ziele auf verschiedenen Aggregationsebenen ermöglichen.

Aus den Untersuchungen kann abgeleitet werden, dass die Erfüllung der Präzisionsanforderungen bei *Ziel 1* unter den Bedingungen der Simulation gesichert erscheint und selbst unter relativ ungünstigen Bedingungen kaum Schwierigkeiten bereiten wird. Probleme können hier nur auftreten, wenn einzelne Melderegister doch wesentlich *schlechter* als erwartet geführt werden. In den Simulationen ergaben sich auf Basis der Karteileichen- und Fehlbestandsmodelle mehrfach deutlich kleinere Korrelationskoeffizienten als der durch Vorgabe angenommene Wert von 0,993. Trotzdem verletzen die durchschnittlichen RRMSEs in diesen Fällen nur knapp die Präzisionsanforderungen. Dies ist insofern erfreulich, da sich durch die politische Vorgabe der erwarteten Präzision bei gleichzeitig minimal zu befragender Anzahl der Bürger ebenso ein anderes Bild hätte zeigen können.

Eine klare Empfehlung für ein bestimmtes Schätzverfahren unter Berücksichtigung von *Ziel 2* gestaltet sich wesentlich schwieriger. Dies liegt vor allem daran, dass es eine Fülle möglicher *Ziel 2* Variablen gibt, deren Varianzstrukturen in Bezug auf die Anschriften sehr unterschiedlich sind. Kritisch sind Merkmale mit hohen Populationsanteilen, für deren geschätztes Merkmalsaufkommen eine höhere (relative) Präzision gefordert wird, oder seltene Ereignisse, also Anteile von seltenen Merkmalsausprägungen.

Die Schätzung von Anteilen bei Merkmalskombinationen, sogenannten Hypercubes, ist aufwändiger als die klassische Schätzung von Anteilen bei nur einem Merkmal. Sind die einzelnen Zellen relativ stark besetzt, so lassen sich akkurate Schätzungen erreichen. Es bleibt aber das Problem der nicht besetzten Stichprobenzellen. In einer Stichprobe kann man kaum zwischen Stichprobennullen und Strukturnullen unterscheiden. Hier ist der Aufwand möglicherweise sehr hoch, wenn genaue Schätzungen erwartet werden. Das Problem von sehr gering besetzten Merkmalszellen ist wohl nur mit Hilfe von speziellen Modellen oder bayesianisch mit informativen a priori-Verteilungen zu lösen. Eine präzise Schätzung von sehr tief gegliederten Hypercubes ist für die zukünftige Forschung von größtem Interesse.

Kohärenz muss sowohl bezüglich Aggregationsstufe und Schätzverfahren (vertikale Kohärenz) als auch bei teilweisen Überschneidungen von Hypercubes (horizontale Kohärenz) gewährleistet werden. Ist man an einem horizontal und vertikal vollständig kohärenten Zensus interessiert, benötigt man einen einzigen Gewichtungsvektor, der alle Schätzungen bezüglich *Ziel 1* und *Ziel 2* als Nebenbedingungen simultan berücksichtigt. Ein erfolgversprechender Ansatz für ein verallgemeinertes Kalibrierungsproblem wird derzeit weiterentwickelt.

Aus den Untersuchungen leiten wir nachfolgende Empfehlungen ab. Diese Empfehlungen berücksichtigen die Robustheit der Verfahren, die Erfüllung der Genauigkeitsanforderungen sowie die einfache Implementierbarkeit.

***Ziel 1*:** Für die Schätzung der amtlichen Einwohnerzahl auf großen SMPs sollte der gruppierte verallgemeinerte Regressionsschätzer (GREG) angewendet werden. Als Gruppen eignen sich Bundesländer und Kreise, mit kleinen Einschränkungen, bedingt durch die etwas reduzierte Robustheit, auch SMPs. Eine auf Schichten bezogene Gruppierung kann indes nicht empfohlen werden. Die Anschriftengröße sollte unbedingt als Hilfsinformation für den GREG verwendet werden. Der klassische Residualvarianzschätzer liefert auf SMP-Ebene akkurate Varianzschätzungen.

KAL/FEB: Für die Schätzung von Karteileichen und Fehlbeständen in der gegebenen Untergliederung erscheint ein kombinierter Schätzer aus χ^2- und GREG-Schätzer am geeignetsten. Als Eingangsgrößen für die Haushaltegenerierung können auch tiefere Untergliederungen herangezogen werden, wenn sie als Orientierungswert ohne Genauigkeitsanforderungen verwendet werden.

***Ziel 2*:** Bei *Ziel 2*-Variablen können der GREG-Schätzer beziehungsweise der YOURAO-Schätzer verwendet werden. Sind sehr unterschiedliche Niveaus der Untersuchungsvariablen zwischen den Areas zu erwarten, ist der YOURAO-Schätzer dem GREG-Schätzer vorzuziehen, da der YOURAO-Schätzer vom Random Effekt profitiert. Ansonsten ist der gruppierte GREG-Schätzer, wie zuvor beschrieben, vorzuziehen. Bei der Varianzschätzung ist der GREG-Schätzer effizienter. Rein Modell-basierte Schätzverfahren können bei sehr kleinen Anteilen Verbesserungen der Schätzungen ermöglichen. Diese sind aber ohne tiefere Kenntnisse der Methodik schwer realisierbar, da hier allgemeine Regeln nicht angegeben werden können.

Hypercubes zu *Ziel 2*: Bei *Ziel 2*-Hypercubes können prinzipiell alle drei Schätzverfahren, der GREG-Schätzer, der YOURAO-Schätzer sowie der χ^2-Schätzer verwendet werden, da auf NUTS2 hinreichend viele Stichprobeneinheiten in den Hypercube-Zellen vorhanden sind. Empfohlen wird wiederum eine Kombination von χ^2- und GREG-Schätzer. Die Kohärenz auf Bundeslandsebene ist hierbei gewährleistet. Mit Hilfe des verallgemeinerten Kalibrierungsansatzes von Münnich et al. (2012), der nicht Gegenstand dieser Forschungsarbeiten war, sollten tiefer gehende Antworten bezüglich horizontaler und vertikaler Kohärenz möglich werden.

Obige Empfehlungen sind eher konservativ, um einen sicheren Einstieg in den Register-gestützten Zensus zu ermöglichen. Tatsächlich können Verbesserungen in den Schätzungen und zwar auf allen Ebenen erreicht werden. Allerdings bedürfen diese eines nicht zu unterschätzenden Aufwands in der zukünftigen Forschung. Des Weiteren muss angemerkt werden, dass konkret im Forschungsprojekt Stichprobendesign und Schätzung im Vordergrund standen. Nachfolgende Forschungsarbeiten sollten sich sicher auf einen ganzheitlichen Ansatz konzentrieren, indem auch weitere Aspekte, wie das Matching von Daten verschiedener Register unter Berücksichtigung von Geheimhaltungsverfahren sowie die Behandlung von Missing Values, integriert werden.

Der Methodik eines Register-gestützten Schätzverfahrens gehört sicherlich die Zukunft. Sie liefert aus derzeitiger Sicht Einsparungspotenziale im Vergleich zu herkömmlichen Volkszählungen sowie gleichzeitig einen Effizienzgewinn im Vergleich zu rein direkten Schätzverfahren. Auf diese Weise könnte man beispielsweise auch die Effizienz des Mikrozensus verbessern. Voraussetzung ist ein sicherer Umgang mit dem Paradigmenwechsel, der eine Abkehr von einer reinen Zählung impliziert sowie die Verwendung von Design-verzerrten Schätzmethoden bei kleinen Nachweiseinheiten erfordert. Dieser beinhaltet tiefer gehende Kenntnisse sowohl in den Methoden als auch in der Umsetzung in Datenanalyse-Programmen. Letztendlich müssen auch die Nutzer in der amtlichen Statistik und Wissenschaft den Hintergrund verstehen, um die Daten in effizienter Weise verwenden zu können. Eingangs wurde in dieser Monographie die Kritik von Gelman an Survey-Gewichten eingehend diskutiert, auch wie sie in diesen Arbeiten berücksichtigt wurden. Von einer zukünftigen Forschung wird man aber auch erwarten, dass Besonderheiten der Survey-Statistik in geeigneter Weise in statistisch-ökonometrischen Modellen berücksichtigt werden, da absehbar ist, dass die Daten immer mehr in hochkomplexen und effizienten Erhebungen gewonnen werden.

5 Chi-Quadrat Verfahren

In (2.4.12) wurde die Matrix G wie folgt definiert:

$$G = \begin{pmatrix} D_{Me} + \frac{E_{II}}{I+J} & M - \frac{E_{IJ}}{I+J} \\ M^{'} - \frac{E_{JI}}{I+J} & D_{e'M} + \frac{E_{JJ}}{I+J} \end{pmatrix}^{-1} \begin{pmatrix} Id_I \\ P' \end{pmatrix} \quad . \tag{5.1}$$

Zunächst können wir G auch schreiben als

$$G = \begin{pmatrix} \Delta_1 \\ 0 \end{pmatrix} + \begin{pmatrix} E_{II}\Delta_2 \\ -E_{JI}\Delta_2 \end{pmatrix} \quad , \tag{5.2}$$

wobei Δ_1 und Δ_2 Diagonalmatrizen sind mit $\Delta_1 = D_{Me}^{-1}$ und $\Delta_2 = -\frac{D_{Me}^{-1}}{I+J} = -\frac{\Delta_1}{I+J}$. Dies kann leicht durch Multiplikation von G von links mit

$$\begin{pmatrix} D_{Me} + \frac{E_{II}}{I+J} & M - \frac{E_{IJ}}{I+J} \\ M^{'} - \frac{E_{JI}}{I+J} & D_{e'M} + \frac{E_{JJ}}{I+J} \end{pmatrix}$$

überprüft werden. Es sei $e_{i,j+I}$ ein Vektor von Nullen außer an der $i-$ten und $(j+I)-$ten Stelle, an denen jeweils eine 1 steht. Aus

$$e'_{i,j+I} \begin{pmatrix} E_{II}\Delta_2 \\ -E_{JI}\Delta_2 \end{pmatrix} = 0$$

folgt

$$\lambda_i + \mu_j = e'_{i,j+I} G * t = \frac{t_i}{m_{i\cdot}}$$

und daher

$$n_{ij}^{opt} = m_{ij}\left(\lambda_i + \mu_j\right) = m_{ij}\frac{t_i}{m_{i\cdot}} = p_{ij}t_i$$

6 Standardmaße und Erläuterungen zu den Abbildungen

Spezielle Abbildungen

Der Violinplot

Der Violinplot, oder auch Vioplot genannt, der von Daniel Adler entwickelt wurde, baut auf dem Boxplot auf, zusätzlich sind in dieser Art von Darstellung noch Informationen über die Dichte der Daten abzulesen. Die Dichte wird dabei durch einen Kern-Schätzer berechnet. Der weiße Punkt in der Mitte zeigt ähnlich wie beim Boxplot den Median der Daten an. Je weiter die Ausdehnung nach rechts und links ist, desto größer ist die Dichte an dieser Stelle. Wenn ein Strich über der Ausdehnung zu sehen ist bedeutet dies, dass Ausreißer in der Verteilung enthalten sind.

Der Mosaikplot

Der Mosaikplot wurde von David Meyer, Achim Zeileis und Kurt Hornik entwickelt. Es handelt sich dabei um eine direkte Visualisierung von Kontingenztabellen. Es werden dabei zweidimensionale Häufigkeitsverteilungen mit Hilfe von Flächen dargestellt. Dabei ist eine Fläche umso größer, je öfter die Merkmalskombination auftritt. Nähere Erklärungen zum Mosaikplot sind in Hofmann (2003) zu finden.

Standardmaße zur Bewertung der Ergebnisse

Zur Bewertung der Schätzungen wird zunächst einmal das klassische Spektrum an Maßen herangezogen. Dazu ist – neben dem relativen Root Mean Square Error – der relative Bias und die relative Dispersion zu zählen. Diese Maße veranschaulichen den Zielkonflikt zwischen Verzerrung und Streuung, der im Rahmen der zusammengesetzten Schätzer von besonderer Bedeutung ist (siehe Abschnitt 2.3).

Zur Berechnung der klassischen Maße werden sowohl der Erwartungswert wie auch die Varianz von $\widehat{\tau}_d$ über die durch die M Stichproben aufgebauten Schätzverteilung wie folgt geschätzt:

$$\widehat{\tau}_{d,mean} = \frac{1}{M}\sum_{j=1}^{M}\widehat{\tau}_{d,j} \quad , \tag{6.1}$$

$$\widehat{\tau}_{d,var} = \frac{1}{M-1}\sum_{j=1}^{M}(\widehat{\tau}_{d,j} - \widehat{\tau}_{d,mean})^2 \quad . \tag{6.2}$$

Mit dem relativen Root Mean Squared Error wird die Abweichung eines Schätzers von dem zu schätzenden Wert TVal_d auf der Skala des zu schätzenden Wertes berechnet. Er kann positive Werte oder den Wert Null annehmen.

$$\mathrm{RRMSE}_d = \frac{\sqrt{\widehat{\tau}_{d,var} + (\widehat{\tau}_{d,mean} - \mathrm{TVal}_d)^2}}{\mathrm{TVal}_d} \tag{6.3}$$

Mit dem relativen Bias wird veranschaulicht, inwieweit die Schätzergebnisse über die Stichproben hinweg verzerrt sind. Die Verzerrung wird in Relation zu den wahren Werten gesetzt, um eine einfache Interpretation zu ermöglichen. Der relative Bias nimmt reelle Werte an und ist wie folgt zu interpretieren: Wenn der relative Bias

negativ ist, unterschätzt der Schätzer im Mittel den wahren Wert,

Null ist, ist er unverzerrt über die Stichproben,

positiv ist, überschätzt der Schätzer im Mittel den wahren Wert.

$$\text{RBias}_d = \frac{\widehat{\tau}_{d,mean} - \text{TVal}_d}{\text{TVal}_d} \tag{6.4}$$

Die relative Dispersion misst, wie weit das 0,05-Quantil vom 0,95-Quantil relativ zum wahren Wert entfernt ist. Somit ist sie mindestens Null und maximal unendlich. Je höher die relative Dispersion ist, desto weiter liegen die geschätzten Werte über die Stichproben auseinander. Mit anderen Worten, je kleiner die relative Dispersion ist, desto enger ist der Bereich der Schätzergebnisse über alle Stichproben.

$$\text{RDisp}_d = \frac{Q(\widehat{\tau}_{d,S=1..M}, 0,95) - Q(\widehat{\tau}_{d,S=1..M}, 0,05)}{\text{TVal}_d} \tag{6.5}$$

Abbildungen zur Darstellung der Simulationsergebnisse

Aufgrund der vielen verschiedenen Kombinationsmöglichkeiten von Designs, Karteileichen- und Fehlbestandsmodellen, Fragestellungen sowie Schätzern ist es ohne Informationsverlust nicht möglich, eine klassische Abbildung zur Veranschaulichung der Ergebnisse zu generieren. Deswegen wird in diesem Abschnitt ausführlich erklärt und an Beispielen dargestellt, wie die verwendeten Abbildungen zu interpretieren sind.

Aufbau und Struktur der Abbildungen

Eine Abbildung enthält grundsätzlich nur die Werte für die im Titel angegebene Fragestellung. Die vertikal übereinander liegenden Blöcke fassen die Kreuzkombinationen der Designs mit den Karteileichen- und Fehlbestandsmodellen für das über dem Block spezifizierte Registerfehlermodell zusammen. Hierbei sind innerhalb dieser Blöcke die Designs horizontal und die Schätzer vertikal angeordnet.

Abbildung 6.1: Inhalt und Legende der Abbildungen

Um eine SMP-Typ (Sampling Point Typ) spezifische Aussage treffen zu können, sind die verschiedenen SMP-Typen farblich, und an der y-Achse versetzt dargestellt. Die farbliche Identifikation ist oben in der Legende aufgeführt (siehe Abbildung 6.1). Falls ein Benchmark gegeben ist, ist dieser mit einer roten vertikalen Linie gekennzeichnet. In der Legende steht die Höhe des Benchmarks.

