Alternative Small–Area–Schätzverfahren am Beispiel des Zensus 2011 in Deutschland

Dissertation

Lucie Dostál

Alternative Small–Area–Schätzverfahren am Beispiel des Zensus 2011 in Deutschland

Die vorliegende Arbeit wurde vom Fachbereich IV, Wirtschafts– und Sozialwissenschaften, Mathematik, Informatik und Wirtschaftsinformatik, der Universität Trier im Jahr 2011 als Dissertation zur Erlangung des Doktors der Wirtschafts– und Sozialwissenschaften (Dr. rer. pol.) angenommen.

Erstgutachter:	Prof. Dr. Ralf Münnich
Zweitgutachter:	PD Dr. Siegfried Gabler
Tag der Disputation:	8. Juli 2011
Vorsitzender:	Prof. Dr. Christian Bauer

Lucie Dostál

Alternative Small–Area–Schätzverfahren am Beispiel des Zensus 2011 in Deutschland

Universität Trier

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

Symbolverzeichnis

Kapitel 2: Theoretische Grundlagen

U	Population von Personen vom Umfang N_P
U_R	Population der Liste R vom Umfang τ_R
U_Z	Population der Liste Z vom Umfang τ_Z
$\#$	Umfang einer Menge
$\hat{}$	Schätzwert

Kapitel 3: Dual–System–Modelle in Deutschland

*	ursprüngliche Anzahl erhöht um eins
a	eine Anschrift
i	eine Person
k	Klasse, d.h. Subpopulation, auf die die Dual–System–Modelle angewendet werden, $k \in \{1, \ldots, K\}$
d	Domain, d.h. Subpopulation, deren Umfang wir schätzen wollen, $d \in \{1, \ldots, D\}$
h	Anschriftenschicht, $h \in \{1, \ldots, H\}$
KAL	Karteileichen – Einträge im Register, deren Personen nicht

	mehr in der Gemeinde wohnhaft sind
FEB	Fehlbestände – Personen, die in der Gemeinde wohnen, aber nicht in der Gemeinde gemeldet sind
τ_R	Anzahl der registrierten Personen
τ_Z	Anzahl der tatsächlich vorhandenen Personen
τ_K	Anzahl der Karteileichen
τ_F	Anzahl der Fehlbestände
τ_{00}	Anzahl der sowohl im Register eingetragenen als auch in der Gesamtheit wohnhaften Personen
p_R	Wahrscheinlichkeit im Register zu sein
p_F	Wahrscheinlichkeit ein Fehlbestand zu sein
p_K	Wahrscheinlichkeit eine Karteileiche zu sein
p_{00}	Wahrscheinlichkeit sowohl im Register erfasst zu werden als auch in der Gesamtheit zu finden zu sein
S	Stichprobe der Anschriften
S_h	Stichprobe der Anschriften innerhalb der Anschriftenschicht h
S_P	Stichprobe der Personen in Gesamtpopulation U
$\bar{S}_P = U \setminus S_P$	Anzahl der nicht in der Stichprobe S_P enthaltenen Personen der Gesamtpopulation U
N	Anzahl aller Anschriften
N_h	Anzahl aller Anschriften innerhalb der Anschriftenschicht h
N_P	Anzahl der Personen in Gesamtpopulation U
$N_{P,h}$	Anzahl aller Personen innerhalb der Anschriftenschicht h
$N_{P,k}$	Anzahl aller Personen innerhalb der Klasse k

n	Umfang der Stichprobe S
n_h	Umfang der Stichprobe S_h
n_P	Umfang der Stichprobe S_P
$\bar{n}_P$=$N_P - n_P$	Anzahl der Personen nicht in der Stichprobe S_P
π_a	Inklusionswahrscheinlichkeit der Anschrift a
$\pi_{a,h}$	Inklusionswahrscheinlichkeit der Anschrift a in der Schicht h
w_a	Designgewicht für Anschrift a
$w_{a,h}$	Designgewicht für Anschrift a in der Schicht h
$\tau_{Z,a}$	Anzahl der tatsächlich vorhandenen Personen an der Anschrift a
$\tau_{Z,d}$	Anzahl der tatsächlich vorhandenen Personen der Domain d
$\tau_{Z,k}$	Anzahl der tatsächlich vorhandenen Personen der Klasse k
$\tau_{Z,a,h}$	Anzahl der tatsächlich vorhandenen Personen an der Anschrift a in der Schicht h
$\tau_{Z,a,k}$	Anzahl der tatsächlich vorhandenen Personen an der Anschrift a in der Klasse k
$\tau_{Z,k,h}$	Anzahl der tatsächlich vorhandenen Personen der Klasse k in der Schicht h
$\tau_{Z,d,a}$	Anzahl der tatsächlich vorhandenen Personen der Domain d an der Anschrift a
$\tau_{Z,d,h}$	Anzahl der tatsächlich vorhandenen Personen der Domain d in der Schicht h
τ_{Z,S_P}	Anzahl der tatsächlich vorhandenen Personen in der Stichprobe S_P
$\tau_{Z,S_P,k}$	Anzahl der tatsächlich vorhandenen Personen der Klasse k in der Stichprobe S_P

$\tau_{Z,S_P,d}$	Anzahl der tatsächlich vorhandenen Personen der Domain d in der Stichprobe S_P
$\tau_{R,a}$	Anzahl der registrierten Personen an der Anschrift a
$\tau_{R,h}$	Anzahl der registrierten Personen in der Schicht h
$\tau_{R,k}$	Anzahl der registrierten Personen in der Klasse k
$\tau_{R,d}$	Anzahl der registrierten Personen in der Domain d
$\tau_{R,a,h}$	Anzahl der registrierten Personen an der Anschrift a in der Schicht h
$\tau_{R,a,k}$	Anzahl der registrierten Personen an der Anschrift a in der Klasse k
$\tau_{R,k,h}$	Anzahl der registrierten Personen der Klasse k in der Schicht h
$\tau_{R,d,a}$	Anzahl der registrierten Personen der Domain d an der Anschrift a
$\tau_{R,d,h}$	Anzahl der registrierten Personen der Domain d in der Schicht h
τ_{R,S_P}	Anzahl der registrierten Personen in der Stichprobe S_P
$\tau_{R,S_P,k}$	Anzahl der Einträge im Register der Klasse k in der Stichprobe S_P
$\tau_{R,S_P,d}$	Anzahl der Einträge im Register der Domain d in der Stichprobe S_P
τ_{F,S_P}	Anzahl der Fehlbestände in der Stichprobe S_P
$\tau_{F,S_P,k}$	Anzahl der Fehlbestände einer Klasse k in der Stichprobe S_P
τ_{K,S_P}	Anzahl der Karteileichen in der Stichprobe S_P
τ_{00,S_P}	Anzahl der in beiden Listen eingetragenen Personen der Stichprobe S_P

DSE	Dual–System–Estimator
$\hat{\tau}_Z^{\text{D-DSE}} = \tau_R \, \frac{\hat{\tau}_Z}{\hat{\tau}_R}$	DSE in Deutschland
$\hat{\tau}_Z^{\text{CHAP}} = (\tau_R + 1) \, \frac{\hat{\tau}_Z + 1}{\hat{\tau}_R + 1} - 1$	Chapman–Schätzer in Deutschland

$\hat{\tau}_{Z,h}^{\text{D-DSE}}$	D–DSE für $\tau_{Z,h}$ in der Anschriftenschicht h
$\hat{\tau}_{Z,k}^{\text{D-DSE}}$	D–DSE für $\tau_{Z,k}$ in der Klasse k
$\hat{\tau}_{Z,k,h}^{\text{D-DSE}}$	D–DSE für $\tau_{Z,k,h}$ in der Klasse k der Anschriftenschicht h
$\hat{\tau}_{Z,h}^{\text{CHAP}}$	Chapman–Schätzer für $\tau_{Z,h}$ in der Anschriftenschicht h
$\hat{\tau}_{Z,k}^{\text{CHAP}}$	Chapman–Schätzer für $\tau_{Z,k}$ in der Klasse k
$\hat{\tau}_{Z,k,h}^{\text{CHAP}}$	Chapman–Schätzer für $\tau_{Z,k,h}$ in der Klasse k der Anschriftenschicht h

Kapitel 4: Small–Area–Schätzung für Dual–System–Modelle

$\pi_{k,d}$	Wahrscheinlichkeit in der Kategorie (k, d) zu sein
$\boldsymbol{\tau}_R$	Vektor der Registerdaten
$\mathbf{X}_1$	Teil der Designmatrix $\mathbf{X}$ für die Haupteffekte $\boldsymbol{\beta}_1$
$\mathbf{X}_2$	Teil der Designmatrix $\mathbf{X}$ für die Interaktionseffekte $\boldsymbol{\beta}_2$
$\boldsymbol{\beta}$	Vektor der unbekannten Parameter in log–linearem–Modell
$\boldsymbol{\beta}_1, \boldsymbol{\beta}_3$	Koeffizienten für die Haupteffekte
$\boldsymbol{\beta}_2, \boldsymbol{\beta}_4$	Koeffizienten für die Interaktionseffekte
$\hat{\boldsymbol{\tau}}_{\text{PW}}$	geschätzter Vektor der Pseudo–Werte
IPF	iterative proportional fitting Algorithmus
SPREE	structure preserving estimator, strukturerhaltender-

	Schätzer
GSPREE	generalized structure preserving estimator, verallgemeinerter–strukturerhaltender–Schätzer
$\hat{\tau}_{Z,d}^{\text{SPREE}} = \frac{\tau_{R,d}}{\tau_{R,k}}\ \hat{\tau}_{Z,k}^{\text{X}}$	SPREE für die Domain d berechnet mittels Schätzer $\hat{\tau}_{Z,k}^{\text{X}}$ für $\tau_{Z,d}$

Kapitel 5: Small–Area–Schätzung für Alternative Modelle

$\boldsymbol{x}_a$	Vektor der Hilfsvariablen an Anschrift a
$\boldsymbol{\tau}_Z$	Vektor der tatsächlich vorhandenen Personen
$\boldsymbol{\tau}_{Z,d}$	Vektor der tatsächlich vorhandenen Personen der Domain d
$\boldsymbol{t}_x = \sum_{a=1}^{N} \boldsymbol{x}_a$	Vektor der Totalwerte
$\mathbf{X}$	Matrix mit n Werten der v Hilfsvariablen
$\mathbf{X}_d$	Matrix mit n Werten der v Hilfsvariablen der Domain d
$\mathbf{X}_{P,d}$	Matrix mit n_P Werten der v Hilfsvariablen der Domain d
$\mathbf{B}$	Vektor der Regressionskoeffizienten
$\hat{\mathbf{B}} = \frac{\hat{\tau}_Z}{\hat{\tau}_R}$	Vektor der geschätzten Regressionskoeffizienten
$\hat{\mathbf{B}}_d = \frac{\hat{\tau}_{Z,d}}{\hat{\tau}_R}$	Vektor der geschätzten Regressionskoeffizienten der Domain d
$\hat{\mathbf{B}}_{d,\text{sep}} = \frac{\hat{\tau}_{Z,d}}{\hat{\tau}_{R,d}}$	Vektor der geschätzten Regressionskoeffizienten für die separate Schätzung der Domain d
$\mathbf{W}$	$n \times n$ Diagonalmatrix der Designgewichte
$\boldsymbol{\Sigma}$	$n \times n$ Diagonalmatrix mit Varianzen
$\boldsymbol{\varepsilon}$	Fehlerterm
y_d	Summe der tatsächlich lebenden Personen in der Domain d

x_d	Summe der registrierten Personen in der Domain d
$\hat{\tau}_Z^{\text{GREG0}} = \frac{\hat{\tau}_Z}{\hat{\tau}_R}\, \tau_R$	Verhältnisschätzer für τ_Z
$\hat{\tau}_{Z,d}^{\text{GREG1}} = \frac{\hat{\tau}_{Z,d}}{\hat{\tau}_R}\, \tau_R$	Verallgemeinerter–Regressionsschätzer 1 in der Domain d für $\tau_{Z,d}$
$\hat{\tau}_{Z,d}^{\text{GREG2}} = \frac{\hat{\tau}_{Z,d}}{\hat{\tau}_{R,d}}\, \tau_{R,d}$	Verallgemeinerter–Regressionsschätzer 2 in der Domain d für $\tau_{Z,d}$
$\hat{\tau}_{Z,d}^{\text{SYN}} = \frac{\hat{\tau}_Z}{\hat{\tau}_R}\, \tau_{R,d}$	Verhältnis–synthetischer Schätzer in der Domain d für $\tau_{Z,d}$
$\hat{\tau}_{Z,d}^{\text{EBLUP}}$	EBLUP, estimated best linear unbiased predictor für $\tau_{Z,d}$
BLUE	best linear unbiased estimator
BLUP	best linear unbiased predictor
REML	restricted maximum likelihood estimator

Kapitel 6: Varianz, Schätzung der Varianz bzw. des MSE

B	Anzahl der Bootstrap–Stichproben
b	eine Bootstrap–Stichprobe
α	Signifikanzniveau
$1-\alpha$	Konfidenzniveau
$\hat{\tau}$	Punktschätzer für τ
$\pi_{aa'}$	Inklusionswahrscheinlichkeit zweiter Ordnung für die Anschriften a und a', $a \neq a'$
$\hat{\tau}_Z^{\text{HT}}$	Horvitz–Thompson–Schätzer für τ_Z der tatsächlich vorhandenen Personen
$\hat{\tau}_{Z,0}$	Taylor–Linearizierung erster Ordnung von $\hat{\tau}_Z$
$\hat{\tau}_{Z,0}^{\text{GREG}}$	Taylor–Linearizierung $\hat{\tau}_Z^{\text{GREG}}$

$E_a = \tau_{Z,a} - \boldsymbol{x}_a^T \, \mathbf{B}$	Residuum der Anschrift a mit $\mathbf{B}$
$e_a = \tau_{Z,a} - \boldsymbol{x}_a^T \, \hat{\mathbf{B}}$	Residuum der Anschrift a mit $\hat{\mathbf{B}}$
$\hat{\tau}^{\text{D-DSE}}_{Z,k,-a}$	D–DSE einer Klasse k berechnet nach Auslassung der Anschrift a
$\hat{\tau}^{\text{D-DSE}\,*}_{Z,k,-a}$	$\hat{\tau}^{\text{D-DSE}}_{Z,k,-a}$ um Bias korrigiert
$\hat{\tau}^{\text{D-DSE}}_{Z,k,a,\text{pseudo}}$	Pseudo–Werte für $\hat{\tau}^{\text{D-DSE}}_{Z,k,a}$
$\hat{\tau}^{\text{D-DSE}\,*}_{Z,k,a,\text{pseudo}}$	$\hat{\tau}^{\text{D-DSE}}_{Z,k,a,\text{pseudo}}$ um Bias korrigiert
$\hat{\bar{\tau}}^{\text{D-DSE}}_{Z,k,\text{jack1}}$	Jackknife–Schätzer für D–DSE der Klasse k bei uneingeschränkter Zufallsauswahl mit Zurücklegen
$\hat{\bar{\tau}}^{\text{D-DSE}\,*}_{Z,k,\text{jack*}}$	Jackknife–Schätzer für D–DSE der Klasse k bei uneingeschränkter Zufallsauswahl ohne Zurücklegen
$\hat{\tau}^{\text{CHAP}}_{Z,k,-a}$	Chapman–Schätzer der Klasse k berechnet nach Auslassung der Anschrift a
$\hat{\tau}^{\text{CHAP}\,*}_{Z,k,-a}$	$\hat{\tau}^{\text{CHAP}}_{Z,k,-a}$ um Bias korrigiert
$\hat{\tau}^{\text{CHAP}}_{Z,k,a,\text{pseudo}}$	Pseudo–Werte für $\hat{\tau}^{\text{CHAP}}_{Z,k,a}$
$\hat{\tau}^{\text{CHAP}\,*}_{Z,k,a,\text{pseudo}}$	$\hat{\tau}^{\text{CHAP}}_{Z,k,a,\text{pseudo}}$ um Bias korrigiert
$\hat{\bar{\tau}}^{\text{CHAP}}_{Z,k,\text{jack1}}$	Jackknife–Schätzer für den Chapman–Schätzer der Klasse k bei uneingeschränkter Zufallsauswahl mit Zurücklegen
$\hat{\bar{\tau}}^{\text{CHAP}\,*}_{Z,k,\text{jack*}}$	Jackknife–Schätzer für den Chapman–Schätzer der Klasse k bei uneingeschränkter Zufallsauswahl ohne Zurücklegen
$\hat{\boldsymbol{\tau}}^{\text{GSPREE,b}}_{Z,d}$	GSPREE der Domain d der Bootstrap–Stichprobe b

Kapitel 7: Aufbau der Simulationen und Ergebnisse

R	Anzahl der Simulationen
r	eine Simulation
RRMSE = $\frac{\sqrt{\text{MSE}}}{\tau_Z}$	relative root mean square error
MSE = $\frac{1}{R}\sum_{r=1}^{R}(\hat{\tau}^{\text{X}}_{Z,r} - \tau_Z)^2$	mean square error, mittlerer quadratischer Fehler des Schätzers $\hat{\tau}^{\text{X}}_Z$ über R Simulationen
RB = $\frac{\frac{1}{R}\sum_{r=1}^{R}\hat{\tau}^{\text{X}}_{Z,r} - \tau_Z}{\tau_Z}$	relative bias, relativer Bias des Schätzers $\hat{\tau}^{\text{X}}_Z$ über R Simulationen
ARB =|RB|	absolute relative bias, absoluter relativer Bias
$\tau_{R,<g>}$	Anzahl der registrierten Personen in der Gemeinde g
$\tau_{R,\ll k\gg}$	Anzahl der registrierten Personen der Gemeinden des Kreises k
Design 01	Uneingeschränkte Zufallsauswahl der Anschriften in Deutschland
Design 05a	Geschichtete Zufallsauswahl der Anschriften in Deutschland
D–DSE	DSE in Deutschland
D–DSE/SPREE	D–DSE mit anschließender Durchführung des SPREEs
CHAP	Chapman–Schätzer
CHAP/SPREE	Chapman–Schätzer mit anschließender Durchführung des SPREEs
CHAP/GSPREE	Chapman–Schätzer im GSPREE–Modell angewendet

CHAP/SPREE/GSPREE	Chapman–Schätzer mit anschließender Durchführung des SPREEs, dann im GSPREE–Modell angewendet
GREG1	Verallgemeinerter–Regressionsschätzer 1
GREG2	Verallgemeinerter–Regressionsschätzer 2
SYN	Verhältnis–synthetischer Schätzer
EBLUP	estimated best linear unbiased predictor

gewidmet
meinem Mann
sowie
unserem noch ungeborenen Kind

Vorwort

Ich möchte von ganzen Herzen meinen Gutachtern, Herrn Prof. Dr. Ralf Münnich und PD Dr. Siegfried Gabler danken, die mit ihren kritischen Anmerkungen diese Dissertation konstruktiv mitgestaltet haben. Meinen Kollegen aus der Universität Trier und GESIS Mannheim, Dr. Matthias Ganninger, Dr. Martin Vogt, Pablo Burgard, Jan-Philipp Kolb, Tobias Enderle, Stefan Zins und Jan Seger danke ich für die große Unterstützung im Programmieren und die stets freundschaftliche Zusammenarbeit – unsere gemeinsame Zeit wird mir immer in schöner Erinnerung bleiben.

Des Weiteren danke ich bei meinem aktuellen Arbeitgeber, dem Deutschen Krebsforschungszentrum, Prof. Dr. Rudolf Kaaks, Dr. Anika Hüsing und Dr. Lars Beckmann für die sehr großzügige Arbeitszeiteinteilung bei der Fertigstellung der Dissertation.

Nicht zuletzt möchte ich mich herzlich bei meiner Familie bedanken, vor allem bei meinem Vater und meiner Mutter, die manche Zeit auf mich verzichten mussten. Mein größter persönlicher Dank gilt meinem Mann, der mir viele Wochenenden und Nächte am Computer nachgesehen hat. Seine uneingeschränkte Unterstützung, das liebevolle Verständnis und viele Ermunterungen, diese Dissertation weiterzuführen, haben ganz wesentlich zum Erfolg dieser Dissertation beigetragen.

1 Einleitung

Die letzte Volkszählung aller Haushalte und Personen fand im früheren Bundesgebiet 1987, in der ehemaligen DDR 1981 statt. In den Jahren 2010, 2011 sollten nach der Europäischen Union (EU) alle Mitgliedstaaten der EU einen Zensus durchführen. Unter anderem ist das Ziel des Zensus zu ermitteln, wie viele Personen in Deutschland tatsächlich leben, wie sie wohnen und arbeiten (siehe `www.zensus2011.de`).

Auch Deutschland wird sich am Zensus 2011 beteiligen. Die aktuellen Bevölkerungszahlen basieren auf Fortschreibungen der letzten Volkszählungen im Jahre 1987 und 1981. Wegen Ungenauigkeiten in der Fortschreibung, aber auch wegen historischer Ereignisse wie der Wiedervereinigung, ist es notwendig, genaue Bevölkerungszahlen für Gesamtdeutschland zu ermitteln. Die Bevölkerungszahlen eines Zensus bilden die Grundlage für die jährliche Fortschreibung des Statistischen Bundesamtes und der zuständigen Landesämter für Statistik. Die Bevölkerungszahlen sind auch wichtig für politische und gesellschaftliche Bereiche, vor allem für den Finanzausgleich zwischen den Bundesländern, für die Aufteilung der Bundestagswahlkreise, für die Planung der Schulen und Krankenhäuser etc. (siehe `www.zensus2011.de`).

Die Voruntersuchungen beim Zensustest 2001 (Statistisches Bundesamt 2004) haben gezeigt, dass eine postalische Gebäude– und Wohnungszählung zuverlässige statistische Daten über die Bevölkerung liefert. Auf Grund dieser Ergebnisse hat am 29. August 2006 die Bundesregierung be-

schlossen, dass in Deutschland 2011 ein *registergestützter Zensus* durchgeführt wird. Im Vergleich zum traditionellen Zensus, bei dem alle Einwohner befragt werden, werden bei dem registergestützten Zensus die Informationen vor allem der Melderegister der Kommunen genutzt. Es gibt in Deutschland kein zentrales Melderegister. Verantwortlich für die Führung des Melderegisters sind die Kommunen, die zum Zensus–Stichtag (im Frühjahr des Jahres 2011) ihre Melderegisterauszüge an das Statistische Landesamt liefern. Mehr Informationen findet man auf der Seite `www.zensus2011.de`.

Da die Melderegisterauszüge Einträge enthalten können, deren Personen nicht mehr in der Gemeinde wohnhaft sind (*Overcoverage* bzw. *Karteileichen*), anderseits in der Gemeinde Personen wohnen können, die nicht in der Gemeinde gemeldet sind (*Undercoverage* bzw. *Fehlbestände*), werden weitere Verfahren zur Sicherung der Datenqualität in Form von Stichprobenerhebungen im Anschluss an den Zensus durchgeführt. Ein geringer Prozentsatz der Bevölkerung wird durch die Fragebögen befragt, um die Daten im Melderegister zu kontrollieren und einige nicht im Melderegister enthaltene Merkmale zu erhalten.

Der Schwerpunkt dieser Dissertation liegt in der Anwendung des *capture–recapture–Modells* am Zensus 2011 in Deutschland. Dieses Modell wird in verschiedenen Ländern (z. B. in der Schweiz, in den USA, in Großbritannien oder in Israel) angewendet, um zu ermitteln, wie gut der Zensus eine Grundgesamtheit abdeckt (*Coverage–Error*). Ziel dieser Dissertation ist zu untersuchen, ob das capture–recapture–Modell am kommenden Zensus auch angewendet werden kann. Dabei ist wichtig, dass in Deutschland in erster Linie nicht der Coverage–Error, sondern die tatsächliche Anzahl der in Deutschland vorhandenen Personen interessiert.

Die capture–recapture–Modelle blicken auf eine lange Entwicklung zurück. Zur Zeit werden sie in verschiedenen Bereichen benutzt, aber erst-

mals angewendet wurden sie in der Ökologie, wo das capture–recapture–Modell *Petersen–Modell* heißt, benannt nach einer Studie von Petersen (1896). In dieser Studie wendet Petersen das capture–recapture–Modell an, um die unbekannte Größe einer Fischpopulation in einem See zu schätzen. Mit Hilfe der ersten Stichprobe (engl. *capture sample*) werden die gefangenen Fische markiert. Dann werden diese Fische wieder zurück zur übrigen Fischpopulation gegeben. Anhand der zweiten Stichprobe (engl. *recapture sample*) ermittelt man dann, wie viele Fische ausschließlich im zweiten Fang oder in beiden Fängen herausgefischt wurden. Auf Basis dieser beiden Werte ist es möglich zu schätzen, wie groß die Fischpopulation in einem See ist.

Die Übertragung auf eine Menschenpopulation erfolgt in einer Studie von Sekar und Deming (1949). In Bezug auf eine Menschenpopulation heißt das capture–recapture–Modell *Dual–System–Modell* oder *Two–Sample–Modell*. Natürlich „fischt" man in diesem Modell keine Menschen, sondern benutzt Listen, die aus unterschiedlichen Quellen stammen. Fienberg (1992) gibt einen umfangreichen Literaturüberblick, der sich mit den capture–recapture–Modellen in Anwendung beim Zensus befasst.

Im Rahmen dieser Dissertation wird im zweiten Kapitel sowohl das capture–recapture–Modell am Beispiel der Fischpopulation F, als auch die Anwendung dieses Modell auf die Bevölkerung U vorgestellt. Es wird ein *Allgemeines Modell* M_g bzw. *Petersen–Modell* M_t definiert. Als Grundlage für Theorie in diesen Kapiteln wird vorzugsweise Wolter (1986) benutzt. In Kapitel 2.3 wird der Dual–System–Estimator (DSE) vorgestellt, der zur Ermittlung des Coverage–Errors benutzt wird.

Im Kapitel 3 wird die Theorie des Dual–System–Modells auf den registergestützten Zensus in Deutschland angewendet. Um die Voraussetzungen des Modells zu erfüllen, wird die Population in geeignete Subpopulationen (*Klassen*) unterteilt und Schätzer in den Subpopulationen berechnet. In der

Regel sind wir aber an einer Schätzung von Subpopulationen (*Domains*), die sich von den Klassen unterscheiden, interessiert. Um eine Schätzung für eine Domain zu gewinnen, werden im Kapitel 4 strukturerhaltende– und verallgemeinerte–strukturerhaltende–Schätzer vorgestellt. Im Kapitel 5 werden alternative Modelle für die Schätzung der Domains herangezogen. Dadurch wird es möglich, die Modelle im Kapitel 3 und im Kapitel 4 mit den alternativen Modellen im Kapitel 5 zu vergleichen.

Die Güte aller in dieser Dissertation betrachteten Schätzer wird unter anderem durch Varianz oder mittlere quadratische Abweichung (MSE) beurteilt. Kapitel 6 stellt unterschiedliche Methoden zur Varianzschätzung und Schätzung von MSE vor.

Der Rest der Dissertation behandelt die Anwendung bisheriger Theorien auf die zur Verfügung stehende Simulations–Population für das Bundesland Saarland, die dem DACSEIS Projekt entnommen wurden. Es werden die uneingeschränkte Zufallsauswahl der Anschriften ohne Zurücklegen und die geschichtete Zufallsauswahl der Anschriften genauer untersucht.

Eine Zusammenfassung dieser Dissertation und ein Ausblick wird im Kapitel 8 vorgestellt. Diese konzentriert sich auf den Vergleich der zwei betrachteten Zufallsauswahlen sowie auf den Vergleich der Subpopulationen.

2 Theoretische Grundlagen

2.1 Capture–Recapture–Modell

In vielen Stichprobenuntersuchung ist der Umfang der Population bekannt. Beim capture–recapture–Modell ist dagegen der Umfang der Population unbekannt und soll geschätzt werden.

Die capture–recapture–Modelle werden oft am Beispiel der Fischpopulation in einem See erklärt, um den unbekannten Umfang N_F der Fischpopulation zu schätzen. Man zieht zuerst eine Stichprobe S_A, markiert die Fische (in unserem Beispiel mit einer 0) und gibt sie zur übrigen Fischpopulation zurück. Nach einem bestimmten Zeitraum zieht man unter gleichen Umständen erneut eine Stichprobe S_B (s. Abbildung 2.1).

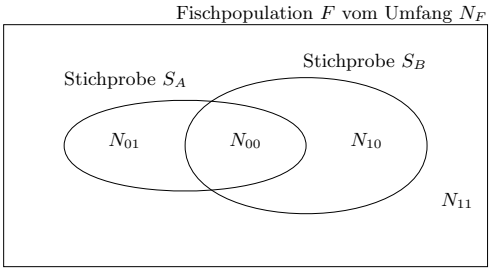


ABBILDUNG 2.1: *Stichproben in Fischpopulation F.*

Die ganze Situation können wir mit der 2×2 Kontingenztabelle (vgl. Agresti 2002, S. 36) darstellen (s. Tabelle 2.1), wobei $\bar{S}_A = F \setminus S_A$, $\bar{S}_B = F \setminus S_B$,

	S_B	$\bar{S}_B$	
S_A	N_{00}	N_{01}	N_A
$\bar{S}_A$	N_{10}	N_{11}	
	N_B		N_F

TABELLE 2.1: *Die Anzahl der Fische in den vier möglichen Kategorien.*

N_{00} die Anzahl der nach der zweiten Stichprobe zweimal markierten Fische ist, N_{10} die Anzahl der nach der zweiten Stichprobe nur einmal (durch zweite Stichprobe S_B) markierten Fische ist. N_A und N_B sind die bekannten Umfänge der Stichprobe S_A bzw. S_B.

Die Anzahl N_{01} kann als Differenz $N_{01} = N_A - N_{00}$ ermittelt werden. Die Anzahl N_{11} ist unbekannt, deswegen ist auch die Gesamtzahl N_F unbekannt.

Unter bestimmten Voraussetzungen können wir aus dem Verhältnis von markierten und nicht markierten Fischen die Größe N_F der Gesamtfischpopulation F bestimmen.

2.2 Petersen–Modell

Im Folgenden wird das capture–recapture–Modell an die Population U übertragen, um die unbekannte Größe N_P der Population zu ermitteln.

Bei einer traditionellen Volkszählung wird versucht, alle Personen der Population U zu ermitteln. Dabei kann es allerdings vorkommen, dass man-

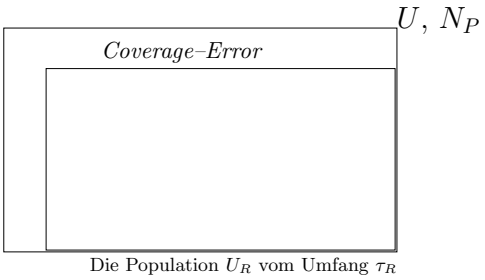


ABBILDUNG 2.2: *Ein traditioneller Zensus in der Population U.*

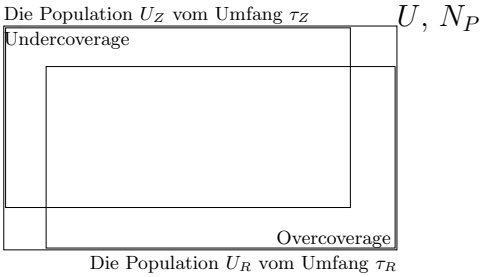


ABBILDUNG 2.3: *Zwei Zensen in der Population U.*

che Personen nicht ermittelt und manche mehr als einmal ermittelt werden. Insgesamt wird eine Population U_R mit dem Umfang τ_R ermittelt (s. Abbildung 2.2). Die Differenz zwischen N_P und τ_R gibt den *Coverage–Error* an. Es sei angemerkt, dass der Umfang τ_R auch größer als N_P sein kann. In der Regel ist τ_R kleiner als N_P (vgl. Wolter 1986).

Um N_P und den Coverage–Error ermitteln zu können, wird neben der Information über die Population U_R weitere Informationen benötigt. Nehmen wir an, eine zweite Volkszählung in der Gesamtheit U wird durchgeführt, bei der wieder manche Personen von U nicht gezählt werden (s. Abbildung 2.3). Insgesamt zählen wir durch die zweite Volkszählung in der

Gesamtheit U eine Stichprobe, bezeichnet als Population U_Z vom Umfang τ_Z . Die kleine Rechtecke rechts oben und links unten werden später (s. S. 15) erläutert.

Am Beispiel der Fischpopulation in einem See entspricht die erste Stichprobe der Fische der ersten Volkszählung in der Gesamtheit U, die zweite Stichprobe der Fische der zweiten Volkszählung.

Die Daten der Population U_R bezeichnen wir als Liste R, die Daten der Population U_Z als Liste Z. Dann können wir schreiben

$$\begin{aligned} U_R &= \{i \in U : i \in R\} \subseteq U \qquad \#U_R = \tau_R \\ U_Z &= \{i \in U : i \in Z\} \subseteq U \qquad \#U_Z = \tau_Z \qquad , \end{aligned}$$

wobei das Symbol $\#$ der Umfang einer Menge ist.

Da R und Z Ergebnisse von Zufallsexperimenten sind, bezeichne für $i \in U$

$$\begin{aligned} p_{Ri} &= P(i \in R) \\ p_{Zi} &= P(i \in Z) \qquad . \end{aligned}$$

Definieren wir nun ein *allgemeines Modell* M_g für N_P und den Coverage–Error, das auf folgenden Anforderungen begründet ist.

(i) Die Population U ist geschlossen und hat die fixe Größe N_P. Es gibt keine Geburten und Sterbefälle.

(ii) Betrachten wir ein Experiment, das nur vier mögliche Ergebnisse hat mit Wahrscheinlichkeiten p_{00i}, p_{Ki}, p_{Fi}, p_{11i}. Dann gibt Tabelle 2.2 mit den Parametern p_{00i}, p_{Ki}, p_{Fi}, p_{11i} an, mit welcher Wahrscheinlichkeit die Person $i \in U$ in die vier möglichen Kategorien fällt.

		Z		
		in	out	
R	in	p_{00i}	p_{Ki}	p_{Ri}
	out	p_{Fi}	p_{11i}	
		p_{Zi}		$p_i = 1$

TABELLE 2.2: *Die Wahrscheinlichkeiten für die Person i im allgemeinen Modell M_g.*

		Z		
		in	out	
R	in	τ_{00}	τ_K	τ_R
	out	τ_F	τ_{11}	
		τ_Z		N_P

TABELLE 2.3: *Die 2×2 Kontingenztabelle für das allgemeine Modell M_g.*

Es gilt

$$\begin{aligned} p_{00i} &= P(i \in R, i \in Z) \\ p_{Ki} &= P(i \in R, i \notin Z) \\ p_{Fi} &= P(i \notin R, i \in Z) \\ p_{11i} &= P(i \notin R, i \notin Z) \quad . \end{aligned}$$

Bei Unabhängigkeit in einer Liste erfasst zu werden, ist

$$\begin{aligned} p_{00i} &= p_{Ri}\, p_{Zi} \\ p_{Ki} &= p_{Ri}\,(1 - p_{Zi}) \\ p_{Fi} &= p_{Zi}\,(1 - p_{Ri}) \\ p_{11i} &= (1 - p_{Ri})\,(1 - p_{Zi}) \qquad . \end{aligned} \tag{2.1}$$

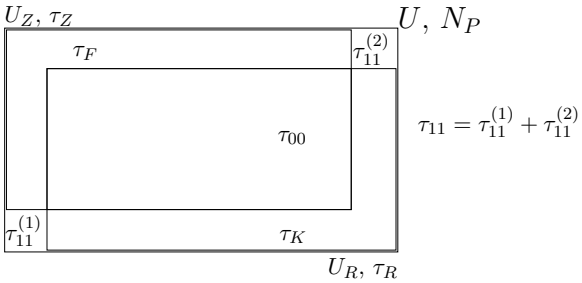


ABBILDUNG 2.4: *Die absoluten Häufigkeiten in der Population U.*

(iii) Wenn das Experiment N_P mal wiederholt wird, sind die absoluten Häufigkeiten der Personen in jeder Zelle wie in der 2×2 Kontingenztabelle 2.3 (vgl. Agresti 2002, S. 36). Die absoluten Häufigkeiten sind ein Ergebnis der N_P unabhängigen Wiederholungen bei multinomialer Verteilung, also

$$\begin{aligned} \tau_{00} &= \sum_{i=1}^{N_P} \tau_{00i} & \tau_K &= \sum_{i=1}^{N_P} \tau_{Ki} & \tau_F &= \sum_{i=1}^{N_P} \tau_{Fi} \\ \tau_{11} &= \sum_{i=1}^{N_P} \tau_{11i} & \tau_R &= \sum_{i=1}^{N_P} \tau_{Ri} & \tau_Z &= \sum_{i=1}^{N_P} \tau_{Zi} \end{aligned} \quad ,$$

wobei

$$\tau_{00i} = \begin{cases} 1 & \text{für } i \in R, i \in Z \\ 0 & \text{sonst} \end{cases} \qquad \tau_{Ki} = \begin{cases} 1 & \text{für } i \in R, i \notin Z \\ 0 & \text{sonst} \end{cases}$$

$$\tau_{Fi} = \begin{cases} 1 & \text{für } i \notin R, i \in Z \\ 0 & \text{sonst} \end{cases} \qquad \tau_{11i} = \begin{cases} 1 & \text{für } i \notin R, i \notin Z \\ 0 & \text{sonst.} \end{cases}$$

Zum besseren Verständnis siehe Abbildung 2.4.