Interpretation der Abbildungen

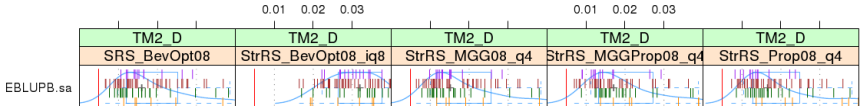


Abbildung 6.2: Vergleich von Designs bzgl. eines Schätzers und eines Karteileichen-/ Fehlbestandsmodells

Die Abbildungen lassen sich in drei Richtungen lesen. Erstens können die Designs bei gegebenen Karteileichen- und Fehlbestandsmodellen und Schätzern verglichen werden. Zweitens lässt sich ein Design mit einem Schätzer über verschiedene Karteileichen- und Fehlbestandsmodelle vergleichen. Drittens kann ein Design mit einem Karteileichen- und Fehlbestandsmodell über verschiedene Schätzer verglichen werden.

Anhand Abbildung 6.2 wird der Vergleich von mehreren Designs mit einem Schätzer und einem Karteileichen- und Fehlbestandsmodell erläutert. Der Schätzer ist in diesem Fall EBLUPB, die Fragestellung ist TotPop02 und das Maß ist der relative Root Mean Square Error (RRMSE). Das Karteileichen-/ Fehlbestandsmodell ist TM2_D.

Die vertikale rote Linie zeigt den vom Auftraggeber vorgegebenen Benchmark an. In Abbildung 6.2 ist zu sehen, dass mit dem EBLUPB unter der Annahme des Karteileichen- und Fehlbestandsmodells TM2_D bei den betrachteten Designs fast keine Gemeinde die erforderte Präzision erreicht, also nicht unterm Benchmark liegt. Weiterhin ist erkennbar, dass die Ergebnisse unter dem Design StrRS_BevOpt08_iq8 stark von den anderen Designs abweichen, die restlichen dargestellten Designs sich aber ähnlich verhalten.

Die blaue Linie kennzeichnet den Kerndichteschätzer des betrachteten Maßes über alle SMPs hinweg. Hierbei wurden für die Kerndichteschätzung die Standardeinstellungen von R verwendet. Im Falle der Verwendung von Abbildungsausschnitten sind Ausreißer oder schiefe Verteilungen der Maße nicht mehr adäquat abgebildet. Dieser Informationsverlust wird aber weitgehend durch die Verwendung der Kerndichteschätzer kompensiert, so dass man einen besseren Überblick über die Ergebnisse bekommt.

Die nächste Frage, die gestellt werden kann, ist, wie eine bestimmte Design-Schätzer-Kombination auf unterschiedliche Karteileichen- und Fehlbestandsmodelle reagiert. Dies wird mit Hilfe von Abbildung 6.3 auf der nächsten Seite erläutert. Auch hier sind wieder Ergebnisse für eine TotPop02 Fragestellung und der RRMSE als Maß abgetragen. Wenn zum Beispiel das Design SRS_BevOpt08 über die verschiedenen Karteileichen- und Fehlbestandsmodelle hinweg verglichen wird, so ist eine klare Änderung der Punktewolken zu erkennen.

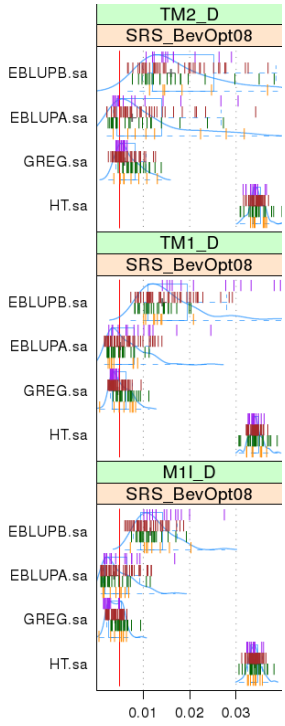


Abbildung 6.3: Vergleich der Registerfehlermodelle und der Schätzer untereinander

Zu sehen ist, dass sowohl der GREG als auch der HT wenig durch die unterschiedliche Modellierung der Karteileichen- und Fehlbestandsmodelle beeinflusst werden. Die Modell-basierten Schätzer EBLUPA und EBLUPB hingegen reagieren recht empfindlich auf höhere Klumpung der Registerfehler (TM2_D) gegenüber homogeneren Registerfehlern (TM1_D und M1I_D). Diese Empfindlichkeit der Modell-basierten Schätzer gegenüber geklumpten Registerfehlern wurde durchgehend über alle Fragestellungen beobachtet.

Die dritte Betrachtungsweise untersucht wie sich die Schätzer untereinander unterscheiden bei gegebenem Karteileichen- und Fehlbestandsmodell und Design. Hierzu kann wieder die Abbildung 6.3 verwendet werden. Hierbei wird nur der obere Abschnitt betrachtet in dem die vier Schätzer auf der y-Achse abgetragen sind.

Man erkennt, dass der GREG mit Abstand die geringsten RRMSEs aufweist. Die beiden EBLUPs schätzen in manchen SMPs besser, in anderen schlechter mit einer großen Variation. Der Horvitz-Thompson-Schätzer hingegen liegt insgesamt sehr hoch mit dem RRMSE, dafür werden alle SMPs mit ähnlichem RRMSE geschätzt. Bei dieser Kombination an Registerfehlermodell und Design wäre also bezüglich der Gesamtbevölkerungsschätzung der GREG zu empfehlen.

Der letzte Punkt, der auf den Abbildungen noch zu sehen ist, ist die verschieden gute Schätzung der unterschiedlichen SMP-Typen. Wie oben erwähnt sind die einzelnen SMP-Typen farblich voneinander zu unterscheiden. Wie zu erwarten war, schneiden die Schätzungen des GREG für den SMP-Typ 0 am Besten ab. Bei den Modell-basierten Schätzern hingegen werden auch die kleineren SMP-Typen noch einigermaßen vernünftig geschätzt.

Auf Grundlage der Phase 1 Grundgesamtheit wurden die schon aus Phase 0 bekannten Monte-Carlo Verfahren mit 1.000 gezogenen Stichproben angewendet, um interessierende Kombinationen aus Schätzern, Auswahlverfahren und KAL/FEB-Modellen im Hinblick auf ihre Qualität (gemessen an Präzision und Verzerrung) zu bewerten. Aufgrund der enormen Datenmengen werden im Folgenden die oben beschriebenen grafischen Darstellungsformen verwendet.

7 Codierung der Bundesländer

Für die Bezeichnung der Schätzer werden Codes verwendet, dabei handelt es sich um eine Nummer 01-16. Die Bundesländer sind nach der geographischen Sortierung durchnummeriert:

Code	Bundesland
01	Schleswig-Holstein
02	Hamburg
03	Niedersachsen
04	Bremen
05	Nordrhein-Westfalen
06	Hessen
07	Rheinland-Pfalz
08	Baden-Württemberg
09	Bayern
10	Saarland
11	Berlin
12	Brandenburg
13	Mecklenburg-Vorpommern
14	Sachsen
15	Sachsen-Anhalt
16	Thüringen

Tabelle 7.1: Kodierung der Bundesländer

8 Ausgewählte Abbildungen in vergrößerter Darstellung

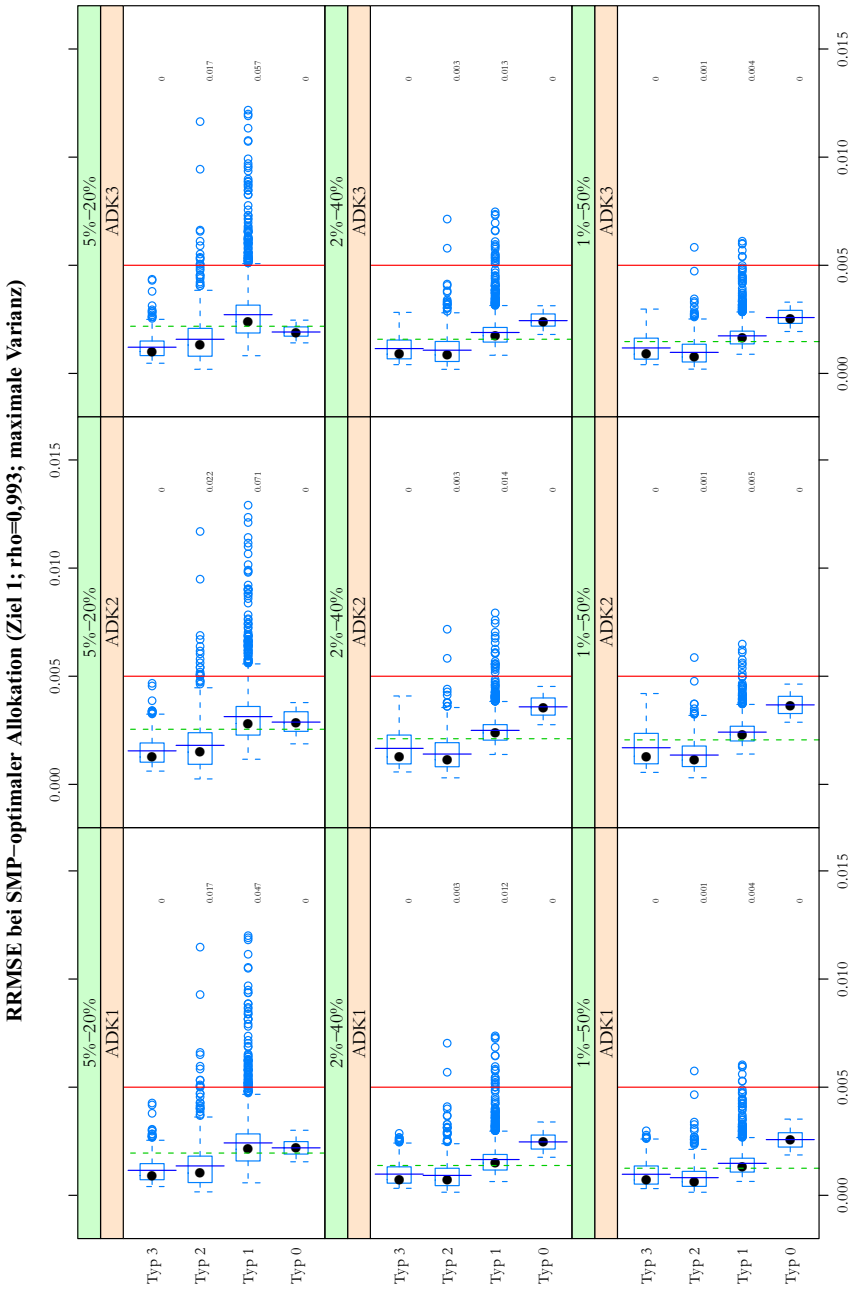


Abbildung 2.3: RRMSE für *Ziel 1* bei $\rho = 0{,}993$ bei drei Schichtungen und drei Entnahmeanteilsvariationen bei SMP-optimaler Allokation

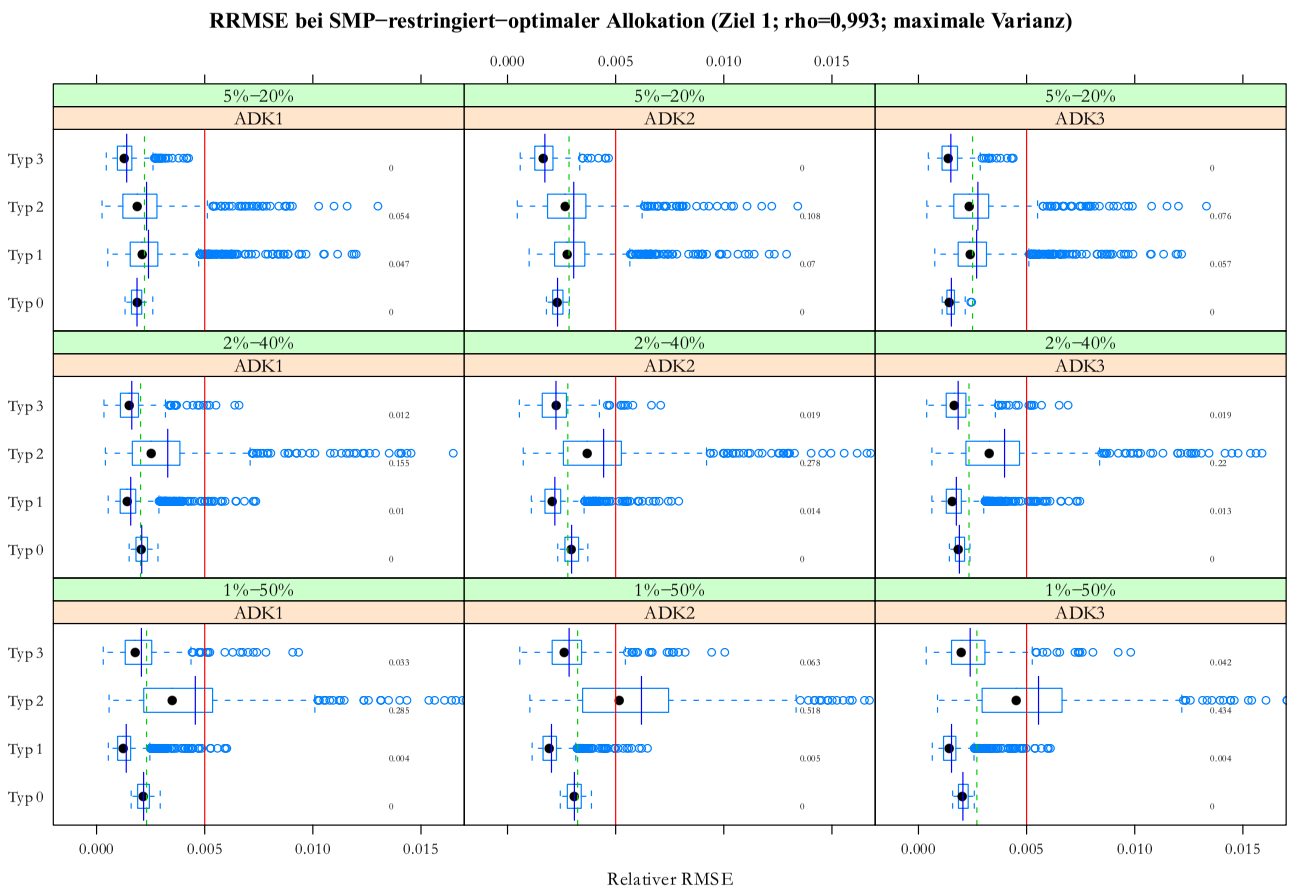


Abbildung 2.4: RRMSE für *Ziel 1* bei $\rho = 0{,}993$ bei drei Schichtungen und drei Entnahmeanteilsvariationen bei eingeschränkter SMP-optimaler Allokation

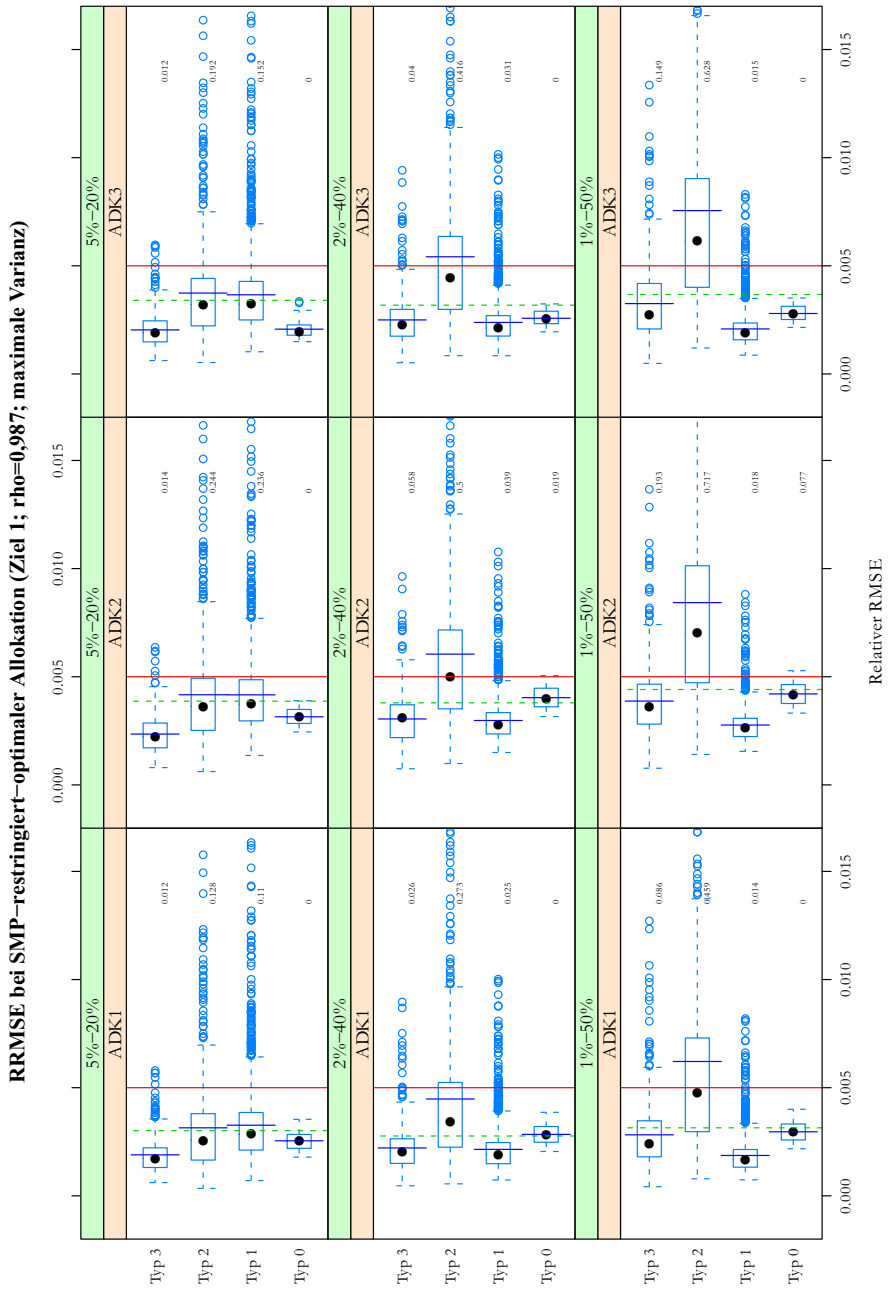


Abbildung 2.5: RRMSE für *Ziel 1* bei $\rho = 0{,}987$ bei drei Schichtungen und drei Entnahmeanteilsvariationen bei eingeschränkter SMP-optimaler Allokation

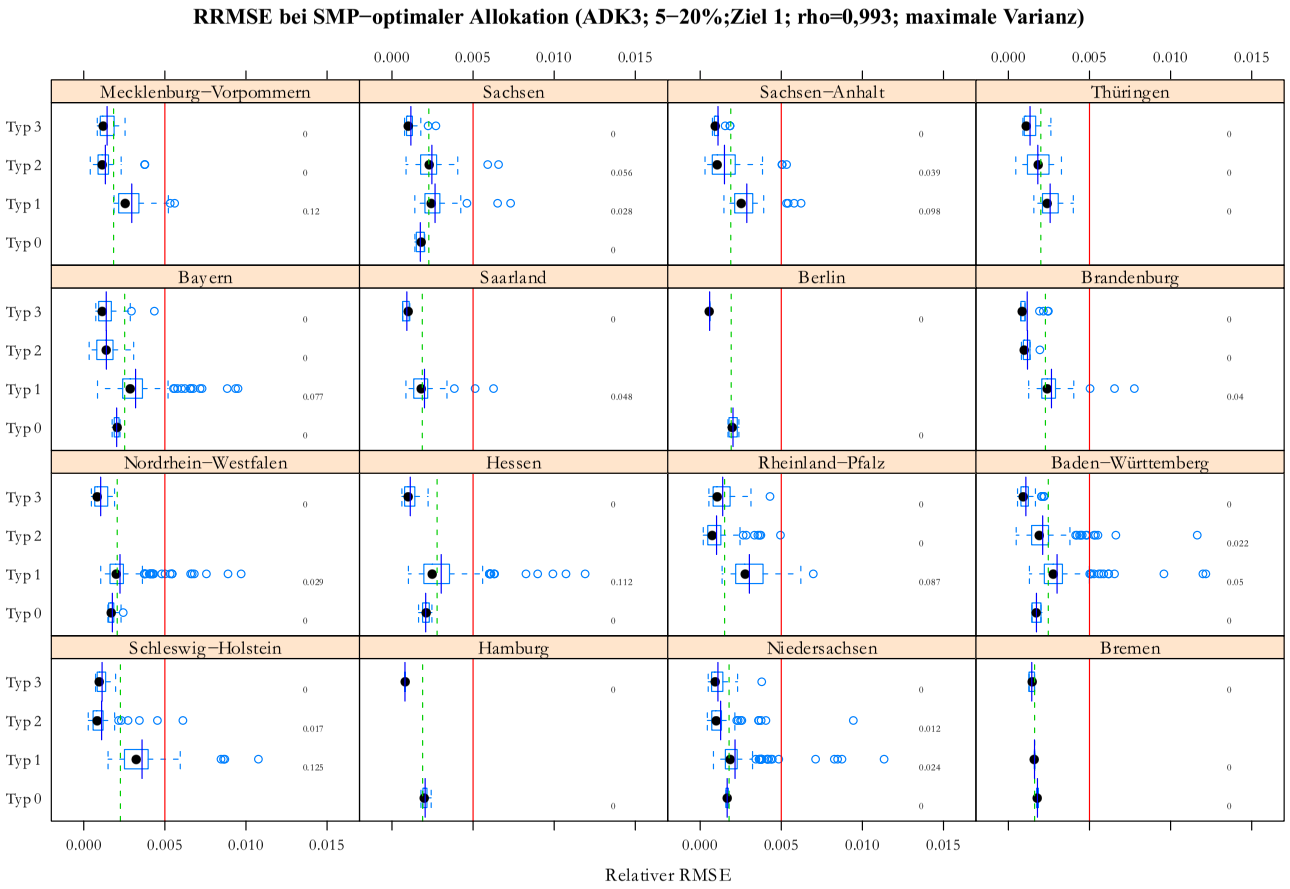


Abbildung 2.6: RRMSE für *Ziel 1* bei $\rho = 0{,}993$ bei ADK3-Schichtungen und Entnahmeanteil 5-20 % bei SMP-optimaler Allokation für die 16 Bundesländer

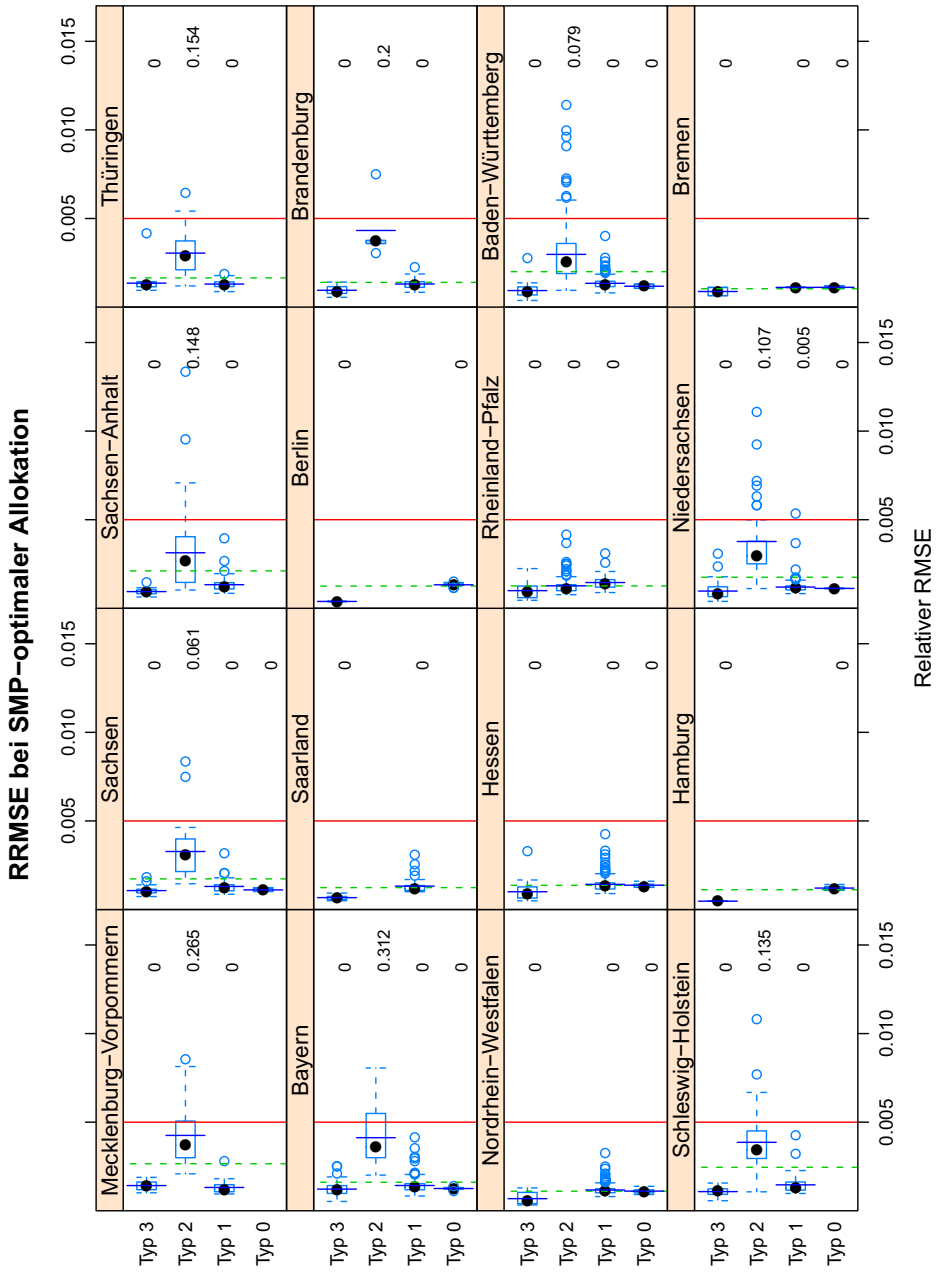


Abbildung 2.7: Theoretische relative RRMSEs in Bezug auf Bundesländer und SMP-Typen

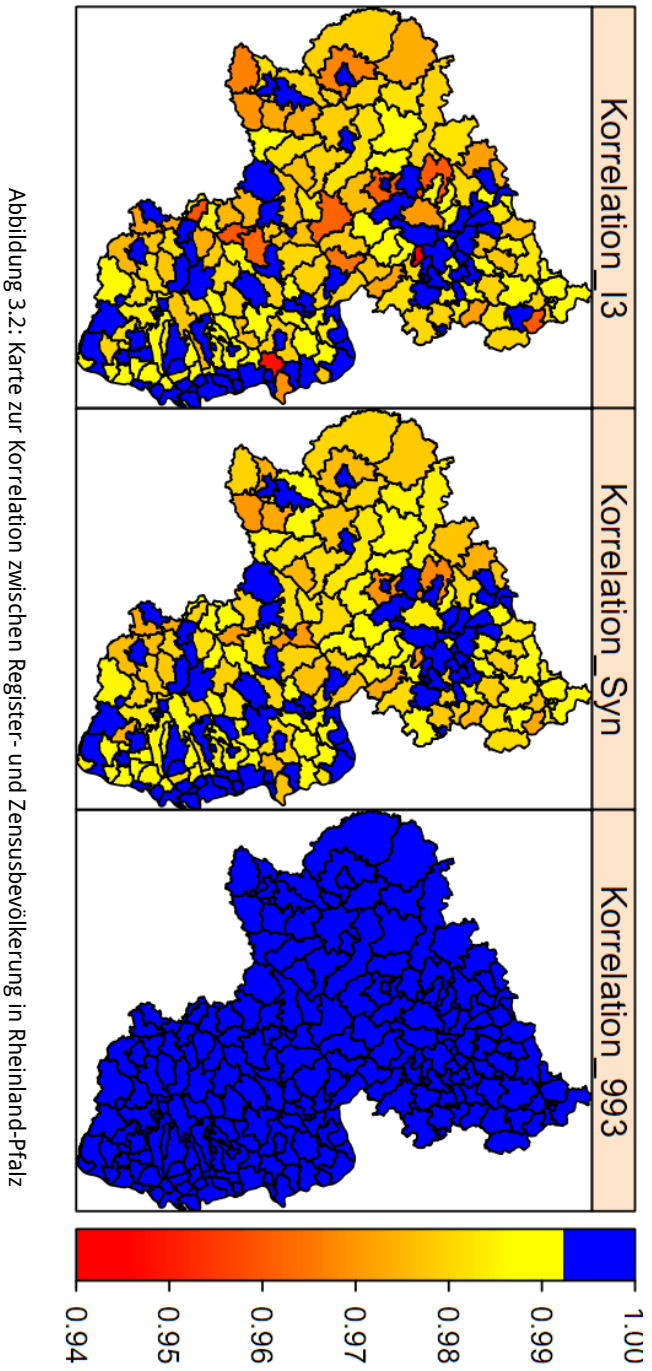


Abbildung 3.2: Karte zur Korrelation zwischen Register- und Zensusbevölkerung in Rheinland-Pfalz