Die absoluten Häufigkeiten $\tau_{00}, \tau_K, \tau_F$ und τ_{11} können wir auch wie folgt beschreiben

$$\begin{aligned} \tau_{00} &= \#(U_R \cap U_Z) \\ \tau_K &= \#(U_R \setminus U_Z) \\ \tau_F &= \#(U_Z \setminus U_R) \\ \tau_{11} &= \#(U \setminus (U_R \cup U_Z)) \qquad . \end{aligned}$$

Die Anzahl τ_R und τ_Z sind bekannt, die Anzahl τ_{00} und τ_F sind auf der Basis der zweiten Volkszählung auch bekannt. Die Anzahl τ_K ist dann $\tau_K = \tau_R - \tau_{00}$. Die Anzahl τ_{11} kann man interpretieren als Anzahl der Personen, die zu keiner der beiden Volkszählungen gehören sollten (z. B. Duplikate oder Personen, die nach einem Referenztag geboren wurden, trotzdem durch die Volkszählung gezählt werden). τ_{11} und N_P sind unbekannt. N_P wird durch das Modell geschätzt.

(iv) Wir können ohne Fehler für Personen, die in Liste Z sind, feststellen, ob sie in Liste R sind oder nicht.

Dieses Modell ist unterdefiniert und hat $3 \times N_P$ Parameter. Es sind weitere

Forderungen notwendig, um die Größe N_P der Population U schätzen zu können.

(v) Unabhängigkeit: Die Inklusion in Liste R ist unabhängig von der Inklusion in Liste Z. Für $i = 1, \ldots, N_P$ also gilt

$$\gamma = \frac{p_{00i}\, p_{11i}}{p_{Fi}\, p_{Ki}} = 1 \qquad .$$

Bell (1993) stellt alternative Modelle vor, in denen eine zusätzliche Information aus der Fortschreibung verwendet wird, um $\gamma \neq 1$ zu betrachten. Aus Platzgründen werden diese alternativen Modelle in dieser Arbeit aber nicht weiter untersucht.

(vi) Homogenität: Für jede Person haben wir die gleiche Wahrscheinlichkeit p_R, dass die Person in Liste R ist, und für jede Person haben wir die gleiche Wahrscheinlichkeit p_Z, dass die Person in Liste Z ist. Für $i = 1, \ldots, N_P$ gilt

$$\begin{aligned} p_{Ri} &= p_R \\ p_{Zi} &= p_Z \qquad . \end{aligned}$$

Dann sind die Wahrscheinlichkeiten des Modells für den Coverage–Error in Tabelle 2.4 eingetragen. Das Modell heißt *Petersen–Modell* M_t.

Die Anzahl τ_F, τ_{00}, τ_K, τ_{11} in Tabelle 2.3 ist die Realisierung eines Zufallsexperiments mit unbekannten Parametern p_F, p_{00}, p_K, p_{11} in Tabelle 2.4. Die Likelihood–Funktion (vgl. Agresti 2002, S. 9) für das Petersen–Modell M_t ist die Funktion dieser unbekannten Parameter

$$L(N_P, p_{00}, p_K, p_F, p_{11}) = \frac{N_P!}{\tau_{00}!\, \tau_K!\, \tau_F!\, \tau_{11}!}\, p_{00}^{\tau_{00}}\, p_K^{\tau_K}\, p_F^{\tau_F}\, p_{11}^{\tau_{11}} \qquad .$$

		Z		
		in	out	
R	in	p_{00}	p_K	p_R
	out	p_F	p_{11}	
		p_Z		1

TABELLE 2.4: *Die Wahrscheinlichkeiten im Petersen–Modell M_t.*

Nach Einsetzung von (2.1) in diese Likelihood–Funktion gilt

$$
\begin{aligned}
&L(N_P, p_{00}, p_K, p_F, p_{11}) = \\
&\frac{N_P!}{\tau_{00}!\,\tau_K!\,\tau_F!\,\tau_{11}!} \times \\
&(p_R\,p_Z)^{\tau_{00}}\,[p_R\,(1-p_Z)]^{\tau_K}\;\,[p_Z\;(1-p_R)]^{\tau_F}\,[(1-p_Z)\,(1-p_R)]^{\tau_{11}} = \\
&\frac{N_P!}{\tau_{00}!\,\tau_K!\,\tau_F!\,\tau_{11}!} \times \\
&p_R^{\tau_{00}+\tau_K}\;p_Z^{\tau_{00}+\tau_F}\;(1-p_Z)^{\tau_K+\tau_{11}}\;\,(1-p_R)^{\tau_F+\tau_{11}} \quad .
\end{aligned}
\tag{2.2}
$$

Aus der Tabelle 2.3 gilt

$$
\begin{aligned}
\tau_Z &= \tau_{00} + \tau_F \\
\tau_R &= \tau_{00} + \tau_K \\
\tau_K + \tau_{11} &= N_P - \tau_Z \\
\tau_F + \tau_{11} &= N_P - \tau_R
\end{aligned}
$$

und daher kann man (2.2) schreiben

$$
\begin{aligned}
&L(N_P, p_R, p_Z) = \\
&\frac{N_P!}{\tau_{00}!\,\tau_K!\,\tau_F!\,\tau_{11}!}\,p_R^{\tau_R}(1-p_R)^{N_P-\tau_R}\,p_Z^{\tau_Z}(1-p_Z)^{N_P-\tau_Z} \quad .
\end{aligned}
\tag{2.3}
$$

Als Maximum–Likelihood–Schätzer wird derjenige Schätzer bezeichnet, der die Likelihood–Funktion maximiert. Es wird häufig die logarithmierte Likelihood–Funktion verwendet, da sie an der selben Stelle wie die Funktion selbst ein Maximum besitzt, jedoch einfacher zu berechnen ist.

Durch Logarithmieren von (2.3) ergibt sich

$$\begin{aligned}\ln L(N_P, p_R, p_Z) = \\ \ln \frac{N_P!}{\tau_{00}!\,\tau_K!\,\tau_F!\,\tau_{11}!} + \ln p_R^{\tau_R} + \ln(1-p_R)^{N_P-\tau_R} + \\ \ln p_Z^{\tau_Z} + \ln(1-p_Z)^{N_P-\tau_Z} \quad .\end{aligned}$$

Ableiten dieser Funktion nach p_R und p_Z und Null setzen ergibt den Maximum–Likelihood–Schätzer des Petersen–Modells M_t für p_R und p_Z

$$\hat{p}_R = \frac{\tau_R}{N_P} \qquad \hat{p}_Z = \frac{\tau_Z}{N_P} \quad . \tag{2.4}$$

Nach Multiplizierung mit N_P der beiden Zeilen der ersten Gleichung in (2.1) erhält man mit Hilfe der Homogenität (vi)

$$N_P\, p_{00} \quad = \quad N_P\, p_R\, p_Z \quad .$$

Nach Einsetzen von (2.4) und $p_{00} = \frac{\tau_{00}}{N_P}$ erhält man schließich

$$\widehat{N}_P = \tau_R\, \frac{\tau_Z}{\tau_{00}} \quad . \tag{2.5}$$

Wie oben erwähnt, ist der Coverage–Error der Population U (die Differenz zwischen N_P und τ_R) für viele Länder interessant. Mittels des Schätzers $\widehat{N}_P$ ist die Schätzung des Coverage–Errors

$$\widehat{\text{Coverage–Error}} = \widehat{N}_P - \tau_R = \tau_R\, \frac{\tau_Z}{\tau_{00}} - \tau_R = \tau_R\, \frac{\tau_Z - \tau_{00}}{\tau_{00}} = \tau_R\, \frac{\tau_F}{\tau_{00}} \quad .$$

2.3 Modell in der Praxis

In der Praxis werden beide Listen (Liste R und Liste Z) verglichen, um τ_{00} und τ_F zu berechnen. Dieser Vergleich kann aufwändig und teuer sein, wenn die Population groß ist. Als Lösung bietet sich an, nur einen Zensus durchzuführen. Statt des zweiten Zensus wird nur eine Stichprobe erhoben und nur die Daten der durch die Stichprobe ausgewählten Einheiten mit dem Zensus verglichen.

2.3.1 Stichprobe in Population U_Z

Eine schematische Darstellung des Zensus und der Stichprobe S_Z aus der Population U_Z zeigt Abbildung 2.5. Der Umfang der Gesamtheit τ_Z wird durch $\hat{\tau}_Z$ geschätzt, auch τ_{00} wird durch die Stichprobenwerte $\hat{\tau}_{00}$ geschätzt. Die folgende Tabelle 2.5 enthält die geschätzten absoluten Häufigkeiten.

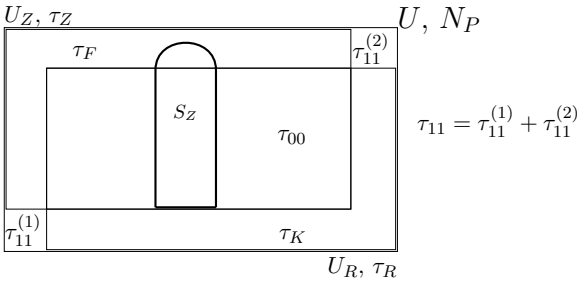


ABBILDUNG 2.5: *Stichprobe S_Z in der Population U_Z.*

		Z		
		in	out	
R	in	$\hat{\tau}_{00}$	τ_R - $\hat{\tau}_{00}$	τ_R
	out	$\hat{\tau}_Z - \hat{\tau}_{00}$	τ_{11}	
		$\hat{\tau}_Z$		$\widehat{N}_P$

TABELLE 2.5: *Die geschätzten absoluten Häufigkeiten bei der Stichprobe S_Z in der Population U_Z.*

Nach Einsetzung der geschätzten $\hat{\tau}_Z$ und $\hat{\tau}_{00}$ in (2.5) ist nun der Maximum–Likelihood–Schätzer für $\widehat{N}_P$

$$\widehat{N}_P = \tau_R \, \frac{\hat{\tau}_Z}{\hat{\tau}_{00}} \qquad . \tag{2.6}$$

Das ganze Modell mit einer Stichprobe in der Population U_Z heißt *Dual–System–Modell*, der Schätzer in (2.6) heißt *Dual–System–Estimator*, kurz DSE.

Bezeichnen wir mit $\hat{\tau}_K = \tau_R - \hat{\tau}_{00}$ und $\hat{\tau}_F = \hat{\tau}_Z - \hat{\tau}_{00}$, so können wir die absoluten Häufigkeiten einfacher schreiben (s. Tabelle 2.6).

		Z		
		in	out	
R	in	$\hat{\tau}_{00}$	$\hat{\tau}_K$	τ_R
	out	$\hat{\tau}_F$	τ_{11}	
		$\hat{\tau}_Z$		$\widehat{N}_P$

TABELLE 2.6: *Die geschätzten absoluten Häufigkeiten in den Listen R und Z.*

Die Schätzung des Coverage–Errors ist

$$\widehat{\text{Coverage–Error}} = \widehat{N}_P - \tau_R = \tau_R \frac{\hat{\tau}_Z}{\hat{\tau}_{00}} - \tau_R = \tau_R \frac{\hat{\tau}_Z - \hat{\tau}_{00}}{\hat{\tau}_{00}} = \tau_R \frac{\hat{\tau}_F}{\hat{\tau}_{00}} \quad .$$

2.3.2 Stichproben in Populationen U_Z und U_R

Häufig wird vorausgesetzt, dass der erste Zensus nicht perfekt ist, da manche Personen mehrfach in der Population N_R erfasst sind. Viele Länder ziehen zusätzlich eine Stichprobe aus der Liste R und suchen nach doppelten Einträgen. Es sei S_R eine Stichprobe aus der Population U_R (s. Abbildung 2.6). Die Anzahl τ_R ist nicht mehr fix, sondern wird durch $\hat{\tau}_R$ geschätzt.

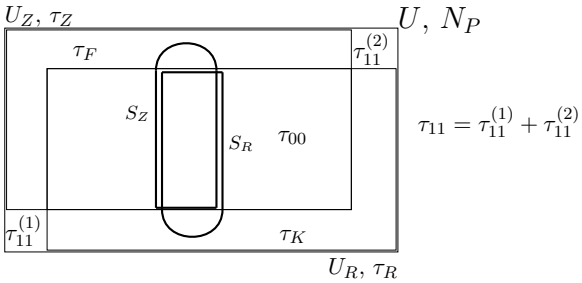


ABBILDUNG 2.6: *Stichproben S_Z und S_R in der Populationen U_Z und U_R.*

Der DSE ist dann

$$\widehat{N}_P = \hat{\tau}_R \frac{\hat{\tau}_Z}{\hat{\tau}_{00}} \quad . \tag{2.7}$$

Die Schweiz, die USA, Großbritannien und Israel benutzen den DSE, um den Coverage–Error zu bestimmen. Die Variante mit zwei Stichproben verwenden zum Beispiel die USA (vgl. Hogan 1993) oder die Schweiz. Im Fol-

genden benutzen wir für neue Begriffe die Nomenklatur der Schweiz, die in Renaud (2004) zu finden ist.

Über die Stichprobe S_Z wird der Umfang τ_Z der Gesamtheit U_Z durch $\hat{\tau}_Z$ geschätzt, τ_{00} wird durch

$$\widehat{M} = \sum_{i \in S_Z} w_{S_Z,i} P_{M,i}$$

geschätzt. $\widehat{M}$ ist eine Schätzung für τ_{00}, dem Totalwert der Personen in Liste Z, die mit einer Person in Liste R übereinstimmen (vgl. Renaud 2004, S. 34). $w_{S_Z,i}$ ist ein Gewicht der Person $i \in S_Z$ und gibt an, wie viele Personen die ausgewählte Person i repräsentiert. Der Status $P_{M,i}$ ist definiert durch

$$P_{M,i} = \begin{cases} 1 & \text{falls } i \in S_Z \text{ mit einer Person aus Liste } R \text{ übereinstimmt} \\ 0 & \text{sonst.} \end{cases}$$

Der erste Teil der geschätzten absoluten Häufigkeiten durch die Stichprobe S_Z ist in Tabelle 2.7 dargestellt.

		Z		
		in	out	
R	in	$\widehat{M}$		
	out	$\hat{\tau}_Z - \widehat{M}$		
		$\hat{\tau}_Z$		

TABELLE 2.7: *Die geschätzten absoluten Häufigkeiten in der Liste Z bei der Stichprobe S_Z.*

Über die zweite Stichprobe S_R wird τ_{00} durch

$$\widehat{CE} = \sum_{i \in S_R} w_{S_R,i} P_{CE,i}$$

geschätzt. $\widehat{CE}$ ist eine Schätzung für τ_{00}, dem Totalwert der *korrekten* Personen in Liste R (vgl. Renaud 2004, S. 27). $w_{S_R,i}$ ist das Gewicht der Person $i \in S_R$. Der Status $P_{CE,i}$ ist definiert durch

$$P_{CE,i} = \begin{cases} 0 & \text{falls } i \in S_R \text{ nicht korrekt ist} \\ 1 & \text{falls } i \in S_R \text{ mit einer Person aus } S_Z \text{ übereinstimmt} \\ \frac{1}{2} & \text{falls } j \text{ Dublette} \\ 1 & \text{sonst.} \end{cases}$$

Durch die Stichprobe S_R wird noch τ_R geschätzt. Die Schätzung ist auf der Korrektur der Gesamtzahl τ_R um die *correct enumeration Rate* $\frac{\widehat{CE}}{\hat{\tau}_R}$ begründet. Der zweite Teil der geschätzten absoluten Häufigkeiten ist Tabelle 2.8 zu entnehmen.

		Z		
		in	out	
R	in	$\widehat{CE}$	$\tau_R \frac{\widehat{CE}}{\hat{\tau}_R} - \widehat{CE}$	$\tau_R \frac{\widehat{CE}}{\hat{\tau}_R}$
	out			

TABELLE 2.8: *Die geschätzten absoluten Häufigkeiten in der Liste R bei der Stichprobe S_R.*

Daher ist wegen der Unabhängigkeit der Stichproben der DSE für N_P wie folgt gegeben

$$\widehat{N}_P = \tau_R \frac{\widehat{CE}}{\hat{\tau}_R} \frac{\hat{\tau}_Z}{\widehat{M}} \quad . \tag{2.8}$$

Beachten wir, dass $\widehat{M}$ und $\hat{\tau}_Z$ durch die Stichprobe S_Z geschätzt ist, $\widehat{CE}$ und $\hat{\tau}_R$ durch die Stichprobe S_R.

Die Schätzung des Coverage–Errors ist

$$\widehat{\text{Coverage–Error}} = \widehat{N}_P - \tau_R \frac{\widehat{CE}}{\hat{\tau}_R} \quad .$$

Wie schon gesagt, werden diese Modelle in vielen Ländern benutzt, um N_P und den Coverage–Error zu bestimmen. In Deutschland sind wir nicht an der Schätzung von N_P und des Coverage–Errors interessiert (s. Kapitel 3). Im Folgenden werden die Dual–System–Modelle umgebaut, um die gewünschten Kennzahlen in Deutschland zu gewinnen.

3 Dual–System–Modelle in Deutschland

Ziel dieser Arbeit ist es zu untersuchen, ob wir die Dual–System–Modelle am kommenden registergestützten Zensus auch anwenden können. In diesem Kapitel werden zwei Alternativen des Dual–System–Modells präsentiert. Im ersten Modell wird die Hauptidee vom obigen Kapitel in einer Gemeinde in Deutschland angewendet. Danach wird das Dual–System–Modell ins Chapman–Modell modifiziert.

3.1 Dual–System–Modell

Bezeichnen wir das Einwohnermelderegister (kurz Register) in Deutschland in der Terminologie des DSEs als Liste R, die Anzahl der in den Einwohnermeldeämtern registrierten Personen als τ_R. Das Register enthält Einträge, deren Personen nicht mehr in einer Gemeinde wohnhaft sind (*Karteileichen*). Die Anzahl dieser Personen sei τ_K.

Es sei Z eine Liste der Population U_Z vom Umfang τ_Z, die in Deutschland tatsächlich zu finden ist. Diese Liste enthält auch Personen, die in einer Gemeinde wohnen, aber nicht dort gemeldet sind (*Fehlbestände*). Die Anzahl dieser Personen sei τ_F.

Es ist

$$N_P = \tau_Z + \tau_K = \tau_R + \tau_F$$

der Umfang der Liste Z und der Karteileichen bzw. die Anzahl aller registrierten Personen und der Fehlbestände.

Das Ziel des Forschungsprojekts für den Zensus 2011 ist, die tatsächlich vorhandenen Personenzahl $\tau_Z = \tau_{00} + \tau_F$ zu ermitteln, wobei τ_{00} die Anzahl der in beiden Listen R und Z vorhandenen Personen ist. Zur Ermittlung werden neben τ_R die Anzahl der Karteileichen τ_K und Anzahl der Fehlbestände τ_F benötigt. Zu beachten ist, dass die Anzahl $\tau_R = \tau_{00} + \tau_K$ der registrierten Personen bekannt ist, nicht aber die beiden Komponenten τ_{00} und τ_K. Die Anzahl der tatsächlich vorhandenen Personen τ_Z ist gegeben durch

$$\tau_Z = \tau_R - \tau_K + \tau_F \qquad , \tag{3.1}$$

welche anschließend geschätzt werden muss.

Die Schätzung in Deutschland erfolgt für Gemeinden. Wir bezeichnen das Register einer Gemeinde g als Liste $R_{<g>}$, die Anzahl der registrierten Personen als $\tau_{R,<g>}$, die Anzahl der Karteileichen in einer Gemeinde als $\tau_{K,<g>}$. Die Population $U_{Z,<g>}$ von Umfang $\tau_{Z,<g>}$ bezeichnet die tatsächlich vorhandenen Personen einer Gemeinde g. Die Anzahl der Fehlbestände sei $\tau_{F,<g>}$. Solange es sich um eine feste Gemeinde g handelt, wird im Folgenden auf das g im Index verzichtet.

Die Situation einer Gemeinde lässt sich wie folgt darstellen (s. Abbildung 3.1). Im Zusammenhang mit dem Kapitel 2 entspricht die Population U_R des Registers der ersten Volkszählung. Die Population U_Z der tatsächlich vorhandenen Personen entspricht der zweiten Volkszählung. Im Gegensatzt zu den USA bzw. Schweiz, nehmen wir an, dass in Deutschland keine Person sowohl Karteileiche als auch Fehlbestand sein kann. Das Modell in Deutschland hat nur drei Kategorien.

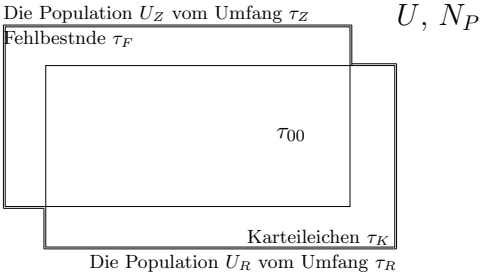


ABBILDUNG 3.1: *Population U in einer Gemeinde.*

Es sei Y eine kategoriale Variable (s. Agresti 2002, S. 1) mit den drei Kategorien

$$Y = \begin{cases} F & \text{für Fehlbestände} \\ 00 & \text{für in beiden Listen } R \text{ und } Z \text{ eingetragenen Personen} \\ K & \text{für Karteileichen} \end{cases} .$$

Eine Person $i \in U$ gehört in genau eine der drei Kategorien F, 00 oder K.

Wir bezeichnen für $i \in U$

$$\begin{aligned} p_{Fi} &= P(i \notin R, i \in Z) \\ p_{00i} &= P(i \in R, i \in Z) \\ p_{Ki} &= P(i \in R, i \notin Z) \end{aligned} .$$

Für die Wahrscheinlichkeiten

$$\begin{aligned} p_{Zi} &= P(i \in Z) \\ p_{Ri} &= P(i \in R) \end{aligned} ,$$

gilt

$$p_{Zi} = p_{00i} + p_{Fi} \qquad p_{Ri} = p_{00i} + p_{Ki} \qquad . \tag{3.2}$$

Die Parameter p_{Fi}, p_{00i}, p_{Ki} geben an, mit welchen Wahrscheinlichkeiten die Person i in die drei Kategorien fällt (s. Tabelle 3.1).

Y			
F	00	K	
p_{Fi}	p_{00i}	p_{Ki}	$p_i = 1$

TABELLE 3.1: *Die Wahrscheinlichkeiten für die Person i in Deutschland.*

Nach N_P maliger Wiederholung sind die absoluten Häufigkeiten der Personen in jedem Feld wie in der Tabelle 3.2, wo gilt

$$\tau_F = \sum_{i=1}^{N_P} \tau_{Fi} \qquad \tau_{00} = \sum_{i=1}^{N_P} \tau_{00i} \qquad \tau_K = \sum_{i=1}^{N_P} \tau_{Ki} \qquad .$$

Y			
F	00	K	
τ_F	τ_{00}	τ_K	N_P

TABELLE 3.2: *Die absoluten Häufigkeiten in der Population U in einer Gemeinde.*

Nehmen wir an, dass jede Person i, für $i = 1, \ldots, N_P$, die gleiche Wahrscheinlichkeit p_F hat, ein Fehlbestand zu sein. Weiter, dass jede Person i die gleiche Wahrscheinlichkeit p_{00} hat, in beiden Listen erfasst zu werden,

und schließlich, dass jede Person i gleiche Wahrscheinlichkeit p_K hat, eine Karteileiche zu sein. Es gilt also

$$\begin{aligned} p_{Fi} &= p_F \\ p_{00i} &= p_{00} \\ p_{Ki} &= p_K \quad . \end{aligned}$$

Wir interpretieren die Anzahl τ_F, τ_{00}, τ_K als Realisierung eines Zufallsexperiments, das von den unbekannten Parametern p_F, p_{00}, p_K abhängt. Mit der Forderung, dass es keine Geburten und Sterbefälle gibt (geschlossene Population U) und dass wir ohne Fehler für die in Liste Z eingetragenen Personen feststellen können, ob sie in Liste R sind oder nicht, ist die Likelihood–Funktion (vgl. Agresti 2002, S. 9) zu unserem Modell

$$L(N_P, p_F, p_{00}, p_K) = \frac{N_P!}{\tau_F!\,\tau_{00}!\,\tau_K!}\, p_F^{\tau_F} p_{00}^{\tau_{00}} p_K^{\tau_K} \quad .$$

Wir suchen ein Maximum dieser Funktion unter der Bedingung

$$p_F + p_{00} + p_K = 1 \qquad p_j \geq 0 \text{ für } j \in \{F, 00, K\} \quad .$$

Durch Logarithmieren der Likelihood–Funktion, Ableiten nach p_F, p_{00} und p_K und Null setzen ist der Maximum–Likelihood–Schätzer für p_F, p_{00} und p_K (vgl. Agresti 2002, S. 21)

$$\hat{p}_F = \frac{\tau_F}{N_P} \qquad \hat{p}_{00} = \frac{\tau_{00}}{N_P} \qquad \hat{p}_K = \frac{\tau_K}{N_P} \quad . \tag{3.3}$$

Nehmen wir an, es wird eine Stichprobe S_P von Personen erhoben (s. Abbildung 3.2). Man beachte, dass in Deutschland keine zusätzliche Stichprobe aus dem Register gezogen wird, um nach doppelten Einträge zu suchen, wie in den USA bzw. der Schweiz (s. Kapitel 2.3.2). Es ist auch wichtig zu betonen, dass uns in Deutschland nicht der Coverage–Error, sondern

die tatsächliche Anzahl der in Deutschland vorhandenen Personen interessiert.

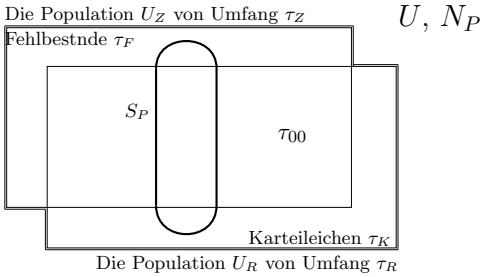


ABBILDUNG 3.2: *Stichprobe S_P in einer Gemeinde.*

Die Personen der Population U haben insgesamt 2×3 Möglichkeiten, in einer Kategorie zu sein (s. Tabelle 3.3), wobei $\bar{S}_P = U \setminus S_P$.

	Y		
	F	00	K
S_P			
$\bar{S}_P$			

TABELLE 3.3: *2×3 Möglichkeiten in einer Kategorie in einer Gemeinde zu sein.*

Tabelle 3.4 gibt die 2×3 Kontingenztabelle (vgl. Agresti 2002, S. 36) an, mit

τ_{F,S_P} als Anzahl der Fehlbestände in der Stichprobe S_P,

τ_{00,S_P} als Anzahl der in beiden Listen eingetragenen Personen der Stichprobe S_P,

τ_{K,S_P} als Anzahl der Karteileichen in der Stichprobe S_P,

$\tau_{j,\bar{S}_P}$ als Anzahl der nicht in der Stichprobe S_P enthaltenen Personen aus der Kategorie j, für $j \in \{F, 00, K\}$,

n_P als Umfang der Stichprobe S_P.

	Y			
	F	00	K	
S_P	τ_{F,S_P}	τ_{00,S_P}	τ_{K,S_P}	n_P
$\bar{S}_P$	$\tau_{F,\bar{S}_P}$	$\tau_{00,\bar{S}_P}$	$\tau_{K,\bar{S}_P}$	$\bar{n}_P = N_P - n_P$
	τ_F	τ_{00}	τ_K	N_P

TABELLE 3.4: *2×3 Kontingenztabelle der Population U in einer Gemeinde.*

Für die Anzahl $\tau_F, \tau_{00}, \tau_K$ gilt

$$\begin{aligned} \tau_F &= \tau_{F,S_P} + \tau_{F,\bar{S}_P} \\ \tau_{00} &= \tau_{00,S_P} + \tau_{00,\bar{S}_P} \\ \tau_K &= \tau_{K,S_P} + \tau_{K,\bar{S}_P} \quad . \end{aligned}$$

Tabelle 3.5 ist die Tafel der Wahrscheinlichkeiten, wobei $p_{j,\bar{S}_P}$ die Wahrscheinlichkeit ist, dass eine Person aus der Kategorie j nicht in der Stichprobe S_P ist, für $j \in \{F, 00, K\}$. Für die Randwahrscheinlichkeiten einer Zeile l, $l \in \{S_P, \bar{S}_P\}$, beziehungsweise Spalte j gilt

$$\sum_j p_{j,S_P} = p_{S_P} \qquad \sum_j p_{j,\bar{S}_P} = p_{\bar{S}_P} \qquad \sum_j p_j = 1$$

$$\sum_l p_{F,l} = p_F \qquad \sum_l p_{00,l} = p_{00} \qquad \sum_l p_{K,l} = p_K \qquad \sum_l p_l = 1 \quad .$$

Wenn es zufällig ist, ob eine Person in die Stichprobe S_P gelangt oder nicht, und der Umfang n_P der Personen in der Stichprobe S_P (und damit $\bar{n}_P$=$N_P - n_P$, weil geschlossene Population U) in Tabelle 3.4 fix ist,

	Y			
	F	00	K	
S_P	p_{F,S_P}	p_{00,S_P}	p_{K,S_P}	p_{S_P}
$\bar{S}_P$	$p_{F,\bar{S}_P}$	$p_{00,\bar{S}_P}$	$p_{K,\bar{S}_P}$	$p_{\bar{S}_P} = 1 - p_{S_P}$
	p_F	p_{00}	p_K	$p = 1$

TABELLE 3.5: *Die Wahrscheinlichkeiten für die Personen in den sechs möglichen Kategorien zu sein.*

dann gilt für alle absoluten Häufigkeiten einer Zeile (vgl. Agresti 2002, S. 40)

$$
\begin{aligned}
&L(n_P, p_{F,S_P}, p_{00,S_P}, p_{K,S_P}) = \\
&\qquad \frac{n_P!}{\tau_{F,S_P}!\,\tau_{00,S_P}!\,\tau_{K,S_P}!} p_{F,S_P}^{\tau_{F,S_P}}\; p_{00,S_P}^{\tau_{00,S_P}}\; p_{K,S_P}^{\tau_{K,S_P}} \\
&L(\bar{n}_P, p_{F,\bar{S}_P}, p_{00,\bar{S}_P}, p_{K,\bar{S}_P}) = \\
&\qquad \frac{\bar{n}_P!}{\tau_{F,\bar{S}_P}!\,\tau_{00,\bar{S}_P}!\,\tau_{K,\bar{S}_P}!} p_{F,\bar{S}_P}^{\tau_{F,\bar{S}_P}}\; p_{00,\bar{S}_P}^{\tau_{00,\bar{S}_P}}\; p_{K,\bar{S}_P}^{\tau_{K,\bar{S}_P}} \quad .
\end{aligned}
\tag{3.4}
$$

Die Likelihood–Funktion zu dem Modell ist das Produkt der Funktionen in (3.4)

$$
\begin{aligned}
&L(n_P, \bar{n}_P, p_{F,S_P}, p_{00,S_P}, p_{K,S_P}, p_{F,\bar{S}_P}, p_{00,\bar{S}_P}, p_{K,\bar{S}_P}) = \\
&\qquad \frac{n_P!\,\bar{n}_P!}{\tau_{F,S_P}!\,\tau_{00,S_P}!\,\tau_{K,S_P}!\,\tau_{F,\bar{S}_P}!\,\tau_{00,\bar{S}_P}!\,\tau_{K,\bar{S}_P}!} \times \\
&\qquad p_{F,S_P}^{\tau_{F,S_P}}\; p_{00,S_P}^{\tau_{00,S_P}}\; p_{K,S_P}^{\tau_{K,S_P}}\; p_{F,\bar{S}_P}^{\tau_{F,\bar{S}_P}}\; p_{00,\bar{S}_P}^{\tau_{00,\bar{S}_P}}\; p_{K,\bar{S}_P}^{\tau_{K,\bar{S}_P}} \quad .
\end{aligned}
$$

Wir analysieren die Daten jeder Zeile der Kontingenztabelle getrennt. Aus

der ersten Zeile der Kontingenztabelle (s. Tabelle 3.4) ist der Maximum–Likelihood–Schätzer für p_{F,S_P}, p_{00,S_P} und p_{K,S_P}

$$\hat{p}_{F,S_P} = \frac{\tau_{F,S_P}}{n_P} \qquad \hat{p}_{00,S_P} = \frac{\tau_{00,S_P}}{n_P} \qquad \hat{p}_{K,S_P} = \frac{\tau_{K,S_P}}{n_P} \qquad . \tag{3.5}$$

Aus der zweiten Zeile ist

$$\hat{p}_{F,\bar{S}_P} = \frac{\tau_{F,\bar{S}_P}}{\bar{n}_P} \qquad \hat{p}_{00,\bar{S}_P} = \frac{\tau_{00,\bar{S}_P}}{\bar{n}_P} \qquad \hat{p}_{K,\bar{S}_P} = \frac{\tau_{K,\bar{S}_P}}{\bar{n}_P} \qquad . \tag{3.6}$$

Nehmen wir an, es soll insgesamt θ 100% der Bevölkerung ausgewählt werden, mit

$$\theta = \frac{n_P}{N_P} \qquad . \tag{3.7}$$

Wir gehen davon aus, dass diese Gesamtstichprobe auch $\theta\ \tau_F$ Fehlbestände, $\theta\ \tau_K$ Karteileichen und $\theta\ \tau_{00}$ der in beiden Listen eingetragenen Personen enthält, also

$$\tau_{F,S_P} = \theta\ \tau_F \qquad \tau_{00,S_P} = \theta\ \tau_{00} \qquad \tau_{K,S_P} = \theta\ \tau_K \qquad . \tag{3.8}$$

Für die Anzahl, der nicht in der Stichprobe S_P enthalten Personen, gilt

$$\tau_{F,\bar{S}_P} = (1-\theta)\ \tau_F \quad \tau_{00,\bar{S}_P} = (1-\theta)\ \tau_{00} \quad \tau_{K,\bar{S}_P} = (1-\theta)\ \tau_K \quad . \tag{3.9}$$

Entsprechend dieser Tatsache sind die absoluten Häufigkeiten der Stichprobe S_P in Tabelle 3.6, die Wahrscheinlichkeiten für die Personen in den sechs möglichen Kategorien zu sein in Tabelle 3.7 gegeben.

Nach Einsetzung von (3.7), (3.8) und (3.9) in (3.5) bzw. (3.6) gilt für den Maximum–Likelihood–Schätzer für p_{F,S_P}, p_{00,S_P} und p_{K,S_P}

$$\hat{p}_{F,S_P} = \frac{\tau_F}{N_P} \qquad \hat{p}_{00,S_P} = \frac{\tau_{00}}{N_P} \qquad \hat{p}_{K,S_P} = \frac{\tau_K}{N_P}$$

	Y			
	F	00	K	
S_P	$\theta\tau_F$	$\theta\tau_{00}$	$\theta\tau_K$	θN_P
$\bar{S}_P$	$(1-\theta)\tau_F$	$(1-\theta)\tau_{00}$	$(1-\theta)\tau_K$	$(1-\theta)N_P$
	τ_F	τ_{00}	τ_K	N_P

TABELLE 3.6: *Die absoluten Häufigkeiten der Stichprobe S_P in Listen R und Z.*

	Y			
	F	00	K	
S_P	p_{F,S_P}	p_{00,S_P}	p_{K,S_P}	θ
$\bar{S}_P$	$p_{F,\bar{S}_P}$	$p_{00,\bar{S}_P}$	$p_{K,\bar{S}_P}$	$1-\theta$
	p_F	p_{00}	p_K	1

TABELLE 3.7: *Die Wahrscheinlichkeiten für die Personen in den sechs möglichen Kategorien zu sein.*

und für $p_{F,\bar{S}_P}, p_{00,\bar{S}_P}, p_{K,\bar{S}_P}$

$$\begin{aligned}
\hat{p}_{F,\bar{S}_P} &= \frac{(1-\theta)\,\tau_F}{\bar{n}_P} = \frac{\tau_F}{N_P} \\
\hat{p}_{00,\bar{S}_P} &= \frac{(1-\theta)\,\tau_{00}}{\bar{n}_P} = \frac{\tau_{00}}{N_P} \\
\hat{p}_{K,\bar{S}_P} &= \frac{(1-\theta)\,\tau_K}{\bar{n}_P} = \frac{\tau_K}{N_P} \qquad .
\end{aligned}$$

Daher ist mit Hilfe der letzten Zeile in Tabelle 3.7, wo

$$\begin{aligned} p_F &= \theta\, p_{F,S_P} + (1-\theta)\, p_{F,\bar{S}_P} \\ p_{00} &= \theta\, p_{00,S_P} + (1-\theta)\, p_{00,\bar{S}_P} \\ p_K &= \theta\, p_{K,S_P} + (1-\theta)\, p_{K,\bar{S}_P} \end{aligned}$$

gilt, der Maximum–Likelihood–Schätzer für p_F, p_{00} und p_K

$$\begin{aligned} \hat{p}_F &= \theta\, \hat{p}_{F,S_P} + (1-\theta)\, \hat{p}_{F,\bar{S}_P} = \frac{\tau_F}{N_P} \\ \hat{p}_{00} &= \theta\, \hat{p}_{00,S_P} + (1-\theta)\, \hat{p}_{00,\bar{S}_P} = \frac{\tau_{00}}{N_P} \\ \hat{p}_K &= \theta\, \hat{p}_{K,S_P} + (1-\theta)\, \hat{p}_{K,\bar{S}_P} = \frac{\tau_K}{N_P} \quad . \end{aligned}$$

Wie oben erwähnt, ist τ_R bekannt. Die Anzahl τ_Z ist zu schätzen. Wegen (3.2) gilt für die entsprechenden Wahrscheinlichkeiten $\hat{p}_Z$, $\hat{p}_R$

$$\begin{aligned} \hat{p}_Z &= \hat{p}_{00} + \hat{p}_F = \frac{\tau_{00} + \tau_F}{N_P} \\ \hat{p}_R &= \hat{p}_{00} + \hat{p}_K = \frac{\tau_{00} + \tau_K}{N_P} \quad . \end{aligned} \tag{3.10}$$

Daher ist

$$\frac{\tau_{00} + \tau_F}{\hat{p}_{00} + \hat{p}_F} = \frac{\tau_{00} + \tau_K}{\hat{p}_{00} + \hat{p}_K}$$

und

$$\tau_Z = \tau_R\, \frac{\hat{p}_{00} + \hat{p}_F}{\hat{p}_{00} + \hat{p}_K} \quad ,$$

wobei mit Hilfe von (3.10) und mit der Erweiterung durch $\frac{n_P}{n_P}$ ist

$$\begin{aligned} \hat{p}_{00} + \hat{p}_F &= \frac{\tau_{00} + \tau_F}{N_P} \frac{n_P}{n_P} = \frac{\tau_{00,S_P} + \tau_{F,S_P}}{n_P} \\ \hat{p}_{00} + \hat{p}_K &= \frac{\tau_{00} + \tau_K}{N_P} \frac{n_P}{n_P} = \frac{\tau_{00,S_P} + \tau_{K,S_P}}{n_P} \quad . \end{aligned}$$

Die Anzahl τ_Z für dieses Modell ist

$$\tau_Z = \tau_R \frac{\tau_{00,S_P} + \tau_{F,S_P}}{\tau_{00,S_P} + \tau_{K,S_P}} = \tau_R \frac{\tau_{Z,S_P}}{\tau_{R,S_P}} \quad . \tag{3.11}$$

Es sei kritisch angemerkt, dass die Voraussetzungen des Stichprobemodells der Realität nicht unbedingt entspricht, da die Unabhängigkeitsannahme verletzt sein dürfte oder die Annahme der identischen Verteidigung für die Personen nicht zutrifft.