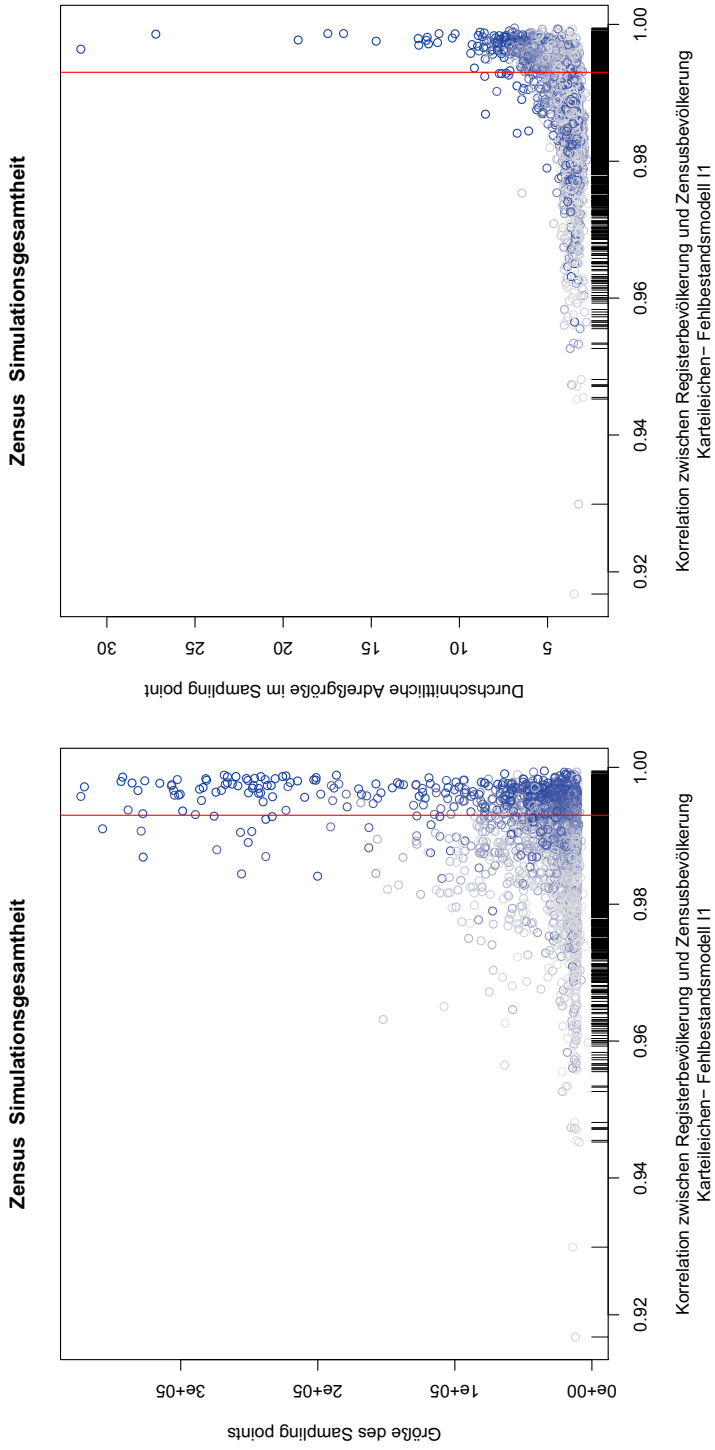


Abbildung 3.3: Korrelation zwischen Register- und Zensusbevölkerung

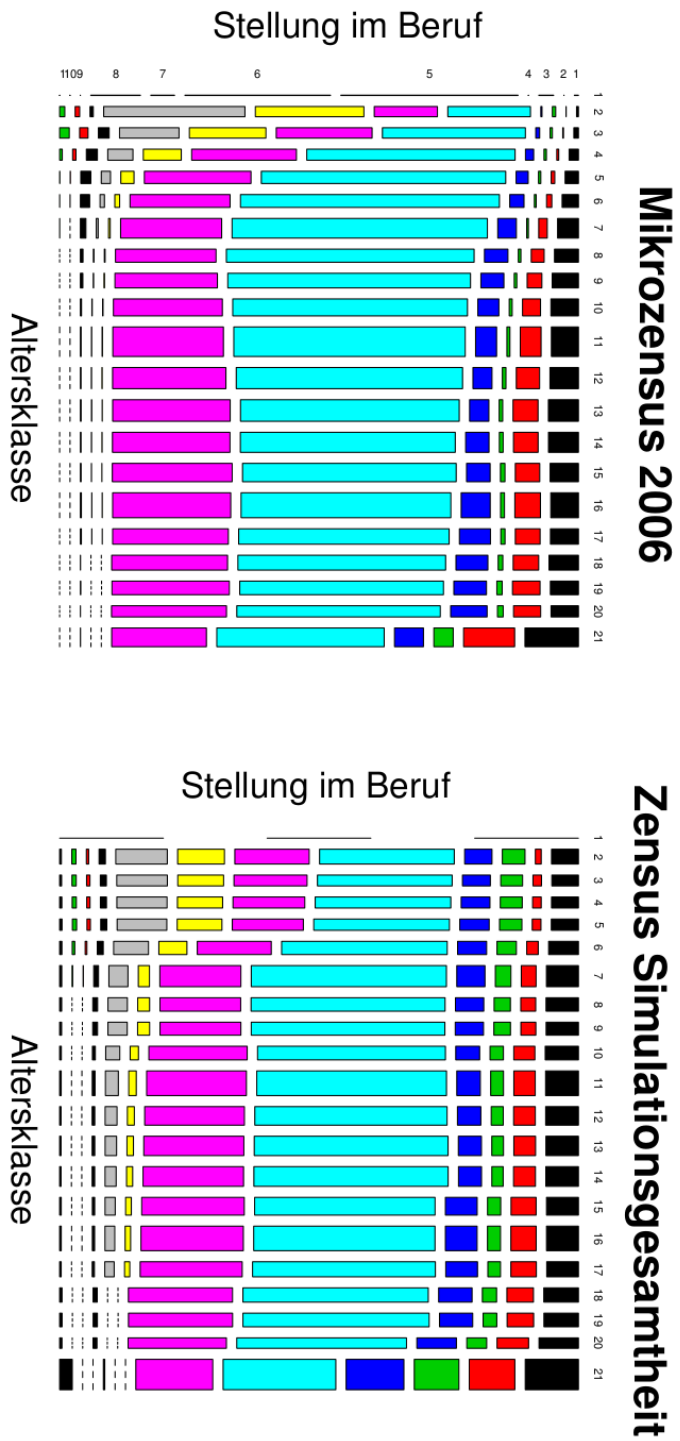


Abbildung 3.4: Zusammenhang zwischen Altersklassen und Stellung im Beruf (Variable EF117)

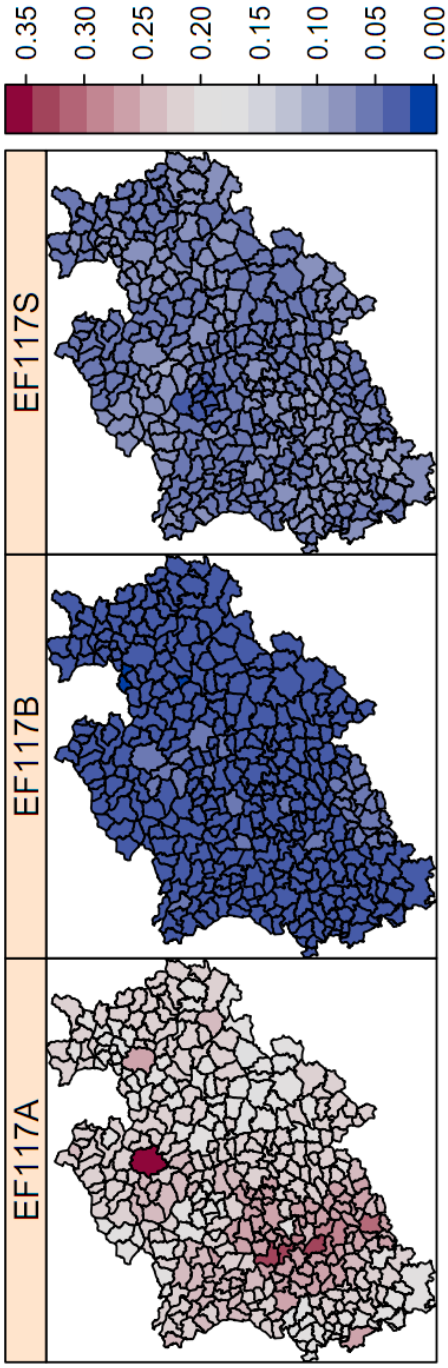


Abbildung 3.5: Räumliche Verteilung der Ausprägungen der Variable EF117 - Stellung im Beruf - in Nordrhein-Westfalen

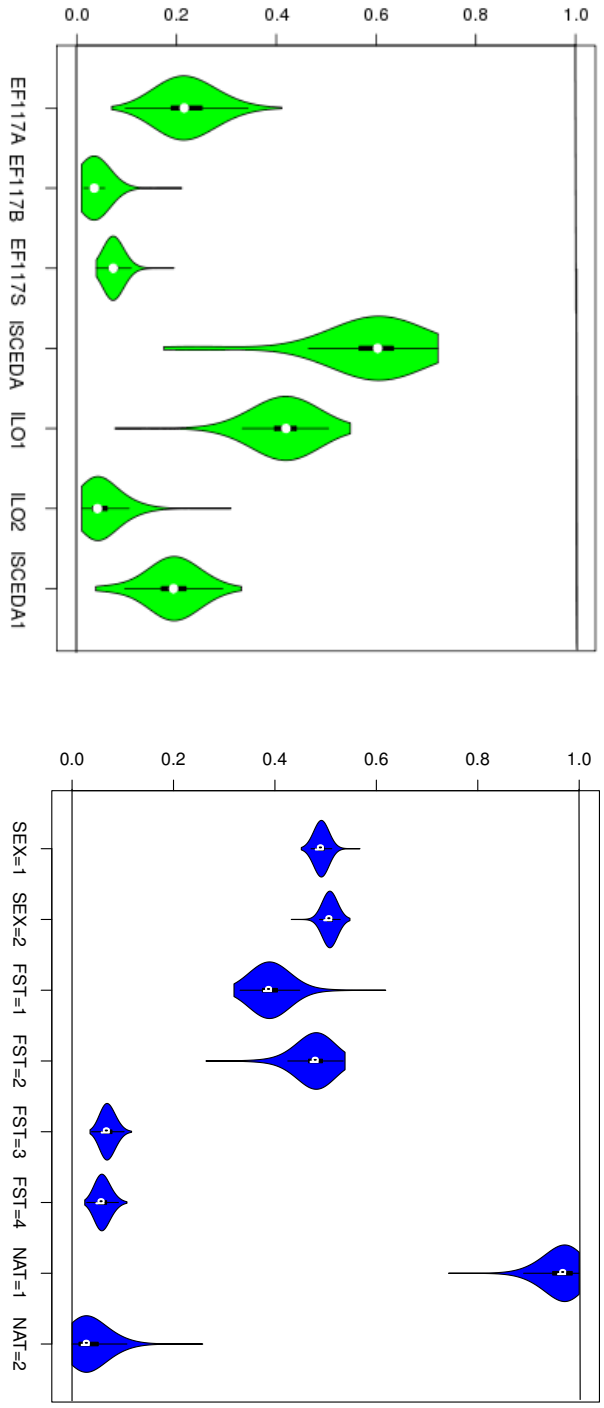


Abbildung 3.7: Vergleich der Variation von Melderegistervariablen und synthetischen Variablen

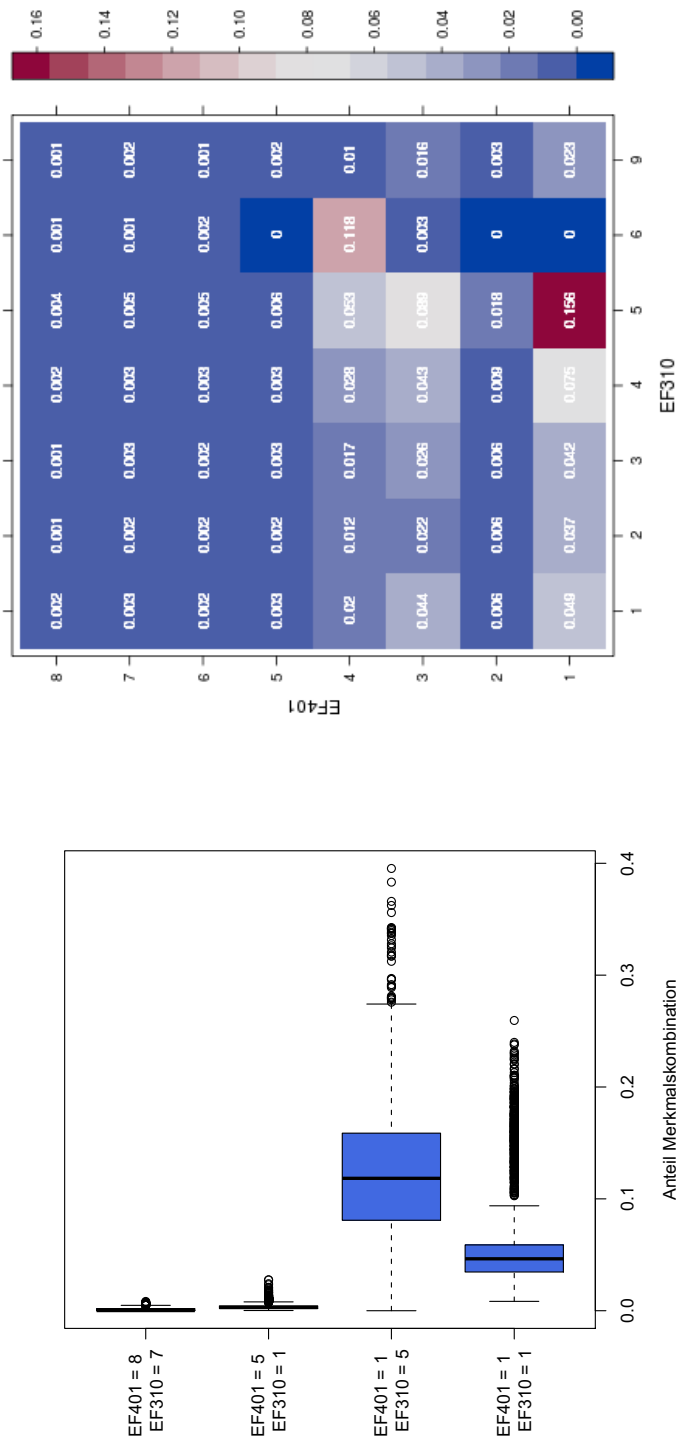


Abbildung 3.8: Relative Häufigkeiten bzgl. der *Ziel 2* Hypercube-Variablen EF310 und EF401 sowie Boxplots

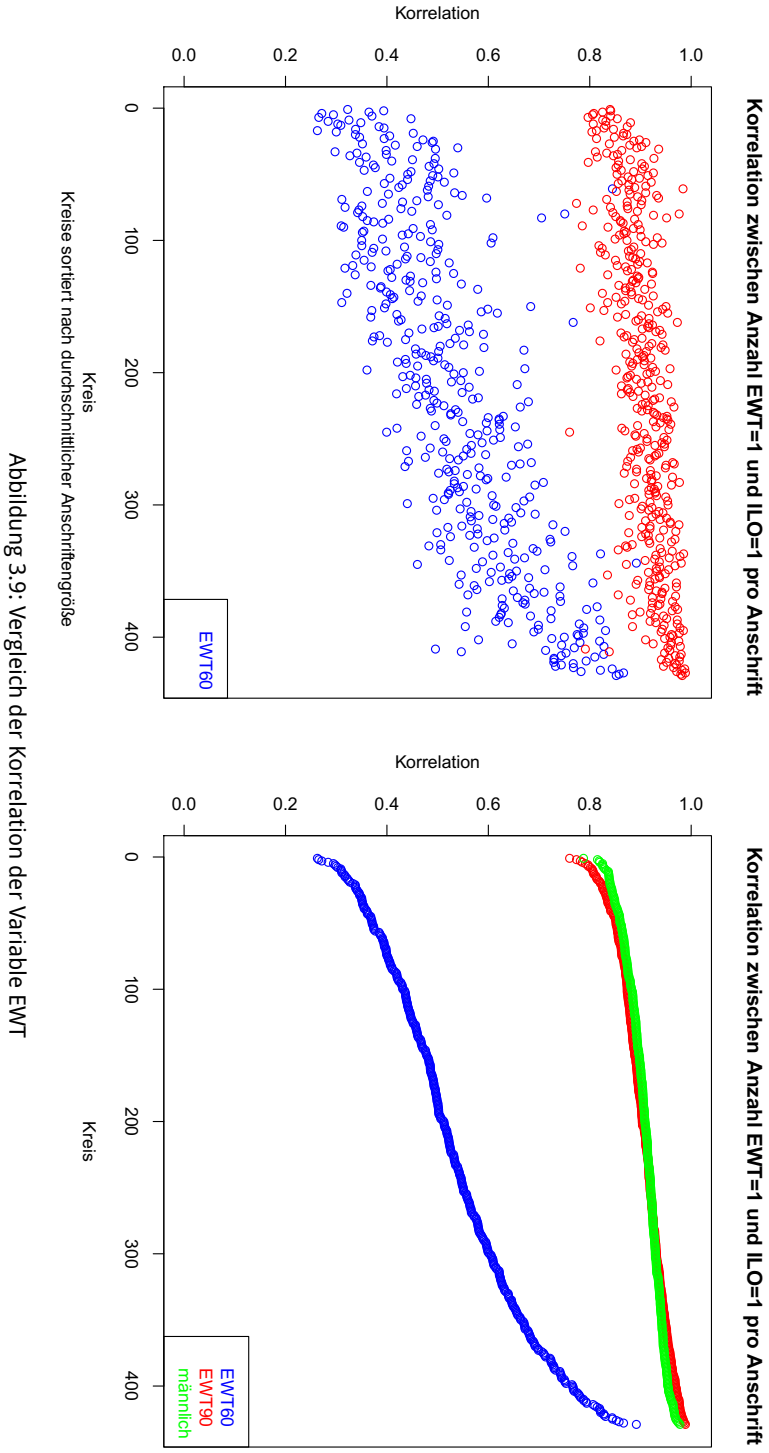


Abbildung 3.9: Vergleich der Korrelation der Variable EWT

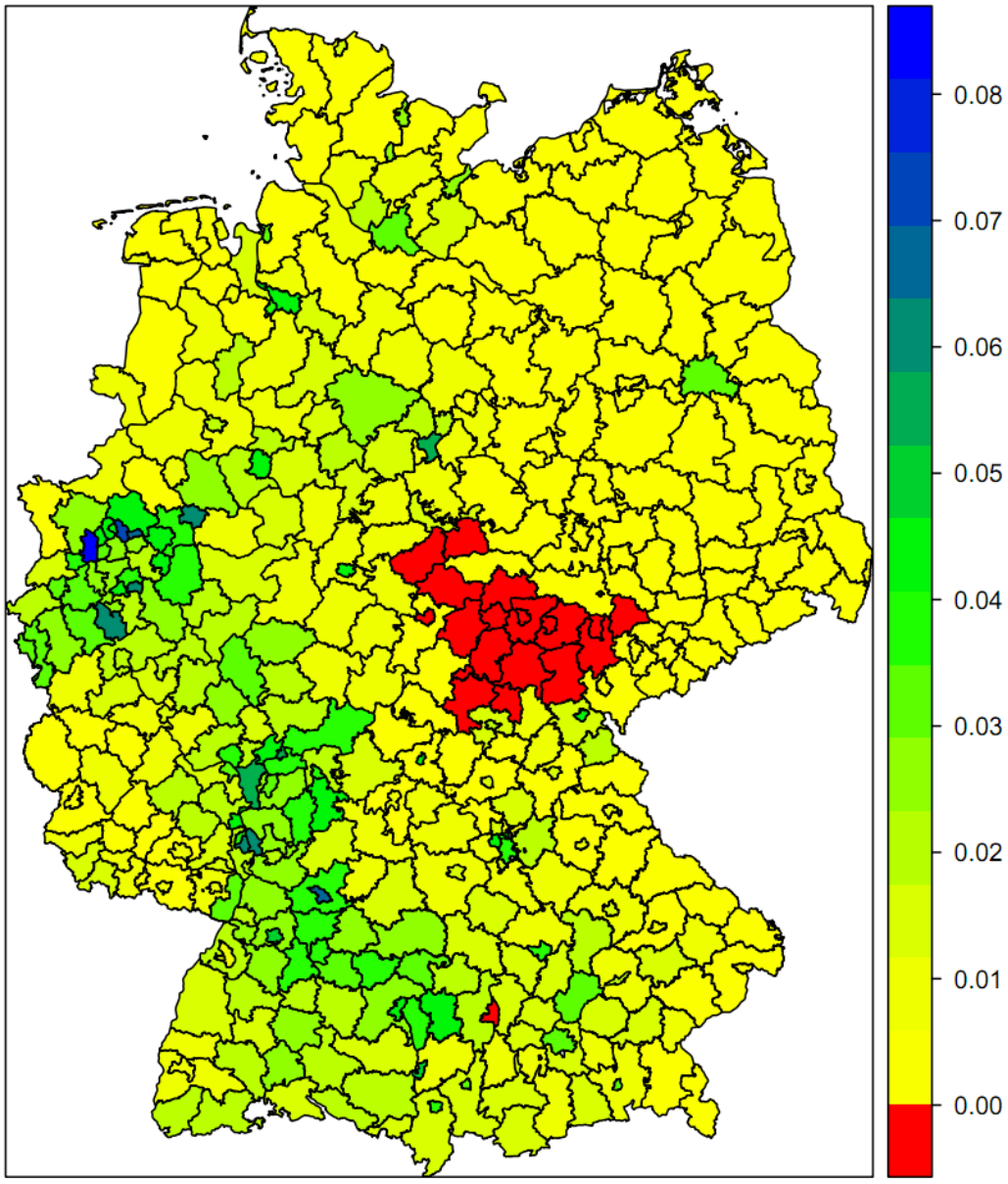


Abbildung 3.10: Anteil der Personen mit türkischer Staatsangehörigkeit an der Gesamtpopulation im Kreis

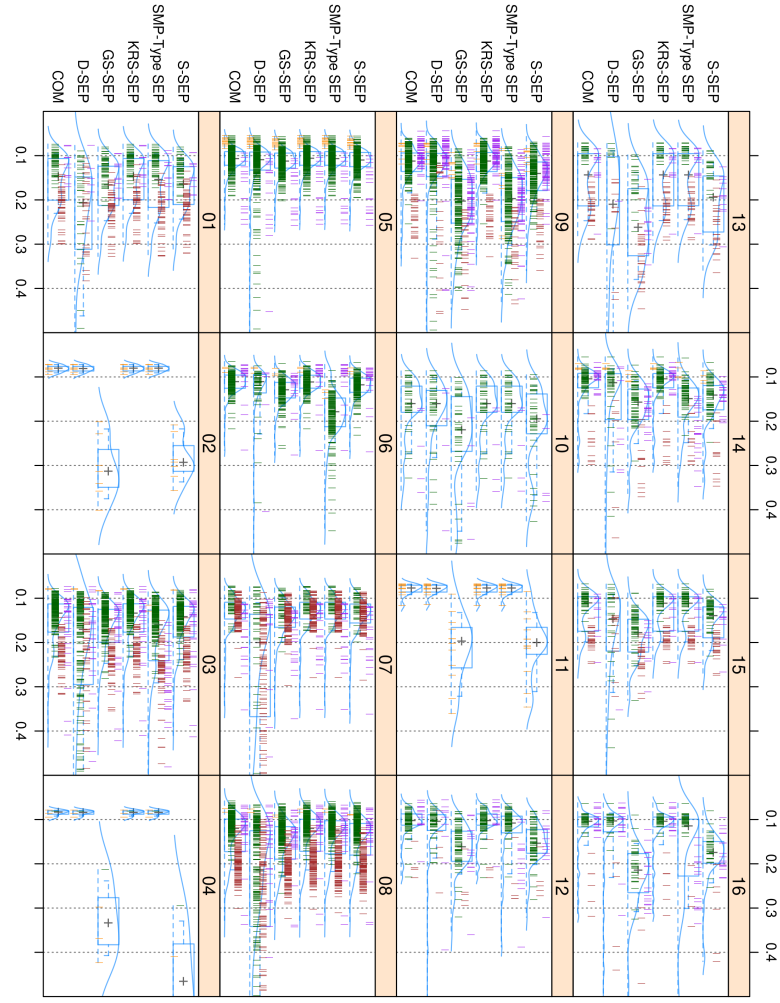


Abbildung 3.17: RRMSE der Karteileichenschätzung in allen 16 Bundesländern nach Modell I1

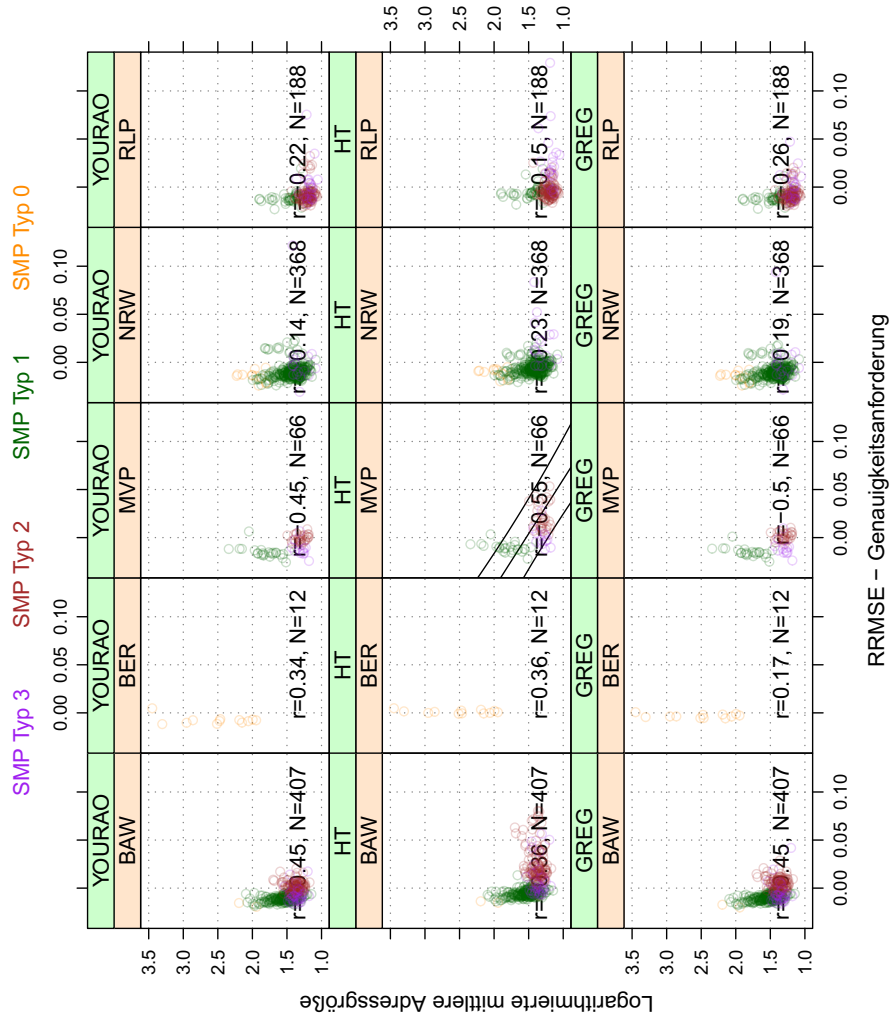


Abbildung 3.19: RRMSE nach SMP Typ und mittlerer Anschriftengröße - Schätzung ISCED