3.1.1 DSE bei uneingeschränkter Zufallsauswahl der Klumpen

Die Stichprobenerhebung für den Zensus 2011 in Deutschland baut auf dem Stichprobenrahmen der Anschriften auf. Auswahleinheiten sind also die Anschriften des Anschriftenregisters einer Gemeinde. An einer Anschrift werden alle Personen erhoben. Es sei N die Anzahl aller Anschriften. Wir gehen davon aus, dass das Anschriftenregister vollständig ist. Da die Personen an einer Anschrift vollständig erhoben werden, spricht man von *Klumpenauswahl*. Die Anschriften heißen Klumpen oder Primäreinheiten (engl. *clusters* oder *primary sampling units*, PSUs). Die Personen heißen Sekundäreinheiten (engl. *secondary sampling units*, SSUs).

Bezeichnen wir mit S eine Stichprobe der Anschriften in einer Gemeinde. Es sei $\pi_a, a = \{1, \ldots, N\}$ die Inklusionswahrscheinlichkeit einer An-

schrift a. Es bezeichnet $w_a = \frac{1}{\pi_a}$ das *Designgewicht* (s. Lohr 1999, S. 103), also die Inverse der Inklusionswahrscheinlichkeit π_a für Anschrift a. Mit S_P wird die Menge aller Personen in der ausgewählten Anschriften der Stichprobe S bezeichnet.

Die Anzahl τ_Z der tatsächlich vorhandenen Personen in einer Gemeinde wird mit Hilfe (3.11) durch den sogenannten D–DSE

$$\hat{\tau}_Z^{\text{D-DSE}} = \tau_R \frac{\sum\limits_{a \in S} w_a \, \tau_{Z,a}}{\sum\limits_{a \in S} w_a \, \tau_{R,a}} = \tau_R \frac{\widehat{\tau}_Z}{\widehat{\tau}_R} \tag{3.12}$$

ermittelt. $\tau_{Z,a}$ ist die Anzahl der tatsächlich vorhandenen Personen an der Anschrift a, $\tau_{R,a}$ ist die Anzahl der registrierten Personen an der Anschrift a.

Bei einer uneingeschränkten Zufallsauswahl der Anschriften ohne Zurücklegen sind alle $\pi_a = \frac{n}{N}, a = \{1, \ldots, N\}$ identisch. Damit kürzen sich w_a in (3.12) weg und ergibt sich wieder die Formel in (3.11).

3.1.2 DSE bei geschichteter Zufallsauswahl der Klumpen

Wie die Voruntersuchungen beim Zensustest 2001 (vgl. Statistisches Bundesamt 2004) gezeigt haben, ist eine geschichtete Zufallsauswahl der Klumpen besonderes interessant. Im Folgenden wird zu jedem Schätzer auch der Schätzer bei uneingeschränkter Zufallsauswahl betrachtet.

Die Menge der Anschriften ist in H Schichten zerlegt, wobei N_h die Anzahl der Anschriften in der h-ten Anschriftenschicht ist, $h = 1, \ldots, H$. Bezeichnen wir mit S_h eine Stichprobe vom Umfang n_h der Anschriften innerhalb der Schicht h. Es sei $\{U_1, \ldots, U_H\}$ eine Zerlegung der Gesamt-

population U definiert in H Anschriftenschichten mit Personenumfängen $\{N_{P,1}, \ldots, N_{P,H}\}$. Es gilt

$$U = \bigcup_{h=1}^{H} U_h \qquad ,$$

wobei U_h definiert ist durch

$$U_h = \{i \in U : \text{Person } i \text{ gehört zu einer Anschrift in der Anschriftenschicht } h\} \subseteq U$$

$$\#U_h = N_{P,h} \qquad .$$

In Analogie zur uneingeschränkten Zufallsauswahl der Anschriften ohne Zurücklegen ist

$$\pi_{a,h} = \frac{n_h}{N_h} \tag{3.13}$$

die Inklusionswahrscheinlichkeit einer Anschrift a der Schicht h. Es bezeichnet $w_{a,h} = \dfrac{1}{\pi_{a,h}} = \dfrac{N_h}{n_h}$ das Designgewicht für Anschrift a der Schicht h.

Die Anzahl τ_Z der tatsächlich vorhandenen Personen in einer Gemeinde wird bei dieser Zufallsauswahl durch

$$\hat{\tau}_Z^{\text{D-DSE,str}} = \sum_{h=1}^{H} \hat{\tau}_{Z,h}^{\text{D-DSE}} = \sum_{h=1}^{H} \tau_{R,h} \, \frac{\sum\limits_{a \in S_h} w_{a,h} \, \tau_{Z,a,h}}{\sum\limits_{a \in S_h} w_{a,h} \, \tau_{R,a,h}} = \sum_{h=1}^{H} \tau_{R,h} \, \frac{\widehat{\tau}_{Z,h}}{\widehat{\tau}_{R,h}} \tag{3.14}$$

ermittelt. $\hat{\tau}_{Z,h}^{\text{D-DSE}}$ ist D–DSE berechnet nach (3.12) in einer Schicht h, $\tau_{Z,a,h}$ ist die Anzahl der tatsächlich vorhandenen Personen an der Anschrift a in der Schicht h, $\tau_{R,a,h}$ ist die Anzahl der registrierten Personen an der Anschrift a in der Schicht h.

3.1.3 DSE in einer Klasse

Die Anforderung, dass alle Personen die gleiche Wahrscheinlichkeit haben, in Liste R oder in Liste Z erfasst zu werden (Homogenität, s. Seite 12 bzw. 25), ist in der Regel nicht erfüllt. Aus Erfahrung und früheren Untersuchungen wissen wir, dass jüngere Personen eine geringere Wahrscheinlichkeit haben, erfasst zu werden als ältere. Oder, dass Menschen in städtischen Regionen auch eine geringere Wahrscheinlichkeit besitzen, erfasst zu werden als in ländlichen Gegenden (siehe z. B. Hogan 1993; Robinson et al. 1993). Diese Anforderung muss modelliert werden.

Es werden Subpopulationen (nennen wir sie *Klassen*) gebildet, die durch die Variablen, die diese heterogene Wahrscheinlichkeiten verursachen, definiert sind. Der D–DSE wird in diesen Klassen berechnet und dann zusammengesetzt, um Schätzwerte für die Gesamtpopulation U zu erhalten.

Die Klassen sollten keinen zu kleinen Umfang besitzen, um bei gegebenem Stichprobenumfang einen zuverlässigen Schätzer für diese Klassen zu verwenden. Klassen müssen sowohl in Liste Z als auch in Liste R definierbar sein. Die Vereinigung der Klassen gibt die Gesamtpopulation U.

Bezeichnen $k = 1, \ldots, K$ die Klassen. Für die Gesamtpopulation U gilt

$$U = \bigcup_{k=1}^{K} U_k \qquad ,$$

wo U_k definiert ist

$$U_k = \{i \in U : \text{Person } i \text{ gehört in Klasse } k\} \subseteq U \qquad \#U_k = N_{P,k} \qquad .$$

Dann schreiben wir bei uneingeschränkter Zufallsauswahl der Anschriften

$$\hat{\tau}_{Z,k}^{\text{D-DSE}} = \tau_{R,k} \, \frac{\widehat{\tau}_{Z,k}}{\widehat{\tau}_{R,k}} \tag{3.15}$$

als D–DSE für $\tau_{Z,k}$ in Klasse k. Für die Gesamtheit U dann gilt

$$\hat{\tau}_Z^{\text{D-DSE}} = \sum_{k=1}^{K} \hat{\tau}_{Z,k}^{\text{D-DSE}} \quad . \tag{3.16}$$

Im Fall der geschichteten Zufallsauswahl der Anschriften gilt

$$\hat{\tau}_{Z,k}^{\text{D-DSE,str}} = \sum_{h=1}^{H} \hat{\tau}_{Z,k,h}^{\text{D-DSE}} \quad ,$$

mit $\hat{\tau}_{Z,k,h}^{\text{D-DSE}}$ als D–DSE für $\tau_{Z,k,h}$ einer Klasse k in einer Anschriftenschicht h. Für die Gesamtheit U dann gilt

$$\hat{\tau}_Z^{\text{D-DSE,str}} = \sum_{k=1}^{K} \hat{\tau}_{Z,k}^{\text{D-DSE,str}} = \sum_{k=1}^{K} \sum_{h=1}^{H} \hat{\tau}_{Z,k,h}^{\text{D-DSE}} \quad .$$

Durch den Einsatz der Klassen können wir sofort den D–DSE sowohl für jede Klasse als auch für jede Vereinigung von Klassen ermitteln. Wenn zum Beispiel die Klassen durch die Variablen Alter und Geschlecht definiert sind, und Alter in sieben Altersklassen, Geschlecht in zwei Geschlechtsklassen klassifiziert ist, erhalten wir die Ergebnisse für alle 14 Kombinationen von Geschlecht und Alter.

3.2 Chapman–Modell

Allgemein ist der Zähler in (3.11) bzw. (3.12) innerhalb einer Stichprobe in der Gesamtheit U immer positiv. Wenn aber ein D–DSE für eine Klasse k angewendet wird, kann es vorkommen, dass kein Eintrag einer Klasse k im Register ist ($\tau_{R,S_P,k} = 0$) und keine Person dieser Klasse k in der Ge-

meinde lebt ($\tau_{Z,S_P,k} = 0$). In diesem Fall teilen wir in (3.15) „0/0“, was in `R` ein `NaN` für die Klasse ergibt.

Darüber hinaus kann es vorkommen, dass eine Person als *Fehlbestand* der Klasse k existiert ($\tau_{F,S_P,k} \neq 0$), aber keine Person aus den Kategorien 00 oder K der Klasse k im Register der Gemeinde vorkommt ($\tau_{R,S_P,k} = 0$). In diesem Fall teilen wir in (3.15) „x/0“, wobei x eine beliebige positive Zahl ist. Dies ergibt in `R` ein `Inf` für die Klasse .

Um diese Fälle zu verhindern, wird in Chapman (1951) ein Schätzer betrachtet, der den DSE modifiziert. Im Folgenden wird die Hauptidee dieses *Chapman–Schätzers* auf die Situation in Deutschland übertragen.

Betrachten wir wieder eine Stichprobe S_P von Personen in der Gesamtpopulation U. Laut Chapman (1951) wird τ_{00,S_P} künstlich um eins erhöht ($\tau^*_{00,S_P} = \tau_{00,S_P} + 1$). Daher erhöhen sich auch die entsprechenden Randhäufigkeiten (s. Tabelle 3.8), wo die Bezeichnung * jeweils die ursprüngliche Anzahl erhöht um eins bedeutet.

	Y			
	F	00	K	
S_P	τ_{F,S_P}	τ^*_{00,S_P}	τ_{K,S_P}	n^*_P
$\bar{S}_P$	$\tau_{F,\bar{S}_P}$	$\tau_{00,\bar{S}_P}$	$\tau_{K,\bar{S}_P}$	$\bar{n}_P$
	τ_F	τ^*_{00}	τ_K	N^*_P

TABELLE 3.8: *2×3 Kontingenztabelle der Population U in einer Gemeinde für den Chapman–Schätzer.*

Die modifizierten Häufigkeiten und die Wahrscheinlichkeiten der Personen sind in den Tabellen 3.9 und 3.10 gegeben und es gilt

$$\theta' = \frac{n^*_P}{N^*_P} \quad . \tag{3.17}$$

	Y			
	F	00	K	
S_P	$\theta'\tau_F$	$\theta'\tau^*_{00}$	$\theta'\tau_K$	$\theta' N^*_P$
$\bar{S}_P$	$(1-\theta')\tau_F$	$(1-\theta')\tau^*_{00}$	$(1-\theta')\tau_K$	$(1-\theta')N^*_P$
	τ_F	τ^*_{00}	τ_K	N^*_P

TABELLE 3.9: *Die modifizierten Häufigkeiten in Listen R und Z für den Chapman–Schätzer .*

	Y			
	F	00	K	
S_P	p'_{F,S_P}	p'_{00,S_P}	p'_{K,S_P}	θ'
$\bar{S}_P$	$p'_{F,\bar{S}_P}$	$p'_{00,\bar{S}_P}$	$p'_{K,\bar{S}_P}$	$1-\theta'$
	p'_F	p'_{00}	p'_K	1

TABELLE 3.10: *Die modifizierten Wahrscheinlichkeiten für die Personen in den sechs möglichen Kategorien zu sein.*

Aus der ersten Zeile der Tabelle 3.8 ergibt sich der Maximum–Likelihood–Schätzer für p'_{F,S_P}, p'_{00,S_P} und p'_{K,S_P} als

$$\hat{p}'_{F,S_P} = \frac{\tau_{F,S_P}}{n^*_P} \qquad \hat{p}'_{00,S_P} = \frac{\tau^*_{00,S_P}}{n^*_P} \qquad \hat{p}'_{K,S_P} = \frac{\tau_{K,S_P}}{n^*_P} \quad . \tag{3.18}$$

Die Maximum–Likelihood–Schätzer der zweiten Zeile sind

$$\hat{p}'_{F,\bar{S}_P} = \frac{\tau_{F,\bar{S}_P}}{\bar{n}_P} \qquad \hat{p}'_{00,\bar{S}_P} = \frac{\tau_{00,\bar{S}_P}}{\bar{n}_P} \qquad \hat{p}'_{K,\bar{S}_P} = \frac{\tau_{K,\bar{S}_P}}{\bar{n}_P} \quad . \tag{3.19}$$

Modifizieren wir (3.8) und (3.9) mit θ' und τ^*_{00} ergibt sich

$$\begin{aligned} \tau_{F,S_P} &= \theta'\tau_F & \tau_{F,\bar{S}_P} &= (1-\theta')\tau_F \\ \tau^*_{00,S_P} &= \theta'\tau^*_{00} & \tau_{00,\bar{S}_P} &= (1-\theta')\tau^*_{00} \\ \tau_{K,S_P} &= \theta'\tau_K & \tau_{K,\bar{S}_P} &= (1-\theta')\tau_K \quad . \end{aligned} \tag{3.20}$$

Weiteres Einsetzen von (3.20) in (3.18) bzw. (3.19) ergibt

$$\hat{p}'_{F,S_P} = \frac{\tau_F}{N^*_P} \qquad \hat{p}'_{00,S_P} = \frac{\tau^*_{00}}{N^*_P} \qquad \hat{p}'_{K,S_P} = \frac{\tau_K}{N^*_P}$$

und

$$\hat{p}'_{F,\bar{S}_P} = \frac{\tau_F}{N^*_P} \qquad \hat{p}'_{00,\bar{S}_P} = \frac{\tau^*_{00}}{N^*_P} \qquad \hat{p}'_{K,\bar{S}_P} = \frac{\tau_K}{N^*_P} \quad .$$

Daher gilt

$$\begin{aligned} \hat{p}'_F &= \theta'\,\hat{p}'_{F,S_P} + (1-\theta')\,\hat{p}'_{F,\bar{S}_P} = \frac{\tau_F}{N^*_P} \\ \hat{p}'_{00} &= \theta'\,\hat{p}'_{00,S_P} + (1-\theta')\,\hat{p}'_{00,\bar{S}_P} = \frac{\tau^*_{00}}{N^*_P} \\ \hat{p}'_K &= \theta'\,\hat{p}'_{K,S_P} + (1-\theta')\,\hat{p}'_{K,\bar{S}_P} = \frac{\tau_K}{N^*_P} \quad . \end{aligned}$$

Für die Maximum–Likelihood–Schätzer der Wahrscheinlichkeiten p'_R und p'_Z gelten

$$\begin{aligned} \hat{p}'_Z &= \hat{p}'_{00} + \hat{p}'_F = \frac{\tau^*_{00} + \tau_F}{N^*_P} \\ \hat{p}'_R &= \hat{p}'_{00} + \hat{p}'_K = \frac{\tau^*_{00} + \tau_K}{N^*_P} \quad . \end{aligned} \tag{3.21}$$

Daher ist

$$\frac{\tau_{00}^* + \tau_F}{\hat{p}'_{00} + \hat{p}'_F} = \frac{\tau_{00}^* + \tau_K}{\hat{p}'_{00} + \hat{p}'_K}$$

und

$$\tau_Z^* = \tau_R^* \, \frac{\hat{p}'_{00} + \hat{p}'_F}{\hat{p}'_{00} + \hat{p}'_K} \qquad . \tag{3.22}$$

Für den letzten Bruch gilt wegen (3.21)

$$\begin{aligned} \hat{p}'_{00} + \hat{p}'_F &= \frac{\tau_{00}^* + \tau_F}{N_P^*} \, \frac{n_P^*}{n_P^*} = \frac{\tau_{00,S_P}^* + \tau_{F,S_P}}{n_P^*} \\ \hat{p}'_{00} + \hat{p}'_K &= \frac{\tau_{00}^* + \tau_K}{N_P^*} \, \frac{n_P^*}{n_P^*} = \frac{\tau_{00,S_P}^* + \tau_{K,S_P}}{n_P^*} \qquad . \end{aligned}$$

Dann schreiben wir (3.22) wie folgt:

$$\tau_Z^* = \tau_R^* \, \frac{\tau_{00,S_P}^* + \tau_{F,S_P}}{\tau_{00,S_P}^* + \tau_{K,S_P}} \qquad .$$

Die Bezeichnung * bedeutet jeweils die ursprüngliche Anzahl erhöht um eins und daher

$$\begin{aligned} \tau_Z &= (\tau_R + 1) \, \frac{(\tau_{00,S_P} + 1) + \tau_{F,S_P}}{(\tau_{00,S_P} + 1) + \tau_{K,S_P}} \\ &= (\tau_R + 1) \, \frac{\tau_{Z,S_P} + 1}{\tau_{R,S_P} + 1} - 1 \qquad . \end{aligned}$$

3.2.1 Chapman–Schätzer bei uneingeschränkter Zufallsauswahl der Klumpen

Wie im Kapitel 3.1.1 beschrieben, sind im Zensus 2011 die Auswahleinheiten die Anschriften einer Gemeinde, an einer Anschrift werden alle Perso-

nen erhoben. Die Anzahl τ_Z der tatsächlich vorhandenen Personen in einer Gemeinde wird dann durch

$$\hat{\tau}_Z^{\text{CHAP}} = (\tau_R + 1) \frac{\sum\limits_{a \in S} w_a \, \tau_{Z,a} + 1}{\sum\limits_{a \in S} w_a \, \tau_{R,a} + 1} - 1 = (\tau_R + 1) \frac{\widehat{\tau}_Z + 1}{\widehat{\tau}_R + 1} - 1 \qquad (3.23)$$

ermittelt, wobei $\tau_{Z,a}$ die Anzahl der tatsächlich vorhandenen Personen an der Anschrift a ist, $\tau_{R,a}$ die Anzahl der registrierten Personen an der Anschrift a ist. S ist die Stichprobe der Anschriften in einer Gemeinde, $w_a = \frac{1}{\pi_a}$ das Designgewicht mit Inklusionswahrscheinlichkeit $\pi_a, a = \{1, \ldots, N\}$.

3.2.2 Chapman–Schätzer bei geschichteter Zufallsauswahl der Klumpen

Beim Zensustest 2011 hat sich gezeigt, dass eine geschichtete Zufallsauswahl der Klumpen eine wesentliche Rolle spielt. In Analogie zum D–DSE bei geschichteter Zufallsauswahl der Klumpen (s. Kapitel 3.1.2) wird die Anzahl τ_Z der tatsächlich vorhandenen Personen in einer Gemeinde durch den Chapman–Schätzer

$$\begin{aligned} \hat{\tau}_Z^{\text{CHAP,str}} &= \sum_{h=1}^{H} \hat{\tau}_{Z,h}^{\text{CHAP}} = \sum_{h=1}^{H} \left[(\tau_{R,h} + 1) \frac{\sum\limits_{a \in S_h} w_{a,h} \, \tau_{Z,a,h} + 1}{\sum\limits_{a \in S_h} w_{a,h} \, \tau_{R,a,h} + 1} - 1 \right] \\ &= \sum_{h=1}^{H} \left[(\tau_{R,h} + 1) \frac{\widehat{\tau}_{Z,h} + 1}{\widehat{\tau}_{R,h} + 1} - 1 \right] \end{aligned} \qquad (3.24)$$

ermittelt. $\hat{\tau}_{Z,h}^{\text{CHAP}}$ ist Chapman–Schätzer berechnet nach (3.23) in einer Schicht h, S_h ist die Stichprobe in der Anschriftenschicht h, $\tau_{Z,a,h}$ ist die Anzahl der tatsächlich vorhandenen Personen an der Anschrift a in der

Schicht h, $\tau_{R,a,h}$ ist die Anzahl der registrierten Personen an der Anschrift a in der Schicht h.

3.2.3 Chapman–Schätzer in einer Klasse

Bei Dual–System–Modellen werden bestimmte Subpopulationen (Klassen) gebildet, in denen der Schätzer berechnet wird. In Analogie zum D–DSE in einer Klasse (s. Kapitel 3.1.3) ist der Chapman–Schätzer in einer Klasse k bei der uneingeschränkten Zufallsauswahl der Anschriften gegeben durch

$$\hat{\tau}_{Z,k}^{\text{CHAP}} = (\tau_{R,k} + 1)\,\frac{\widehat{\tau}_{Z,k} + 1}{\widehat{\tau}_{R,k} + 1} - 1 \qquad . \tag{3.25}$$

Für die Gesamtheit U dann gilt

$$\hat{\tau}_{Z}^{\text{CHAP}} = \sum_{k=1}^{K} \hat{\tau}_{Z,k}^{\text{CHAP}} = \sum_{k=1}^{K} \left[(\tau_{R,k} + 1)\,\frac{\widehat{\tau}_{Z,k} + 1}{\widehat{\tau}_{R,k} + 1} - 1 \right] \qquad . \tag{3.26}$$

Bei der geschichteten Zufallsauswahl der Anschriften ist der Chapman–Schätzer einer Klasse k

$$\hat{\tau}_{Z,k}^{\text{CHAP,str}} = \sum_{h=1}^{H} \hat{\tau}_{Z,k,h}^{\text{CHAP}} \qquad , \tag{3.27}$$

mit $\hat{\tau}_{Z,k,h}^{\text{CHAP}}$ als Chapman–Schätzer für $\tau_{Z,k,h}$ einer Klasse k in einer Anschriftenschicht h. Für die Gesamtheit U gilt

$$\hat{\tau}_{Z}^{\text{CHAP,str}} = \sum_{k=1}^{K} \sum_{h=1}^{H} \hat{\tau}_{Z,k,h}^{\text{CHAP}} \qquad . \tag{3.28}$$

Wie wir sehen können, sind sowohl Nenner als auch Zähler in (3.25) po-

sitiv, was die Situation vermeidet, in der kein Eintrag einer Klasse k im Register ist oder keine Person dieser Klasse k in der Gemeinde lebt.

4 Small–Area–Schätzung für Dual–System–Modelle

Das Dual–System–Modell im Kapitel 3 geht von gleichen Wahrscheinlichkeiten der Personen aus, in Liste R oder in Liste Z erfasst zu werden. Es werden Subpopulationen (Klassen) gebildet, die durch die Variablen, die die heterogenen Wahrscheinlichkeiten verursachen, definiert sind. Das Dual–System–Modell wird dann in diesen Klassen angewendet.

Einen besonderen Schwerpunkt in der amtlichen Statistik bildet die Schätzung von Subpopulationen (nennen wir sie *Domains*), die demographisch (z. B. Frauen im Alter von 17-20 Jahren) abgegrenzt sein können. Die interessierenden Subpopulationen können auch geographisch (z. B. ein Stadtteil) abgegrenzt sein, sie werden dann *Areas* genannt. Nach der Zerlegung der Stichprobe in die Domains kann es vorkommen, dass die Domains zu schwach besetzt sind, um eine valide Schätzung durchführen zu können. In der Literatur wird das Problem allgemein im Bereich der *Small–Area–Schätzung* behandelt.

Im Folgenden legen wir einen wesentlichen Akzent auf den Unterschied zwischen:

Klasse	eine Subpopulation, auf die Dual–System–Modelle angewendet werden, aber nicht unbedingt die interessierende Subpopulation ist.
Domain	eine Subpopulation, an deren geschätzten Umfang wir interessiert sind.

In diesem Kapitel werden strukturerhaltende– und verallgemeinerte–strukturerhaltende–Schätzer vorgestellt, die D–DSE und Chapman–Schätzer in den Klassen verwenden, um Schätzungen für die interessierenden Domains abzuleiten. Eine schematische Übersicht ist im Anhang F zu finden.

4.1 Log–lineare–Modelle für zweidimensionale Tabellen

Betrachten wir eine $K \times D$ Kontingenztabelle (vgl. Agresti 2002, S. 36) mit Poisson–verteilten Häufigkeiten. Die Poisson–Verteilung ist in Agresti (2002, S. 7) beschrieben. Es sei n_{P_1} die zufällige Gesamtzahl in der Tafel.

Eine Kontingenztabelle kann mit einem *log–linearen–Modell* untersucht werden, d.h. mit einem *verallgemeinerten–linearen–Modell* (engl. *generalized linear model*) mit der log–Link–Funktion. Eine Einführung in die verallgemeinerten–linearen–Modelle gibt zum Beispiel Agresti (2002, S. 116), die log–linearen–Modelle für die Kontingenztabellen sind ausführlich im Kapitel 8 (vgl. Agresti 2002, S. 314) beschrieben.

Es sei $\pi_{k,d,i}$ die Wahrscheinlichkeit für eine Person i, in der Kategorie (k, d) der Kontingenztabelle zu sein. Nehmen wir an, $\pi_{k,d,i} = \pi_{k,d}$ für $i = 1, \ldots, n_{P_1}$. Wenn die Häufigkeiten in den Zellen unabhängig voneinander sind, gilt für die erwarteten Häufigkeiten in den Zellen

$$\hat{\tau}_{R,k,d} = n_{P_1} \pi_{k,d} = n_{P_1} \pi_k \pi_d \qquad ,$$

mit $\sum_{k=1}^{K} \pi_k = \sum_{d=1}^{D} \pi_d = 1$. Es ist üblich, den Logarithmus zu betrachten. Daher ist

$$\begin{aligned} \log \hat{\tau}_{R,k,d} &= \log n_{P_1} + \log \pi_k + \log \pi_d \\ &= \beta^R + \beta_k^{R,K} + \beta_d^{R,D} \end{aligned} \tag{4.1}$$

und die Analogie zum linearen Modell wird deutlich.

Allgemein sind die Häufigkeiten in den Zellen einer Kontingenztabelle nicht unabhängig und in das Modell (4.1) muss ein Interaktionseffekt $\beta_{k,d}^{R,KD}$ eingeführt werden. Agresti (2002, S. 316) gibt ein *saturiertes Modell*

$$\log \tau_{R,k,d} = \beta^R + \beta_k^{R,K} + \beta_d^{R,D} + \beta_{k,d}^{R,KD} \tag{4.2}$$

an, mit den Bedingungen

$$\sum_{k=1}^{K} \beta_k^{R,K} = \sum_{d=1}^{D} \beta_d^{R,D} = \sum_{k=1}^{K} \beta_{k,d}^{R,KD} = \sum_{d=1}^{D} \beta_{k,d}^{R,KD} = 0 \qquad . \tag{4.3}$$

4.2 Structure Preserving Estimator

Purcell and Kish (1980) präsentieren eine „neue“ Methode, bei der die log–linearen–Modelle auf zwei Datenquellen angepasst sind, um Schätzungen in den interessierenden Domains zu berechnen. Damit ist eine Assoziation zwischen den beiden Daten ermöglicht.

Die ersten Daten werden mittels Variablen in die Zellen einer Kontigenztabelle geteilt. Dabei wird vorausgesetzt, dass diese Variablen einen Einfluss auf die spätere Schätzung der interessierenden Domains haben. Purcell and Kish (1980) nennen diese Variablen *associated variables* und die Kontingenztabelle des Zusammenhangs zwischen den Variablen als *association structure*. Im Forschungsprojekt für den Zensus 2011 stellen die ersten Daten das Register vor, die Registerdaten können mit dem Modell (4.2) und Bedingungen (4.3) beschrieben werden.

Zweitens stehen die Randhäufigkeiten der Klassen $\hat{\tau}_{Z,k}$, $k = 1, \ldots, K$ als die Summen über unbekannte aber interessierende Domains zur Verfügung (mit anderen Worten eine Klasse überdeckt mehrere Domains). Oft

sind diese Randhäufigkeiten mit späterem Zeitpunkt als das Register mittels einer Stichprobe gewonnen. Der Zusammenhang zwischen den Randhäufigkeiten und den interessierenden Domains ist in Purcell and Kish (1980) als *allocation structure* genannt. Dies kann wieder mit dem log–linearen–Modell

$$\log \tau_{Z,k,d} = \beta^{Z} + \beta_{k}^{Z,K} + \beta_{d}^{Z,D} + \beta_{k,d}^{Z,KD}$$

und mit den Bedingungen

$$\sum_{k=1}^{K} \beta_{k}^{Z,K} = \sum_{d=1}^{D} \beta_{d}^{Z,D} = \sum_{k=1}^{K} \beta_{k,d}^{Z,KD} = \sum_{d=1}^{D} \beta_{k,d}^{Z,KD} = 0$$

beschrieben werden.

Die Idee in Purcell and Kish (1980) ist dann, die Häufigkeiten in der Kontingenztabelle des Registers (association structure) anzupassen, um die aktuellen Randhäufigkeiten in der allocation structure festzuhalten. Dabei wird gefordert, dass die Interaktionseffekte in der association structure und in der allocation structure gleich sind, also $\beta_{k,d}^{Z,KD} = \beta_{k,d}^{R,KD}, \forall k, d$ gilt.

Es sind also die Schätzer der Domains gesucht, die den Abstand zwischen bekannten und gesuchten Werten der Domains unter den Nebenbedingungen, dass gewisse Randhäufigkeiten festgehalten werden, minimieren. Um dieses Problem zu lösen, bietet sich die *Lagrange–Methode* an (s. Fahrmeir and Hamerle 1984, S. 720). Mathematisch muss die *Lagrange–Funktion*

$$F(\tau_{Z,k,d}) = \sum_{k=1}^{K} \sum_{d=1}^{D} \frac{\left(\tau_{R,k,d} - \tau_{Z,k,d}\right)^2}{\tau_{R,k,d}} - \sum_{k=1}^{K} \lambda_k \left(\sum_{d=1}^{D} \tau_{Z,k,d} - \hat{\tau}_{Z,k} \right)$$

minimiert werden.

Nach Null setzen der Ableitung von $F(\tau_{Z,k,d})$ nach $\tau_{Z,k,d}$ ergibt sich

$$\hat{\tau}_{Z,k,d} = \frac{\tau_{R,k,d}}{\tau_{R,k}} \hat{\tau}_{Z,k} \qquad .$$

Summiert über die Klassen lässt sich als Schätzer der tatsächlich in Domain d lebenden Personen

$$\hat{\tau}_{Z,d}^{\text{SPREE}} = \sum_{k=1}^{K} \frac{\tau_{R,k,d}}{\tau_{R,k}} \hat{\tau}_{Z,k}^{\text{X}} \tag{4.4}$$

verwenden, mit $\tau_{R,k,d}$ als Schnittmenge $k \cap d$ des Registers R in Klasse k und des Registers R in Domain d. X in $\hat{\tau}_{Z,k}^{\text{X}}$ bezeichnet entweder D–DSE oder CHAP, d.h. $\hat{\tau}_{Z,k}^{\text{D-DSE}}$ bzw. $\hat{\tau}_{Z,k}^{\text{CHAP}}$. Es liegt in (4.4) ein strukturerhaltender–Schätzer (engl. *structure preserving estimator*, SPREE) vor, der auf voller Ausnutzung des zuverlässigen Schätzers der Klassen $\hat{\tau}_{Z,k}^{\text{X}}$ begründet ist (vgl. Rao 2003, S. 53).

Im einfachsten Fall, wo nur eine Klasse k mehrere Domains d, $d = 1, \ldots, D$ überdeckt, ist der Schätzer (4.4) einer Domain d von der Form

$$\hat{\tau}_{Z,d}^{\text{SPREE}} = \frac{\tau_{R,k,d}}{\tau_{R,k}} \hat{\tau}_{Z,k}^{\text{X}} \qquad , \tag{4.5}$$

wobei $\tau_{R,k,d}$ der Durchschnitt der Registeranzahl in Domain d mit der Registeranzahl in einer Klasse k ist. $\hat{\tau}_{Z,k}^{\text{X}}$ wird mittels Registeranzahl der Domains und Klassen proportional aufgeteilt, um die Schätzer $\hat{\tau}_{Z,d}^{\text{SPREE}}$ der Domains d, $d = 1, \ldots, D$ zu gewinnen.

Als Beispiel definieren wir die Klasse k als 18-95 jährige Männer, die mittels der Stichprobe mit dem Chapman–Schätzer bzw. D–DSE geschätzt wird. Wir gewinnen $\hat{\tau}_{Z,k}^{\text{D-DSE}}$ bzw. $\hat{\tau}_{Z,k}^{\text{CHAP}}$. Man ist an der Schätzung der Untergruppe der 25-29 jährigen Männer (Domain $\tau_{Z,d}$) interessiert. Dafür wird die Registeranzahl $\tau_{R,k,d}$ der 25-29 jährigen Männer sowie die Registeranzahl $\tau_{R,k}$ der 18-95 jährigen Männer benötigt, um den Chapman–Schätzer

bzw. D–DSE proportional aufzuteilen und $\hat{\tau}_{Z,d}^{\text{SPREE}}$ der 25-29 jährigen Männer (Domain $\tau_{Z,d}$) zu gewinnen.

Wie die Voruntersuchungen beim Zensustest 2001 (vgl. Statistisches Bundesamt 2004) gezeigt haben, ist eine geschichtete Zufallsauswahl der Klumpen besonderes interessant. Bei geschichteter Zufallsauswahl, wo nur eine Klasse mehrere Domains überdeckt, wird

$$\hat{\tau}_{Z,d}^{\text{SPREE,str}} = \frac{\tau_{R,k,d}}{\tau_{R,k}} \sum_{h=1}^{H} \hat{\tau}_{Z,k,h}^{\text{X}} \tag{4.6}$$

verwendet.

4.3 Generalized Structure Preserving Estimator

In Zhang and Chambers (2004) sowie Saei et al. (2005) werden log–lineare–Modelle betrachtet, die das SPREE–Modell im Kapitel 4.2 verallgemeinern. Um diese Modelle zu erklären und später auch zu programmieren, wird das log–lineare–Modell in Matrixalgebra geschrieben.

Nehmen wir an, wir haben die Registerdaten $\tau_{R,k,d}$ für $K \times D$ Domains zur Verfügung. Der Vektor der Registerdaten für die Domains, $\boldsymbol{\tau}_R = (\tau_{R,1,1}, \ldots, \tau_{R,1,D}, \ldots, \tau_{R,K,D})^T$, lässt sich mit dem *generalized linear model* (GLM)

$$\log \boldsymbol{\tau}_R = \mathbf{X}\boldsymbol{\beta}$$

analysieren, wo $\mathbf{X}$ die *Designmatrix* (siehe Andress et al. 1997, S. 167) und $\boldsymbol{\beta}$ der Vektor der unbekannten Parameter für das saturierte Modell sind.

Ein Beispiel für zwei Klassen und zwei Domains beschrieben mit dem saturierten Modell (4.2) wäre

$$\begin{pmatrix} \log \tau_{R,1,1} \\ \log \tau_{R,1,2} \\ \log \tau_{R,2,1} \\ \log \tau_{R,2,2} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix} \times \begin{pmatrix} \beta^R \\ \beta_1^{R,K} \\ \beta_2^{R,K} \\ \beta_1^{R,D} \\ \beta_2^{R,D} \\ \beta_{1,1}^{R,KD} \\ \beta_{1,2}^{R,KD} \\ \beta_{2,1}^{R,KD} \\ \beta_{2,2}^{R,KD} \end{pmatrix}$$

mit den Nebenbedingungen

$$\sum_{k=1}^{2} \beta_k^{R,K} = \sum_{d=1}^{2} \beta_d^{R,D} = \sum_{k=1}^{2} \beta_{k,d}^{R,KD} = \sum_{d=1}^{2} \beta_{k,d}^{R,KD} = 0 \qquad .$$

Man kann $\mathbf{X}\boldsymbol{\beta}$ in der Form

$$\mathbf{X}\boldsymbol{\beta} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2$$

schreiben, wo $\boldsymbol{\beta}_1$ die Koeffizienten für die Haupteffekte und $\boldsymbol{\beta}_2$ für die Interaktionseffekte sind. In unserem Beispiel sind $\beta^R, \beta_1^{R,K}, \beta_2^{R,K}, \beta_1^{R,D}, \beta_2^{R,D}$ die Haupteffekte und die entsprechende Matrix $\mathbf{X}_1$ wird durch die ersten fünf Spalten der Matrix $\mathbf{X}$ gebildet. $\mathbf{X}_2$ bzw. $\boldsymbol{\beta}_2$ ist der Rest der Matrix $\mathbf{X}$ bzw. des Vektors $\boldsymbol{\beta}$.