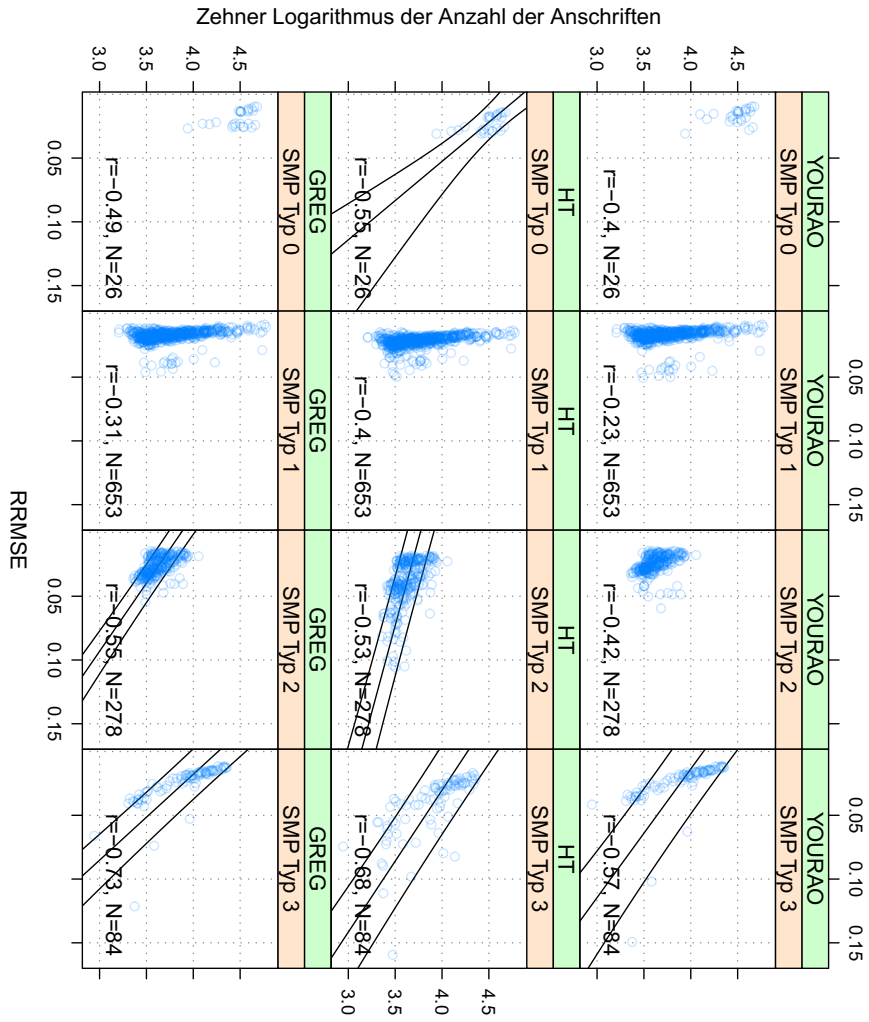


Abbildung 3.20: RRMSE nach SMP Typ und Zahl der Anschriften - Schätzung ISCED

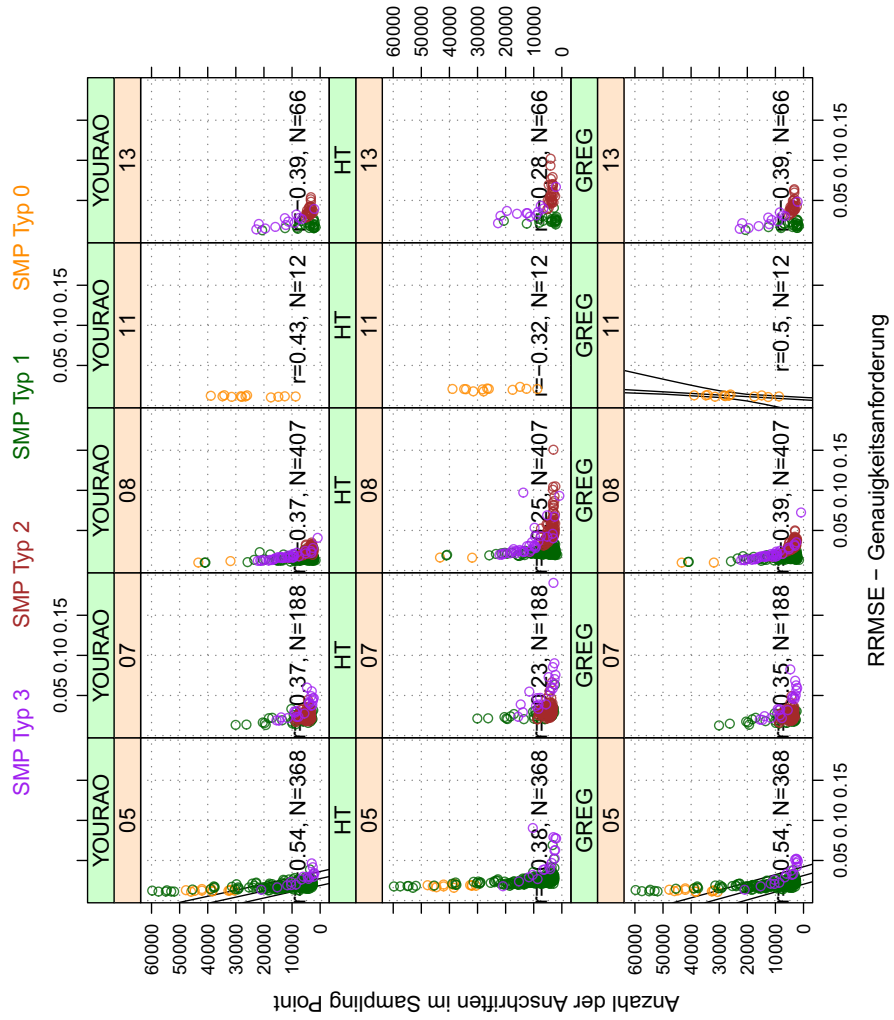


Abbildung 3.22: RRMSE für die Schätzung der Fragestellung ILO1

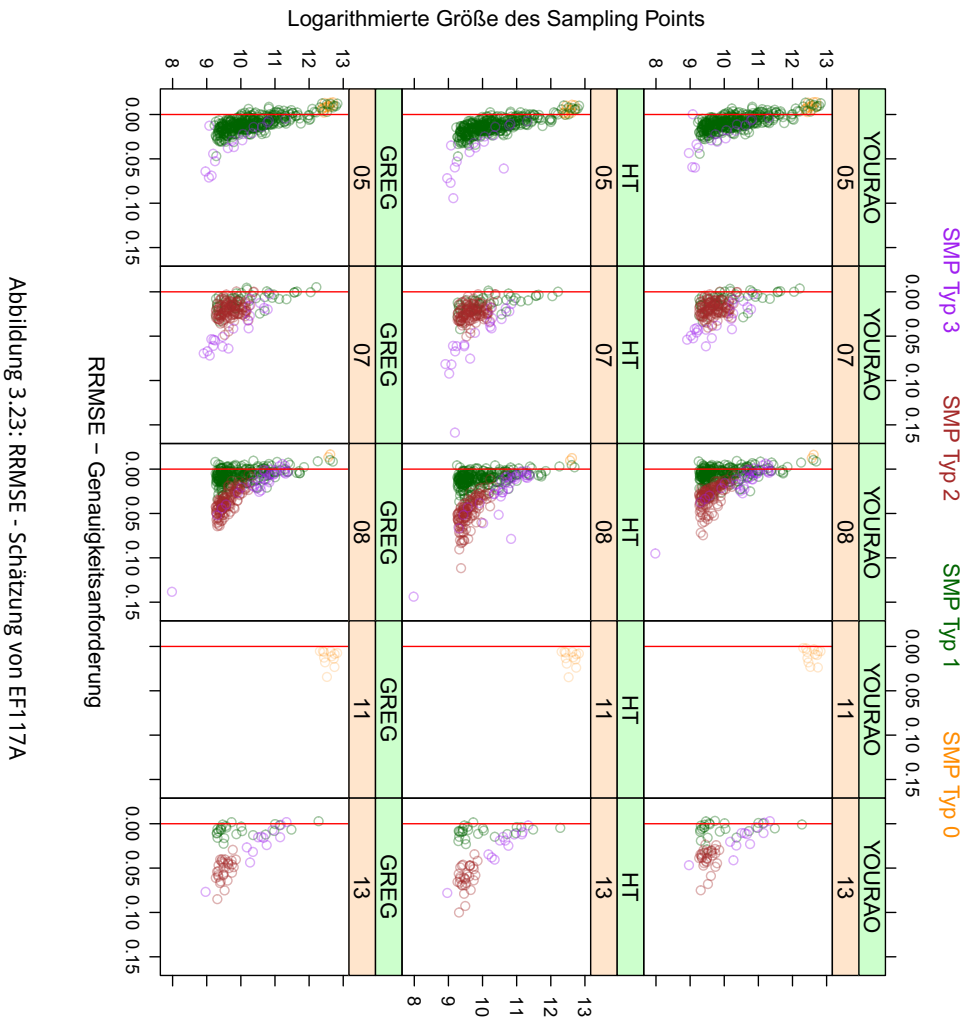


Abbildung 3.23: RRMSE - Schätzung von EF117A

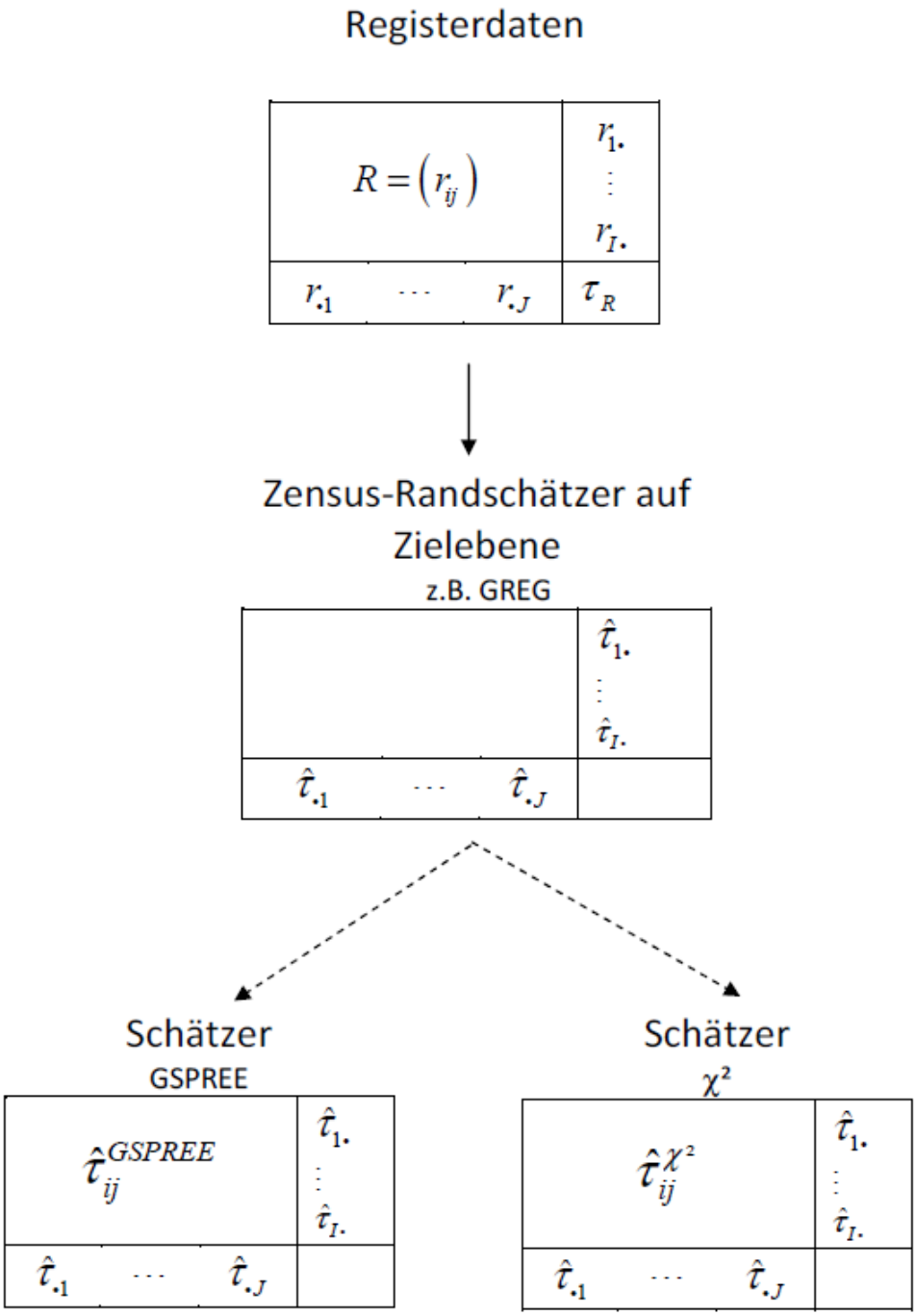


Abbildung 3.28: Schematischer Ablaufplan bei Register als Zusammenhangsstruktur

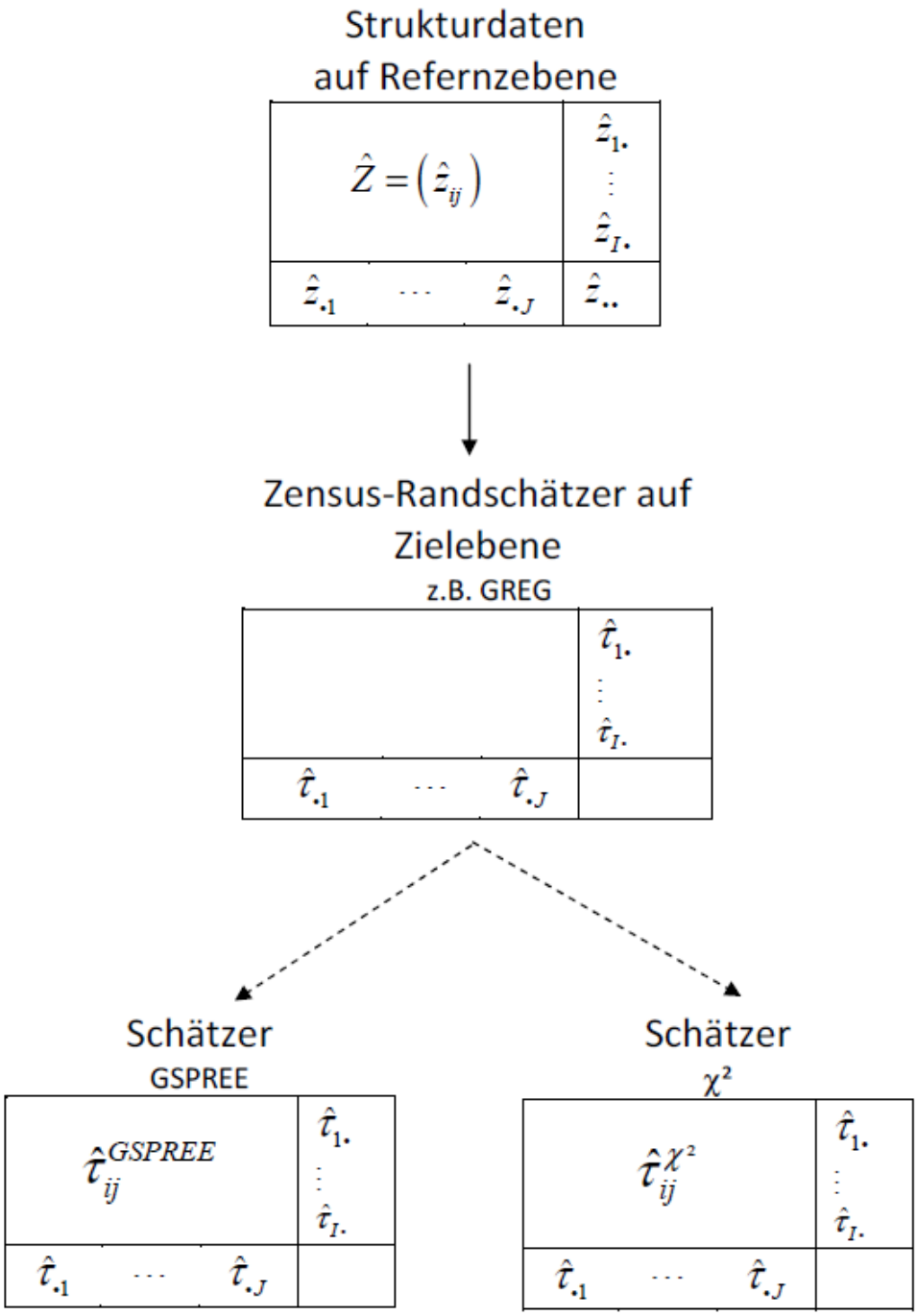


Abbildung 3.29: Schematischer Ablaufplan bei Schätzung auf Bundesland-Ebene als Zusammenhangsstruktur

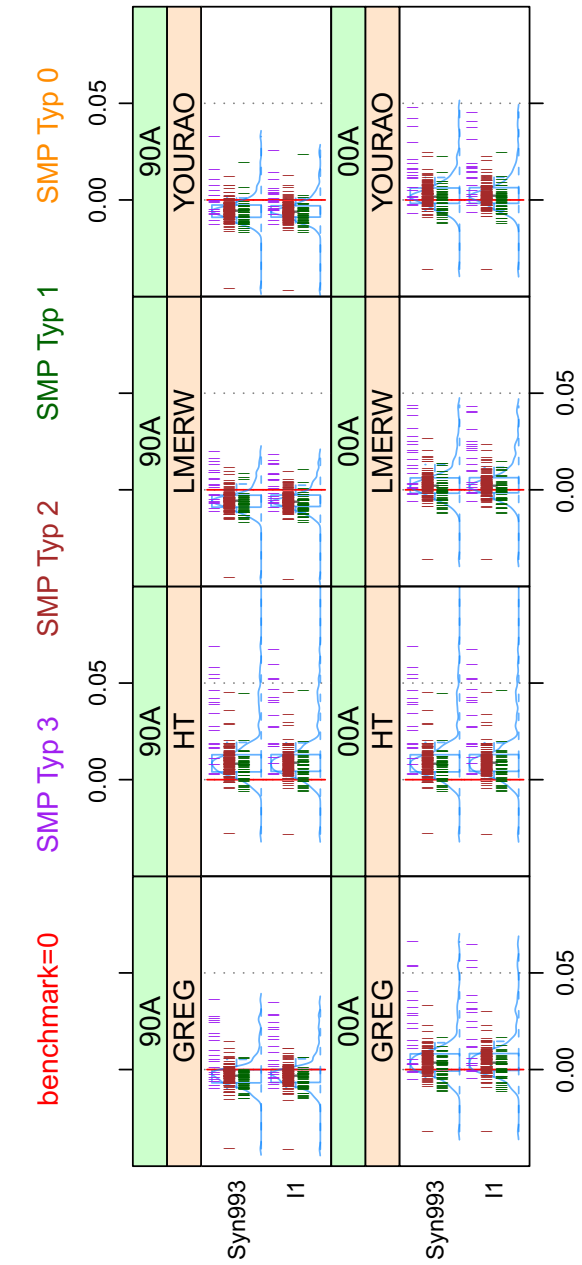


Abbildung 3.41: Schätzung von UBS=1 mit zwei verschiedenen Karteileichen- und Fehlbestandsmodellierungen

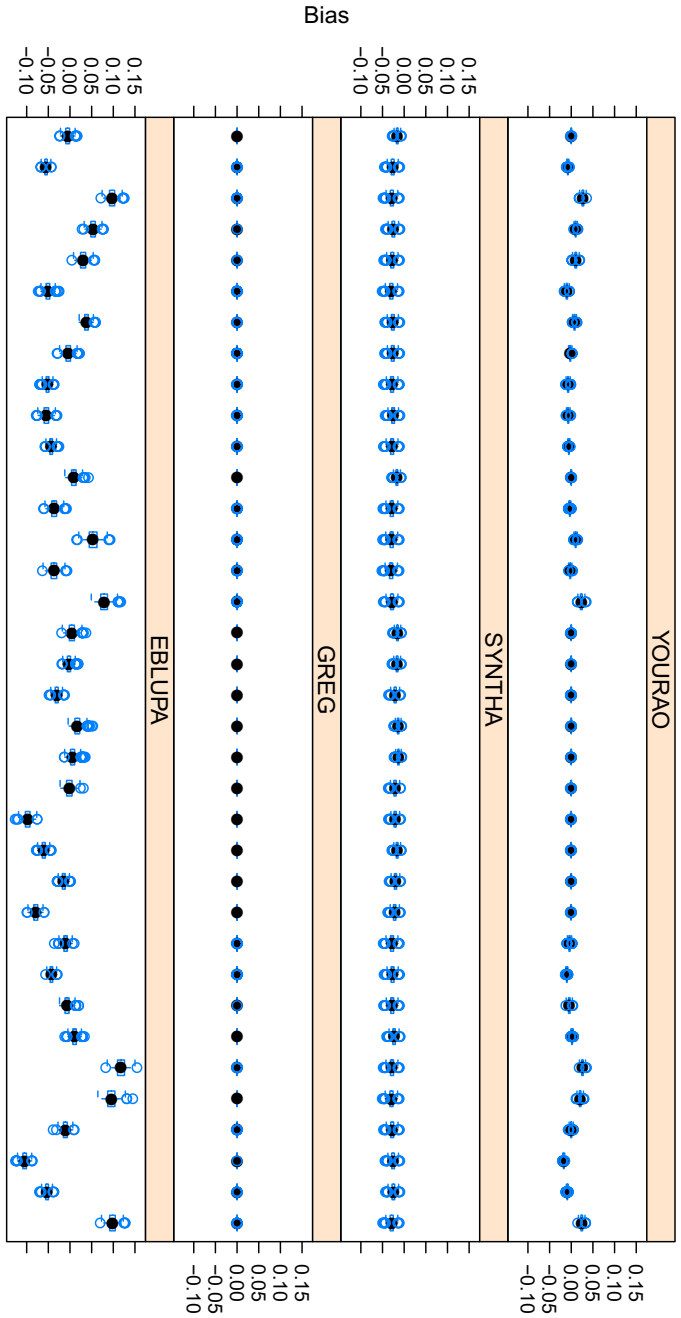


Abbildung 3.42: Kohärenz von ISCED in KRS (Syn993)