Das saturierte Modell für die Registerdaten liefert $\hat{\boldsymbol{\beta}}_1 = \boldsymbol{\beta}_1$ und $\hat{\boldsymbol{\beta}}_2 = \boldsymbol{\beta}_2$. Damit ist es möglich $\mathbf{X}_2\hat{\boldsymbol{\beta}}_2$ zu ermitteln.

Außer den Registerdaten haben wir auch die sogenannten Pseudo–Werte, $\hat{\tau}_{\mathrm{PW},k,d}$ für die gleichen $K \times D$ Domains zur Verfügung, die aus der Stichprobe berechnet werden. Der Vektor dieser Pseudo–Werte, $\hat{\boldsymbol{\tau}}_{\mathrm{PW}} = (\hat{\tau}_{\mathrm{PW},1,1}, \ldots, \hat{\tau}_{\mathrm{PW},K,D})^T$ lässt sich wieder mit GLM als

$$\log \hat{\boldsymbol{\tau}}_{\mathrm{PW}} = \mathbf{X}\boldsymbol{\beta}^{'} = \mathbf{X}_1\boldsymbol{\beta}_3 + \mathbf{X}_2\boldsymbol{\beta}_4$$

modellieren, wobei $\boldsymbol{\beta}_3$ und $\boldsymbol{\beta}_4 = \delta\ \boldsymbol{\beta}_2$ die Vektoren der unbekannten Parameter für dieses Modell sind. Wieder enthält $\boldsymbol{\beta}_3$ die Koeffizienten für die Haupteffekte und $\boldsymbol{\beta}_4$ die Koeffizienten für die Interaktionseffekte dieses zweiten Modells. Es sei angemerkt, dass die (geschätzten) Interaktionseffekte $\hat{\boldsymbol{\beta}}_2$ aus dem ersten Modell mit einer unbekannten Konstanten δ multipliziert werden, um zeitliche oder andere Veränderungen zwischen der Erhebung der Registerdaten und der Zensusstichprobe zu berücksichtigen. Im SPREE–Modell im Kapitel 4.2 sind demgegenüber die konstanten Interaktionseffekte ($\delta = 1$) betrachtet.

Das Modell für die Pseudo–Werte ist demnach

$$\log \hat{\boldsymbol{\tau}}_{\mathrm{PW}} = \mathbf{X}_1\boldsymbol{\beta}_3 + \delta\mathbf{X}_2\hat{\boldsymbol{\beta}}_2 \qquad .$$

Aus diesem Modell erhalten wir die Werte $\hat{\boldsymbol{\beta}}_3$ und $\hat{\delta}$.

Um den verallgemeinerten–strukturerhaltenden–Schätzer $\hat{\boldsymbol{\tau}}_Z^{\mathrm{GSPREE}}$ (engl. *generalized structure preserving estimator*, GSPREE) für alle Domains zu berechnen, wird

$$\log \hat{\boldsymbol{\tau}}_Z^{\mathrm{GSPREE}} = \mathbf{X}_1\hat{\boldsymbol{\beta}}_3 + \hat{\delta}\mathbf{X}_2\hat{\boldsymbol{\beta}}_2 \qquad . \tag{4.7}$$

verwendet.

Die Pseudo–Werte für die Domains können unterschiedlich aus der Stichprobe gewonnen werden. Für die in der Simulationsstudie verwendeten Schätzwerte und den Zusammenhang mit dem Zensus verweisen wir auf Kapitel 7.

5 Small–Area–Schätzung für Alternative Modelle

Im Kapitel 3 wurden Klassen gebildet, in denen die Dual–System–Modelle angewendet wurden, um die Schätzungen für die Klassen zu gewinnen. Demgegenüber wurden im Kapitel 4 Modelle für die Schätzung der interessierenden Subpopulationen (Domains) vorgestellt, die D–DSE und Chapman–Schätzer der Klassen verwenden.

In diesem Kapitel werden alternative Modelle für die Schätzung in Domains präsentiert, wie sie in Münnich et al. (2008) zu finden sind. Bei jedem Modell wird nachfolgend auch die Variante des Schätzers für die geschichteten Zufallsauswahl vorgestellt. Damit wird es möglich, die Modelle im Kapitel 3 und Kapitel 4 mit den alternativen Modellen in diesem Kapitel in unserem konkreten Anwendungsfall zu vergleichen. Eine Übersicht über die alternativen Schätzer ist in Tabelle 5.1 gegeben.

5.1 Verallgemeinerte Regressionsmodelle

Nehmen wir an, es ist für jede Anschrift a, $a \in \{1, \ldots, N\}$ ein Vektor $\boldsymbol{x}_a = (x_{1,a}, \ldots, x_{v,a})^T$ der v Hilfsvariablen bekannt. Neben der Stichprobe S beobachten wir zusätzlich für jede Anschrift a, $a \in S$ auch die Anzahl der tatsächlich vorhandenen Personen an der Anschrift a, $\tau_{Z,a}$.

Name des Schätzers	Beschreibung
GREG1	verallgemeinerter Regressionsschätzer mit Berücksichtigung der Populationsanzahl des Registers für die Gemeinde
GREG2	verallgemeinerter Regressionsschätzer mit Berücksichtigung der Populationsanzahl des Registers für die Domain der Gemeinde
SYN	Verhältnis–synthetischer Schätzer
EBLUP	Schätzer basiert auf dem Unit–level Modell

TABELLE 5.1: *Bezeichnung der verschiedenen verwendeten alternativen Schätzer.*

Bezeichnen wir mit $\boldsymbol{\tau}_Z = (\tau_{Z,1}, \ldots, \tau_{Z,a}, \ldots, \tau_{Z,n})^T$ den Vektor der beobachteten Anzahl der tatsächlich vorhandenen Personen an der Anschrift a, $a \in S$. Um Doppelindizes zu vermeiden wird im Folgenden $(\tau_{Z,1}, \ldots, \tau_{Z,n})^T$ statt $(\tau_{Z,a_1}, \ldots, \tau_{Z,a_n})^T$, $S = \{a_1, \ldots, a_n\}$ geschrieben.

Der Vektor $\boldsymbol{\tau}_Z$ kann in der Matrixschreibweise mit einem *Regressionsmodell* (vgl. Särndal et al. 1992, S. 226)

$$\boldsymbol{\tau}_Z = \mathbf{XB} + \boldsymbol{\varepsilon} \tag{5.1}$$

dargestellt werden, wobei $\mathbf{X}$ eine $n \times v$ Matrix ist. Eine Zeile der Matrix $\mathbf{X}$ ist ein Vektor der bekannten v Hilfsvariablen für die Anschrift a in der Form $\boldsymbol{x}_a = (x_{1,a}, \ldots, x_{v,a})^T$. Weiter ist in (5.1) $\mathbf{B} = (B_1, \ldots, B_v)^T$ ein $v \times 1$ Vektor der Regressionskoeffizienten und $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_a, \ldots, \varepsilon_n)$ ein Fehlerterm. Es wird angenommen, $\varepsilon_a, a \in S$ sind unabhängig normalverteilt mit $E(\varepsilon_a) = 0$ und $\mathrm{Var}(\varepsilon_a) = \sigma_a^2$.

Die Hilfsvariablen haben einen Einfluss auf die untersuchten Variablen, der durch ein lineares Modell näherungsweise beschrieben werden kann.

Üblicherweise werden daraus folgende Schätzer abgeleitet (vgl. Särndal et al. 1992, S. 222).

Der *Differenzen–Schätzer*

$$\hat{\tau}_Z^{\text{Diff}} = \sum_{a=1}^{N} \boldsymbol{x}_a^T \mathbf{A} + \sum_{a \in S} (\hat{\tau}_{Z,a} - \hat{\boldsymbol{x}}_a^T \mathbf{A}) \qquad , \tag{5.2}$$

wobei $\mathbf{A}$ als bekannt vorausgesetzt wird und $\hat{\tau}_{Z,a} = w_a \tau_{Z,a}$ ist. Analog ist $\hat{\boldsymbol{x}}_a = w_a \boldsymbol{x}_a$ als gewichteter Vektor der Hilfsvariablen der Anschrift a zu verstehen. Man kann den Differenzen–Schätzer als

$$\hat{\tau}_Z^{\text{Diff}} = \hat{\tau}_Z + (\boldsymbol{t}_x - \hat{\boldsymbol{t}}_x)^T \mathbf{A} \tag{5.3}$$

schreiben, wobei $\hat{\tau}_Z = \sum_{a \in S} w_a \tau_{Z,a}$, $\boldsymbol{t}_x = \sum_{a=1}^{N} \boldsymbol{x}_a$ ein bekannter Vektor der Totalwerte der v Hilfsvariable ist und $\hat{\boldsymbol{t}}_x = \sum_{a \in S} w_a \boldsymbol{x}_a$ ein Vektor der geschätzten Totalwerte der v Hilfsvariablen ist.

Wenn die unbekannten Koeffizienten $A_1, \ldots, A_v$ geschätzt werden, werden sie in der Literatur häufig mit $\hat{B}_1, \ldots, \hat{B}_v$ als mit $\hat{A}_1, \ldots, \hat{A}_v$ bezeichnet. Der *verallgemeinerte–Regressionsschätzer*, der sogenannte GREG–Schätzer (engl. *generalized regression estimator*, Lohr 1999, S. 373), wird dargestellt durch

$$\hat{\tau}_Z^{\text{GREG}} = \hat{\tau}_Z + (\boldsymbol{t}_x - \hat{\boldsymbol{t}}_x)^T \, \hat{\mathbf{B}} \qquad , \tag{5.4}$$

mit dem $n \times 1$ Vektor der geschätzten Regressionskoeffizienten

$$\hat{\mathbf{B}} = (\mathbf{X}^T \mathbf{W} \, \boldsymbol{\Sigma}^{-1} \mathbf{X})^{-1} \, \mathbf{X}^T \mathbf{W} \, \boldsymbol{\Sigma}^{-1} \, \boldsymbol{\tau}_Z \qquad . \tag{5.5}$$

Dabei ist
$\mathbf{X}$ die $n \times v$ Matrix mit n Werten der v Hilfsvariablen in den Spalten 1 bis v,
$\mathbf{W}$die $n \times n$ Diagonalmatrix der Designgewichte w_a, $a \in S$,

$\boldsymbol{\Sigma}$ die $n \times n$ Diagonalmatrix mit σ_a^2,
$\boldsymbol{\tau}_Z = (\tau_{Z,1}, \ldots, \tau_{Z,n})^T$ der $n \times 1$ Vektor der beobachteten Werte in der Stichprobe S.

Beim verallgemeinerten–Regressionsschätzer handelt es sich um ein *modellunterstütztes* Verfahren, da die designbasierte Schätzung $\hat{\tau}_Z$ mit Hilfe eines Regressionsmodells korrigiert wird. In die designbasierte Schätzung gehen die Inklusionswahrscheinlichkeiten ein, mit denen die Elemente der Gesamtheit bei gegebenem Auswahlverfahren in die Stichprobe gelangen. Im Gegensatz dazu hängt ein *modellbasierter* Schätzer nur vom Modell ab (z. B. EBLUP)

Definieren wir für den Zensus 2011 in Deutschland ein Modell, wo als einzige Hilfsvariable die bekannte Populationsanzahl des Registers zur Verfügung steht und ε_a unabhängig und normalverteilt ist, mit $E(\varepsilon_a) = 0$, $\text{Var}(\varepsilon_a) = \sigma^2 \tau_{R,a}$ (vgl. Särndal et al. 1992, S. 226, Modell (6.4.5)).

Genauer ist $\boldsymbol{x}_a = \tau_{R,a}$, $\boldsymbol{t}_x = \sum_{a=1}^{N} \tau_{R,a} = \tau_R$ und $\hat{\boldsymbol{t}}_x = \sum_{a \in S} w_a \tau_{R,a} = \hat{\tau}_R$, wobei $\tau_{R,a}$ die Anzahl der registrierten Personen an der Anschrift a ist. Weiter ist $\mathbf{X} = (\tau_{R,1}, \ldots, \tau_{R,n})^T$,

$$\mathbf{W} = \begin{pmatrix} w_1 & 0 & \ldots & 0 \\ 0 & w_2 & \ddots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \ldots & w_n \end{pmatrix} \qquad \boldsymbol{\Sigma}^{-1} = \frac{1}{\sigma^2} \begin{pmatrix} \tau_{R,1}^{-1} & 0 & \ldots & 0 \\ 0 & \tau_{R,2}^{-1} & \ddots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \ldots & \tau_{R,n}^{-1} \end{pmatrix}$$

und $\boldsymbol{\tau}_Z = (\tau_{Z,1}, \ldots, \tau_{Z,n})^T$, wobei $\tau_{Z,a}$ die Anzahl der tatsächlich vorhandenen Personen an der Anschrift a ist.

Für $\hat{\mathbf{B}}$ gilt

$$\hat{\mathbf{B}} = \frac{\sum\limits_{a \in S} w_a \tau_{Z,a}}{\sum\limits_{a \in S} w_a \tau_{R,a}} = \frac{\hat{\tau}_Z}{\hat{\tau}_R} \tag{5.6}$$

und daher

$$\hat{\tau}_Z^{\text{GREG0}} = \frac{\hat{\tau}_Z}{\hat{\tau}_R}\, \tau_R \qquad . \tag{5.7}$$

Wir erhalten als Spezialfall für den in (5.4) definierten GREG–Schätzer den bekannten *Verhältnisschätzer* (engl. *ratio estimator*).

5.1.1 Zerlegung der Population in Domains

Es sei die Gesamtheit U in D Domains $U_1, \ldots, U_d, \ldots, U_D$ zerlegt. Es gilt

$$U = \bigcup_{d=1}^{D} U_d \qquad U_d \cap U_d' = \emptyset \quad \text{für } d \neq d'$$

$$N_P = \sum_{d=1}^{D} N_{P,d} \qquad ,$$

wo U_d definiert ist

$U_d = \{i \in U : \text{Person } i \text{ gehört zur Domain } d\} \subseteq U \qquad \#U_d = N_{P,d} \qquad .$

Wir sind an der Schätzung $\tau_{Z,d}$ der tatsächlich lebenden Personen in einer Domain d interessiert.

5.1.2 Schätzer in einer Domain

Es ist möglich, den GREG–Schätzer für die Schätzung in einer Domain zu verwenden. Im Fall der Berücksichtigung der Populationsanzahl des Registers für die Gesamtheit hat (5.4) die Form

$$\hat{\tau}_{Z,d}^{\text{GREG1}} = \hat{\tau}_{Z,d} + (\tau_R - \hat{\tau}_R)^T \, \hat{\mathbf{B}}_d \tag{5.8}$$

mit

$$\hat{\mathbf{B}}_d = (\mathbf{X}^T \mathbf{W} \, \boldsymbol{\Sigma}^{-1} \mathbf{X})^{-1} \, \mathbf{X}^T \mathbf{W} \, \boldsymbol{\Sigma}^{-1} \, \boldsymbol{\tau}_{Z,d} \qquad ,$$

wobei $\boldsymbol{\tau}_{Z,d} = (\tau_{Z,d,1}, \ldots, \tau_{Z,d,n})^T$ und $\tau_{Z,d,a}$ die Anzahl der tatsächlich vorhandenen Personen der Domain d an der Anschrift a, $a \in S$ ist.

Dann ist

$$\hat{\mathbf{B}}_d = \frac{\sum\limits_{a \in S} w_a \tau_{Z,d,a}}{\sum\limits_{a \in S} w_a \tau_{R,a}} = \frac{\hat{\tau}_{Z,d}}{\hat{\tau}_R}$$

und der GREG1–Schätzer (vgl. Rao 2003, S. 17)

$$\hat{\tau}_{Z,d}^{\text{GREG1}} = \frac{\hat{\tau}_{Z,d}}{\hat{\tau}_R} \, \tau_R \qquad . \tag{5.9}$$

Bei geschichteter Zufallsauswahl der Anschriften nimmt (5.9) die Form

$$\hat{\tau}_{Z,d}^{\text{GREG1,str}} = \sum_{h=1}^{H} \frac{\hat{\tau}_{Z,d,h}}{\hat{\tau}_{R,h}} \, \tau_{R,h} \qquad , \tag{5.10}$$

mit

$$\begin{aligned} \hat{\tau}_{Z,d,h} &= \sum_{a \in S_h} w_a \tau_{Z,d,a} \\ \hat{\tau}_{R,h} &= \sum_{a \in S_h} w_a \tau_{R,a} \qquad . \end{aligned}$$

$\tau_{R,h}$ ist die Anzahl der registrierten Personen in der Schicht h und S_h bezeichnet die Stichprobe in der Schicht h.

Falls wir die Populationsanzahl des Registers für jede Domain kennen, kann separat in jeder Domain ein Regressionsmodell betrachtet werden, das nicht von den Werten aus anderen Domains abhängt. Wir können (5.4) als

$$\hat{\tau}_{Z,d}^{\text{GREG2}} = \hat{\tau}_{Z,d} + (\tau_{R,d} - \hat{\tau}_{R,d})^T \, \hat{\mathbf{B}}_{d,\text{sep}} \tag{5.11}$$

schreiben, mit

$$\hat{\mathbf{B}}_{d,\text{sep}} = (\mathbf{X}_d^T \mathbf{W} \, \boldsymbol{\Sigma}^{-1} \mathbf{X}_d)^{-1} \, \mathbf{X}_d^T \mathbf{W} \, \boldsymbol{\Sigma}^{-1} \, \boldsymbol{\tau}_{Z,d} \qquad ,$$

wobei

$$\hat{\tau}_{R,d} = \sum_{a \in S} w_a \tau_{R,d,a} \qquad .$$

$\tau_{R,d,a}$ ist die Anzahl der registrierten Personen der Domain d an der Anschrift a. Weiter ist $\tau_{R,d}$ die Anzahl der registrierten Personen der Domain d und $\mathbf{X}_d = (\tau_{R,d,1}, \ldots, \tau_{R,d,n})^T$.

Nach der Einsetzung $\hat{\mathbf{B}}_{d,\text{sep}}$ in (5.11) benutzen wir für die Schätzung in einer Domain d den GREG2–Schätzer (vgl. Rao 2003, S. 18)

$$\hat{\tau}_{Z,d}^{\text{GREG2}} = \frac{\hat{\tau}_{Z,d}}{\hat{\tau}_{R,d}} \, \tau_{R,d} \qquad . \tag{5.12}$$

Der Schätzer in (5.12) ist der Verhältnisschätzer in einer Domain d. Es ist interessant zu erwähnen, dass D–DSE (3.15) in einer Klasse k genau die gleiche Form wie der GREG2–Schätzer (5.12) in einer Domain d hat.

Bei geschichteter Zufallsauswahl der Anschriften nimmt (5.12) die Form

$$\hat{\tau}_{Z,d}^{\text{GREG2,str}} = \sum_{h=1}^{H} \frac{\hat{\tau}_{Z,d,h}}{\hat{\tau}_{R,d,h}} \, \tau_{R,d,h} \tag{5.13}$$

an, wobei

$$\hat{\tau}_{R,d,h} \quad = \quad \sum_{a \in S_h} w_a \tau_{R,d,a} \tag{5.14}$$

und $\tau_{R,d,h}$ die Anzahl der registrierten Personen der Domain d in der Schicht h ist.

5.2 Regressions–synthetischer Schätzer

Alternativ kann man den verallgemeinerten–Regressionsschätzer (5.4) schreiben als

$$\hat{\tau}_Z^{\text{GREG}} = \boldsymbol{t}_x^T \, \hat{\mathbf{B}} + (\hat{\tau}_Z - \hat{\boldsymbol{t}}_x^T \, \hat{\mathbf{B}}) \qquad . \tag{5.15}$$

Falls die Hilfsvariablen für eine Domain zur Verfügung stehen, kann der erste Term als *Regressions–synthetischer Schätzer* (engl. *regression synthetic estimator*) für die Schätzung in einer Domain d verwendet werden (vgl. Rao 2003, S. 46)

$$\hat{\tau}_{Z,d}^{\text{REG-SYN}} = \boldsymbol{t}_{x,d}^T \, \hat{\mathbf{B}} \qquad , \tag{5.16}$$

wobei $\hat{\mathbf{B}}$ in (5.5) definiert ist und $\boldsymbol{t}_{x,d}$ der Vektor der Totalwerte der v Hilfsvariablen in der Domain d ist. $\hat{\mathbf{B}}$ ist der Vektor der geschätzten Regressionskoeffizienten für die Gesamtheit.

Im Fall einer Hilfsvariable in Form der bekannten Populationsanzahl des Registers der Domain ist $\boldsymbol{t}_{x,d} = \tau_{R,d}$. Nach der Einsetzung (5.6) in (5.16) bekommen wir den *Verhältnis–synthetischen Schätzer* (engl. *ratio synthetic estimator*, Rao 2003, S. 47)

$$\hat{\tau}_{Z,d}^{\text{SYN}} = \frac{\hat{\tau}_Z}{\hat{\tau}_R} \, \tau_{R,d} \qquad . \tag{5.17}$$

Bei geschichteter Zufallsauswahl der Anschriften wird

$$\hat{\tau}_{Z,d}^{\text{SYN,str}} = \sum_{h=1}^{H} \frac{\hat{\tau}_{Z,h}}{\hat{\tau}_{R,h}} \tau_{R,d,h}$$

verwendet, wobei

$$\hat{\tau}_{Z,h} = \sum_{a \in S_h} w_a \tau_{Z,a} \qquad .$$

5.3 Linear Mixed Model

Bisher haben wir nur Modelle betrachtet, bei denen die folgenden Annahmen über die Verteilung der Zufallsvariablen ε_a gelten: Unabhängigkeit, $E(\varepsilon_a) = 0$, $\text{Var}(\varepsilon_a) = \sigma^2 \tau_{R,a}$ (siehe Seite 58). Durch das Modell werden die unbekannten Koeffizienten, oft auch *fixed effects* benannt, geschätzt.

Lineare–gemischte–Modelle (engl. *linear mixed models*) erweitern die linearen–Modelle und ermöglichen, flexiblere Annahmen für das Modell in Betracht zu ziehen. Vor allem finden die linearen–gemischten–Modelle ihre Anwendung, wenn in der Population Gruppenzugehörigkeit einen Einfluß auf die Beobachtungen hat. Die Gruppen sind auch als *random effect* bezeichnet.

Die linearen–gemischten–Modelle sind ausführlich in Faraway (2006) oder Rao (2003) beschrieben. Rao (2003, S. 107) gibt auch eine spezielle Form des Modells für die Schätzung in Domains.

Betrachten wir das *Unit–level Modell* (in Battese et al. 1988 oder Datta and Lahiri 2000 auch *nested error regression model* genannt, s. auch Rao 2003, S. 79)

$$\mathbf{y}_d = \mathbf{X}_{P,d}\mathbf{B} + \alpha_d \mathbf{1}_{n_{P,d}} + \boldsymbol{\varepsilon}_d \qquad d = 1, \ldots, D \qquad , \tag{5.18}$$

womit die Beobachtungen $\mathbf{y}_d = (y_{d,1}, \ldots, y_{d,i}, \ldots, y_{d,n_{P,d}})^T$ mit fixed effects $\mathbf{B}$ und random effect α_d der Domain d beschrieben werden. Es sei $\mathbf{X}_{P,d}$ eine $n_{P,d} \times v$ Matrix der v Hilfsvariablen der Domain d, mit $n_{P,d}$ als Anzahl der ausgewählten Personen in der Domain d. Vektor $\boldsymbol{\varepsilon}_d$ sei ein Fehlerterm, α_d von $\boldsymbol{\varepsilon}_d$, $d = 1, \ldots, D$ sind unabhängig.

Im Forschungsprojekt für den Zensus 2011 werden Anschriften gezogen, deren Merkmalseigenschaften auf Personenebene aggregiert werden. Interpretieren wir $y_{d,i}$ auf der Personenebene als

$$y_{d,i} = \begin{cases} 1 & \text{für die tatsächlich lebende Person } i \text{ in der Domain } d \\ 0 & \text{sonst.} \end{cases}$$

Wir sind an einer Schätzung der tatsächlich lebenden Personen $\tau_{Z,d} = \sum_{i \in U_d} y_{d,i}$ als die lineare Kombination der fixed effects $\mathbf{B}$ und des random effect α_d

$$\tau_{Z,d} = \boldsymbol{t}_{x,d}^T \mathbf{B} + \alpha_d \tag{5.19}$$

interessiert.

Nehmen wir an, α_d, $\boldsymbol{\varepsilon}_d$ sind im Modell (5.18) normalverteilt mit $E(\alpha_d) = 0$, $E(\boldsymbol{\varepsilon}_d) = \mathbf{0}$, $\mathbf{G}_d = \text{Var}(\alpha_d) = \sigma_\alpha^2$ und $\mathbf{R}_d = \sigma_\varepsilon^2 \mathbf{I}_{n_{P,d}}$, wobei $\mathbf{I}_{n_{P,d}}$ die $n_{P,d} \times n_{P,d}$ Einheitsmatrix ist. Laut Rao (2003, S. 96) ist die Varianz des Modells (5.18)

$$\text{Var}(\mathbf{y}_d) = \mathbf{V}_d = \text{Var}(\alpha_d \mathbf{1}_{n_{P,d}}) + \text{Var}(\boldsymbol{\varepsilon}_d) = \sigma_\alpha^2 \mathbf{1}_{n_{P,d}} \mathbf{1}_{n_{P,d}}^T + \sigma_\varepsilon^2 \mathbf{I}_{n_{P,d}} \quad .$$

Bei bekannten Varianzen σ_α^2 und σ_ε^2 ist BLUE (engl. *best linear unbiased estimator*) für die fixed effects $\mathbf{B}$ im Modell (5.18) gegeben durch

$$\dot{\mathbf{B}} = \left(\sum_d \mathbf{X}_d^T \mathbf{V}_d^{-1} \mathbf{X}_d \right)^{-1} \left(\sum_d \mathbf{X}_d^T \mathbf{V}_d^{-1} \mathbf{y}_d \right) \quad ,$$

wobei $\left(\sum_d \mathbf{X}_d^T \mathbf{V}_d^{-1} \mathbf{X}_d \right)^{-1}$ die Varianz–Kovarianz Matrix von $\dot{\mathbf{B}}$ ist (vgl. Rao 2003, S. 137).

Nehmen wir an, es steht im Zensus 2011 in Deutschland als einzige Hilfsvariable die Information im Register zur Verfügung. Bei der bekannten Anzahl N_d ist dann die Zahl der tatsächlich lebenden Personen in einer Domain d (vgl. Rao 2003, S. 136)

$$\dot{\tau}_{Z,d} = \tau_{R,d} \dot{\mathbf{B}} + \gamma_d \left(\tau_{Z,S_P,d} - \tau_{R,S_P,d} \dot{\mathbf{B}} \right) \quad , \tag{5.20}$$

mit

$$\gamma_d = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + {\sigma_\varepsilon^2}/{n_{P,d}}} \quad ,$$

$\tau_{R,d} = \sum_{i \in U_d} x_{d,i}$, $\tau_{Z,S_P,d} = \sum_{i \in S_P} y_{d,i}$ und $\tau_{R,S_P,d} = \sum_{i \in S_P} x_{d,i}$. Es bezeichnet S_P die mittels Anschriftenstichprobe S gezogenen Personen. $x_{d,i}$ wird im Forschungsprojekt für den Zensus 2011 interpretiert als

$$x_{d,i} = \begin{cases} 1 & \text{für die registrierte Person } i \text{ der Domain } d \\ 0 & \text{sonst.} \end{cases}$$

$\dot{\tau}_{Z,d}$ in (5.20) ist von den in der Regel unbekannten Parametern σ_α^2 und σ_ε^2 abhängig. Nach Einsetzen der geschätzten Werte $\hat{\sigma}_\alpha^2$ und $\hat{\sigma}_\varepsilon^2$ in γ_d und $\mathbf{V}_d$ bekommen wir den *estimated best linear unbiased predictor* EBLUP (vgl. Rao 2003, S. 137).

Der EBLUP wird geschrieben

$$\hat{\tau}_{Z,d}^{\text{EBLUP}} = \hat{\gamma}_d \left[\tau_{Z,S_P,d} + (\tau_{R,d} - \tau_{R,S_P,d})\,\check{\mathbf{B}}\right] + (1 - \hat{\gamma}_d)\,\tau_{R,d}\check{\mathbf{B}} \qquad ,$$

mit

$$\check{\mathbf{B}} = \left(\sum_d \mathbf{X}_d^T \widehat{\mathbf{V}}_d^{-1} \mathbf{X}_d\right)^{-1} \left(\sum_d \mathbf{X}_d^T \widehat{\mathbf{V}}_d^{-1} \mathbf{y}_d\right)$$

$$\widehat{\mathbf{V}}_d = \hat{\sigma}_\alpha^2 \mathbf{1}_{n_{P,d}} \mathbf{1}_{n_{P,d}}^T + \hat{\sigma}_\varepsilon^2 \mathbf{I}_{n_{P,d}}$$

$$\hat{\gamma}_d = \frac{\hat{\sigma}_\alpha^2}{\hat{\sigma}_\alpha^2 + \hat{\sigma}_\varepsilon^2 / n_{P,d}} \qquad . \tag{5.21}$$

Die Schätzung der Varianzen σ_α^2 und σ_ε^2 erfolgt im Allgemeinen durch Maximum Likelihood (ML) oder Restricted Maximum Likelihood (REML) Methoden, wie sie in der Literatur vorgeschlagen werden (vgl. Goldstein 2003 und Münnich et al. 2012).

6 Varianz, Schätzung der Varianz bzw. des MSE

In der EU–Verordnung 763/2008 (The European Parliament and the Council 2008) ist für die Durchführung von Zensen auch eine Qualitätsmessung vorgehen, welche auch die Angaben der Genauigkeit der Schätzwerte umfasst. Ein Maß für die Güte des Schätzers basiert auf der mittleren quadratischen Abweichung des Schätzers (engl. *mean square error*, MSE; siehe Wolter 1986, S. 227 oder Särndal et al. 1992, S. 40). Dieser MSE ist durch

$$\begin{aligned} \text{MSE}(\hat{\tau}) &= E\left[(\hat{\tau}-\tau)^2\right] \\ &= \text{Var}(\hat{\tau}) + [\text{Bias}(\hat{\tau})]^2 \end{aligned} \tag{6.1}$$

definiert, also als die Summe von der *Varianz* $\text{Var}(\hat{\tau})$ und der quadrierten *Verzerrung* $\text{Bias}(\hat{\tau})$, die durch

$$\text{Bias}(\hat{\tau}) = E(\hat{\tau}) - \tau \tag{6.2}$$

definiert ist.

Ein Schätzer $\hat{\tau}$ ist *unverzerrt* für τ, wenn

$$\text{Bias}(\hat{\tau}) = 0 \qquad . \tag{6.3}$$

Für solche Schätzer gilt offensichtlich $\text{MSE}(\hat{\tau}) = \text{Var}(\hat{\tau})$.

Im Folgenden wird bei einem unverzerrten Schätzer die Varianz $\mathrm{Var}(\hat{\tau})$ des Schätzers $\hat{\tau}$ behandelt. Anderseits wird bei verzerrten Schätzern (wie z. B. EBLUP) der MSE des Schätzers verwendet.

Bei einer einfachen Zufallsauswahl und für einfache Schätzfunktionen lässt sich die Varianz durch die Stichprobenvarianz schätzen. Bei komplexen Stichprobenverfahren sind andere Varianzschätzer anzuwenden. Um die Genauigkeit von sowohl D–DSE, Chapman–Schätzer, SPREE, GSPREE als auch den alternativen Schätzern zu messen, werden im Folgenden verschiedene Varianzschätzmethoden vorgestellt.

6.1 Direkte Varianzschätzung

Der erste Teil des GREG–Schätzer (5.4) ist ein *Horvitz–Thompson–Schätzer* (Särndal et al. 1992, S. 42) und ist definiert durch

$$\hat{\tau}_Z^{\mathrm{HT}} = \sum_{a \in S} \frac{\tau_{Z,a}}{\pi_a} = \sum_{a \in S} w_a \tau_{Z,a} \tag{6.4}$$

wobei π_a die Inklusionswahrscheinlichkeit einer Anschrift a ist. Es bezeichnet $w_a = \dfrac{1}{\pi_a}$ das Designgewicht. Es sei angemerkt, dass w_a von der Stichprobe unabhängig ist.

Die Varianz dieses Schätzers lässt sich darstellen als (Särndal et al. 1992, S. 43)

$$\mathrm{Var}_{\mathrm{HT}}(\hat{\tau}_Z^{\mathrm{HT}}) = \sum_{a=1}^{N} \sum_{a'=1}^{N} \left(\pi_{aa'} - \pi_a \pi_{a'}\right) \frac{\tau_{Z,a}}{\pi_a} \frac{\tau_{Z,a'}}{\pi_{a'}} \quad , \tag{6.5}$$

wobei π_a die Inklusionswahrscheinlichkeit erster Ordnung und $\pi_{aa'}$ für $a \neq a'$ die Inklusionswahrscheinlichkeit zweiter Ordnung bedeutet. a bzw.

a' bezeichnet eine Anschrift. Diese Varianz kann im Falle $\pi_{aa'} > 0$ für alle a, a' direkt geschätzt werden durch

$$\widehat{\mathrm{Var}}_{\mathrm{HT}}(\hat{\tau}_Z^{\mathrm{HT}}) = \sum_{a \in S} \sum_{a' \in S} \frac{\pi_{aa'} - \pi_a \pi_{a'}}{\pi_{aa'}} \frac{\tau_{Z,a}}{\pi_a} \frac{\tau_{Z,a'}}{\pi_{a'}} \quad . \tag{6.6}$$

Im Fall der uneingeschränkten Zufallsstichprobe ohne Zurücklegen ist die Inklusionswahrscheinlichkeit erster Ordnung

$$\pi_a \quad = \quad \frac{n}{N} \quad . \tag{6.7}$$

Für die Inklusionswahrscheinlichkeit zweiter Ordnung erhält man

$$\pi_{aa'} \quad = \quad \frac{n(n-1)}{N(N-1)} \quad a \neq a' \quad , \tag{6.8}$$

wobei für $a = a'$ gilt $\pi_{aa} = \pi_a$.

Nach Einsetzen dieser Inklusionswahrscheinlichkeiten in (6.6) ist die Varianzschätzung für den Horvitz–Thompson–Schätzer wie folgt

$$\widehat{\mathrm{Var}}_{\mathrm{HT}}(\hat{\tau}_Z^{\mathrm{HT}}) = \frac{N^2}{n} \left(1 - \frac{n}{N}\right) s^2 \quad , \tag{6.9}$$

mit

$$s^2 = \frac{1}{n-1} \sum_{a \in S} (\tau_{Z,a} - \bar{\tau}_Z)^2$$

$$\bar{\tau}_Z = \frac{1}{n} \sum_{a \in S} \tau_{Z,a}$$

(Särndal et al. 1992, S. 46).

6.2 Linearisierungsmethoden

Es sei $\hat{\tau}_Z = f(\hat{\tau}_1, \ldots, \hat{\tau}_q)$ eine *nicht lineare* Funktion der Schätzer $\hat{\tau}_1, \ldots, \hat{\tau}_q$. Bei solchen Schätzer sind meist keine speziellen Varianzschätzer verfügbar. Eine einfache Linearisierung ist die Taylor–Approximation erster Ordnung

$$\hat{\tau}_Z = \hat{\tau}_{Z,0} \doteq \tau_Z + \sum_{t=1}^{q} \frac{\partial f(\hat{\tau}_1, \ldots, \hat{\tau}_q)}{\partial \hat{\tau}_t} (\hat{\tau}_t - \tau_t) \qquad , \tag{6.10}$$

für die dann eine Varianzschätzung berechnet wird. Die *Taylor–Linearisierung* ist zum Beispiel in Wolter (1986) oder Särndal et al. (1992) beschrieben.

Münnich (2008) gibt eine Varianzschätzung für (6.10). Allerdings wird vorausgesetzt, dass der Stichprobenumfang groß ist. Sonst ist die Varianzschätzung durch Linearisierung verzerrt (Bias$(\hat{\tau}_Z) \neq 0$). Siehe auch Wolter (1986, S. 174).

Der GREG–Schätzer (5.4) ist ein Beispiel eines nicht linearen Schätzers. Mit der Taylor–Linearisierung $\hat{\tau}_{Z,0}^{\text{GREG}}$ approximieren wir den GREG–Schätzer $\hat{\tau}_Z^{\text{GREG}}$ wie folgt

$$\hat{\tau}_Z^{\text{GREG}} \doteq \hat{\tau}_{Z,0}^{\text{GREG}} = \hat{\tau}_Z + (\boldsymbol{t}_x - \hat{\boldsymbol{t}}_x)^T \, \mathbf{B} = \sum_{a=1}^{N} \boldsymbol{x}_a^T \mathbf{B} + \sum_{a \in S} \hat{E}_a \qquad , \tag{6.11}$$

mit

$$\hat{E}_a = \hat{\tau}_{Z,a} - \hat{\boldsymbol{x}}_a^T \, \mathbf{B} \qquad .$$

Es ist $\hat{E}_a = w_a E_a = \dfrac{E_a}{\pi_a}$, wobei

$$E_a = \tau_{Z,a} - \boldsymbol{x}_a^T \, \mathbf{B} \tag{6.12}$$

auch *Residuum* der Anschrift a genannt wird. Die partiellen Ableitungen

für die Bildung der Taylor–Approximation erster Ordnung sind in Wolter (1986, S. 236) herleitet.

Mit Hilfe der Varianz (6.5) des Horvitz–Thompson–Schätzers (6.4) gilt für die Taylor–Linearizierung (6.11)

$$\begin{aligned} \mathrm{Var}(\hat{\tau}_{Z,0}^{\mathrm{GREG}}) &= \mathrm{Var}\left(\sum_{a=1}^{N} \hat{E}_a\right) \\ &= \sum_{a=1}^{N}\sum_{a'=1}^{N} (\pi_{aa'} - \pi_a\pi_{a'}) \frac{E_a}{\pi_a}\frac{E_{a'}}{\pi_{a'}} \quad . \end{aligned} \tag{6.13}$$

Es sei angemerkt, dass der erste Term in (6.11) nach dem letzten Gleichheitszeichen eine Konstante ist. Daher ist die Varianz von diesem Term Null.

Betrachten wir nun den GREG–Schätzer (5.4) mit $\hat{\mathbf{B}}$ gegeben in (5.5) und schreiben

$$\begin{aligned} \hat{\tau}_Z^{\mathrm{GREG}} &= \hat{\tau}_Z + (\boldsymbol{t}_x - \hat{\boldsymbol{t}}_x)^T \left[(\mathbf{X}^T\mathbf{W}\,\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\,\mathbf{X}^T\mathbf{W}\,\boldsymbol{\Sigma}^{-1}\,\boldsymbol{\tau}_Z\right] \\ &= \sum_{a\in S} w_a\tau_{Z,a} + (\boldsymbol{t}_x - \hat{\boldsymbol{t}}_x)^T\,(\mathbf{X}^T\mathbf{W}\,\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\sum_{a\in S}\frac{\boldsymbol{x}_a}{\sigma_a^2}\,w_a\tau_{Z,a} \\ &= \sum_{a\in S} g_{a,S}\,w_a\,\tau_{Z,a} \quad , \end{aligned} \tag{6.14}$$

wobei

$$g_{a,S} = 1 + (\boldsymbol{t}_x - \hat{\boldsymbol{t}}_x)^T\,(\mathbf{X}^T\mathbf{W}\,\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\frac{\boldsymbol{x}_a}{\sigma_a^2} \tag{6.15}$$

und $\boldsymbol{x}_a$=$(x_{1,a},\ldots,x_{v,a})^T$ ein Vektor der bekannten v Hilfsvariablen für die Anschrift a ist.

Ausgehend von (6.12) ist $\tau_{Z,a} = \boldsymbol{x}_a^T \mathbf{B} + E_a$ und nach Einsetzen in (6.14) ist

$$\begin{aligned} \hat{\tau}_Z^{\text{GREG}} &= \sum_{a \in S} g_{a,S}\, w_a \left(\boldsymbol{x}_a^T \mathbf{B} + E_a\right) \\ &= \sum_{a=1}^{N} \boldsymbol{x}_a^T \mathbf{B} + \sum_{a \in S} w_a\, g_{a,S}\, E_a \qquad . \end{aligned} \tag{6.16}$$

Die letzte Gleichung ist in Särndal et al. (1992, S. 233) erklärt.

Die Varianz des GREG–Schätzer in (6.16) ist näherungsweise durch (6.13) gegeben. Für die Varianzschätzung des Schätzers wendet Münnich et al. (2003) die *Woodruff–Methode* an, wobei $g_{a,S}$ in (6.16) als unabhängig von einer Stichprobe betrachtet wird und $e_a = \tau_{Z,a} - \boldsymbol{x}_a^T \hat{\mathbf{B}}$ statt der unbekannten E_a verwendet wird. Damit ergibt sich als Varianzschätzer des GREG–Schätzers

$$\widehat{\text{Var}}(\hat{\tau}_Z^{\text{GREG}}) = \sum_{a \in S} \sum_{a' \in S} \frac{\pi_{aa'} - \pi_a \pi_{a'}}{\pi_{aa'}} \, \frac{g_{a,S}\, e_a}{\pi_a} \, \frac{g_{a',S}\, e_{a'}}{\pi_{a'}} \qquad . \tag{6.17}$$

Die Varianzschätzung des GREG–Schätzer (6.17) erhalten wir einfach durch Verwendung der gewichteten Residuen $g_{a,S}\, e_a$ statt der Totalwerte $\tau_{Z,a}$ in (6.6). Diese Methode wird auch *Residual–Methode* genannt (s. Deville 1999).

6.2.1 Varianz und Varianzschätzung für den GREG–Schätzer in Deutschland

Für das Modell in Deutschland, wo nur eine Hilfsvariable in Form der bekannten Populationsanzahl des Registers zur Verfügung steht und $\varepsilon_a \sim N(0, \sigma^2 \tau_{R,a})$ angenommen wird (s. Seite 58), ist

$$g_{a,S} = 1 + (\tau_R - \hat{\tau}_R) \left(\frac{1}{\sigma^2} \sum_{a \in S} \tau_{R,a} w_a \right)^{-1} \frac{\tau_{R,a}}{\sigma^2 \tau_{R,a}} = \frac{\tau_R}{\hat{\tau}_R}$$

unabhängig von der ausgewählten Anschrift a. Der Term $g_{a,S}$ wird in der Literatur (vgl. Lohr 1999) oft gleich eins gesetzt.

Die Varianz des GREG0–Schätzers (5.7) bei uneingeschränkten Zufallsauswahl der Anschriften ohne Zurücklegen ist gegeben durch

$$\text{Var}(\hat{\tau}_Z^{\text{GREG0}}) = N^2 \, \frac{1}{n} \, \left(1 - \frac{n}{N}\right) \frac{1}{N-1} \sum_{a=1}^{N} \left(\tau_{Z,a} - \frac{\tau_Z}{\tau_R} \tau_{R,a} \right)^2 \quad . \quad (6.18)$$

Beim GREG0–Schätzer handelt es sich um einen Verhältnisschätzer. Für diese Schätzfunktion wird die Varianzschätzung fast in jeder Literatur betrachtet. Lohr (1999, S. 145) zum Beispiel gibt als Varianzschätzung

$$\widehat{\text{Var}}(\hat{\tau}_Z^{\text{GREG0}}) = N^2 \, \frac{1}{n} \, \left(1 - \frac{n}{N}\right) \frac{1}{n-1} \sum_{a \in S} \left(\tau_{Z,a} - \frac{\hat{\tau}_Z}{\hat{\tau}_R} \tau_{R,a} \right)^2 \quad (6.19)$$

an.

Im Folgenden betrachten wir auch beim GREG1 und GREG2–Schätzer nur die vereinfachte Formel für die Varianzschätzung, nämlich mit $g_{a,S} = 1$.

Die Varianz bzw. die Varianzschätzung des GREG1–Schätzers (5.9) lässt

sich wieder mit (6.5) bzw. (6.6) berechnen. Laut Rao (2003, S. 17) setzen wir bei der Varianzschätzung in (6.6) für $\tau_{Z,a}$ die Residuen

$$e_{d,a} = \tau_{Z,d,a} - \hat{\mathbf{B}}_d \, \tau_{R,d,a} = \tau_{Z,d,a} - \frac{\sum\limits_{a \in S} w_a \, \tau_{Z,d,a}}{\sum\limits_{a \in S} w_a \, \tau_{R,a}} \, \tau_{R,d,a} \qquad , \tag{6.20}$$

mit $\tau_{Z,d,a}$ als die Anzahl der tatsächlich vorhandenen Personen der Domain d an der Anschrift a und $\tau_{R,a}$ als die Anzahl der registrierten Personen. Daher ist

$$\text{Var}(\hat{\tau}_{Z,d}^{\text{GREG1}}) = N^2 \, \frac{1}{n} \, \left(1 - \frac{n}{N}\right) \frac{1}{N-1} \sum_{a=1}^{N} \left(\tau_{Z,d,a} - \frac{\tau_{Z,d}}{\tau_R} \, \tau_{R,d,a}\right)^2 \tag{6.21}$$

$$\widehat{\text{Var}}(\hat{\tau}_{Z,d}^{\text{GREG1}}) = N^2 \, \frac{1}{n} \, \left(1 - \frac{n}{N}\right) \frac{1}{n-1} \sum_{a \in S} \left(\tau_{Z,d,a} - \frac{\hat{\tau}_{Z,d}}{\hat{\tau}_R} \, \tau_{R,d,a}\right)^2 \tag{6.22}$$

Man beachte, dass für eine Anschrift a'' mit keiner tatsächlich lebenden Person in der Domain d ($\tau_{Z,d,a''} = 0$) gilt

$$e_{d,a''} = -\tau_{R,d,a''} \frac{\sum\limits_{a \in S} w_a \, \tau_{Z,d,a}}{\sum\limits_{a \in S} w_a \, \tau_{R,a}} \qquad , \tag{6.23}$$

was zu einer großen Varianz dieses Schätzers führt.

Der GREG2–Schätzer (5.12) ist eigentlich ein GREG0–Schätzer innerhalb einer Domain d. Die Varianz und die Varianzschätzung des GREG2–Schätzers ist daher mit den Formeln

$$\text{Var}(\hat{\tau}_{Z,d}^{\text{GREG2}}) = N^2 \, \frac{1}{n} \, \left(1 - \frac{n}{N}\right) \frac{1}{N-1} \sum_{a=1}^{N} \left(\tau_{Z,d,a} - \frac{\tau_{Z,d}}{\tau_{R,d}} \tau_{R,d,a}\right)^2 \tag{6.24}$$

$$\widehat{\text{Var}}(\hat{\tau}_{Z,d}^{\text{GREG2}}) = N^2 \, \frac{1}{n} \, \left(1 - \frac{n}{N}\right) \frac{1}{n-1} \sum_{a \in S} \left(\tau_{Z,d,a} - \frac{\hat{\tau}_{Z,d}}{\hat{\tau}_{R,d}} \tau_{R,d,a} \right)^2 \quad (6.25)$$

innerhalb einer Domain d berechenbar.

Bei geschichteter Zufallsauswahl der Anschriften nehmen die Varianzschätzung des GREG1–Schätzers und GREG2–Schätzers innerhalb einer Domain d die Form

$$\widehat{\text{Var}}(\hat{\tau}_{Z,d}^{\text{GREG1,str}}) = \sum_{h=1}^{H} N_h^2 \, \frac{1}{n_h} \, \left(1 - \frac{n_h}{N_h}\right) \frac{1}{n_h - 1} \sum_{a \in S_h} \left(\tau_{Z,d,a} - \frac{\hat{\tau}_{Z,d,h}}{\hat{\tau}_{R,h}} \, \tau_{R,d,a} \right)^2 \quad (6.26)$$

und

$$\widehat{\text{Var}}(\hat{\tau}_{Z,d}^{\text{GREG2,str}}) = \sum_{h=1}^{H} N_h^2 \, \frac{1}{n_h} \, \left(1 - \frac{n_h}{N_h}\right) \frac{1}{n_h - 1} \sum_{a \in S_h} \left(\tau_{Z,d,a} - \frac{\hat{\tau}_{Z,d,h}}{\hat{\tau}_{R,d,h}} \tau_{R,d,a} \right)^2 . \quad (6.27)$$

6.2.2 Varianz und Varianzschätzung für den D–DSE und Chapman–Schätzer

Wie im Kapitel 3.1.3 erwähnt wird D–DSE in den Klassen berechnet, um die heterogenen Wahrscheinlichkeiten zu modellieren. Im wesentlichen ist D–DSE einer Klasse in (3.12) ein Verhältnisschätzer. Nach kleiner Umformulierung ist für die Varianzschätzung des D–DSEs einer Klasse k die Formel für die Varianzschätzung (6.25) einer Domain d anwendbar. Für die geschichtete Zufallsauswahl der Anschriften gilt die Formel in (6.27). Im Folgenden wird diese Varianzschätzung als direkte Varianzschätzung für D–DSE bezeichnet.

Die Situation im Chapman–Schätzer ist etwas unterschiedlich, da im Nen-

ner und Zähler die Schätzwerte um eins erhöht stehen. Wir können aber auch den Chapman–Schätzer (3.23) schreiben

$$\hat{\tau}_Z^{\text{CHAP}} = (\tau_R + 1)\,\frac{\widehat{\tau}_Z + 1}{\widehat{\tau}_R + 1} - 1 = (\tau_R + 1)\,\frac{\sum\limits_{a\in S} w_a \left(\tau_{Z,a} + \dfrac{1}{w_a n}\right)}{\sum\limits_{a\in S} w_a \left(\tau_{R,a} + \dfrac{1}{w_a n}\right)} - 1 \qquad .$$

Daher ist klar, dass für den Chapman–Schätzer im Modell in Deutschland

$$\mathbf{X} = \left(\tau_{R,1} + \frac{1}{w_1 n}, \ldots, \tau_{R,n} + \frac{1}{w_n n}\right)^T \qquad ,$$

$$\hat{\mathbf{B}} = \frac{\sum\limits_{a\in S} w_a \left(\tau_{Z,a} + \dfrac{1}{w_a n}\right)}{\sum\limits_{a\in S} w_a \left(\tau_{R,a} + \dfrac{1}{w_a n}\right)} = \frac{\hat{\tau}_Z + 1}{\hat{\tau}_R + 1}$$

gilt. Darüber hinaus folgt

$$\begin{aligned} g_{a,S} &= 1 + [\tau_R + 1 - (\hat{\tau}_R + 1)] \times \\ &\quad \left[\frac{1}{\sigma^2} \sum_{a\in S} \left(\tau_{R,a} + \frac{1}{w_a n}\right) w_a\right]^{-1} \frac{\tau_{R,a} + \dfrac{1}{w_a n}}{\sigma^2 \left(\tau_{R,a} + \dfrac{1}{w_a n}\right)} \\ &= \frac{\tau_R + 1}{\hat{\tau}_R + 1} \qquad . \end{aligned}$$

Nach Einsetzen von $\left(\tau_{Z,a} + \dfrac{1}{w_a n}\right)$ für $\tau_{Z,a}$ und von $\left(\tau_{R,a} + \dfrac{1}{w_a n}\right)$ für $\tau_{R,a}$ in (6.19) kann die Varianz für den Chapman–Schätzer

$$\text{Var}(\hat{\tau}_Z^{\text{CHAP}}) = N^2\,\frac{1}{n}\,\left(1 - \frac{n}{N}\right)\frac{1}{N-1}\,\Phi \tag{6.28}$$

durch

$$\widehat{\text{Var}}(\hat{\tau}_Z^{\text{CHAP}}) = N^2 \, \frac{1}{n} \, \left(1 - \frac{n}{N}\right) \frac{1}{n-1} \, \Phi_s \tag{6.29}$$

geschätzt werden, mit

$$\Phi = \sum_{a=1}^{N} \left(\tau_{R,a} + \frac{1}{w_a n}\right)^2 \left(\frac{w_a n \tau_{Z,a} + 1}{w_a n \tau_{R,a} + 1} - \frac{\tau_Z + 1}{\tau_R + 1}\right)^2 \tag{6.30}$$

$$\Phi_s = \sum_{a \in S} \left(\tau_{R,a} + \frac{1}{w_a n}\right)^2 \left(\frac{w_a n \tau_{Z,a} + 1}{w_a n \tau_{R,a} + 1} - \frac{\hat{\tau}_Z + 1}{\hat{\tau}_R + 1}\right)^2 \quad . \tag{6.31}$$

Innerhalb einer Klasse k ist die Varianzschätzung wie folgt

$$\widehat{\text{Var}}_2(\hat{\tau}_{Z,k}^{\text{CHAP}}) = N_k^2 \, \frac{1}{n_k} \, \left(1 - \frac{n_k}{N_k}\right) \frac{1}{n_k - 1} \, \Phi_{s_k} \quad , \tag{6.32}$$

wobei Φ_{s_k} analog zu (6.31) definiert ist mit der Summe über die Stichprobe innerhalb der Klasse k.

Nach dem Ableiten der Formeln für die Varianzschätzung des D–DSEs, GREG1 bzw. GREG2–Schätzers bei geschichteter Zufallsauswahl ist die Formel für den Chapman–Schätzer bei geschichteter Zufallsauswahl herzuleiten. Im Folgenden wird diese Varianzschätzung als direkte Varianzschätzung für den Chapman–Schätzer bezeichnet.

6.3 Resampling Methoden für D–DSE und Chapman–Schätzer

Für nicht lineare Schätzfunktionen werden in Shao and Tu (1995) auch die *Resampling Methoden* diskutiert. Das Prinzip bei diesen Methoden ist, wie-

derholte Substichproben aus der bereits ermittelten Stichprobe zu ziehen und die Schätzfunktion auf Basis der Substichproben wiederholt zu berechnen. Eine bekannte Methode ist das *Jackknife–Verfahren*.

Bei *Delete–1–Jackknife* wird wiederholt genau eine Einheit der ursprünglichen Stichprobe ausgelassen und die Schätzfunktion auf die Substichprobe angewendet.

D–DSE bzw. Chapman–Schätzer sind Schätzer des Totalwertes der tatsächlich lebenden Personen in Deutschland τ_Z basierend auf einer Stichprobe S der Anschriften, wo Personen an einer Anschrift vollständig erhoben werden. Bei solchem Klumpendesign, wo die Anschriften die Klumpen bilden und die Auswahl der Klumpen durch uneingeschränkte Zufallsauswahl vom Umfang n gegeben ist, sollte in jedem Schritt des Delete–1–Jackknife jeweils ein Klumpen eliminiert werden. Für die Jackknife–Varianzschätzungen wird die Technik wie beim uneingeschränkten Design benutzt.

Es sei $\hat{\tau}^{\text{D-DSE}}_{Z,k,-a}$ bzw. $\hat{\tau}^{\text{CHAP}}_{Z,k,-a}$ der gleiche Schätzer wie $\hat{\tau}^{\text{D-DSE}}_{Z,k}$ bzw. $\hat{\tau}^{\text{CHAP}}_{Z,k}$, berechnet nach Auslassung der Anschrift a aus der ursprünglichen Stichprobe, k ist die Klasse.

Wir definieren

$$\begin{aligned} \hat{\tau}^{\text{D-DSE}}_{Z,k,a,\text{pseudo}} &= n\,\hat{\tau}^{\text{D-DSE}}_{Z,k} - (n-1)\,\hat{\tau}^{\text{D-DSE}}_{Z,k,-a} \\ \hat{\tau}^{\text{CHAP}}_{Z,k,a,\text{pseudo}} &= n\,\hat{\tau}^{\text{CHAP}}_{Z,k} - (n-1)\,\hat{\tau}^{\text{CHAP}}_{Z,k,-a} \quad , \end{aligned} \tag{6.33}$$

wobei n die Anzahl der Anschriften in der Stichprobe S ist. Dann sind

$$\begin{aligned} \hat{\bar{\tau}}^{\text{D-DSE}}_{Z,k,\text{jack1}} &= \frac{1}{n}\sum_{a\in S}\hat{\tau}^{\text{D-DSE}}_{Z,k,a,\text{pseudo}} \\ \hat{\bar{\tau}}^{\text{CHAP}}_{Z,k,\text{jack1}} &= \frac{1}{n}\sum_{a\in S}\hat{\tau}^{\text{CHAP}}_{Z,k,a,\text{pseudo}} \end{aligned} \tag{6.34}$$

die Jackknife–Schätzer für D–DSE bzw. Chapman–Schätzer (Wolter 1986, S. 154). Die Werte in (6.33) werden auch Pseudo–Werte genannt. Diese Pseudo–Werte sind andere als die Pseudo–Werte bei GSPREE im Kapitel 4.3.

Die Jackknife–Schätzer in (6.34) sind die Mittelwerte der Pseudo–Werten $\hat{\tau}^{\text{D-DSE}}_{Z,k,a,\text{pseudo}}$ bzw. $\hat{\tau}^{\text{CHAP}}_{Z,k,a,\text{pseudo}}$. Daher gilt für die Varianzschätzung beim Ziehen mit Zurücklegen

$$\begin{aligned}
\widehat{\text{Var}}\,(\hat{\bar{\tau}}^{\text{D-DSE}}_{Z,k,\text{jack1}}) &= \frac{1}{n\,(n-1)} \sum_{a\in S} (\hat{\tau}^{\text{D-DSE}}_{Z,k,a,\text{pseudo}} - \hat{\bar{\tau}}^{\text{D-DSE}}_{Z,k,\text{jack1}})^2 \\
\widehat{\text{Var}}\,(\hat{\bar{\tau}}^{\text{CHAP}}_{Z,k,\text{jack1}}) &= \frac{1}{n\,(n-1)} \sum_{a\in S} (\hat{\tau}^{\text{CHAP}}_{Z,k,a,\text{pseudo}} - \hat{\bar{\tau}}^{\text{CHAP}}_{Z,k,\text{jack1}})^2 \quad .
\end{aligned} \tag{6.35}$$

In der Praxis werden diese Schätzer nicht nur für die Varianzschätzung des Jackknife–Schätzers in (6.34) genutzt, sondern auch für die Varianzschätzung $\hat{\tau}^{\text{D-DSE}}_{Z,k}$ bzw. $\hat{\tau}^{\text{CHAP}}_{Z,k}$ (Wolter 1986, S. 155), d.h.

$$\begin{aligned}
\widehat{\text{Var}}_{\text{jack1}}(\hat{\tau}^{\text{D-DSE}}_{Z,k}) &= \widehat{\text{Var}}(\hat{\bar{\tau}}^{\text{D-DSE}}_{Z,k,\text{jack1}}) \\
\widehat{\text{Var}}_{\text{jack1}}(\hat{\tau}^{\text{CHAP}}_{Z,k}) &= \widehat{\text{Var}}(\hat{\bar{\tau}}^{\text{CHAP}}_{Z,k,\text{jack1}}) \quad .
\end{aligned}$$

6.3.1 Jackknife bei uneingeschränkter Zufallsauswahl ohne Zurücklegen

Beim uneingeschränkten Ziehen ohne Zurücklegen sind die Schätzungen in (6.35) nicht mehr unverzerrt und überschätzen die Varianz (Wolter 1986, S. 168). Um diesen Bias zu korrigieren, wird mit

$$\begin{aligned}
\hat{\tau}^{\text{D-DSE}\,*}_{Z,k,-a} &= \hat{\tau}^{\text{D-DSE}}_{Z,k} + \sqrt{1-\frac{n}{N}}\,(\hat{\tau}^{\text{D-DSE}}_{Z,k,-a} - \hat{\tau}^{\text{D-DSE}}_{Z,k}) \\
\hat{\tau}^{\text{CHAP}\,*}_{Z,k,-a} &= \hat{\tau}^{\text{CHAP}}_{Z,k} + \sqrt{1-\frac{n}{N}}\,(\hat{\tau}^{\text{CHAP}}_{Z,k,-a} - \hat{\tau}^{\text{CHAP}}_{Z,k})
\end{aligned} \tag{6.36}$$

gearbeitet (Wolter 1986, S. 169), wobei N die Anzahl aller Anschriften bezeichnet. Die Pseudo–Werte sind dann

$$\begin{aligned} \hat{\tau}^{\text{D-DSE}\,*}_{Z,k,a,\text{pseudo}} &= n\,\hat{\tau}^{\text{D-DSE}}_{Z,k} - (n-1)\,\hat{\tau}^{\text{D-DSE}\,*}_{Z,k,-a} \\ \hat{\tau}^{\text{CHAP}\,*}_{Z,k,a,\text{pseudo}} &= n\,\hat{\tau}^{\text{CHAP}}_{Z,k} - (n-1)\,\hat{\tau}^{\text{CHAP}\,*}_{Z,k,-a} \quad . \end{aligned} \tag{6.37}$$

Die Jackknife–Varianzschätzer sind

$$\begin{aligned} \widehat{\text{Var}}(\hat{\bar{\tau}}^{\text{D-DSE}\,*}_{Z,k,\text{jack}^*}) &= \frac{1}{n\,(n-1)} \sum_{a\in S} (\hat{\tau}^{\text{D-DSE}\,*}_{Z,k,a,\text{pseudo}} - \hat{\bar{\tau}}^{\text{D-DSE}\,*}_{Z,k,\text{jack}^*})^2 \\ \widehat{\text{Var}}(\hat{\bar{\tau}}^{\text{CHAP}\,*}_{Z,k,\text{jack}^*}) &= \frac{1}{n\,(n-1)} \sum_{a\in S} (\hat{\tau}^{\text{CHAP}\,*}_{Z,k,a,\text{pseudo}} - \hat{\bar{\tau}}^{\text{CHAP}\,*}_{Z,k,\text{jack}^*})^2 \quad , \end{aligned} \tag{6.38}$$

wobei

$$\begin{aligned} \hat{\bar{\tau}}^{\text{D-DSE}\,*}_{Z,k,\text{jack}^*} &= \frac{1}{n} \sum_{a\in S} \hat{\tau}^{\text{D-DSE}\,*}_{Z,k,a,\text{pseudo}} \\ \hat{\bar{\tau}}^{\text{CHAP}\,*}_{Z,k,\text{jack}^*} &= \frac{1}{n} \sum_{a\in S} \hat{\tau}^{\text{CHAP}\,*}_{Z,k,a,\text{pseudo}} \quad . \end{aligned}$$

Für die Varianzschätzung von D–DSE bzw. Chapman–Schätzer beim Ziehen ohne Zurücklegen werden die Varianzschätzer $\widehat{\text{Var}}\,(\hat{\bar{\tau}}^{\text{D-DSE}\,*}_{Z,k,\text{jack}^*})$ bzw. $\widehat{\text{Var}}\,(\hat{\bar{\tau}}^{\text{CHAP}\,*}_{Z,k,\text{jack}^*})$ verwendet, d.h.

$$\begin{aligned} \widehat{\text{Var}}_{\text{jack}^*}(\hat{\tau}^{\text{D-DSE}}_{Z,k}) &= \widehat{\text{Var}}(\hat{\bar{\tau}}^{\text{D-DSE}\,*}_{Z,k,\text{jack}^*}) \\ \widehat{\text{Var}}_{\text{jack}^*}(\hat{\tau}^{\text{CHAP}}_{Z,k}) &= \widehat{\text{Var}}(\hat{\bar{\tau}}^{\text{CHAP}\,*}_{Z,k,\text{jack}^*}) \quad . \end{aligned}$$

6.3.2 Jackknife bei geschichteter Zufallsauswahl

Die Delete–1–Jackknife–Varianzschätzung kann Schwierigkeiten bei geschichteten Designs bringen. Auf eine Delete–1–Jackknife–Varianzschätzung sollte in diesem Design verzichtet werden (Wolter 1986, S. 175).

Es sei S_h eine uneingeschränkte zufällig ohne Zurücklegen gezogene Anschriftenstichprobe vom Umfang n_h innerhalb der Schicht h. Die Auswahlen zwischen den Schichten sind unabhängig voneinander. Wir bezeichnen mit $\hat{\tau}^{\text{D-DSE}}_{Z,k,h,-a}$ bzw. $\hat{\tau}^{\text{CHAP}}_{Z,k,h,-a}$ den D–DSE bzw. Chapman–Schätzer der tatsächlich lebenden Personen einer Klasse k der Schicht h in Deutschland, berechnet nach Auslassung der Anschrift a in der h-ten Schicht der ursprünglichen Stichprobe. Wir definieren die Pseudo–Werte

$$\begin{aligned} \hat{\tau}^{\text{D-DSE}}_{Z,k,h,a,\text{pseudo}} &= (Hw_h + 1)\,\hat{\tau}^{\text{D-DSE}}_{Z,k} - Hw_h\,\hat{\tau}^{\text{D-DSE}}_{Z,k,h,-a} \\ \hat{\tau}^{\text{CHAP}}_{Z,k,h,a,\text{pseudo}} &= (Hw_h + 1)\,\hat{\tau}^{\text{CHAP}}_{Z,k} - Hw_h\,\hat{\tau}^{\text{CHAP}}_{Z,k,h,-a} \quad , \end{aligned} \tag{6.39}$$

mit

$$w_h = (n_h - 1)\left(1 - \frac{n_h}{N_h}\right) \quad .$$

Die Jackknife–Schätzer für $\hat{\tau}^{\text{D-DSE}}_{Z,k}$ und $\hat{\tau}^{\text{CHAP}}_{Z,k}$ sind dann (Wolter 1986, S. 176)

$$\begin{aligned} \hat{\bar{\tau}}^{\text{D-DSE}}_{Z,k,\text{jack2}} &= \frac{1}{H}\sum_{h=1}^{H}\frac{1}{n_h}\sum_{a\in S_h}\hat{\tau}^{\text{D-DSE}}_{Z,k,h,a,\text{pseudo}} \\ &= \left(1 + \sum_{h=1}^{H} w_h\right)\hat{\tau}^{\text{D-DSE}}_{Z,k} - \sum_{h=1}^{H} w_h\,\hat{\tau}^{\text{D-DSE}}_{Z,k,h-} \\ \hat{\bar{\tau}}^{\text{CHAP}}_{Z,k,\text{jack2}} &= \frac{1}{H}\sum_{h=1}^{H}\frac{1}{n_h}\sum_{a\in S_h}\hat{\tau}^{\text{CHAP}}_{Z,k,h,a,\text{pseudo}} \\ &= \left(1 + \sum_{h=1}^{H} w_h\right)\hat{\tau}^{\text{CHAP}}_{Z,k} - \sum_{h=1}^{H} w_h\,\hat{\tau}^{\text{CHAP}}_{Z,k,h-} \quad , \end{aligned} \tag{6.40}$$

wobei

$$\begin{aligned} \hat{\tau}^{\text{D-DSE}}_{Z,k,h-} &= \frac{1}{n_h} \sum_{a \in S_h} \hat{\tau}^{\text{D-DSE}}_{Z,k,h,-a} \\ \hat{\tau}^{\text{CHAP}}_{Z,k,h-} &= \frac{1}{n_h} \sum_{a \in S_h} \hat{\tau}^{\text{CHAP}}_{Z,k,h,-a} \quad . \end{aligned} \tag{6.41}$$

Die Jackknife–Varianzschätzer für die geschichteten Designs sind (Wolter 1986, S. 178)

$$\begin{aligned} \widehat{\text{Var}}(\hat{\bar{\tau}}^{\text{D-DSE}}_{Z,k,\text{jack2}}) &= \sum_{h=1}^{H} \frac{w_h}{n_h} \sum_{a \in S_h} (\hat{\tau}^{\text{D-DSE}}_{Z,k,h,-a} - \hat{\tau}^{\text{D-DSE}}_{Z,k,h-})^2 \\ \widehat{\text{Var}}(\hat{\bar{\tau}}^{\text{CHAP}}_{Z,k,\text{jack2}}) &= \sum_{h=1}^{H} \frac{w_h}{n_h} \sum_{a \in S_h} (\hat{\tau}^{\text{CHAP}}_{Z,k,h,-a} - \hat{\tau}^{\text{CHAP}}_{Z,k,h-})^2 \quad . \end{aligned} \tag{6.42}$$

Für Varianzschätzung für D–DSE bzw. Chapman–Schätzer bei geschichteter Zufallsauswahl wird $\widehat{\text{Var}}\,(\hat{\bar{\tau}}^{\text{D-DSE}}_{Z,k,\text{jack2}})$ bzw. $\widehat{\text{Var}}\,(\hat{\bar{\tau}}^{\text{CHAP}}_{Z,k,\text{jack2}})$ verwendet, d.h.

$$\begin{aligned} \widehat{\text{Var}}_{\text{jack2}}(\hat{\tau}^{\text{D-DSE,str}}_{Z,k}) &= \widehat{\text{Var}}(\hat{\bar{\tau}}^{\text{D-DSE}}_{Z,k,\text{jack2}}) \\ \widehat{\text{Var}}_{\text{jack2}}(\hat{\tau}^{\text{CHAP,str}}_{Z,k}) &= \widehat{\text{Var}}(\hat{\bar{\tau}}^{\text{CHAP}}_{Z,k,\text{jack2}}) \quad . \end{aligned}$$

Die Delete–1–Jackknife–Varianzschätzung kann nur verwendet werden, wenn $n \geq 2$. Vor allem beim geschichteten Design kann es vorkommen, dass in einzelnen Schichten die Bedingung $n_h \geq 2$ nicht erfüllt ist. In solchen Fällen werden die Schichten erst zusammengefasst, damit es nur Schichten mit Stichprobenumfang mindestens $n_h \geq 2$ gibt. Die Delete–1–Jackknife–Varianzschätzung wird dann auf die neue Schichtung angewendet.

6.4 Varianz und Varianzschätzung für Regressions–synthetischer Schätzer

Für den Fall einer bekannten Hilfsvariablen für die Domains wurde im Kapitel 5.2 der Verhältnis–synthetische Schätzer (5.17) als eine alternative zum GREG–Schätzer vorgestellt. Dieser Schätzer ist eigentlich der GREG2–Schätzer (5.12), mit $\hat{\mathbf{B}} = \frac{\hat{\tau}_Z}{\hat{\tau}_R}$ statt $\hat{\mathbf{B}}_{d,\text{sep}} = \frac{\hat{\tau}_{Z,d}}{\hat{\tau}_{R,d}}$ (siehe Seiten 59, 61 und 62).

Daher gehen die Varianz und die Varianzschätzung für den Verhältnis–synthetischen Schätzer von den Formeln (6.24) und (6.25) des GREG2–Schätzers aus und sind durch

$$\text{Var}(\hat{\tau}^{\text{SYN}}_{Z,d}) = N^2 \, \frac{1}{n} \, \left(1 - \frac{n}{N}\right) \frac{1}{N-1} \sum_{a=1}^{N} \left(\tau_{Z,d,a} - \frac{\tau_Z}{\tau_R}\tau_{R,d,a}\right)^2 \qquad (6.43)$$

mit Varianzschätzer

$$\widehat{\text{Var}}(\hat{\tau}^{\text{SYN}}_{Z,d}) = N^2 \, \frac{1}{n} \, \left(1 - \frac{n}{N}\right) \frac{1}{n-1} \sum_{a \in S} \left(\tau_{Z,d,a} - \frac{\hat{\tau}_Z}{\hat{\tau}_R}\tau_{R,d,a}\right)^2 \qquad (6.44)$$

gegeben.

Bei geschichteter Zufallsauswahl ist

$$\widehat{\text{Var}}(\hat{\tau}^{\text{SYN,str}}_{Z,d}) \quad = \quad \sum_{h=1}^{H} N_h^2 \, \frac{1}{n_h} \, \left(1 - \frac{n_h}{N_h}\right) \frac{1}{n_h - 1} \sum_{a \in S_h} \left(\tau_{Z,d,a} - \frac{\hat{\tau}_{Z,h}}{\hat{\tau}_{R,h}}\tau_{R,d,a}\right)^2 \qquad (6.45)$$

Wenn der Klammerterm in (5.15) ungleich Null ist, ist der Regressions–synthetische Schätzer (5.16) verzerrt mit dem Bias

$$\text{Bias}(\hat{\tau}^{\text{REG-SYN}}_{Z,d}) = \boldsymbol{t}^T_{x,d} \, \hat{\mathbf{B}} - \tau_{Z,d} \neq 0 \qquad .$$

Bei einem verzerrten Schätzer $\hat{\tau}$ sollte der MSE($\hat{\tau}$) berechnet werden. Eine Schätzung des MSE für den Verhältnis–synthetischen Schätzer ist in Rao (2003, S. 52) gegeben. Diese Schätzung liefert in unseren Simulationen häufig negative Ergebnisse und wird daher in dieser Arbeit nicht weiter untersucht. Wir werden Varianzschätzungen in (6.44) bzw. (6.45) ermitteln.

Da der Vektor $\hat{\mathbf{B}}$ der geschätzten Regressionskoeffizienten auf der Gemeindeebene gewonnen ist und damit sehr präzis ist, ist die Varianz der Regressions–synthetischen Schätzer meistens sehr klein. Der Schätzer wird häufig in sehr kleinen Domains benutzt (Särndal et al. 1992, S. 399). Anderseits kann der Regressions–synthetische Schätzer einen großen Bias aufweisen.

6.5 MSE und Schätzung von MSE für EBLUP

Im Kapitel 5.3 werden die linearen–gemischten–Modelle und EBLUP betrachtet. Da der EBLUP einen verzerrten Verhältnis–synthetischen Schätzer beinhaltet, ist EBLUP auch verzerrt. Prasad and Rao (1990) oder Rao (2003, S. 139) geben die Formel für den MSE des EBLUPs im Unit–level Modell an. Der EBLUP dient nur als alternativer Schätzer zu Vergleichszwecken mit Dual–System–Modellen. Daher wird im Weiteren die Formel zur Schätzung des MSE vorgestellt.