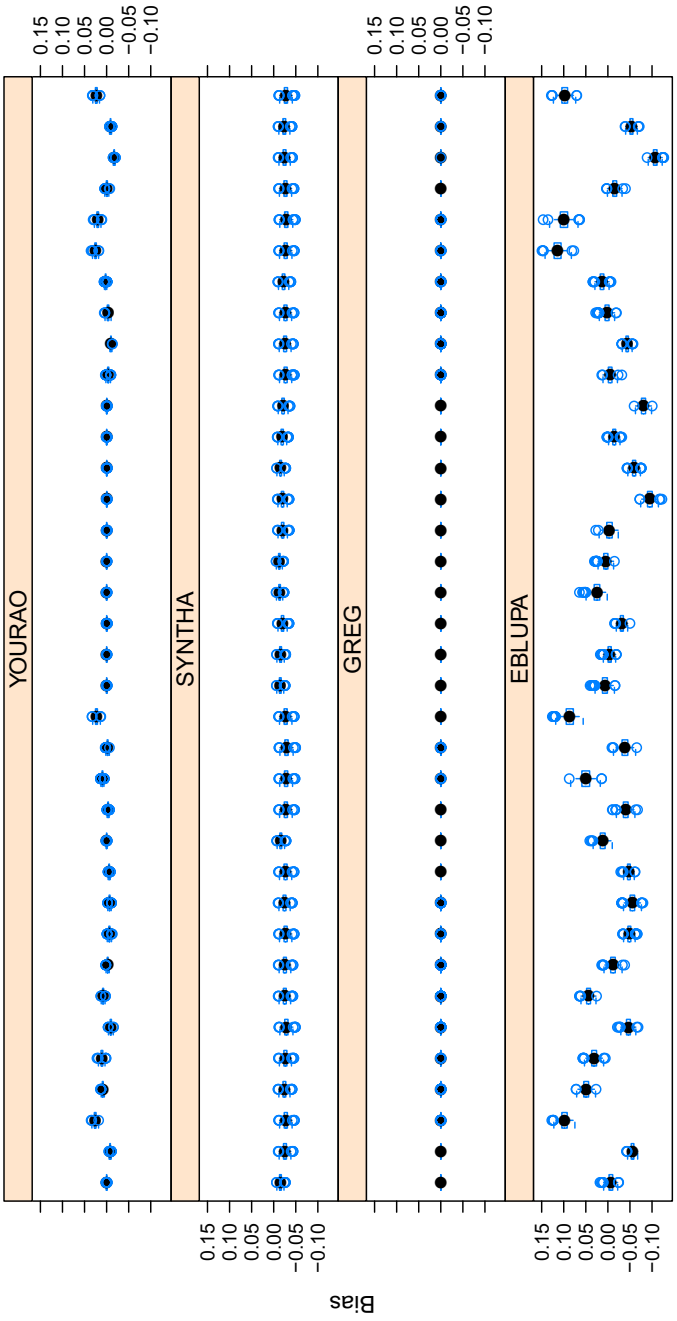


Abbildung 3.43: Kohärenz von ISCED in KRS (I1)

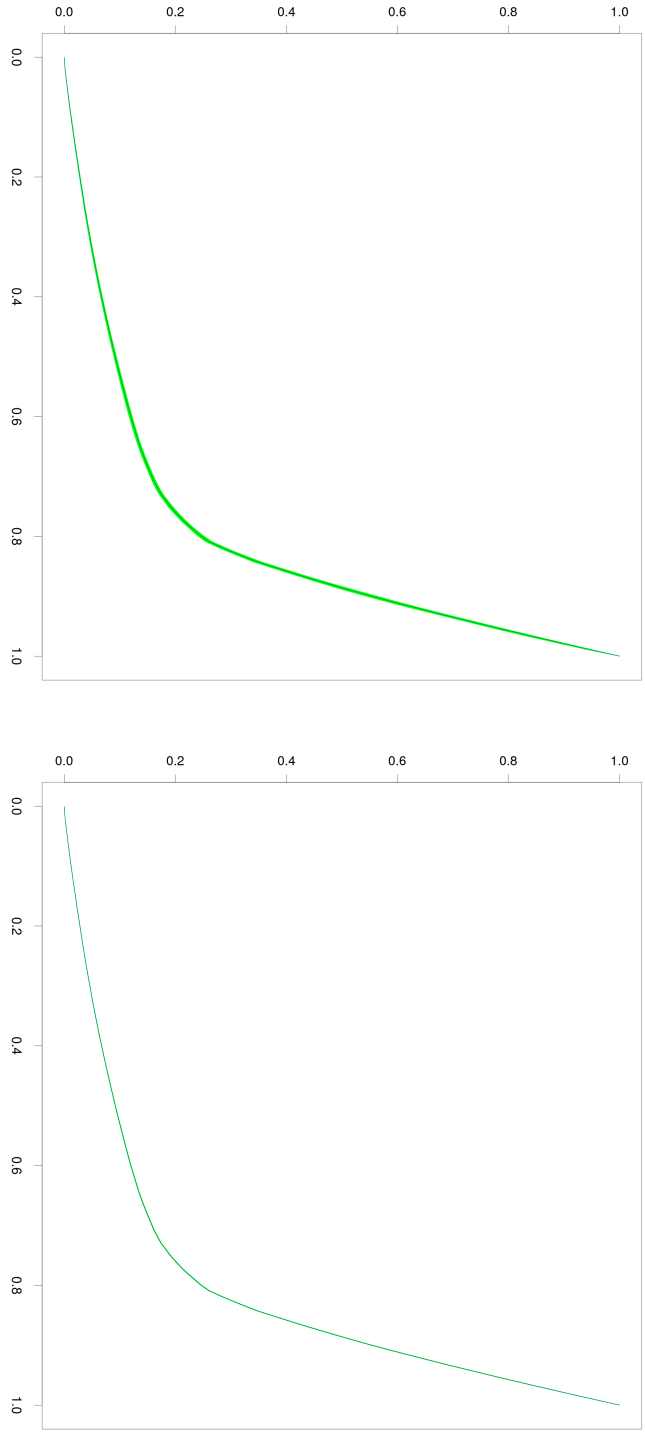


Abbildung 3.44: Lorenzinferenzkurven für SRS (HT (links) und GREG (rechts))

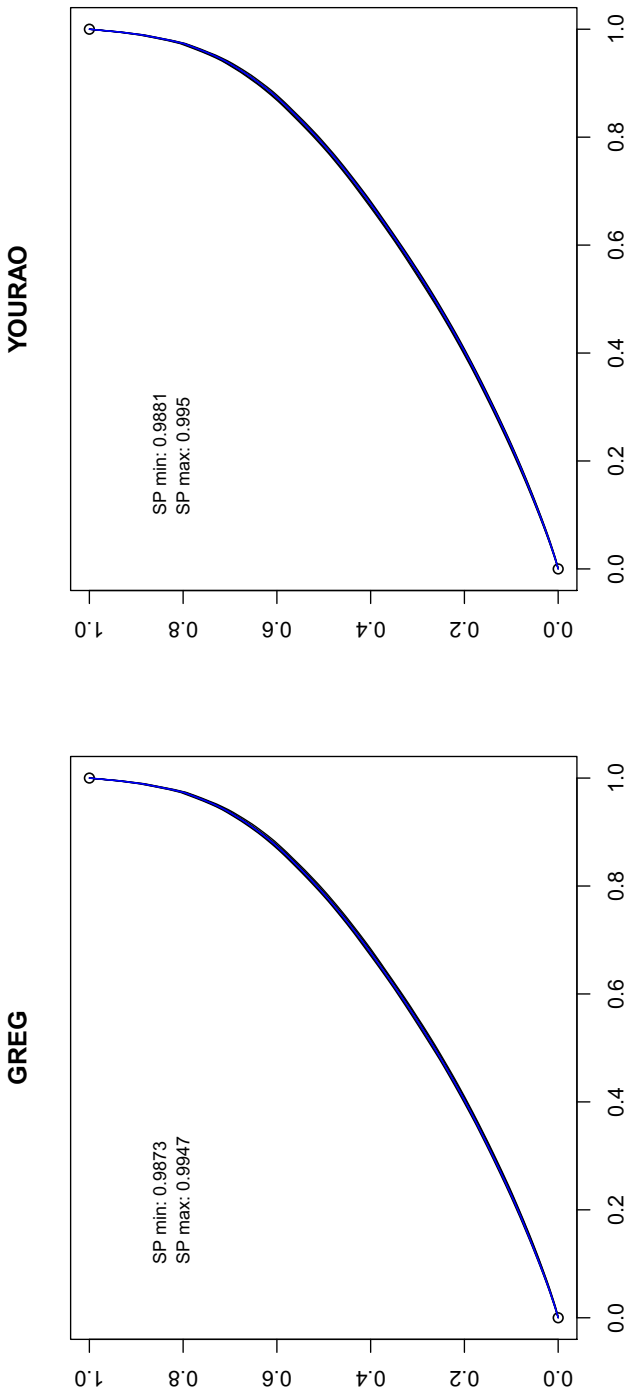


Abbildung 3.46: Disparität: EF117A (I)

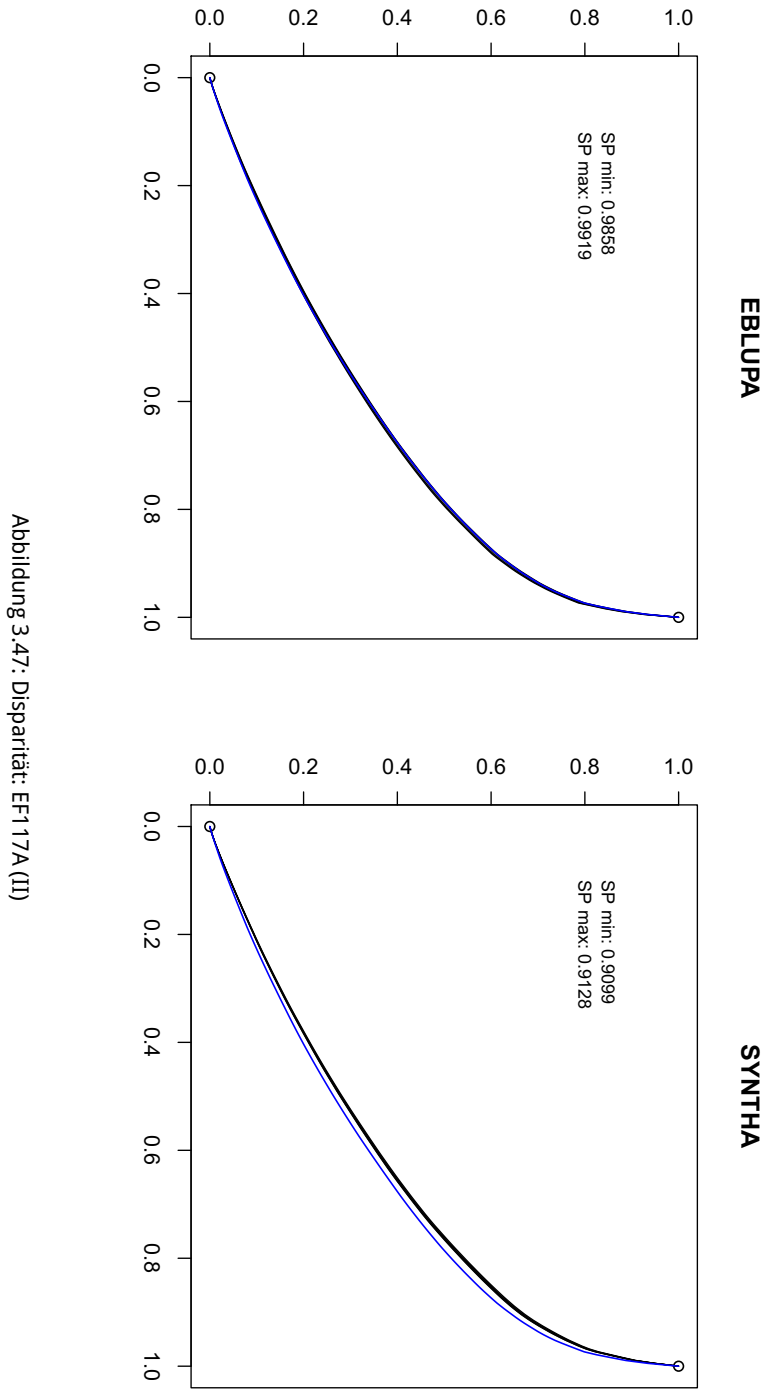


Abbildung 3.47: Disparität: EF117A (II)

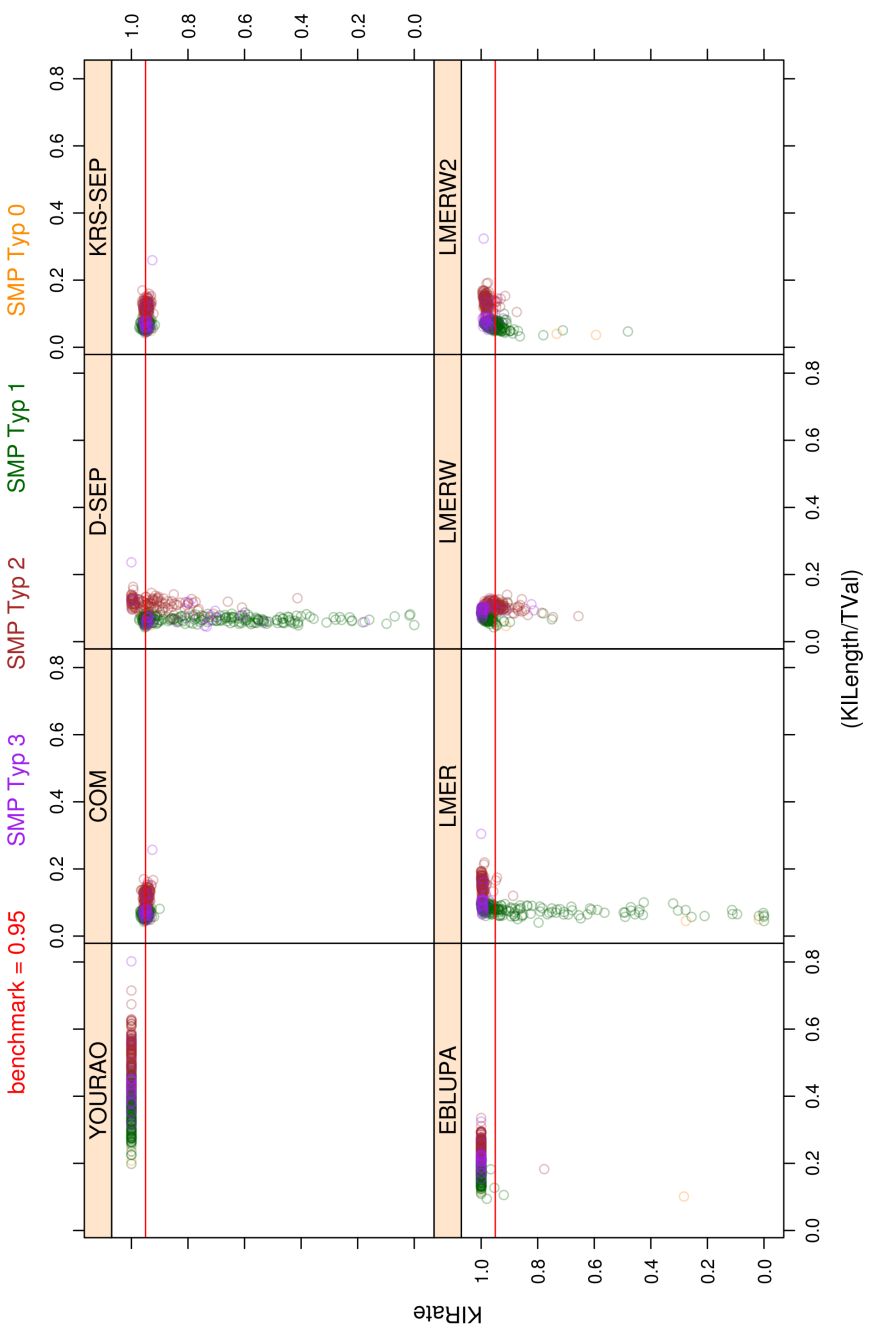


Abbildung 3.48: Konfidenzintervallüberdeckungsrate zur Begutachtung der Varianzschätzung für die Fragestellung ISCEDA in BAW

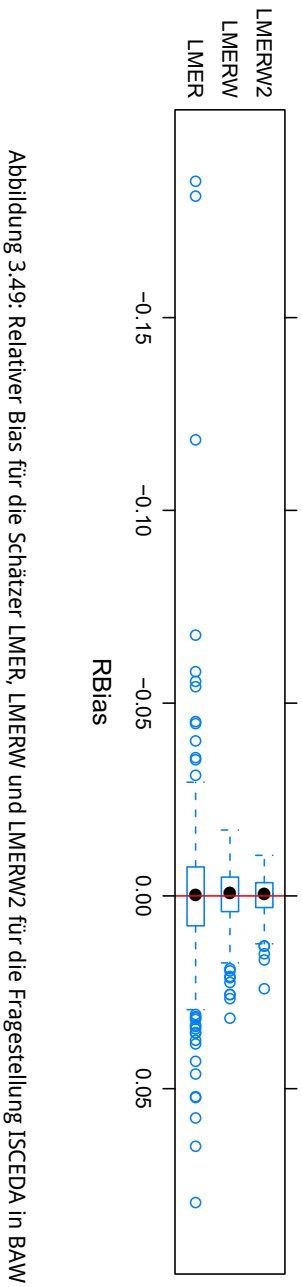


Abbildung 3.49: Relativer Bias für die Schätzer LMER, LMERW und LMERW2 für die Fragestellung ISCEDA in BAW