Der MSE für den EBLUP ist gegeben (s. Rao 2003, S. 139) durch

$$\widehat{\mathrm{MSE}}(\hat{\tau}_{Z,d}^{\mathrm{EBLUP}}) = g_{1,d}(\sigma_\alpha^2, \sigma_\varepsilon^2) + g_{2,d}(\sigma_\alpha^2, \sigma_\varepsilon^2) + g_{3,d}(\sigma_\alpha^2, \sigma_\varepsilon^2)$$

mit

$$g_{1,d}(\sigma_\alpha^2, \sigma_\varepsilon^2) = (1 - \gamma_d)\sigma_\alpha^2$$

$$g_{2,d}(\sigma_\alpha^2, \sigma_\varepsilon^2) = (\tau_{R,d} - \gamma_d \, \tau_{R,S_P,d}) \left(\sum_d \mathbf{X}_d^T \mathbf{V}_d^{-1} \mathbf{X}_d \right)^{-1} (\tau_{R,d} - \gamma_d \, \tau_{R,S_P,d})$$

$$g_{3,d}(\sigma_\alpha^2, \sigma_\varepsilon^2) = n_{P,d}^{-2} \left(\sigma_\alpha^2 + \frac{\sigma_\varepsilon^2}{n_{P,d}} \right)^{-3} \left[\sigma_\varepsilon^4 \, I_{\alpha\alpha} + \sigma_\alpha^4 I_{\varepsilon\varepsilon} - 2\sigma_\varepsilon^2 \sigma_\alpha^2 \, I_{\alpha\varepsilon} \right] \quad ,$$

wobei γ_d in (5.21) definiert ist. Weiter ist

$$I_{\alpha\alpha} = \frac{2}{t} \sum_{d=1}^{D} \frac{n_{P,d} - 1}{\sigma_\varepsilon^4} + \frac{1}{u_d^2} \qquad I_{\varepsilon\varepsilon} = \frac{2}{t} \sum_{d=1}^{D} \frac{n_{P,d}^2}{u_d^2} \qquad I_{\alpha\varepsilon} = -\frac{2}{t} \sum_{d=1}^{D} \frac{n_{P,d}}{u_d^2}$$

mit

$$t = \left[\sum_{d=1}^{D} \frac{n_{P,d}^2}{u_d^2} \right] \left[\sum_{d=1}^{D} \left(\frac{n_{P,d} - 1}{\sigma_\varepsilon^4} + \frac{1}{u_d^2} \right) \right] - \left[\sum_{d=1}^{D} \frac{n_{P,d}}{u_d^2} \right]^2$$

und

$$u_d = \sigma_\varepsilon^2 + n_{P,d} \sigma_\alpha^2 \quad .$$

Eine REML–Schätzung des MSE für den EBLUP geben Datta and Lahiri (2000) oder Rao (2003, S. 140) als

$$\widehat{\mathrm{MSE}}(\hat{\bar{\tau}}_{Z,d}^{\mathrm{EBLUP}}) = g_{1,d}(\hat{\sigma}_\alpha^2, \hat{\sigma}_\varepsilon^2) + g_{2,d}(\hat{\sigma}_\alpha^2, \hat{\sigma}_\varepsilon^2) + 2g_{3,d}(\hat{\sigma}_\alpha^2, \hat{\sigma}_\varepsilon^2)$$

an. Es kann gezeigt werden, dass die einzelnen Schätzer sehr kleinen Bias aufweisen (vgl. Prasad and Rao 1990).

6.6 SPREE–Varianz und SPREE–Varianzschätzung

Betrachten wir den Schätzer $\hat{\tau}_t$ der Zufallsvariablen τ_t und Zahlen $a_t \in \mathbb{Z}$, $t \in 1, \ldots, T$. Für Linearkombinationen der unabhängigen Variablen gilt

$$\begin{aligned} \widehat{\text{Var}}\,(a_t\ \hat{\tau}_t) &= a_t^2\ \widehat{\text{Var}}\,(\hat{\tau}_t) \\ \widehat{\text{Var}}\left(\sum_{t=1}^{T} a_t\ \hat{\tau}_t\right) &= \sum_{k=t}^{T} a_t^2\ \widehat{\text{Var}}\,(\hat{\tau}_t) \qquad . \end{aligned} \tag{6.46}$$

Im Kapitel 4.2 betrachten wir SPREE (4.4) als eine Linearkombination der $\hat{\tau}^{\text{X}}_{Z,k}$, wobei X in $\hat{\tau}^{\text{X}}_{Z,k}$ D–DSE bzw. CHAP ist, $k \in 1, \ldots, K$. Mit Hilfe von (6.46) lässt sich die Varianz des SPREEs durch

$$\text{Var}\,(\hat{\tau}^{\text{SPREE}}_{Z,d}) = \sum_{k=1}^{K} \frac{\tau^2_{R,k,d}}{\tau^2_{R,k}}\ \text{Var}\,(\hat{\tau}^{\text{X}}_{Z,k}) \tag{6.47}$$

angeben, mit Varianzschätzer

$$\widehat{\text{Var}}\,(\hat{\tau}^{\text{SPREE}}_{Z,d}) = \sum_{k=1}^{K} \frac{\tau^2_{R,k,d}}{\tau^2_{R,k}}\ \widehat{\text{Var}}\,(\hat{\tau}^{\text{X}}_{Z,k}) \qquad , \tag{6.48}$$

wobei $\text{Var}\,(\hat{\tau}^{\text{X}}_{Z,k})$ bzw. $\widehat{\text{Var}}\,(\hat{\tau}^{\text{X}}_{Z,k})$ die Varianz bzw. Varianzschätzung des Schätzers $\hat{\tau}^{\text{X}}_{Z,k}$ ist. $\tau_{R,k,d}$ ist der Umfang der Schnittmenge $k \cap d$ des Registers R in Klasse k und des Registers R in Domain d.

Im einfachsten Fall, wo nur eine Klasse mehrere Domains überdeckt und SPREE mit (4.5) definiert ist, lässt sich die Varianzschätzung des SPREEs mithilfe (6.46) mit

$$\widehat{\text{Var}}\,(\hat{\tau}^{\text{SPREE}}_{Z,d}) = \frac{\tau^2_{R,k,d}}{\tau^2_{R,k}}\ \widehat{\text{Var}}\,(\hat{\tau}^{\text{X}}_{Z,k}) \tag{6.49}$$

gewinnen.

Die Varianzschätzung für SPREE einer Domain d in (4.6) ist einfach herzuleiten. Es gilt

$$\widehat{\text{Var}}\left(\hat{\tau}_{Z,d}^{\text{SPREE,str}}\right) = \sum_{k=1}^{K} \frac{\tau_{R,k,d}^2}{\tau_{R,k}^2} \widehat{\text{Var}}\left(\hat{\tau}_{Z,k}^{\text{X,str}}\right) \qquad , \tag{6.50}$$

mit

$$\widehat{\text{Var}}\left(\hat{\tau}_{Z,k}^{\text{X,str}}\right) = \sum_{h=1}^{H} \widehat{\text{Var}}\left(\hat{\tau}_{Z,k,h}^{\text{X}}\right) \qquad .$$

6.7 GSPREE–Varianz und GSPREE–Bootstrap–Varianzschätzung

Die Varianz des GSPREEs ist in keiner Literatur behandelt. Um die Varianz des GSPREEs zu schätzen, benutzen Zhang and Chambers (2004) die *Bootstrap–Methode*. Bootstrap ist in Shao (1996), Lohr (1999, S. 306) sowie in Särndal et al. (1992, S. 277) erläutert. Im Hintergrund dieser Methode steht die ursprüngliche Stichprobe vom Umfang n, aus denen B mal wiederholt eine *Bootstrap–Stichprobe* (in der Regel) vom Umfang n mit Zurücklegen gezogen wird.

In unserem konkreten Anwendungsfall der uneingeschränkten Zufallsauswahl ziehen wir aus der ursprünglichen Stichprobe vom Umfang n eine Bootstrap–Stichprobe b, um D–DSE und Chapman–Schätzer der Klassen zu gewinnen. Weiter wird das SPREE–Modell verwendet, um die Schätzungen für die interessierenden Domains zu ermitteln. Zum Schluss wird das GSPREE–Modell angewendet und $\hat{\boldsymbol{\tau}}_{Z,d}^{\text{GSPREE,b}}$ für die Bootstrap–Stichprobe b berechnet.

Laut Thompson (1992, S. 190) ist die Varianzschätzung

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\tau}}_{Z,d}^{\mathrm{GSPREE}}) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\boldsymbol{\tau}}_{Z,d}^{\mathrm{GSPREE,b}} - \hat{\bar{\boldsymbol{\tau}}}_{Z,d}^{\mathrm{GSPREE}}\right)^2 \quad ,$$

wobei

$$\hat{\bar{\boldsymbol{\tau}}}_{Z,d}^{\mathrm{GSPREE}} = \frac{1}{B}\sum_{b=1}^{B}\hat{\boldsymbol{\tau}}_{Z,d}^{\mathrm{GSPREE,b}} \quad . \tag{6.51}$$

Die Bootstrap–Methode für die Varianzschätzung des GSPREEs der geschichteten Zufallsauswahl ist genauer in Zhang and Chambers (2004, Sektion 5) beschrieben. Hier wird in der Schicht h eine Bootstrap–Stichprobe b vom Umfang $n_h - 1$ mit Zurücklegen gezogen. Als Begründung wird die asymptotische Validität angeführt. Die Bootstrap–Designgewichte sind dann definiert durch

$$w_{a,h}^{B} = \frac{w_{a,h}\, n_h}{n_h - 1} \quad ,$$

wobei $w_{a,h}$ das Designgewicht für Anschrift a der Schicht h und n_h die Anzahl der ausgewählten Anschriften innerhalb einer Schicht h sind.

Diese Prozedur wird unabhängig in jeder Schicht durchgeführt. In jeder Schicht werden die D–DSE und Chapman–Schätzer mit Bootstrap–Designgewichten berechnet und dies über die Schichten summiert, um die Schätzung einer Klasse zu gewinnen. Dann wird das SPREE–Modell verwendet und schließlich GSPREE angewendet. Für die Varianzschätzung des GSPREEs wird die oben erwähnte Formel benutzt.

Offensichtlich muss in allen Schichten die Bedingung $n_h \geq 2$ erfüllt sein, damit die Bootstrap–Stichprobe gezogen werden kann. Falls diese Bedingung nicht erfüllt ist, werden die Schichten erst zusammengefasst und die Bootstrap–Methode auf die neue Schichtung angewendet.

7 Aufbau der Simulationen und Ergebnisse

7.1 Simulations–Population

Um die Schätzverfahren zu testen, steht für diese Arbeit die Simulations–Population für das Bundesland Saarland zur Verfügung, die die Grundlage im Forschungsprojekt für den Zensus 2011 war und vom DACSEIS Projekt übernommen werden. Für dieses Bundesland gibt es den Datensatz SAL. In der folgenden Tabelle 7.1 sind alle Variablen aufgeführt, die für diese Arbeit verfügbar sind.

Es sind ein paar Regeln zu beachten. Die Länge der Vektoren muss übereinstimmen, jeder Vektor ist 1.057.915 Zeilen lang. Die Variablen GEM, ADR und KRS sind von eins bis zur Maximalanzahl durchnummeriert. Das Alter darf minimal 0, maximal 95 sein. Eine Person darf nicht sowohl Karteileiche als auch Fehlbestand sein. Die möglichen Kombinationen einer Person sind daher

KAL	FEB	Beschreibung
1	0	Karteileiche
0	0	weder Karteileiche noch Fehlbestand
0	1	Fehlbestand

Variable	Name	Codierung
AGE	Alter	1, . . . , 95
SEX	Geschlecht	0: Männlich
		1: Weiblich
NAT	Nationalität	0: Deutsch
		1: EU–Ausländer
		2: Ausländer außerhalb der EU–Länder
GEM	Gemeinde	1, . . . , 52
ADR	Anschriftsgröße	1, . . . , 248.832
KRS	Kreis	1, . . . , 6
FEB	Fehlbestand	0: kein Fehlbestand
		1: Fehlbestand
KAL	Karteileiche	0: keine Karteileiche
		1: Karteileiche

TABELLE 7.1: *Variablenübersicht.*

Die Information, ob es sich bei der Person um eine Karteileiche oder einen Fehlbestand handelt, ergibt sich aus Logit–Modellen. Dabei wird jeweils ein Vektor für Karteileichen (KAL) und für Fehlbestände (FEB) erzeugt. Die Logit–Modelle basieren auf den Daten des DACSEIS Projekts, wobei die tatsächliche Information zu Karteileichen und Fehlbeständen aus dem Zensustest 2001 des Statistisches Bundesamts modelliert wird. Bei der Modellierung waren die Ausprägungen der Variablen Geschlecht, Staatsangehörigkeit, Anschriftengröße und Alter klassifiziert und Dummy kodiert (Münnich et al. 2008).

Weiter wird die Nationalität im Vektor NAT der Simulations–Population leicht umkodiert. Bei dieser Umkodierung handelt es sich um die Verbindung der zwei unterschiedlichen nicht deutschen Nationalitäten NAT=1

und NAT=2 zu einer Gruppe. Die einheitliche Bezeichnung der Nationalität in dieser neuen Gruppe ist NAT=1.

In der Abbildung 7.1 sind mit einem sogenannten *Violinplot* (Hintze and Nelson 1998) die Anteile der Personen mit bestimmten Ausprägungen in 52 Gemeinden des Bundeslandes Saarland dargestellt. Der Violinplot baut auf dem Boxplot auf, der mit dem schwarzen Quadrat und nach oben und unten senkrechten Linien gezeichnet ist. Der weiße Punkt zeigt den Median an. Zusätzlich stellt der Violinplot an beiden Seiten des Boxplots die relative Verteilung der Häufigkeiten geglättet dar.

Es ist zu sehen, dass die Anteile der Männer (SEX=0) und Frauen (SEX=1) in allen 52 Gemeinden bei ca. 50% liegen. Demgegenüber ist der Anteil der nicht deutschen Population (NAT=1) sehr gering. Es handelt sich um ein seltenes Ereignis. Der Prozentsatz der nicht deutschen Personen in den Gemeinden ist in Tabelle A.1 zu finden.

R–Programme

Für das DSE–Verfahren wird ein R–Programm geschrieben, das für das Bundesland Saarland den D–DSE und Chapman–Schätzer für bestimmte Klassen nach (3.15) und (3.25) berechnet. Die Ergebnisse dieser Klassen dienen weiter zur Berechnung der interessierenden Subpopulationen (Domains). Diese Domains sind über Geschlecht, Nationalität und Altersgliederung definiert. Diese Altersgliederung ist die gleiche, wie die Altersklassen bei der Logit–Modellierung der Karteileichen und Fehlbestände. Die Altersgliederung der sieben AGE-Domains ist in Tabelle 7.2 beschrieben.

Das R–Programm ist durch die Schätzung für diese durch zwei Nationalitätskategorien, zwei Geschlechtskategorien und sieben Altersklassen gebildeten Domains ergänzt. Um die Schätzungen der Domains zu ermitteln,

AGE-Domain	Alter
1	0 - 17
2	18 - 24
3	25 - 29
4	30 - 39
5	40 - 49
6	50 - 64
7	65 - 95

TABELLE 7.2: *AGE-Domains der Bevölkerung in Deutschland.*

NAT=0	
AGE-Domains	Alter
1	0 - 17
2	18 - 24
3	25 - 29
4	30 - 39
5	40 - 49
6	50 - 64
7	65 - 95

NAT=1	
AGE-Domain	Alter
8	0 - 95

TABELLE 7.3: *AGE-Domains der deutschen und nicht deutschen Bevölkerung.*

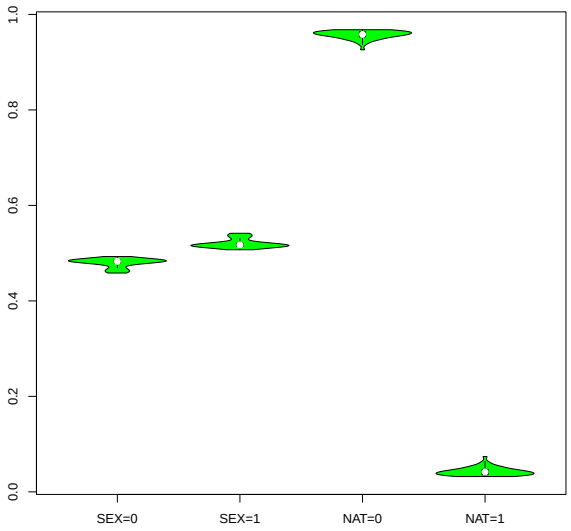


ABBILDUNG 7.1: *Violinplot für die Anteile der Frauen (SEX=1), Männer (SEX=0), deutsche Population (NAT=0) und nicht deutsche Population (NAT=1) über die 52 Gemeinden.*

wird SPREE nach (4.4) angewendet, nachdem der D–DSE und Chapman–Schätzer in den Klassen verwendet wurden.

Wegen der geringen Anzahl von Nichtdeutschen ist im R–Programm die Altersgliederung bei den Nichtdeutschen nicht so detailliert wie bei der deutschen Population. Die AGE-Domain der nicht deutschen Population ist über die ganze nicht deutsche Population verbreitet (s. Tabelle 7.3).

Der Vorteil des Chapman–Schätzers, immer ein Ergebnis für eine Klasse (und damit durch SPREE auch ein Ergebnis der Domain) zu bekommen, wird weiter ausgenutzt. Die SPREE absoluten Häufigkeiten jeder Domain,

aufgebaut auf Chapman Schätzungen der Klassen, werden als Pseudo–Werte in GSPREE (4.7) benutzt.

Als nächstes werden im R–Programm alle im Kapitel 5 angeführten alternativen Schätzer berechnet. D–DSE, Chapman–Schätzer, SPREE und GSPREE können so verglichen werden.

Weil es sich um eine Simulations–Population handelt, ist es möglich, die Ergebnisse einer Domain mit der Anzahl $\tau_{Z,d}$ der tatsächlich vorhandenen Personen in einer Domain d zu vergleichen. Die Güte des Schätzers wird durch die Wurzel aus dem mittleren quadratischen Fehler ermittelt, relativiert durch die wahre Anzahl $\tau_{Z,d}$ (engl. *relative root mean square error*, RRMSE). Für Domain d ist RRMSE definiert (Rao 2003, S. 62) durch

$$\text{RRMSE}_d\left(\hat{\tau}^{\text{X}}_{Z,d,r}\right) = \frac{\sqrt{\text{MSE}_d\left(\hat{\tau}^{\text{X}}_{Z,d,r}\right)}}{\tau_{Z,d}} \quad , \tag{7.1}$$

wobei der *mittlere quadratische Fehler* (engl. *mean square error*, MSE) für Domain d durch

$$\text{MSE}_d\left(\hat{\tau}^{\text{X}}_{Z,d,r}\right) = \frac{1}{R}\sum_{r=1}^{R}(\hat{\tau}^{\text{X}}_{Z,d,r} - \tau_{Z,d})^2 \tag{7.2}$$

definiert ist. R ist die Anzahl der Simulationen. X in $\hat{\tau}^{\text{X}}_{Z,d,r}$ ist D–DSE, CHAP, SPREE, GSPREE, GREG1, GREG2, SYN oder EBLUP. $\hat{\tau}^{\text{X}}_{Z,d,r}$ ist ein entsprechender Schätzer für die Anzahl $\tau_{Z,d,r}$ in der Domain d der Simulation r.

Die Güte des Schätzers wird auch durch den *absoluten–relativen–Bias* (engl. *absolute relative bias*, ARB) ermittelt. Definieren wir für Domain d den *rela-*

tiven–Bias (engl. *relative bias*, RB) als Bias relativiert durch die wahre Anzahl $\tau_{Z,d}$. Laut Rao (2003, S. 62) ist

$$\mathrm{RB}_d\left(\hat{\tau}^{\mathrm{X}}_{Z,d,r}\right) = \frac{\frac{1}{R}\sum_{r=1}^{R}\hat{\tau}^{\mathrm{X}}_{Z,d,r} - \tau_{Z,d}}{\tau_{Z,d}} = \frac{1}{R}\sum_{r=1}^{R}\left(\frac{\hat{\tau}^{\mathrm{X}}_{Z,d,r}}{\tau_{Z,d}} - 1\right) \quad .$$

Der absolute–relative–Bias ist dann

$$\mathrm{ARB}_d\left(\hat{\tau}^{\mathrm{X}}_{Z,d,r}\right) = \left|\mathrm{RB}_d\left(\hat{\tau}^{\mathrm{X}}_{Z,d,r}\right)\right| \quad . \tag{7.3}$$

Eine weitere Möglichkeit zur Auswahl von Schätzern bietet sich beim Einsatz von Konfidenzintervallschätzungen. Das *Konfidenzintervall* (Agresti 2002, S. 13) schließt einen Bereich um den Schätzern ein, der mit einer vorgegebenen Wahrscheinlichkeit $1-\alpha$ den untersuchten Schätzer überdeckt. Vorausgesetzt, dass ein unverzerrter Schätzer $\hat{\tau}_{Z,d}$ approximativ normalverteilt mit Erwartungswert $\tau_{Z,d}$ und Varianz $\mathrm{Var}(\hat{\tau}_{Z,d})$ ist. Man erhält das $100(1-\alpha)\%$ Konfidenzintervall

$$\left[\hat{\tau}_{Z,d} - z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\hat{\tau}_{Z,d})}\;;\; \hat{\tau}_{Z,d} + z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\hat{\tau}_{Z,d})}\right] \quad , \tag{7.4}$$

wobei $z_{1-\alpha/2}$ das $(1-\alpha/2)$–Quantil der Normalverteilung ist. Für einen verzerrten Schätzer gibt Särndal et al. (1992, S. 166) das Konfidenzintervall

$$\left[\hat{\tau}_{Z,d} - z_{1-\alpha/2}\sqrt{\widehat{\mathrm{MSE}}(\hat{\tau}_{Z,d})}\;;\; \hat{\tau}_{Z,d} + z_{1-\alpha/2}\sqrt{\widehat{\mathrm{MSE}}(\hat{\tau}_{Z,d})}\right] \tag{7.5}$$

an.

Bei der Simulationen ist es möglich zu untersuchen, wie viele ermittelten Konfidenzintervalle den wahren Wert tatsächlich überdecken. Diese *Überdeckungsrate* sollte mit dem theoretischen Konfidenzniveau $1-\alpha$ übereinstimmen. Im Folgenden wird der Wert $\alpha = 0,05$ verwendet. Damit erwar-

ten wir, dass 95% der ermittelten Konfidenzintervalle den wahren Wert überdecken.

7.2 Uneingeschränkte Zufallsauswahl

Nachdem aus der Grundgesamtheit eine Stichprobe gezogen und die Stichprobendaten erhoben wurden, lassen sich Schätzer testen. In Münnich et al. (2008, S. 29) sind unterschiedliche Stichprobendesigns beschrieben. Alle beziehen sich auf die Ziehung von Anschriften einer Gemeinde.

Im Folgenden richten wir unsere Aufmerksamkeit auf die uneingeschränkte Zufallsauswahl. Pro Gemeinde g werden

$$n_{<g>} = \begin{cases} 550 & \text{in der Gemeinde mit } \tau_{R,<g>} \geq 10.000 \\ 550 \, \dfrac{\tau_{R,<g>}}{\tau_{R,\ll k \gg}} & \text{sonst} \end{cases} \tag{7.6}$$

Anschriften durch uneingeschränkte Zufallsauswahl ohne Zurücklegen ausgewählt. $\tau_{R,<g>}$ bezeichnet die Anzahl der im Einwohnermeldeamt registrierten Personen der Gemeinde g, $\tau_{R,\ll k \gg}$ die Anzahl der im Einwohnermeldeamt registrierten Personen der Gemeinden des Kreises k, zu dem die Gemeinde g gehört.

Eine Übersicht $\tau_{R,<g>}$ aller Gemeinden im Datensatz SAL gibt Tabelle A.1 auf Seite 180 an. Die Anzahl der im Einwohnermeldeamt registrierten Personen eines Kreises k, $\tau_{R,\ll k \gg}$ sind in Tabelle A.2 auf der Seite 180 zu finden.

Zum Beispiel, gehört die Gemeinde Nr. 45 mit $\tau_{R,<45>} = 7.953$ registrierten Personen zum Kreis Nr. 6 mit $\tau_{R,\ll 6\gg} = 90.194$. Daher ist

$$n_{<45>} \quad = \quad 550\,\frac{\tau_{R,<45>}}{\tau_{R,\ll 6\gg}} \quad = \quad 550\,\frac{7.953}{90.194} \quad \doteq \quad 49 \qquad .$$

In Gemeinde Nr. 45 werden also 49 Anschriften durch uneingeschränkte Zufallsauswahl ohne Zurücklegen gezogen.

Im Rahmen des DACSEIS Projekts wurde die uneingeschränkte Zufallsauswahl ohne Zurücklegen als *Design 01* bezeichnet. Um die Grafiken, Tabellen und Beschriftung der Tabellen in dieser Arbeit übersichtlich zu machen, wird im Folgenden die Bezeichnung Design 01 für uneingeschränkte Zufallsauswahl ohne Zurücklegen benutzt.

Aus dem DACSEIS Projekt sind 100 Vektoren der Länge 248.832 vorhanden, die eine 1 an der Stelle der gezogenen Anschriften haben, falls diese Anschriften durch uneingeschränkte Zufallsauswahl ohne Zurücklegen ausgewählt waren. An der Stelle der nicht ausgewählten Anschriften haben die Vektoren eine 0. Diese 100 Vektoren (`SAL.GEM.design01.ADR.00r.RData`, wobei `r` von 00 bis 99 läuft) zusammen mit den Vektoren aus Tabelle 7.1 werden im `R`–Programm für die Berechnungen verwendet.

Weiter ist im Datensatz SAL der Vektor `SAL.IIP.design01.dat` der Länge 248.832 vorhanden, der die Designgewichte der Anschriften, w_a enthält. Bei uneingeschränkter Zufallsauswahl der Anschriften ohne Zurücklegen sind alle w_a in der Gemeinde identisch. Für die Designgewichte aller Anschriften in der Gemeinde Nr. 45 gilt

$$w_{a,<45>} \quad = \quad \frac{1}{n_{<45>}/N_{<45>}} \quad = \quad \frac{2.134}{49} \quad = \quad 43{,}55102 \qquad .$$

Die Designgewichte für alle Gemeinden sind im Anhang auf der Seite 181 aufgelistet.

Es ist auch der Vektor `SAL.IIP.design01.P.dat` der Länge 1.057.915 vorhanden, der die Designgewichte auf Personen–Ebene enthält. Designgewichte sind für alle Personen einer Anschrift konstant.

Solange es vom Kontext her klar ist, um welche Gemeinde es sich handelt, wird im Folgenden auf das g im Index verzichtet.

Um die Klassen für D–DSE und Chapman–Schätzer zu bilden, ist eine Schichtung der Population nötig. Die Schichtung der Population nach NAT=0 und NAT=1 sowie die Schichtung nach SEX=0 und SEX=1 ist im `R`–Programm festgelegt. Die Gruppierung des Alters kann man im `R`–Programm durch den Vektor `AGK.de` für die deutsche und `AGK.nd` für die nicht deutsche Population festlegen.

Bei diesem Design (Design 01: Uneingeschränkte Zufallsauswahl) werden drei unterschiedliche Versionen der Vektoren `AGK.de` und `AGK.nd` präsentiert. Die grafische Darstellungen der Versionen sind im Anhang G zu sehen.

7.2.1 Version 1

Bezüglich der deutschen und nicht deutschen Population wird zunächst die Altersgliederung der Klassen (AGE-Klassen) genauso klassifiziert, wie die AGE-Domains in Tabelle 7.3. Eine Übersicht der Klassen in der Population gibt Tabelle 7.4.

Weil die Definition der Klassen und der Domains in dieser Version übereinstimmt, liefert SPREE für die interessierenden Domains genau die gleichen Ergebnisse wie D–DSE und der Chapman–Schätzer für diese Klassen. In diesem Fall wird also keine anschließende Durchführung des SPREEs für die Domains nötig.

NAT	SEX	AGE-Klassen						
		1	2	3	4	5	6	7
0	1							
	0							

NAT	SEX	AGE-Klasse
		8
1	1	
	0	

TABELLE 7.4: *Die 16 Klassen der deutschen und nicht deutschen Bevölkerung in Version 1.*

D–DSE

Es kam leider in dieser Version bei dieser Altersgliederung in Gemeinden unter 10.000 Einwohnern oft vor, dass keine Person in einigen Klassen ausgewählt war. In diesem Fall teilen wir in D–DSE (3.15) „0/0“, was ein `NaN` für die Klasse gibt (und damit ein `NaN` für die Domain).

Als Beispiel nennen wir die Klasse bzw. Domain 0-1-2. Die Nummerierung 0-1-2 dient als Schlüssel und bezeichnet eine Klasse bzw. Domain, die durch NAT=0, SEX=1, AGE-Domain=2 definiert ist, d.h. deutsche Frauen im Alter von 18-24.

Bei der zwölften Stichprobe in der Gemeinde Nr. 45 wohnte niemand der Klasse 0-1-2 an den ausgewählten Anschriften. Dazu war auch niemand

der Klasse 0-1-2 an diesen ausgewählten Anschriften im Register R. $\forall a \in S$ ist

$$\begin{aligned} \tau_{Z,\text{0-1-2},a} &= 0 \\ \tau_{R,\text{0-1-2},a} &= 0 \quad . \end{aligned}$$

Daraus folgt

$$\begin{aligned} \hat{\tau}_{Z,\text{0-1-2}} &= \sum_{a \in S} \tau_{Z,\text{0-1-2},a}\, w_a = 0 \cdot 43{,}55102 = 0 \\ \hat{\tau}_{R,\text{0-1-2}} &= \sum_{a \in S} \tau_{R,\text{0-1-2},a}\, w_a = 0 \cdot 43{,}55102 = 0 \quad . \end{aligned}$$

Die Anzahl der registrierten Personen der Klasse 0-1-2 an allen Anschriften der Gemeinde Nr. 45 ist $\tau_{R,\text{0-1-2}} = 205$. Daher gibt (3.15)

$$\hat{\tau}^{\text{D-DSE}}_{Z,\text{0-1-2}} = \tau_{R,\text{0-1-2}} \frac{\widehat{\tau}_{Z,\text{0-1-2}}}{\widehat{\tau}_{R,\text{0-1-2}}} = 205\, \frac{0}{0} = \texttt{NaN} \tag{7.7}$$

eine Schätzung für die Zahl der deutschen Frauen der Gemeinde Nr. 45 im Alter von 18-24 an.

Es kam auch vor, dass in einer Anschrift eine Person als Fehlbestand der Klasse gefunden wurde, aber keine Person aus den Kategorien 00 oder K der Klasse an dieser Anschrift wohnt. In diesem Fall teilen wir in D–DSE (3.15) „x/0“, wo x beliebige positive Zahl der Fehlbestände ist, was `Inf` für die Klasse bzw. für die Domain ergibt.

Zum Beispiel wohnte in der Stichprobe $r = 72$ in der Gemeinde Nr. 37, wo $w_a = \frac{N_{<37>}}{n_{<37>}} = \frac{2.232}{22} = 101{,}4545$ (s. Tabelle auf der Seite 180), eine Person der Klasse 0-0-3 (deutsche Männer im Alter von 25-29) an der aus-

gewählten Anschrift $a = 75.422$, war aber nicht im Register R (ist also ein Fehlbestand). $\forall a \in S$ gilt

$$\begin{aligned} \tau_{Z,\text{0-0-3},a} &= \begin{cases} 1 & \text{für } a = 75.422 \\ 0 & \text{sonst} \end{cases} \\ \tau_{R,\text{0-0-3},a} &= 0 \quad . \end{aligned}$$

Daraus folgt

$$\begin{aligned} \hat{\tau}_{Z,\text{0-0-3}} &= \sum_{a \in S} \tau_{Z,\text{0-0-3},a}\, w_a = 1 \cdot 101{,}4545 = 101{,}4545 \\ \hat{\tau}_{R,\text{0-0-3}} &= \sum_{a \in S} \tau_{R,\text{0-0-3},a}\, w_a = 0 \cdot 101{,}4545 = 0 \quad . \end{aligned}$$

Die Anzahl der registrierten Personen der Klasse 0-0-3 an allen Anschriften der Gemeinde Nr. 37 ist $\tau_{R,\text{0-0-3}} = 212$. Daher gibt (3.15)

$$\hat{\tau}^{\text{D-DSE}}_{Z,\text{0-0-3}} = \tau_{R,\text{0-0-3}}\, \frac{\widehat{\tau}_{Z,\text{0-0-3}}}{\widehat{\tau}_{R,\text{0-0-3}}} = 212\, \frac{101{,}4545}{0} = \texttt{Inf} \tag{7.8}$$

eine Schätzung für die deutschen Männer der Gemeinde Nr. 37 im Alter von 25-29 an.

Die D–DSE Spalte der Tabelle 7.5 (s. Seite 103) gibt die genaue Anzahl aller `NaN` und `Inf` der Population pro Domain an, ausgehend von jeweils 100 Stichproben in allen 52 Gemeinden. Als Beispiel schauen wir die Klasse mit Schlüssel 0-1-2 (deutsche Frauen im Alter von 18-24) genauer an. Von insgesamt 5.200 Stichproben (jeweils 100 Stichproben in 52 Gemeinden) werden 47 mal keine Frau aus dieser Klasse ausgewählt, was ein `NaN` für die Klasse 0-1-2 und damit für die Domain 0-1-2 ergibt. Es passierte nur einmal in der Klasse 0-1-2, dass bei der Stichprobe Frauen als Fehlbestände der Klasse 0-1-2 gefunden wurden, aber keine Frau der Klasse 0-1-2 an den

ausgewählten Anschriften registriert ist, was ein `Inf` für die Klasse 0-1-2 und damit für die Domain 0-1-2 ergibt.

Chapman–Schätzer

Im Kapitel 3.2 haben wir die Theorie zum Chapman–Schätzer entwickelt, die die Situationen bereinigt, wenn beim D–DSE im Nenner und/oder Zähler eine „0“ steht. In der CHAP Spalte der Tabelle 7.5 (s. Seite 103) ist zu sehen, dass es bei der Verwendung dieses Schätzers keine Probleme mehr gibt, die auf das Vorliegen von `NaN` oder `Inf` zurückzuführen sind.

In der Gemeinde Nr. 45, Klasse 0-1-2 (deutsche Frauen im Alter von 18-24), in der D–DSE versagt, ist der Chapman–Schätzer (3.25)

$$\begin{aligned}\hat{\tau}^{\text{CHAP}}_{Z,\text{0-1-2}} &= (\tau_{R,\text{0-1-2}}+1)\,\frac{\widehat{\tau}_{Z,\text{0-1-2}}+1}{\widehat{\tau}_{R,\text{0-1-2}}+1}-1 \quad = \quad (205+1)\,\frac{0+1}{0+1}-1\\ &= 205 \quad .\end{aligned}$$

Im Vergleich zum D–DSE kommt bei dieser Schätzung kein `NaN` vor, in solchen Situationen ist der Chapman–Schätzer gleich der Anzahl der registrierten Personen, $\hat{\tau}^{\text{CHAP}}_{Z,k} = \tau_{R,k}$.

Auch im Fall, in dem D–DSE ein `Inf` liefert, liefert der Chapman–Schätzer ein Ergebnis. Im Beispiel der Klasse 0-0-3 in der Gemeinde Nr. 37 ist der Chapman–Schätzer

$$\begin{aligned}\hat{\tau}^{\text{CHAP}}_{Z,\text{0-0-3}} &= (\tau_{R,\text{0-0-3}}+1)\,\frac{\widehat{\tau}_{Z,\text{0-0-3}}+1}{\widehat{\tau}_{R,\text{0-0-3}}+1}-1\\ &= (212+1)\,\frac{101{,}4545+1}{0+1}-1\\ &= 21.821{,}8085 \quad .\end{aligned} \tag{7.9}$$

Domain	D–DSE		CHAP		CHAP/ GSPREE		GREG1		GREG2		SYN		EBLUP	
	NaN	Inf	NaN	Inf	NaN	Inf	NaN	Inf	NaN	Inf	NaN	Inf	NaN	Inf
0-0-1	0	0	0	0	0	0	0	0	0	0	0	0	8	0
0-1-1	2	0	0	0	0	0	0	0	2	0	0	0	10	0
0-0-2	29	1	0	0	0	0	0	0	29	1	0	0	36	0
0-1-2	47	1	0	0	0	0	0	0	47	1	0	0	53	0
0-0-3	35	2	0	0	0	0	0	0	35	2	0	0	41	0
0-1-3	28	1	0	0	0	0	0	0	28	1	0	0	35	0
0-0-4	0	0	0	0	0	0	0	0	0	0	0	0	8	0
0-1-4	0	0	0	0	0	0	0	0	0	0	0	0	8	0
0-0-5	0	0	0	0	0	0	0	0	0	0	0	0	8	0
0-1-5	2	0	0	0	0	0	0	0	2	0	0	0	10	0
0-0-6	0	0	0	0	0	0	0	0	0	0	0	0	8	0
0-1-6	0	0	0	0	0	0	0	0	0	0	0	0	8	0
0-0-7	0	0	0	0	0	0	0	0	0	0	0	0	8	0
0-1-7	0	0	0	0	0	0	0	0	0	0	0	0	8	0
1-0-8	78	5	0	0	0	0	0	0	78	5	0	0	84	0
1-1-8	108	6	0	0	0	0	0	0	108	6	0	0	114	0

TABELLE 7.5: *Anzahl der* `NaN` *und* `Inf` *pro Domain in der Gesamtpopulation. Design 01, Version 1. Die Nummerierung einer Domain dient als Schlüssel. Zum Beispiel: 0-0-2 bezeichnet eine Domain, die durch NAT=0, SEX=0, AGE-Domain=2 definiert ist.*

Domain	Gemeinde									Total
	15	17	36	37	40	45	47	49	50	
0-0-1	0	0	0	0	0	0	0	0	0	0
0-1-1	0	0	1	0	1	0	0	0	0	2
0-0-2	2	2	9	7	1	1	0	0	7	29
0-1-2	5	5	14	9	9	1	1	1	2	47
0-0-3	3	5	7	13	2	1	1	0	3	35
0-1-3	3	1	8	10	3	0	0	0	3	28
0-0-4	0	0	0	0	0	0	0	0	0	0
0-1-4	0	0	0	0	0	0	0	0	0	0
0-0-5	0	0	0	0	0	0	0	0	0	0
0-1-5	0	0	2	0	0	0	0	0	0	2
0-0-6	0	0	0	0	0	0	0	0	0	0
0-1-6	0	0	0	0	0	0	0	0	0	0
0-0-7	0	0	0	0	0	0	0	0	0	0
0-1-7	0	0	0	0	0	0	0	0	0	0
1-0-8	6	5	19	15	12	6	3	6	6	78
1-1-8	6	8	33	18	15	7	4	7	10	108

TABELLE 7.6: *Anzahl der* `NaN` *in D–DSE pro Domain und pro Gemeinde. Design 01, Version 1.*

Domain	Gemeinde									Total
	15	17	36	37	40	45	47	49	50	
0-0-2	0	0	0	0	1	0	0	0	0	1
0-1-2	0	0	0	0	0	1	0	0	0	1
0-0-3	0	0	1	1	0	0	0	0	0	2
0-1-3	0	0	0	0	1	0	0	0	0	1
1-0-8	0	0	0	1	1	2	0	1	0	5
1-1-8	0	1	2	1	0	1	0	0	1	6

TABELLE 7.7: *Anzahl der* `Inf` *in D–DSE pro Domain und pro Gemeinde. Design 01, Version 1.*

Der Chapman–Schätzer liefert zwar ein Ergebnis, diese Schätzung ist aber infolge der Gewichtung w_a weit weg von der Realität (vgl. mit $\tau_{Z,0\text{-}0\text{-}3} = 206$). Diese extremen Ergebnisse zeigen sich auch in der grafischen Darstellung auf den nächsten Seiten.

CHAP/GSPREE

Im GSPREE–Modell werden als Pseudo–Werte die absoluten Häufigkeiten aller Domains in allen 52 Gemeinden verwendet, die mittels Chapman–Schätzer berechnet sind. Bezeichnen wir also diese Schätzer als CHAP/GSPREE, damit es klar ist, welcher Schätzer zur GSPREE Berechnung benutzt ist. Eine Übersicht über die verwendeten Hilfsinformationen der Schätzer findet man in Tabelle D.1 auf der Seite 193.

Bei den Berechnungen des Chapman–Schätzers ist in dieser Arbeit die Schichtung der Population nach NAT=0 und NAT=1 sowie die Schichtung nach SEX=0 und SEX=1 festgelegt (s. Text auf Seite 98). Auch bei dem GSPREE–Modell wird diese Idee beibehalten. Erstens wird das GSPREE–Modell auf die geschätzten absoluten Häufigkeiten aller AGE-Domains der deutschen Frauen (NAT=0, SEX=1) in allen 52 Gemeinden angewendet. Zweitens wird das GSPREE–Modell auf die geschätzten absoluten Häufigkeiten aller AGE-Domains der deutschen Männer (NAT=0, SEX=0) in allen 52 Gemeinden anwendet. Wegen nur einer AGE-Domain bei den nicht deutschen Frauen und einer AGE-Domain bei der nicht deutschen Männer ist hier das GSPREE–Modell nicht anwendbar. Bei der nicht deutschen Population wird als CHAP/GSPREE der CHAP benutzt.

In der Spalte CHAP/GSPREE treten keine Probleme mit `NaN` oder `Inf` für diesen Schätzer auf, vor allem, weil für diese Schätzung die Ergebnisse des Chapman–Schätzers verwendet wurden und dieser Schätzer diese Probleme nicht hat. In die CHAP/GSPREE Schätzungen gingen aber die

Chapman–Schätzer von allen 52 Gemeinden ein. Daher wirken sich die extremen Chapman–Schätzer der kleinen Gemeinden auch bei CHAP/GSPREE in den großen Gemeinden aus.

Alternative Schätzer

Unter den alternativen Schätzern ist die Situation mit `NaN` und `Inf` unterschiedlich. Es ist interessant, dass der GREG2–Schätzer die Situation in D–DSE wiederspiegelt, vor allem, weil bei der gleichen Definition der Domains und Klassen (Version 1) der GREG2–Schätzer gleich dem D–DSE ist (siehe auch Seite 61).

Grafische Darstellung allgemein

Für die grafische Präsentation der Ergebnisse werden für diese Arbeit nur zwei Gemeinden mit $\tau_{R,<g>} < 10.000$ und zwei Gemeinden mit $\tau_{R,<g>} \geq 10.000$ registrierten Einwohnern genauer untersucht.

Die zwei linken Boxen der Abbildung H.1 bzw. H.2 auf Seite 203 bzw. 204 zeigen für große bzw. kleine Gemeinden die Mittelwerte der RRMSEs für die 16 Domains einer Gemeinde auf der x–Achse versus der Varianz der 100 Stichprobenmittelwerte der RRMSEs für die 16 Domains einer Gemeinde auf der y–Achse. Am Besten sollten die Schätzungen der einzelnen Gemeinen einen kleinen Mittelwert der RRMSE und gleichzeitig eine kleine Varianz der RRMSE aufweisen. Daher untersuchen wir detaillierter die Gemeinde Nr. 37 und die Gemeinde Nr. 45 (beide unter 10.000 Einwohner). Um die Ergebnisse in Gemeinden über 10.000 Einwohner zu untersuchen, werden die Gemeinde Nr. 1 und die Gemeinde Nr. 35 ausgewählt. Tabelle 7.8 gibt eine Übersicht über die Anzahl der im Einwohnermeldeamt registrierten Personen in diesen vier Gemeinden.

GEM	τ_R
1	157.205
35	11.150
37	8.221
45	7.953

TABELLE 7.8: *Übersicht der vier untersuchten Gemeinden.*

Die Gemeinden mit über 10.000 registrierten Einwohnern gelten als *große* Gemeinden, die Gemeinden unter 10.000 registrierten Einwohnern als *kleine* Gemeinden.

Grafische Darstellung der Boxplots

Die grafischen Darstellungen der Verteilung aller Schätzer befindet sich auf den Seiten 108–109, wo jeweils *Boxplots* für alle sieben Schätzer in allen 16 Domains der Gemeinden Nr. 1, 35, 37 und 45 dargestellt sind. Die Schätzungen der Domains sind durch den Mittelwert der 100 Stichproben in der entsprechenden Domain normiert.

Jeder Boxplot besteht aus einem Rechteck und zwei Linien. Das Rechteck umfasst die mittleren 50% der Daten und zeigt den Median als einen schwarzen Punkt innerhalb des Rechtecks. Das Rechteck ist durch das obere und das untere Quartil begrenzt. Jede äußere Linie (Whisker) ist maximal 1,5 mal die Länge des Rechtecks. Die Werte außerhalb der Whisker werden als *Ausreißer* bezeichnet. Um die Boxplots besser vergleichen zu können, ist die Skala innerhalb einer Version in einem Design für die zwei

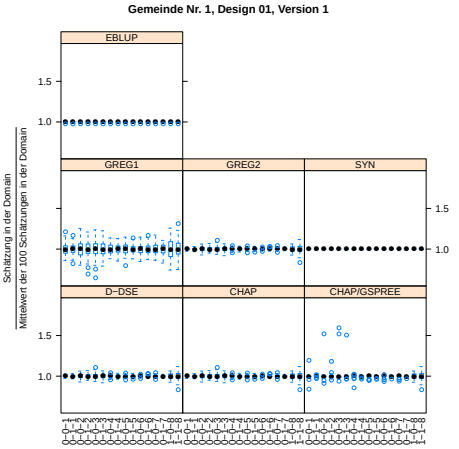


ABBILDUNG 7.2: *Boxplots der geschätzten Totalwerte der Domains in der Gemeinde Nr. 1. Design 01, Version 1.*

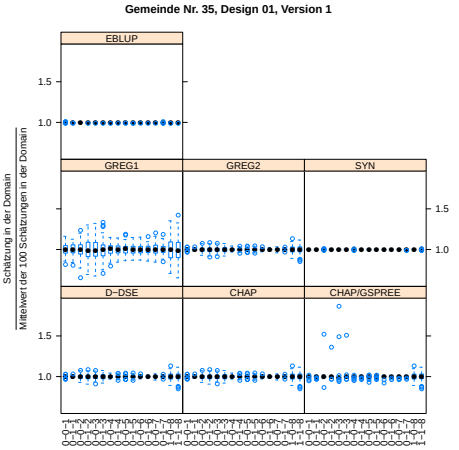


ABBILDUNG 7.3: *Boxplots der geschätzten Totalwerte der Domains in der Gemeinde Nr. 35. Design 01, Version 1.*

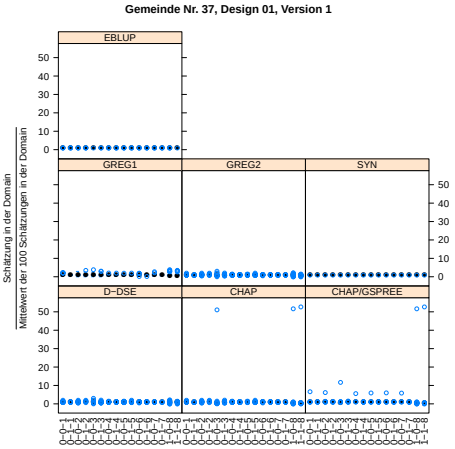


ABBILDUNG 7.4: *Boxplots der geschätzten Totalwerte der Domains in der Gemeinde Nr. 37. Design 01, Version 1.*

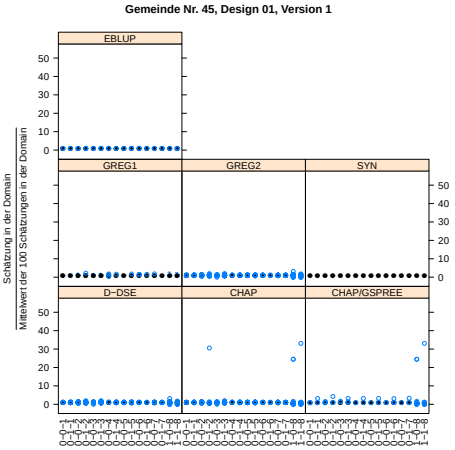


ABBILDUNG 7.5: *Boxplots der geschätzten Totalwerte der Domains in der Gemeinde Nr. 45. Design 01, Version 1.*

großen Gemeinden fest. Für die zwei kleinen Gemeinden gibt es eine andere Skala, die innerhalb einer Version in einem Design auch fest ist.

Man kann sehen, dass die Streuung beim CHAP/GSPREE in den großen Gemeinden (Seite 108) und beim Chapman–Schätzer, CHAP/GSPREE in den kleinen Gemeinden (Seite 109) nicht gleichmäßig auf beiden Seiten des Medians verteilt ist. Wegen der Ausreißer sind auch die Boxplots der anderen Schätzer nach unten gedrückt, in den kleinen Gemeinden sogar so deutlich, dass die Boxplots nicht mehr lesbar sind. Man beachte, dass in den kleinen Gemeinden deutlich mehr Ausreißer als in den großen Gemeinden, vorkommen. Weiter kann man sehen, dass in den großen Gemeinden die Boxplots des D–DSEs und Chapman–Schätzers vergleichbar sind. Anderseits haben in den kleinen Gemeinden die Boxplots des Chapman–Schätzers einige positive Ausreißer im Vergleich mit dem D–DSE. Die Ursache für die Ausreißer wurde schon erläutert.

Grafische Darstellung der RRMSEs

Die Güte sowohl der untersuchten als auch der alternativen Schätzer für jede Domain der Gemeinden Nr. 1, 35, 37 und 45 wird mittels des RRMSE dargestellt. Die vier Grafiken in Abbildung 7.6 präsentieren die RRMSEs des Stichprobendesigns 01 in vier Gemeinden für die Version 1. Die Beschreibung im Folgenden ist aber auch für weitere Stichprobendesigns und Versionen gültig.

An der x–Achse jeder Grafik befinden sich alle Schätzer, die wir in dieser Arbeit in Betracht ziehen. An der y–Achse links ist die Skala des RRMSEs. Diese Skala wird im Rahmen einer Version in einem Design für die zwei großen Gemeinden festgehalten. Für die zwei kleinen Gemeinden gibt es eine andere Skala, die innerhalb einer Version in einem Design fest ist.

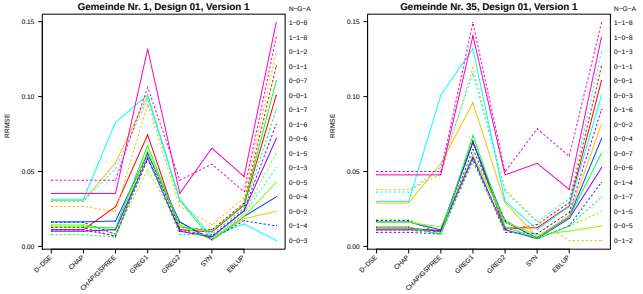
Jede Grafik enthält genau 16 Linien mit unterschiedlichen Farben und unterschiedlichen Arten. Jede Farbe entspricht einen von acht AGE-Domains, die Farbstellung einer Domain bleibt in dieser Arbeit fest. Die Art der Linien ist so ausgewählt, dass die durchgezogenen Linien die Frauen (SEX=1) darstellen, die gestrichelten Linien die Männer (SEX=0).

Die 16 Linien repräsentieren 16 Domains, 14 in der deutschen Population und zwei in der nicht deutschen Population (s. Tabelle 7.3, die die Altersgliederung für ein Geschlecht definiert).

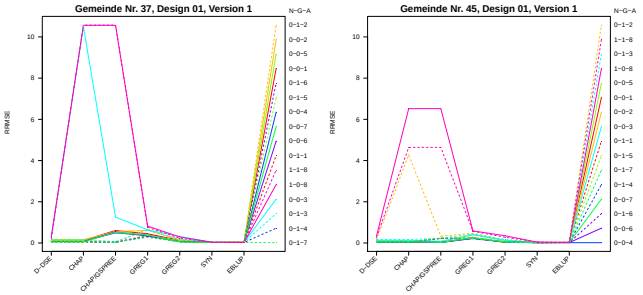
Alle Linien gehen von links nach rechts über D–DSE, Chapman–Schätzer, CHAP/GSPREE, GREG1–Schätzer, GREG2–Schätzer, Verhältnis–synthetischen Schätzer, EBLUP und enden an der y–Achse rechts. Hier rechts befindet sich zu jeder Linie eine Beschriftung dieser Linie, die eine Domain genau definiert.

Die erste Ziffer in der Beschriftung ist die Nationalität, 0 für deutsch, 1 für nicht deutsch. Die zweite Ziffer ist das Geschlecht, 0 für die Männer, 1 für die Frauen. Letzte Ziffer ist die AGE-Domain, 1 bis 7 für die AGE-Domains in der deutschen Population, AGE-Domain 8 ist für die nicht deutsche Population. Zum Beispiel bedeutet die Beschriftung 0-0-2 eine Domain, die durch NAT=0, SEX=0, AGE-Domain=2 definiert ist, d.h. deutsche Männer im Alter von 18-24.

Die Probleme, die sich aus dem Vorhandensein von `NaN` und `Inf` ergeben, wirken sich auf die Berechnung der RRMSEs aller 100 Stichproben aus. Aus diesem Grund basieren die RRMSEs jeder Domain jedes Schätzers auf der Anzahl der Stichproben, in denen weder `NaN` noch `Inf` vorkommen. Tabelle 7.5 gibt die gesamten Anzahl der `NaN` und `Inf` pro Domain im Saarland. Die ausführlichen Anzahlen der `NaN` bzw. `Inf` pro Domain pro Gemeinde findet man in Tabelle 7.6 bzw. Tabelle 7.7. Durch Vergleich dieser zwei Tabellen mit der Tabelle A.1 auf Seite 180 bemerkt man, dass die



(a) RRMSEs in der Gemeinde Nr. 1. (b) RRMSEs in der Gemeinde Nr. 35.



(c) RRMSEs in der Gemeinde Nr. 37. (d) RRMSEs in der Gemeinde Nr. 45.

ABBILDUNG 7.6: *RRMSEs für die Domains in den Gemeinden Nr. 1, 35, 37 und 45. Design 01, Version 1. 16 Linien in jeder Grafik repräsentieren 14 Domains für die deutsche Population und zwei für die nicht deutsche Population. Die Beschriftung jeder Linie definiert genau eine Domain. Die erste Ziffer in der Beschriftung ist die Nationalität, 0 für deutsch, 1 für nicht deutsch. Die zweite Ziffer ist das Geschlecht, 0 für die Männer, 1 für die Frauen. Letzte Ziffer ist die AGE-Domain, 1 bis 7 für die AGE-Domains in der deutschen Population, AGE-Domain 8 ist für die nicht deutsche Population. Zum Beispiel bedeutet die Beschriftung 0-0-2 eine Domain, die durch NAT=0, SEX=0, AGE-Domain=2 definiert ist, d.h. deutsche Männer im Alter von 18-24.*

Probleme mit `NaN` und `Inf` nur in kleinen Gemeinden vorkommen. Daher sind die RRMSEs für D–DSE in allen Domains der Gemeinden Nr. 1 und 35 (große Gemeinden) über 100 Stichproben gerechnet, die RRMSEs für D–DSE zum Beispiel in der Domain 0-0-3 der Gemeinde Nr. 37 über 100-13-1=86 Stichproben.

Die Abbildung 7.6 gibt eine Übersicht für alle RRMSEs der Version 1 in den Gemeinden Nr. 1, 35, 37 und 45 an, wobei die ersten zwei Grafiken die großen Gemeinden repräsentieren, die letzten zwei die kleinen Gemeinden. Hingewiesen sei auf den maximalen gemeinsamen RRMSE Wert 0,15 in den großen Gemeinden in (a) und (b), sowie den maximalen gemeinsamen RRMSE Wert 10 für die kleinen Gemeinden in (c) und (d).

Man kann sehen, dass die RRMSEs des GREG1–Schätzers in großen Gemeinden in allen 16 Domains deutlich am höchsten sind. Ohne einige extreme RRMSEs in CHAP oder CHAP/GSPREE wären die RRMSEs des GREG1–Schätzers in kleinen Gemeinden auch am höchsten. Im Vergleich mit dem GREG2–Schätzer ist eine Verbesserung der RRMSEs bei diesem Schätzer deutlich sichtbar. Daraus folgt, dass das Verhältnis $\frac{\tau_{R,d}}{\hat{\tau}_{R,d}}$ im GREG2–Schätzer (5.12) besser die Realität widerspiegelt, als das Verhältnis $\frac{\tau_R}{\hat{\tau}_R}$ im GREG1–Schätzer (5.9).

Man kann in den großen Gemeinden sehen, dass die zwei Linien für die nicht deutsche Population (die Beschriftung an der rechten Seite beginnt mit 1) ganz oben liegen, mit paar Ausnahmen in RRMSEs für CHAP/GSPREE den anderen Domains. Also sind alle Schätzer der Domains für die nicht deutsche Population sehr schlecht.

In der kleinen Gemeinde Nr. 37 sind die RRMSEs für CHAP und CHAP/GSPREE sehr hoch für die nicht deutschen Männer (Domain 1-0-8) und für die deutschen Männer im Alter von 25-29 (Domain 0-0-3), in der kleinen Gemeinde Nr. 45 für die nicht deutsche Population (Domains 1-0-8, 1-1-

8) und für die deutschen Frauen im Alter von 18-24 (Domain 0-1-2). Diese Schwankungen werden durch die Gewichtung der Anschriften und die Anzahl der `Inf` in den einzelnen Domains der Gemeinden Nr. 37 und 45 verursacht (s. Beispiel in (7.9) auf der Seite 102 und Tabelle 7.7).

Auch die hohen Schwankungen in den RRMSEs für CHAP/GSPREE in den großen Gemeinden sind durch `Inf` in den kleinen Gemeinden erklärbar. CHAP/GSPREE verwenden die Chapman Schätzungen aller 52 Gemeinden, um die GSPREE für alle Domains zu berechnen. Da bei manchen Stichproben die sehr schlechten Chapman Schätzungen der Domains in den kleinen Gemeinden verwendet werden, sind auch die schlechten Ergebnisse für GSPREE in den großen Gemeinden die Folge. In zwei Domains der nicht deutschen Population bleibt CHAP/GSPREE konstant, weil dieser Schätzer in diesen Domains nicht anwendbar war. An diesen Stellen wird der Chapman–Schätzer wieder verwendet.

Man sieht auch, dass in allen vier Gemeinden die RRMSEs des Chapman–Schätzers praktisch identisch und damit vergleichbar mit den RRMSEs des D–DSE sind, außer die Domains 1-0-8, 0-0-3 in der Gemeinde Nr. 37 und 1-0-8, 1-1-8, 0-1-2 in der Gemeinde Nr. 45.

Unter allen untersuchten Schätzern hat der Verhältnis–synthetische Schätzer einen kleinen RRMSE.

Absoluter–relativer–Bias

Die Tabellen auf Seite 214 enthalten die ARBs in jeder Domain der Gemeinden Nr. 1, 35, 37 und 45. Für die ARB Definition siehe (7.3). Die Mittelwerte über alle Domains in einzelnen Gemeinden sind in Tabelle 7.9 zu sehen. Sowohl in kleinen als auch in großen Gemeinden sind ARB für D–DSE bzw. GREG2–Schätzer am kleinsten. ARB für Chapman–Schätzer in den großen

Gemeinden ist vergleichbar mit dem ARB für D–DSE, in den kleinen Gemeinden ist der ARB schlechter.

GEM	D–DSE	CHAP	CHAP/ GSPREE	GREG1	GREG2	SYN	EBLUP
1	0,00156	0,00156	0,00341	0,00630	0,00156	0,01218	0,02430
35	0,00211	0,00211	0,00619	0,00575	0,00211	0,01494	0,02248
37	0,01171	0,20479	0,16858	0,02877	0,01171	0,01312	0,01779
45	0,00643	0,11478	0,09602	0,03607	0,00643	0,01186	0,02114

TABELLE 7.9: *Mittelwert der ARBs über alle Domains in den Gemeinden Nr. 1, 35, 37 und 45. Design 01, Version 1.*

Varianzschätzung und MSE–Schätzung

Die Grafiken im Anhang H auf den Seiten 206-207 zeigen die relativen Verzerrungen der Varianzschätzungen auf Gemeindebasis, d.h. die Summe der Varianzschätzungen des Schätzers $\hat{\tau}^{\mathrm{X}}_{Z,d,g,r}$ über die Domains $d = 1, \ldots, D$ einer Gemeinde g für die Stichprobe $r = 1, \ldots, 100$ dividiert durch die Summe der Varianzen des Schätzers. Mathematisch geschrieben

$$\frac{\sum_{d=1}^{D} \widehat{\mathrm{Var}}(\hat{\tau}^{\mathrm{X}}_{Z,d,g,r})}{\sum_{d=1}^{D} \mathrm{Var}(\hat{\tau}^{\mathrm{X}}_{Z,d,g})} \quad , \tag{7.10}$$

wobei X in $\hat{\tau}^{\mathrm{X}}_{Z,d,g,r}$ bzw. $\hat{\tau}^{\mathrm{X}}_{Z,d,g}$ entweder D–DSE, CHAP, SPREE, GREG1, GREG2 oder SYN ist. Im Fall des EBLUPs ersetzt $\widehat{\mathrm{MSE}}(\hat{\tau}^{\mathrm{EBLUP}}_{Z,d,g,r})$ den $\widehat{\mathrm{Var}}(\hat{\tau}^{\mathrm{X}}_{Z,d,g,r})$ bzw. $\mathrm{MSE}(\hat{\tau}^{\mathrm{EBLUP}}_{Z,d,g})$ die $\mathrm{Var}(\hat{\tau}^{\mathrm{X}}_{Z,d,g})$. Die Varianz des GSPREEs ist in keiner Literatur behandelt. Beim GSPREE wird daher im Nenner (7.10)

der Mittelwert der 100 Varianzschätzungen berechnet, d.h. $\frac{1}{100}\sum_{r=1}^{R}\sum_{d=1}^{D}\widehat{\text{Var}}(\hat{\tau}^{\text{GSPREE}}_{Z,d,g,r})$.

Wegen der unterschiedlichen Skala beim Jackknife für Chapman–Schätzer in den großen und kleinen Gemeinden sind die Boxplots getrennt abgebildet für die großen Gemeinden (Abbildung H.3) und für die kleinen Gemeinden (Abbildung H.4). Die Boxplots für Bootstrap–Varianzschätzung von CHAP/GSPREE sind für alle Gemeinden in Abbildung H.5 dargestellt.

Allgemein können wir über die Boxplots in großen Gemeinden sagen, dass der geschätzte MSE des EBLUPs sehr weit weg von dem wahren MSE liegt. In den kleinen Gemeinden überschätzt der Jackknife des Chapman–Schätzers die wahre Varianz.

Die Abbildungen 7.7-7.10 auf den nächsten Seiten zeigen die relative Verzerrung der Varianzschätzungen auf Domainbasis, d.h. die Varianzschätzungen des Schätzers $\hat{\tau}^{X}_{Z,d,g,r}$ jeder Domain d der Gemeinde g für die Stichprobe $r = 1, \ldots, 100$ dividiert durch ihre Varianz in dieser Domain. Mathematisch geschrieben

$$\frac{\widehat{\text{Var}}(\hat{\tau}^{X}_{Z,d,g,r})}{\text{Var}(\hat{\tau}^{X}_{Z,d,g})} \quad , \tag{7.11}$$

wobei X in $\hat{\tau}^{X}_{Z,d,g,r}$ bzw. $\hat{\tau}^{X}_{Z,d,g}$ entweder D–DSE, CHAP, SPREE, GREG1, GREG2 oder SYN ist. Im Fall des EBLUPs ersetzt $\widehat{\text{MSE}}(\hat{\tau}^{\text{EBLUP}}_{Z,d,g,r})$ den $\widehat{\text{Var}}(\hat{\tau}^{X}_{Z,d,g,r})$ bzw. $\text{MSE}(\hat{\tau}^{\text{EBLUP}}_{Z,d,g})$ die $\text{Var}(\hat{\tau}^{X}_{Z,d,g})$. Beim GSPREE wird im Nenner (7.11) der Mittelwert der 100 Varianzschätzungen berechnet, d.h. $\frac{1}{100}\sum_{r=1}^{R}\widehat{\text{Var}}(\hat{\tau}^{\text{GSPREE}}_{Z,d,g,r})$.

Jede Abbildung ist für eine der Gemeinden Nr. 1, 35, 37 oder 45. Man beachte die sehr unterschiedlichen Skalierung in den großen Gemeinden

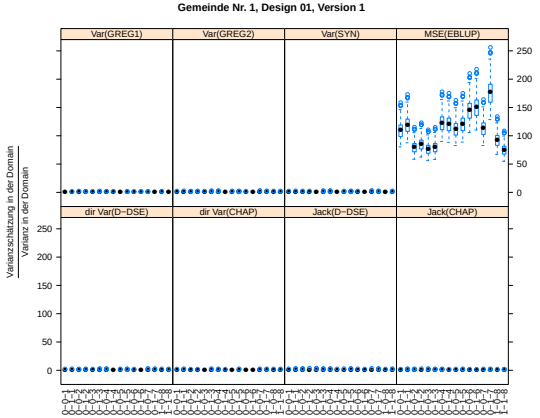


ABBILDUNG 7.7: *Relative Verzerrung der Varianzschätzungen auf Domainbasis in der Gemeinde Nr. 1. Design 01, Version 1.*

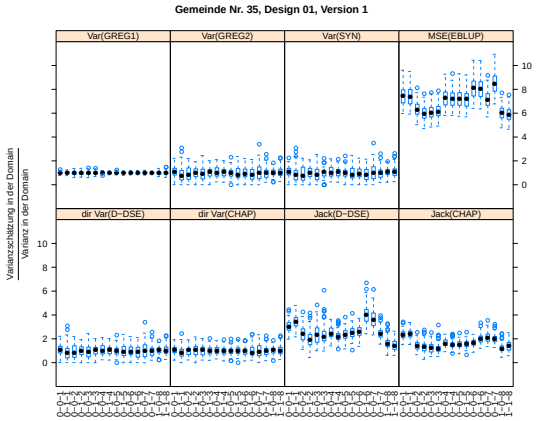


ABBILDUNG 7.8: *Relative Verzerrung der Varianzschätzungen auf Domainbasis in der Gemeinde Nr. 35. Design 01, Version 1.*

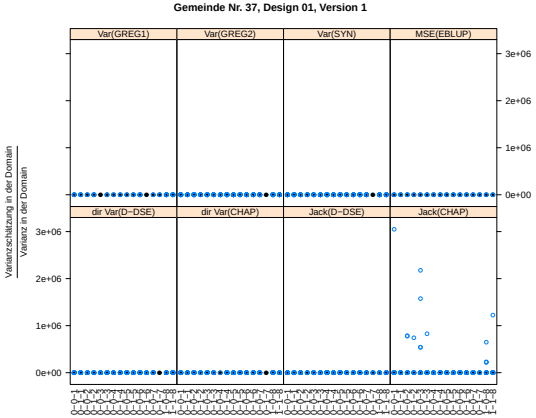


ABBILDUNG 7.9: *Relative Verzerrung der Varianzschätzungen auf Domainbasis in der Gemeinde Nr. 37. Design 01, Version 1.*

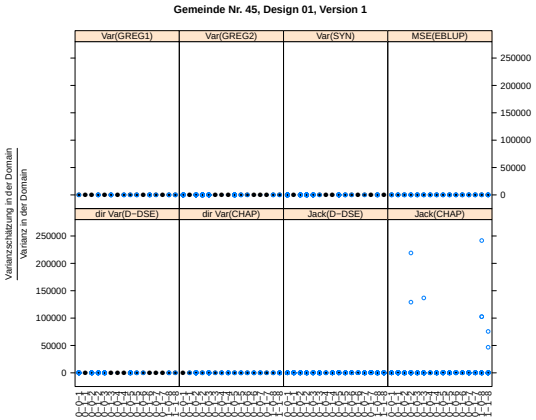


ABBILDUNG 7.10: *Relative Verzerrung der Varianzschätzungen auf Domainbasis in der Gemeinde Nr. 45. Design 01, Version 1.*

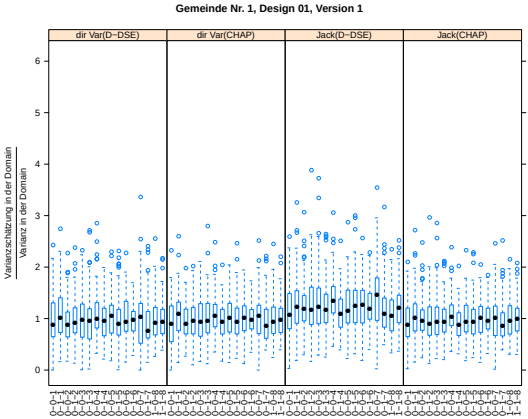


ABBILDUNG 7.11: *Relative Verzerrung der Varianzschätzungen auf Domainbasis in der Gemeinde Nr. 1. Nur direkte Varianzschätzung und Jackknife. Design 01, Version 1.*

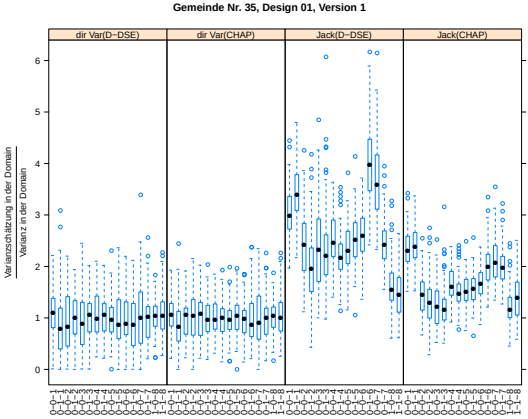


ABBILDUNG 7.12: *Relative Verzerrung der Varianzschätzungen auf Domainbasis in der Gemeinde Nr. 35. Nur direkte Varianzschätzung und Jackknife. Design 01, Version 1.*

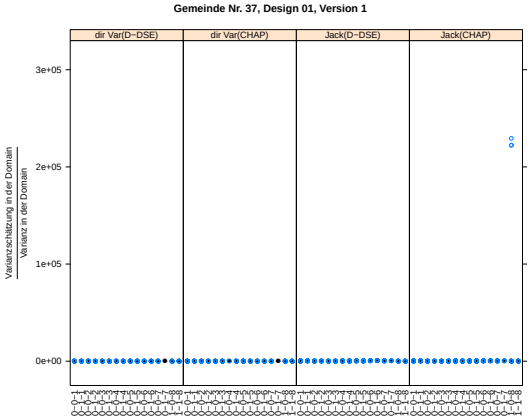


ABBILDUNG 7.13: *Relative Verzerrung der Varianzschätzungen auf Domainbasis in der Gemeinde Nr. 37. Nur direkte Varianzschätzung und Jackknife. Design 01, Version 1.*

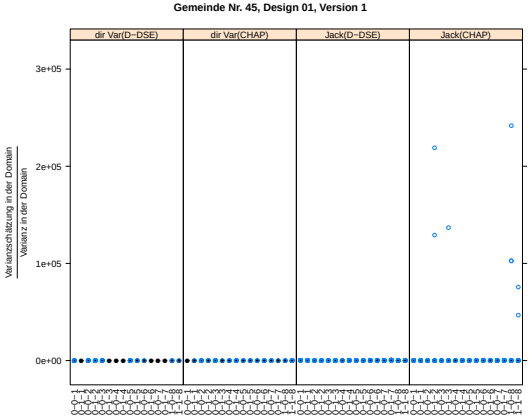


ABBILDUNG 7.14: *Relative Verzerrung der Varianzschätzungen auf Domainbasis in der Gemeinde Nr. 45. Nur direkte Varianzschätzung und Jackknife. Design 01, Version 1.*

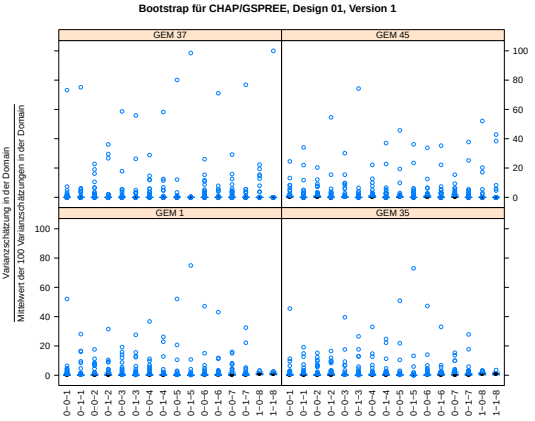


ABBILDUNG 7.15: *Verzerrung der Bootstrap–Varianzschätzungen des CHAP/GSPREEs auf Domainbasis in Gemeinden Nr. 1, 35, 37 und 45. Design 01, Version 1.*

(Abbildung 7.7 und Abbildung 7.8). Die MSE–Schätzung des EBLUPs in der Gemeinde Nr. 1 ist deutlich am schlimmsten. In den kleinen Gemeinden (Abbildung 7.9 und Abbildung 7.10) ist der Jackknife des Chapman–Schätzers am schlimmsten. Die Ergebnisse der Verhältnis–synthetischen Schätzers können nur mit Einschränkungen verwendet werden, da in der Praxis auftretende Verzerrungen unberücksichtigt geblieben sind (s. S. 84).

Bei der Berechnung einer Jackknife–Varianzschätzung für D–DSE kann es vorkommen, dass es aus den bekannten Gründen (siehe (7.7) und (7.8)) kein Ergebnis gibt. Im Folgenden werden daher in Boxplots für die Jackknife–Varianzschätzung von D–DSE nur die Stichproben abgebildet, in denen weder `NaN` noch `Inf` vorkommen.

Für einen besseren Vergleich sind auf den Seiten 119-120 nur die relativen Verzerrungen des D–DSEs und Chapman–Schätzers abgebildet. Die Skala bei den beiden großen Gemeinden ist gleich (Abbildung 7.11 und 7.12).

Die Skala bei den beiden kleinen Gemeinden ist ebenfalls gleich (Abbildung 7.13 und 7.14)

Betrachten wir den Jackknife für den Chapman–Schätzer in der kleinen Gemeinde Nr. 45, Domain 0-1-2 (deutsche Frauen im Alter von 18-24). Die Anzahl der registrierten Personen in dieser Domain ist $\tau_{R,\text{0-1-2}} = 205$ und die Anzahl der tatsächlich lebenden Personen dieser Domain ist $\tau_{Z,\text{0-1-2}} = 210$. Der Ausreißer in dieser Domain (s. Abbildung 7.14) kommt bei der Stichprobe Nr. 39 vor.

Ursprünglich sind insgesamt in der Gemeinde $n_{<45>} = 49$ Anschriften ausgewählt (s. Tabelle A.1 auf der Seite 180). Nach der Auslassung einer von 49 Anschriften kann es vorkommen, dass $\hat{\tau}^{\text{CHAP}}_{Z,\text{0-1-2},-a}$ sehr hoch ist (s. Beispiel in (7.9)). Daher ist $\hat{\tau}^{\text{CHAP}\,*}_{Z,\text{0-1-2},-a}$ auch sehr hoch und sowohl $\hat{\tau}^{\text{CHAP}\,*}_{Z,\text{0-1-2},a,\text{pseudo}}$ als auch $\hat{\bar{\tau}}^{\text{CHAP}\,*}_{Z,\text{0-1-2}_{\text{jack}^*}}$ sind negativ. Das hat zur Folge, dass $\widehat{\text{Var}}_{\text{jack}^*}(\hat{\tau}^{\text{CHAP}}_{Z,\text{0-1-2}})$ extrem hoch ist und als Ausreißer der Domain 0-1-2 in Abbildung 7.14 erscheint. Für die Definitionen $\hat{\tau}^{\text{CHAP}\,*}_{Z,\text{0-1-2},-a}$, $\hat{\tau}^{\text{CHAP}\,*}_{Z,\text{0-1-2},a,\text{pseudo}}$, $\hat{\bar{\tau}}^{\text{CHAP}\,*}_{Z,\text{0-1-2}_{\text{jack}^*}}$ und $\widehat{\text{Var}}_{\text{jack}^*}(\hat{\tau}^{\text{CHAP}}_{Z,\text{0-1-2}})$ siehe Kapitel 6.3.1.