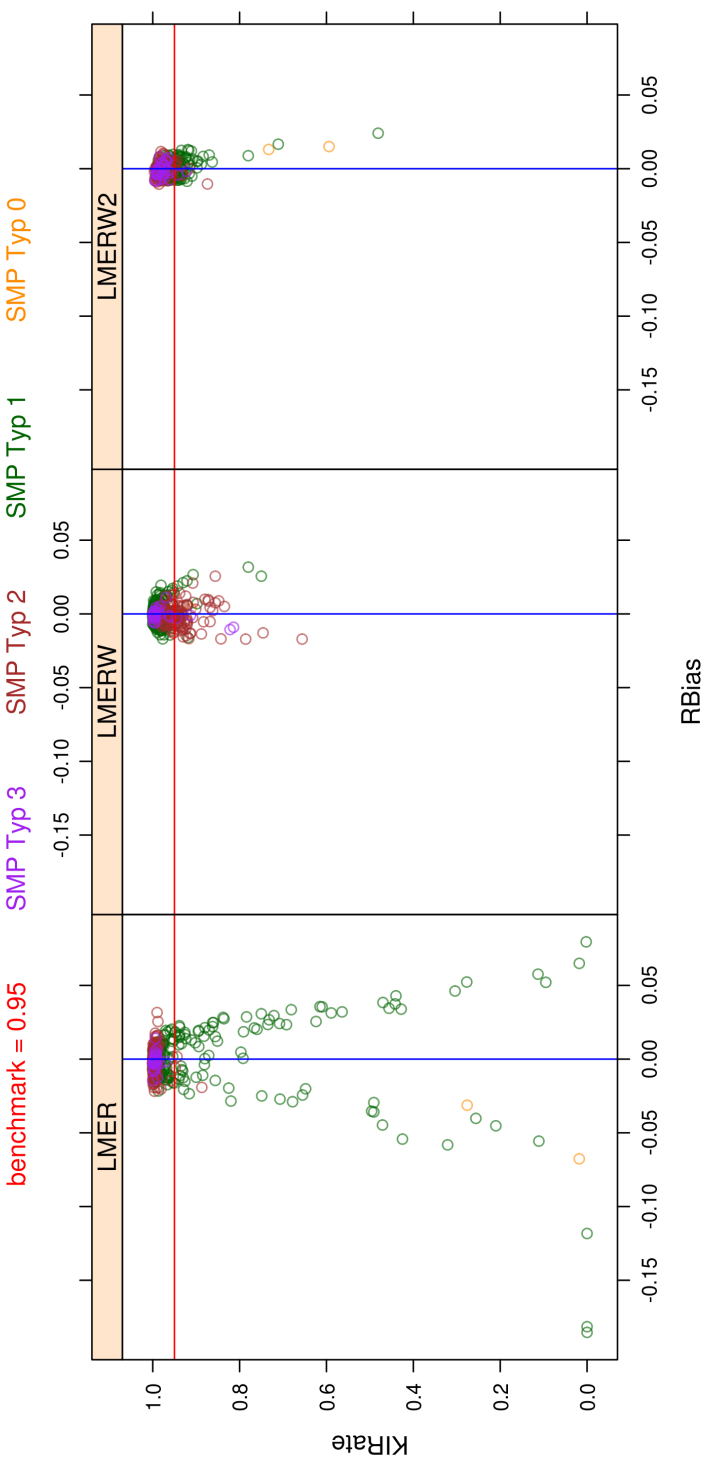


Abbildung 3.50: Konfidenzintervallüberdeckungsrate zur Begutachtung der Varianzschätzung für die Fragestellung ISCEDA in BAW

Literaturverzeichnis

Alfons, A., Burgard, J. P., Filzmoser, P., Hulliger, B., Kolb, J.-P., Kraft, S., Münnich, R., Schoch, T. und Templ, M. (2011): The AMELI Simulation Study, Research Project Report WP6 – D6.1, FP7-SSH-2007-217322 AMELI.
http://ameli.surveystatistics.net

Alfons, A., Filzmoser, P., Hulliger, B., Kolb, J.-P., Kraft, S., Münnich, R. und Templ, M. (2011): Synthetic Data Generation of SILC data, Research Project Report WP6 – D6.2, FP7-SSH-2007-217322 AMELI.
http://ameli.surveystatistics.net

Battese, G. E., Harter, R. M. und Fuller, W. A. (1988): „An error-components model for prediction of county crop areas using survey and satellite data.“, *Journal of the American Statistical Association* 83 (401), S. 28 – 36.

Berg, A. und Bihler, W. (2011): „Das Stichprobendesign der Haushaltsstichprobe des Zensus 2011“, *Wirtschaft und Statistik* 4, S. 317–329.

Bierau, D. (2001): „Neue Methode der Volkszählung – Der Test eines registergestützten Zensus“, *Wirtschaft und Statistik* 5, S. 333 ff.

Burgard, J. P. und Münnich, R. T. (2010): „Modelling over- and undercounts for design-based Monte Carlo studies in small area estimation: An application to the German register-assisted Census“, *Computational Statistics and Data Analysis* .
DOI:10.1016/j.csda.2010.11.002

Chapman, D. (1951): *Some properties of the hypergeometric distribution with applications to zoological censuses*, University of California publications in statistics.

Cochran, W. G. (1977): *Sampling Techniques*, John Wiley and Sons.

Cox, L. (1987): „A constructive procedure for unbiased controlled rounding“, *Journal of American Statistical Association* 82(398), S. 520–524.

Datta, G. S. und Lahiri, P. (2000): „A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems“, *Statistica Sinica* 10, S. 613–627.

Deville, J.-C. und Sarndal, C.-E. (1992): „Calibration estimators in survey sampling“, *Journal of the American Statistical Association* 87(418), S. 376–382.

Devroye, L. (1986): *Non-Uniform Random Variate Generation*, Springer, New York.

Doerr, B., Friedrich, T., Klein, C. und Osbild, R. (2006): Unbiased matrix rounding, *in* „Lecture Notes in Computer Science“, Algorithm theory – SWAT 2006 : 10th Scandinavian Workshop on Algorithm Theory.

Dostal, L. (2012): Alternative Small Area Schätzverfahren am Beispiel des Zensus 2011 in Deutschland, Dissertation, Universität Trier.

Drew, J., Singh, M. und Chaoudhry, G. (1982), „Evaluation of Small-Area Estimation Techniques for the Canadian Labour Force Survey“, *Survey Methodology* 8, S. 17–47.

Fay, R. E. und Herriot, R. A. (1979): „Estimates of income for small places: An application of James-Stein procedures to census data“, *Journal of the American Statistical Association* 74(366), S. 269–277.

Fürnrohr, M., Rimmelspacher, B. und von Roncador, T. (2002): „Zusammenführung von Datenbeständen ohne numerische Identifikatoren – ein Verfahren im Rahmen der Testuntersuchungen zu einem registergestützten Zensus“, *Bayern in Zahlen* 58, S. 308–321.

Fürnrohr, M. und König, M. (1999): „Möglichkeiten einer Haushaltegenerierung im Rahmen der Zusammenführung von Einzeldaten aus Melderegisters und primärstatistisch gewonnenen Wohnungsdaten“, *Bayern in Zahlen* 4, S. 161 ff.

Gabler, S. (1990): „An identity for reflexive g-inverses in the context with BLU estimators“, *Statistical Papers* 31, S. 225–231.

Gabler, S., Ganninger, M. und Münnich, R. (2012): „Optimal allocation of the sample size to strata under box constraints“, *Metrika* 75(2), S. 151–161.

Gelman, A. (2007): „Struggles with survey weighting and regression modeling“, *Statistical Science* 22(2), S. 153–164.

Ghosh, M. und Rao, J. N. K. (1994): „Small Area Estimation: An Appraisal“, *Statistical Science* 9(1), S. 55–76.

Giessing, S. (2008): Europäische Zusammenarbeit im Bereich der Tabellengeheimhaltung, Technical report, Statistisches Bundesamt, Wiesbaden.

Goldstein, H. (2003): *Multilevel Statistical Models*, 3rd edition, Oxford University Press.

González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D. und Santamaría, L. (2007): „Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model“, *Computational Statistics & Data Analysis* 51, S. 2720–2733.

Grohmann, H. (2009): „Von der Volkszählung zum Registerzensus: Paradigmenwechsel in der deutschen amtlichen Statistik“, *AStA Wirtschafts- und Sozialstatistisches Archiv* 3(1), S. 3–23.

Hofmann, H. (2003): „Constructing and reading mosaicplots“, *Computational Statistics & Data Analysis* 43, S. 565–580.

Hogan, H. (2003): „The accuracy and coverage evaluation: Theory and design“, *Survey Methodology* 29(2), S. 129–138.

Jiang, J. und Lahiri, P. (2001): „Empirical best prediction for small area inference with binary data“, *Annals of the Institute of Statistical Mathematics* 53, S. 217–243.

Kleber, B., Maldonado, A., Scheuregger, D. und Ziprik, K. (2009): „Aufbau des Anschriften- und Gebäuderegisters für den Zensus 2011“, *Wirtschaft und Statistik* 7, S. 629–641.

Knobelspies, M. und Münnich, R. (2008): „Variablenselektion bei gebundener Hochrechnung“, *Austrian Journal of Statistics* 37(3&4), S. 335–347.

Kolb, J.-P. (2012): Methoden zur Erzeugung synthetischer Simulationsgesamtheiten, Dissertation, Universität Trier.

Lehtonen, R. und Veijanen, A. (2009): „Chapter 31 – Design-based Methods of Estimation for Domains and Small Areas“, *Handbook of Statistics Sample Surveys: Inference and Analysis* 29(Part B), S. 219 – 249. Editor C.R. Rao.

Linzer, D. A. und Lewis., J. (2007): „polca: Polytomous variable latent class analysis. R package version 1.1.“

Lohr, S. L. (1999): *Sampling: Design and Analysis*, Duxbury Press.

Meng, X.-L., Duan, N., Chen, C.-n. und Alegria, M. (2009): Power-shrinkage: An alternative method for dealing with excessive weights, *in* „Invited Paper Sessions at the Joint Statistical Meeting in Washington".

Münnich, R. (1997): *Gebundene Hochrechnung bei Stichprobenerhebungen mit Hilfe von Splines*, number 43 *in* „Angewandte Statistik und Ökonometrie", Vandenhoeck & Ruprecht, Göttingen.

Münnich, R. (2008): „Varianzschätzung in komplexen Erhebungen", *Austrian Journal of Statistics* 37(3&4), S. 319–334.

Münnich, R., Burgard, J. P. und Vogt, M. (2009): Small area estimation for population counts in the German Census 2011, *in* „Section on Survey Research Methods JSM 2009".

Münnich, R., Gabler, S., Ganninger, M., Burgard, J. P. und Kolb, J.-P. (2010): „Das Stichprobendesign des registergestützten Zensus", *Methoden, Daten, Analysen* 5(1), S. 37–61.

Münnich, R., Sachs, E. und Wagner, M. (2011*a*): „Calibration of estimator-weights via semismooth newton method", *Journal of Global Optimization* 52(3), S. 1–15.

Münnich, R., Sachs, E. und Wagner, M. (2011*b*): „Numerical solution of optimal allocation problems in stratified sampling under box constraints", *AStA Advances in Statistical Analysis* online first, S. 1–16.

Münnich, R., Sachs, E. und Wagner, M. (2012): Calibration benchmarking for small area estimates: an application to the german census 2011, *in* „Fields Institute Symposium on the Analysis of Survey Data and Small Area Estimation in honour of the 75th Birthday of Professor J.N.K. Rao". Invited paper.

Münnich, R. und Burgard, J. P. (2012): „On the influence of sampling design on Small Area Estimates", *Journal of the Indian Society of Agricultural Statistics* 66(1), S. 145–156. Special Issue on Small Area Estimation.

Münnich, R. und Wiegert, R. (2001): „The DACSEIS project", *DACSEIS research paper series* 1.

Neyman, J. (1934): „On the two different apsects of the representative method: The method of stratified sampling and the method of purposive selection", *Journal of the Royal Statistical Society* 97, S. 558–606.

Petersen, C. G. J. (1896): „The yearly immigration of young plaice into the Limfjord from the German sea", *Report of Danish Biological Station* 6, S. 1–48.

Prasad, N. G. N. und Rao, J. N. K. (1990): „The estimation of the mean squared error of small-area estimators", *Journal of the American Statistical Association* 85(409), S. 163–171.

Purcell, N. J. und Kish, L. (1980): „Postcensal estimates for local areas (or domains)", *International Statistical Review* 48, S. 3–18.

Rao, J. N. K. (2003): *Small Area Estimation*, Wiley Series in Survey Methodology, John Wiley and Sons, New York.

Renaud, A. (2001): Methodology of the swiss census 2000 coverage survey, *in* „Proceedings of the Survey Research Methods Section", American Statistical Association.

Renaud, A. (2004): Coverage estimation for the Swiss population Census 2000, estimation methodology and results, Technical report, Swiss Federal Statistical Office.

Renaud, A. (2007): „Estimation of the coverage of the 2000 census of population in Switzerland: Methods and results“, *Survey Methodology* 33, S. 199–210.

Robinson, J. G., Ahmed, B., Gupta, P. D. und Woodrow, K. A. (1993): „Estimation of population coverage in the 1990 united states census based on demographics analysis“, *Journal of the American Statistical Association* 88, S. 1061–1071.

Salazar-González, J. (2002): Controlled rounding and cell perturbation: Statistical disclosure limitation methods for tabular data, Technical report, University of La Laguna, Tenerife, Spain.

Salazar-González, J. und Schoch, M. (2004), A new tool for applying controlled rounding to a statistical table in microsoft excel, *in* „Privacy in Statistical Databases“, Springer.

Särndal, C. E., Swensson, B. und Wretman, J. (1992): *Model Assisted Survey Sampling*, Springer, New York.

Schäfer, J. (2004): „Ergänzende Verfahren für einen künftigen registergestützten Zensus“, *Statistische Analysen und Studien NRW* 17, S. 20–27.

Statistische Ämter des Bundes und der Länder (2004): „Ergebnisse des Zensustests“, *Wirtschaft und Statistik* 8/2004, S. 813 ff.

Statistische Ämter des Bundes und der Länder (2011): „Das registergestützte Verfahren beim Zensus 2011“.

Statistisches Bundesamt (2006): *Qualitätsstandards in der amtlichen Statistik*, Statistische Ämter des Bundes und der Länder.

Statistisches Bundesamt (2010): „Haushaltebefragung beim Zensus 2011 – Erläuterungen zum Stichprobenverfahren“.

Stenger, H. und Gabler, S. (2005): „Combining random sampling and census strategies – justification of inclusion probabilities equal to 1“, *Metrika* 61, S. 137–156.

Tillé, Y. (2006): *Sampling Algorithms*, Springer Verlag.

Torabi, M. und Rao, J. N. K. (2010): „Mean squared error estimators of small area means using survey weights“, *Canadian Journal of Statistics* 38(4), S. 598–608.

Tschuprov, A. A. (1923): „On the mathematical expectation of the moments of frequency distributions in the case of correlated observations“, *Metron* 2, S. 461–493 und 646–683.

Wagner, G. (2010): „Zensus 2010/11 – Eine längst überfällige Erhebung“, *Wochenbericht* 77(4), S. 11–14.

Wiegert, R. und Münnich, R. (2004): „German register data for regression estimation in survey sampling – a study on the German microcensus respecting for data protection“, *Jahrücher für Nationalökonomie und Statistik* 224 (1/2), S. 247 – 259.

You, Y. und Rao, J. N. K. (2002): „A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights“, *The Canadian Journal of Statistics* 30(3), S. 431–439.

Zaslavsky, A. M. (1989): Multiple-system methods for census coverage evaluation, *in* „Proceedings of the Survey Research Methods Section“, American StatisticalAssociation.

Zhang, L.-C. und Chambers, R. L. (2004): „Small area estimates for cross-classifications“, *Journal of the Royal Statistical Society, Series B* 66 (Part 2), S. 479–496.