Wegen der im Gegensatz zu Abbildungen 7.7-7.10 unterschiedlichen Skalierung wird im Folgenden die Bootstrap–Varianzschätzung von CHAP/GSPREE für alle untersuchten Gemeinden separat abgebildet. Die Abbildung 7.15 zeigt für die Gemeinden Nr. 1, 35, 37 und 45 die Boxplots der Bootstrap–Varianzschätzung von CHAP/GSPREE jeder Domain normiert durch den Mittelwert der 100 Varianzschätzungen in der Domain. Für die Formel der Varianzschätzung siehe Seite 88. In der Simulation wurde B=20 mal wiederholt eine Bootstrap–Stichprobe mit Zurücklegen aus der ursprünglichen Stichprobe gezogen. Allgemein können wir sehen, dass es keine großen Unterschiede in Varianzschätzungen von GSPREE in kleinen und großen Gemeinden gibt wie bei den anderen Schätzern. Vor allem, weil das GSPREE–Modell auf alle Chapman–Schätzer der 52 Gemeinden angewendet wird.

Als Beispiel erklären wir den Ausreißer in der Gemeinde Nr. 1, Domain 0-1-5 (deutsche Frauen im Alter von 40-49). Dieser kommt von drei Bootstrap–Stichproben her, alle aus der ursprünglichen Stichprobe r=95. Der Chapman–Schätzer in der Gemeinde Nr. 37, Domain 0-1-5 ist in diesen drei Bootstrap–Stichproben sehr hoch (s. Beispiel (7.9)), was nach der Anwendung vom GSPREE–Modell auf alle Gemeinden auch hohe Schätzungen für CHAP/GSPREE in der Gemeinde Nr. 1 liefert. Diese wirkt durch (6.51) auf die Varianzschätzung von CHAP/GSPREE. Da alle drei Bootstrap–Stichproben aus einer ursprünglichen Stichprobe (r=95) sind, kommt der Ausreißer in der Gemeinde Nr. 37, Domain 0-1-5 auch aus dieser Stichprobe her.

Fazit

In dieser Version haben wir die Altersgliederung der Klassen (AGE-Klassen) genauso klassifiziert, wie die Altersgliederung der Domains (AGE-Domains). Es hat sich bei den Schätzungen und Varianzschätzungen gezeigt, dass diese Definition sehr fein ist. Es kommt vor allem in den Domains 0-0-2, 0-1-2, 0-0-3, 0-1-3 (s. Tabelle 7.5) oft vor, dass keine Person ausgewählt war bzw. eine Person als Fehlbestand gefunden wurde, was bei der Schätzung ein `NaN` bzw. `Inf` gibt. Im Folgenden untersuchen wir andere Definitionen der AGE-Klassen, wo die Altersgliederung vergröbert wird und damit weniger Probleme mit `NaN` und `Inf` gibt.

7.2.2 Version 2

Damit es in D–DSE weniger Probleme mit `NaN` und `Inf` gibt (s. die Spalte D–DSE in Tabelle 7.5), werden die AGE-Klassen in der deutschen Popu-

lation geändert. Weil es für die nicht deutsche Population nur eine AGE-Klasse gibt, müssen wir uns mit den Ergebnissen dort zufrieden geben.

Von der Anzahl `NaN` und `Inf` der Spalte D–DSE in Tabelle 7.5 kann man sehen, dass bei der deutschen Population die „problematischen“ Domains bzw. Klassen 0-0-2, 0-1-2, 0-0-3 und 0-1-3 sind, die durch AGE-Klassen Nr. 2 und 3 definiert sind. Deswegen werden diese zwei in einer AGE-Klasse zusammengefasst. Damit gibt es nun sechs AGE-Klassen für die deutsche Population und eine AGE-Klasse für die nicht deutsche Population wie vorher. Die Definitionen für die AGE-Klassen und dadurch definierten Klassen stehen in Tabellen 7.10 und 7.11.

NAT=0	
AGE-Klassen	Alter
1	0 - 17
2	18 - 29
3	30 - 39
4	40 - 49
5	50 - 64
6	65 - 95

NAT=1	
AGE-Klasse	Alter
7	0 - 95

TABELLE 7.10: *Das Alter der AGE-Klassen der deutschen und nicht deutschen Bevölkerung in der Version 2.*

D–DSE, Chapman–Schätzer, SPREE und GSPREE

Wie in Version 1 sind fast alle Klassen identisch definiert wie die Domains. Deswegen lassen sich nach der Applikation des D–DSE und Chapman–Schätzers sofort die Ergebnisse der interessierenden Domains ermitteln.

NAT	SEX	AGE-Klassen					
		1	2	3	4	5	6
0	1						
	0						

NAT	SEX	AGE-Klasse
		7
1	1	
	0	

TABELLE 7.11: *Die 14 Klassen der deutschen und nicht deutschen Bevölkerung in Version 2.*

Jedoch ist AGE-Klasse=2 breiter als in der Version 1 und beinhaltet nun zwei AGE-Domains. Um die Ergebnisse für diese zwei AGE-Domains zu erhalten, wird SPREE nach (4.4) bzw. (4.5) verwendet.

Im Folgenden bezeichnen wir mit D–DSE/SPREE den Schätzer in (4.5) mit $\hat{\tau}^{\mathrm{X}}_{Z,k} = \hat{\tau}^{\text{D-DSE}}_{Z,k}$, d.h.

$$\hat{\tau}^{\text{D-DSE/SPREE}}_{Z,d} = \frac{\tau_{R,d}}{\tau_{R,k}}\, \hat{\tau}^{\text{D-DSE}}_{Z,k} \qquad .$$

Ähnlich bezeichnen wir mit CHAP/SPREE den Schätzer in (4.5) mit $\hat{\tau}^{\mathrm{X}}_{Z,k} = \hat{\tau}^{\text{CHAP}}_{Z,k}$, d.h.

$$\hat{\tau}^{\text{CHAP/SPREE}}_{Z,d} = \frac{\tau_{R,d}}{\tau_{R,k}}\, \hat{\tau}^{\text{CHAP}}_{Z,k} \qquad .$$

Diese beiden Kombinationen von D–DSE oder Chapman–Schätzer mit SPREE liefern die gewünschten Ergebnisse für die Domains. Eine Über-

sicht über die Bezeichnungen der Schätzer findet man in Tabelle C.1 auf der Seite 191.

In Tabelle 7.12 sieht man, dass die Zusammenfassung der beiden AGE-Klassen eine Verringerung der Anzahl `NaN` oder `Inf` gebracht hat (als Vergleich s. Tabelle 7.5). Es werden insgesamt 29 `NaN` beim D–DSE in der Domain 0-0-2 bzw. 35 `NaN` beim D–DSE in der Domain 0-0-3 auf vier `NaN` bei der Kombination D–DSE/SPREE in diesen zwei Domains reduziert. In Subpopulationen der deutschen Frauen, Domains 0-1-2 und 0-1-3, werden 47 bzw. 28 `NaN` in D–DSE auf zwei `NaN` bei der Kombination D–DSE/SPREE reduziert. Nach der Zusammenfassung der beiden AGE-Klassen gibt es `Inf` in D–DSE/SPREE nicht mehr, obwohl in Version 1 ein bzw. zwei `Inf` in D–DSE für Subpopulationen der deutschen Männer oder Frauen auftauchen. Der Rest der `NaN` und `Inf` in D–DSE hat sich nicht geändert, vor allem weil keine andere Zusammenfassung der AGE-Klassen realisiert wurde.

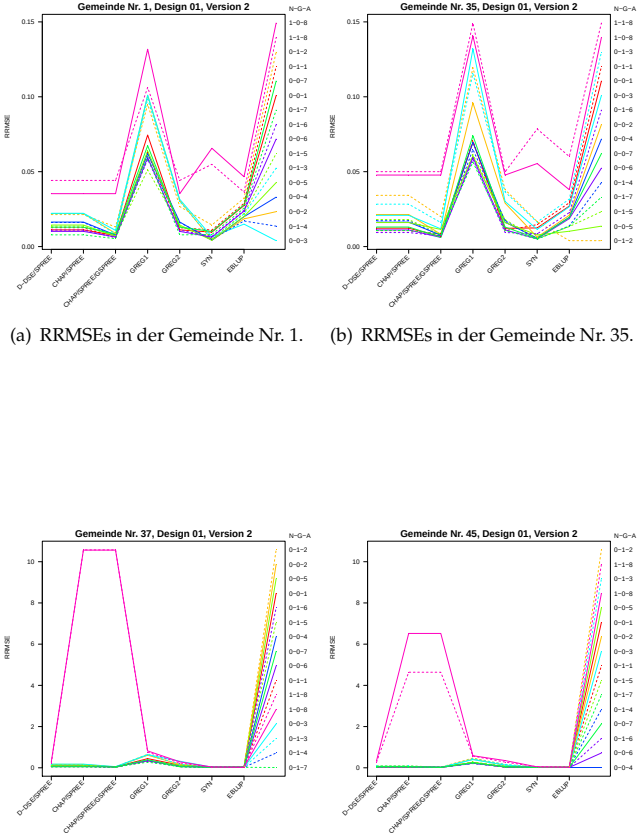
Im GSPREE–Modell werden als Pseudo–Werte die Ergebnisse des CHAP/SPREEs aller Domains in allen Gemeinden verwendet. Wir bezeichnen dann diese Schätzung als CHAP/SPREE/GSPREE. In der entsprechenden Spalte der Tabelle 7.12 kann man sehen, dass es keine Probleme mit `NaN` oder `Inf` gibt, weil die Kombination CHAP/SPREE/GSPREE vom Chapman–Schätzer für alle Domains ausgeht. Der Chapman–Schätzer hat bekanntlich keine Probleme mit `NaN` oder `Inf`. Nach anschließender Durchführung des SPREEs und GSPREEs kommen diese Probleme auch nicht vor.

Alternative Schätzer

GREG1–Schätzer, GREG2–Schätzer, Verhältnis–synthetischer Schätzer und EBLUP für alle Domains der Gemeinden sind unabhängig von dem Auf-

Domain	D–DSE/SPREE		CHAP/SPREE		CHAP/SPREE/GSPREE		GREG1		GREG2		SYN		EBLUP	
	NaN	Inf	NaN	Inf	NaN	Inf	NaN	Inf	NaN	Inf	NaN	Inf	NaN	Inf
0-0-1	0	0	0	0	0	0	0	0	0	0	0	0	8	0
0-1-1	2	0	0	0	0	0	0	0	2	0	0	0	10	0
0-0-2	4	0	0	0	0	0	0	0	29	1	0	0	36	0
0-1-2	2	0	0	0	0	0	0	0	47	1	0	0	53	0
0-0-3	4	0	0	0	0	0	0	0	35	2	0	0	41	0
0-1-3	2	0	0	0	0	0	0	0	28	1	0	0	35	0
0-0-4	0	0	0	0	0	0	0	0	0	0	0	0	8	0
0-1-4	0	0	0	0	0	0	0	0	0	0	0	0	8	0
0-0-5	0	0	0	0	0	0	0	0	0	0	0	0	8	0
0-1-5	2	0	0	0	0	0	0	0	2	0	0	0	10	0
0-0-6	0	0	0	0	0	0	0	0	0	0	0	0	8	0
0-1-6	0	0	0	0	0	0	0	0	0	0	0	0	8	0
0-0-7	0	0	0	0	0	0	0	0	0	0	0	0	8	0
0-1-7	0	0	0	0	0	0	0	0	0	0	0	0	8	0
1-0-8	78	5	0	0	0	0	0	0	78	5	0	0	84	0
1-1-8	108	6	0	0	0	0	0	0	108	6	0	0	114	0

TABELLE 7.12: *Anzahl der* `NaN` *und* `Inf` *pro Domain in der Gesamtpopulation. Design 01, Version 2. Die Nummerierung einer Domain dient als Schlüssel. Zum Beispiel: 0-0-2 bezeichnet eine Domain, die durch NAT=0, SEX=0, AGE-Domain=2 definiert ist.*



(a) RRMSEs in der Gemeinde Nr. 1. (b) RRMSEs in der Gemeinde Nr. 35.

(c) RRMSEs in der Gemeinde Nr. 37. (d) RRMSEs in der Gemeinde Nr. 45.

ABBILDUNG 7.16: *RRMSEs für die Domains in den Gemeinden Nr. 1, 35, 37 und 45. Design 01, Version 2. 16 Linien in jeder Grafik repräsentieren 14 Domains für die deutsche Population und zwei für die nicht deutsche Population. Die Beschriftung jeder Linie definiert genau eine Domain. Die erste Ziffer in der Beschriftung ist die Nationalität, 0 für deutsch, 1 für nicht deutsch. Die zweite Ziffer ist das Geschlecht, 0 für die Männer, 1 für die Frauen. Letzte Ziffer ist die AGE-Domain, 1 bis 7 für die AGE-Domains in der deutschen Population, AGE-Domain 8 ist für die nicht deutsche Population. Zum Beispiel bedeutet die Beschriftung 0-0-2 eine Domain, die durch NAT=0, SEX=0, AGE-Domain=2 definiert ist, d.h. deutsche Männer im Alter von 18-24.*

bau der Klassen berechnet. Deswegen sind die RRMSEs dieser Schätzer überall gleich wie in Version 1.

Grafische Darstellung der RRMSEs

Die zwei mittleren Boxen der Abbildungen H.1 bzw. H.2 auf Seite 203 bzw. 204 zeigen für große bzw. kleine Gemeinden die Mittelwerte der RRMSEs für die 16 Domains einer Gemeinde auf der x–Achse versus der Varianz der 100 Stichprobenmittelwerte der RRMSEs für die 16 Domains einer Gemeinde auf der y–Achse. Wir können sehen, dass alle Gemeinden nach links geschoben sind (vgl. mit zwei linken Boxen), also die RRMSEs für jede Gemeinde in Version 2 besser geworden sind.

Die Boxplots des Schätzers präsentieren wir hier nicht, nur die RRMSEs jeder Domain in den vier untersuchten Gemeinden Nr. 1, 35, 37 und 45. Die Grafiken in Abbildung 7.16 liefern eine Übersicht über die RRMSEs der untersuchten Schätzer in den vier Gemeinden der Version 2.

Die Ergebnisse in D–DSE/SPREE und CHAP/SPREE für die Domains mit AGE-Domains Nr. 1, 4, 5, 6, 7 und 8 sind wie in Version 1 fest geblieben. AGE-Domains Nr. 2 und 3 werden in Version 2 von der breiter definierten AGE-Klasse Nr. 2 mithilfe SPREE berechnet. Deshalb kann man in den großen Gemeinden in (a) und (b) deutlich sehen, dass die RRMSEs des D–DSE/SPREEs und CHAP/SPREEs für die Domains 0-0-2, 0-1-2, 0-0-3 und 0-1-3 im Vergleich zur Version 1 (s. Seite 112) verbessert sind. In den kleinen Gemeinden in (c) und (d) ist einige Verbesserung auch erkennbar, den Rest der Domains kann man wegen der breiten y–Achse nur numerisch vergleichen (s. Tabelle H.2 auf der Seite 211 für die Version 1 und Tabelle H.3 auf der Seite 212 für die Version 2). Zum Beispiel ist der RRMSE für den Chapman–Schätzer der Gemeinde Nr. 37 in der Domain 0-1-3 von 0,16993 auf 0,14670 gesunken.

In die Berechnung von CHAP/SPREE/GSPREE sind die CHAP/SPREE Schätzungen aller Domains in allen Gemeinden eingegangen. Deswegen hat sich CHAP/SPREE/GSPREE in jeder Domain leicht geändert.

Wie oben erwähnt bleiben die Ergebnisse bei den alternativen Schätzern ungeändert zur Version 1, weil die Berechnung der Domains von dem Aufbau der Klassen nicht abhängig ist.

Absoluter–relativer–Bias, Varianzschätzung und MSE–Schätzung

Vollständigkeitshalber sind die ARBs für jede Domain der Gemeinden Nr. 1, 35, 37 und 45 auf der Seite 215 präsentiert. Die Mittelwerte der ARBs über alle Domains in einzelnen Gemeinden, Varianzschätzung und MSE–Schätzung präsentieren wir bei dieser Version nicht.

Fazit

Allgemein kann man sagen, dass die Zusammenfassung der AGE-Klassen Nr. 2 und 3 in eine AGE-Klasse die RRMSE Ergebnisse für die entsprechenden Domains positiv beeinflusst. Dieses Resultat nützen wir in der nächsten Version 3.

7.2.3 Version 3

Wie in Version 2, wird die Zusammenlegung von Klassen weitergeführt. In Version 3 werden alle AGE-Klassen für die deutsche Population zu einer AGE-Klasse zusammengefasst. Für die nicht deutsche Population gibt es nur eine AGE-Klasse (s. Tabelle 7.13). Die entsprechenden Klassen sind in einer Übersicht in Tabelle 7.14.

NAT=0		NAT=1	
AGE-Klasse	Alter	AGE-Klasse	Alter
1	0 - 95	2	0 - 95

TABELLE 7.13: *Das Alter der AGE-Klassen der deutschen und nicht deutschen Bevölkerung in der Version 3.*

D–DSE, Chapman–Schätzer, SPREE und GSPREE

In dieser Version werden sowohl D–DSE als auch Chapman–Schätzer in den vier Klassen berechnet. Es ist also nötig, danach einen SPREE für Schätzungen aller 16 interessierenden Domains zu verwenden.

Die Definition der Klassen in Version 3 hat zur Folge, dass es bei der deutschen Population keine Probleme mit `NaN` und `Inf` gibt. Damit gibt es nach anschließender Durchführung des SPREEs auch kein `NaN` und `Inf` in den Domains (s. die Spalte D–DSE/SPREE oder CHAP/SPREE in Tabelle 7.15). Die `NaN` und `Inf` bei der nicht deutschen Population in der D–DSE/SPREE Spalte sind nicht mehr zu verbessern.

Als Pseudo–Werte für GSPREE werden die Schätzungen aller Gemeinden verwendet. Da die Schätzungen mittels des Chapman–Schätzers innerhalb einer Gemeinde mit anschließender SPREE Aufteilung auf die einzelnen Domains gerechnet sind, bezeichnen wir den Schätzer mit CHAP/SPREE/GSPREE. Eine Übersicht über die Bezeichnungen der Schätzer findet man in Tabelle C.1 auf der Seite 191. Eine Übersicht über die verwendeten Hilfsinformationen der Schätzer findet man in Tabelle D.1 auf der Seite 193.

NAT	SEX	AGE-Klasse
		1
0	1	
	0	

NAT	SEX	AGE-Klasse
		2
1	1	
	0	

TABELLE 7.14: *Die vier Klassen der deutschen und nicht deutschen Bevölkerung in Version 3.*

Alternative Schätzer

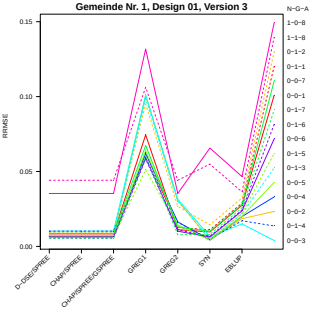
Die Anwendung dieser Schätzer ist wie in der Version 1 und Version 2. Die Anzahl `NaN` und `Inf` in GREG1, GREG2, Verhältnis–synthetischen Schätzer und EBLUP ist identisch mit der Anzahl in Tabelle 7.5 oder in Tabelle 7.12.

Grafische Darstellung der RRMSEs

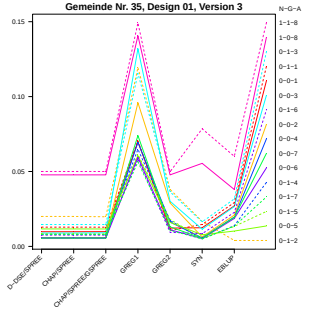
Die zwei rechten Boxen der Abbildungen H.1 bzw. H.2 auf Seite 203 bzw. 204 zeigen für große bzw. kleine Gemeinden die Mittelwerte der RRMSEs für die 16 Domains einer Gemeinde auf der x–Achse versus der Varianz der 100 Stichprobenmittelwerte der RRMSEs für die 16 Domains einer

Domain	D–DSE/SPREE		CHAP/SPREE		CHAP/SPREE/GSPREE		GREG1		GREG2		SYN		EBLUP	
	NaN	Inf	NaN	Inf	NaN	Inf	NaN	Inf	NaN	Inf	NaN	Inf	NaN	Inf
0-0-1	0	0	0	0	0	0	0	0	0	0	0	0	8	0
0-1-1	0	0	0	0	0	0	0	0	2	0	0	0	10	0
0-0-2	0	0	0	0	0	0	0	0	29	1	0	0	36	0
0-1-2	0	0	0	0	0	0	0	0	47	1	0	0	53	0
0-0-3	0	0	0	0	0	0	0	0	35	2	0	0	41	0
0-1-3	0	0	0	0	0	0	0	0	28	1	0	0	35	0
0-0-4	0	0	0	0	0	0	0	0	0	0	0	0	8	0
0-1-4	0	0	0	0	0	0	0	0	0	0	0	0	8	0
0-0-5	0	0	0	0	0	0	0	0	0	0	0	0	8	0
0-1-5	0	0	0	0	0	0	0	0	2	0	0	0	10	0
0-0-6	0	0	0	0	0	0	0	0	0	0	0	0	8	0
0-1-6	0	0	0	0	0	0	0	0	0	0	0	0	8	0
0-0-7	0	0	0	0	0	0	0	0	0	0	0	0	8	0
0-1-7	0	0	0	0	0	0	0	0	0	0	0	0	8	0
1-0-8	78	5	0	0	0	0	0	0	78	5	0	0	84	0
1-1-8	108	6	0	0	0	0	0	0	108	6	0	0	114	0

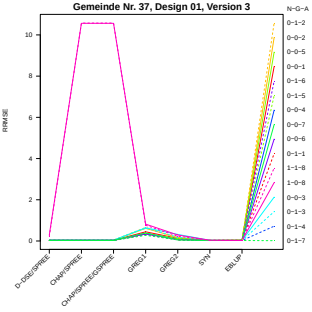
TABELLE 7.15: *Anzahl der* `NaN` *und* `Inf` *pro Domain in der Gesamtpopulation. Design 01, Version 3. Die Nummerierung einer Domain dient als ein Schlüssel. Zum Beispiel: 0-0-2 bezeichnet eine Domain, die durch NAT=0, SEX=0, AGE-Domain=2 definiert ist.*



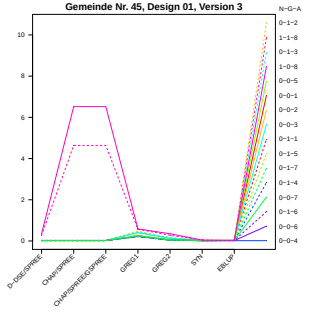
(a) RRMSEs in der Gemeinde Nr. 1.



(b) RRMSEs in der Gemeinde Nr. 35.

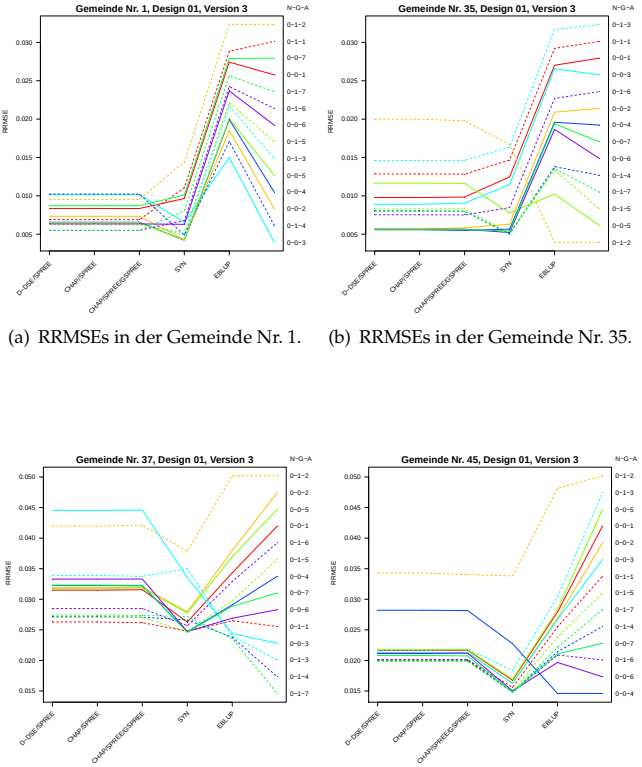


(c) RRMSEs in der Gemeinde Nr. 37.



(d) RRMSEs in der Gemeinde Nr. 45.

ABBILDUNG 7.17: *RRMSEs für die Domains in den Gemeinden Nr. 1, 35, 37 und 45. Design 01, Version 3. 16 Linien in jeder Grafik repräsentieren 14 Domains für die deutsche Population und zwei für die nicht deutsche Population. Die Beschriftung jeder Linie definiert genau eine Domain. Die erste Ziffer in der Beschriftung ist die Nationalität, 0 für deutsch, 1 für nicht deutsch. Die zweite Ziffer ist das Geschlecht, 0 für die Männer, 1 für die Frauen. Letzte Ziffer ist die AGE-Domain, 1 bis 7 für die AGE-Domains in der deutschen Population, AGE-Domain 8 ist für die nicht deutsche Population. Zum Beispiel bedeutet die Beschriftung 0-0-2 eine Domain, die durch NAT=0, SEX=0, AGE-Domain=2 definiert ist, d.h. deutsche Männer im Alter von 18-24.*



(a) RRMSEs in der Gemeinde Nr. 1.

(b) RRMSEs in der Gemeinde Nr. 35.

(c) RRMSEs in der Gemeinde Nr. 37.

(d) RRMSEs in der Gemeinde Nr. 45.

ABBILDUNG 7.18: *RRMSEs des D–DSE/SPREEs, CHAP/SPREEs, CHAP/SPREE/GSPREEs, Verhältnis–synthetischen Schätzers und EBLUPs für die Domains der deutschen Population (die Beschriftung beginnt mit 0) in den Gemeinden Nr. 1, 35, 37 und 45. Design 01, Version 3.*

Gemeinde auf der y–Achse. Von allen drei Versionen sind die RRMSEs der Version 3 offensichtlich am kleinsten (vgl. mit linken und mittleren Boxen).

Die vier Grafiken in Abbildung 7.17 stellen die RRMSEs des Designs 01 in den vier untersuchten Gemeinden Nr. 1, 35, 37 und 45 für Version 3 dar. Die RRMSEs für D–DSE/SPREE sind in jeder Domain (ausschließlich die Domains 1-0-8 und 1-1-8 der nicht deutschen Population) durch 100 Stichproben berechnet (s. Tabellle 7.15). Im Vergleich mit den RRMSEs der Version 1 (s. Seite 112) oder Version 2 (s. Seite 128) sind die RRMSEs der Kombinationen D–DSE/SPREE, CHAP/SPREE und CHAP/SPREE/GSPREE der Version 3 überall sehr niedrig.

Die Ergebnisse bei den alternativen Schätzern bleiben ungeändert zur Version 1 und Version 2 (s. Seite 112 bzw. Seite 128), weil die Berechnung der Domains von dem Aufbau der Klassen nicht abhängig ist.

Wegen des Ergebnises in der Gemeinde Nr. 37 für die nicht deutschen Männer (Domain 1-0-8) ist die Skala der y–Achse in Grafiken (c) und (d) sehr breit und die Höhe den RRMSEs nicht ablesbar. Für einen besseren Vergleich der drei Kombinationen mit dem Verhältnis–synthetischen Schätzer und EBLUP innerhalb der Version 3 ist in Abbildung 7.18 ein Teil der Abbildung 7.17 dargestellt, wo nur die RRMSEs dieser fünf Schätzer für die deutsche Population (die Beschriftung der Domains beginnt nur mit 0) abgebildet sind.

Man kann sehen, dass die RRMSEs der D–DSE/SPREE, CHAP/SPREE und CHAP/SPREE/GSPREE der deutschen Population in allen Gemeinden parallel verlaufen. Wir können daher sagen, dass alle drei Schätzer stabil sind. Der Verhältnis–synthetische Schätzer in den großen Gemeinden Nr. 1 und Nr. 35 ist nicht leicht einzuordnen. Für manche Domains ist er besser als D–DSE/SPREE, CHAP/SPREE oder CHAP/SPREE/GSPREE. Es gibt aber auch Domains, wo er nicht besser ist.

In der deutschen Population der kleinen Gemeinden ist der Verhältnis–synthetische Schätzer immer besser als D–DSE/SPREE, CHAP/SPREE oder CHAP/SPREE/GSPREE mit der Ausnahme in der Domain 0-1-3 der Gemeinde Nr. 37. Im Gegenteil dazu ist der EBLUP (mit ein paar Ausnahmen) oft schlechter als D–DSE/SPREE, CHAP/SPREE oder CHAP/SPREE/GSPREE. Es ist aber auch wichtig zu betonen, dass unterschiedliche Hilfsinformationen in jedem Schätzer benutzt wurden, was die Ergebnisse beeinflusst hat (s. Tabelle D.1 auf der Seite 193).

Für einen numerischen Vergleich der RRMSEs siehe Tabelle H.4 auf der Seite 213.

Absoluter–relativer–Bias

In der Tabelle auf der Seite 216 ist der ARB in jeder Domain der Gemeinden Nr. 1, 35, 37 und 45 zu sehen. Die Mittelwerte über alle Domains in einzelnen Gemeinden zeigt Tabelle 7.16. Die ARBs für D–DSE/SPREE, CHAP/SPREE oder CHAP/SPREE/GSPREE in den großen Gemeinden sind gleich. In den kleinen Gemeinden sind von diesen drei Schätzern die ARBs für D–DSE/SPREE am kleinsten. Man beachte, dass die RRMSEs für den Verhältnis–synthetischen Schätzer in den kleinen Gemeinden sehr niedrig sind, die ARBs für diesen Schätzer im Vergleich mit ARB für D–DSE/SPREE aber nicht am kleinsten sind.

Varianzschätzung und MSE–Schätzung

Die Grafiken im Anhang H auf den Seiten 208-209 zeigen die relativen Verzerrungen der Varianzschätzungen auf Gemeindebasis, d.h. die Summe der Varianzschätzungen des Schätzers $\hat{\tau}^{X}_{Z,d,g,r}$ über die Domains $d =$

GEM	D–DSE/ SPREE	CHAP/ SPREE	CHAP/ SPREE/ GSPREE	GREG1	GREG2	SYN	EBLUP
1	0,00398	0,00398	0,00398	0,00630	0,00156	0,01218	0,02430
35	0,00652	0,00652	0,00652	0,00575	0,00211	0,01494	0,02248
37	0,01267	0,14024	0,14026	0,02877	0,01171	0,01312	0,01779
45	0,00879	0,09052	0,09051	0,03607	0,00643	0,01186	0,02114

TABELLE 7.16: *Mittelwerte der ARBs über alle Domains in den Gemeinden Nr. 1, 35, 37 und 45. Design 01, Version 3.*

$1, \ldots, D$ einer Gemeinde g für die Stichprobe $r = 1, \ldots, 100$ dividiert durch die Summe der Varianzen des Schätzers. Die Formel ist in (7.10) zu finden.

Abbildung H.6 stellt die großen Gemeinden dar, Abbildung H.7 die kleinen Gemeinden. Die Bootstrap–Varianzschätzung des CHAP/SPREE/GSPREEs für alle Gemeinden ist in Abbildung H.8 gegeben. Die Skala der y–Achse ist mit den Grafiken der Version 1 identisch (s. Seite 206-207).

Die Boxplots für GREG1–Schätzer, GREG2–Schätzer, Verhältnis–synthetischen Schätzer und EBLUP haben sich im Vergleich mit Version 1 nicht geändert, weil der Aufbau von unterschiedlichen Klassen keinen Einfluss auf die alternativen Schätzer und deren Varianzschätzungen hat. Die Bilder zeigen, dass eine Zusammenlegung von Klassen sehr positiv die Bootstrap–Varianzschätzung von CHAP/SPREE/GSPREE beeinflusst (vgl. mit Bootstrap–Varianzschätzung von CHAP/GSPREE der Version 1 auf der Seite 207).

Bei der Bootstrap–Varianzschätzung wird wiederholt eine Bootstrap–Stichprobe mit Zurücklegen aus der ursprünglichen Stichprobe gezogen und das GSPREE–Modell auf alle Chapman–Schätzer der Klassen der 52 Gemeinden angewendet. Weil die Klassen in der Version 1 sehr fein definiert sind, ist der Chapman–Schätzer in einigen Klassen sehr hoch (s. Bei-

spiel (7.9)), was durch (6.51) auf die Bootstrap–Varianzschätzung von CHAP/GSPREE wirkt. Daher ist die relative Verzerrung der Bootstrap–Varianzschätzung in der Abbildung H.5 auf der Seite 207 breiter als in der Abbildung H.8 auf der Seite 209. In der Version 3 werden die Klassen zusammengefasst. Durch die gröbere Klasseneinteilung liefert der Chapman–Schätzer keine auffälligen Schätzungen mehr. Ebenso ist die relative Verzerrung der Bootstrap–Varianzschätzung von CHAP/SPREE/GSPREE nicht so breit.

Einen Überblick über die relativen Verzerrung der Varianzschätzungen von D–DSE/SPREE, CHAP/SPREE, GREG1–Schätzer, GREG2–Schätzer, Verhältnis–synthetischen Schätzer und EBLUP in jeder Domain der Gemeinden Nr. 1, 35, 37 und 45 dividiert durch die Varianz des Schätzers in der Domain geben die Abbildungen auf den Seiten 140-141 (mathematisch geschrieben in (7.11)).

Für die alternativen GREG1–Schätzer, GREG2–Schätzer, Verhältnis–synthetischen Schätzer und EBLUP sind die Boxplots gleich wie in der Version 1, weil der Aufbau von unterschiedlichen Klassen keinen Einfluss auf die alternativen Schätzer und deren Varianzschätzungen hat. Wegen der extrem breiten relativen Verzerrungen des geschätzten MSEs für EBLUP sind die anderen Boxplots nicht gut lesbar. In den Abbildungen auf der Seite 142 sind nur die relativen Varianzschätzungen für D–DSE bzw. Chapman–Schätzer dargestellt. Man erkennt die gleiche Form der Boxplots für direkte Varianzschätzung und Jackknife–Varianzschätzung für D–DSE/SPREE, CHAP/SPREE für deutsche Frauen (Domains 0-1-x, wobei x eine Zahl zwischen 1 und 7 ist) und die gleiche Form der Boxplots für deutsche Männer (Domains 0-0-x, wobei x eine Zahl zwischen 1 und 7 ist). Dies kommt vom Aufbau der Klassen in Version 3 und der Art der Abbildung der relativen Varianzschätzung.

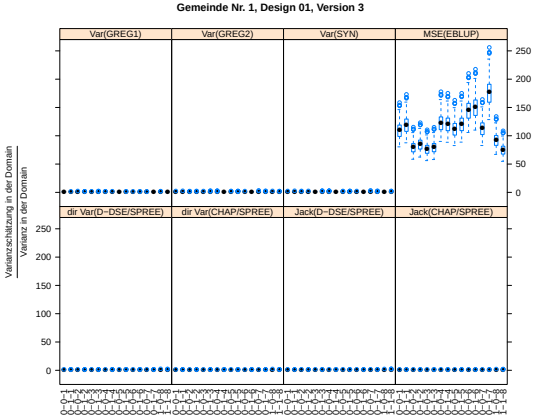


ABBILDUNG 7.19: *Relative Verzerrung der Varianzschätzungen auf Domainbasis in der Gemeinde Nr. 1. Design 01, Version 3.*

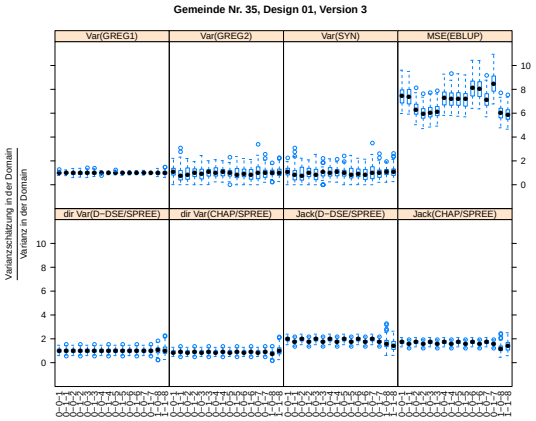


ABBILDUNG 7.20: *Relative Verzerrung der Varianzschätzungen auf Domainbasis in der Gemeinde Nr. 35. Design 01, Version 3.*

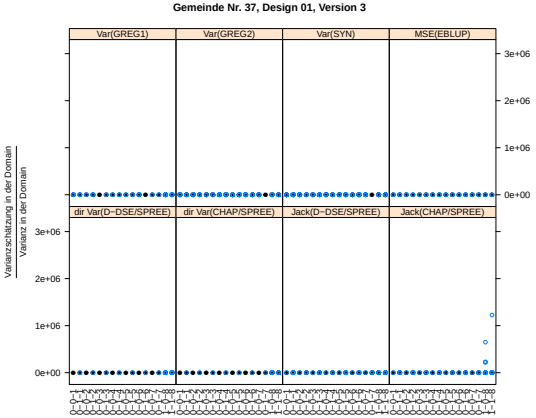


ABBILDUNG 7.21: *Relative Verzerrung der Varianzschätzungen auf Domainbasis in der Gemeinde Nr. 37. Design 01, Version 3.*

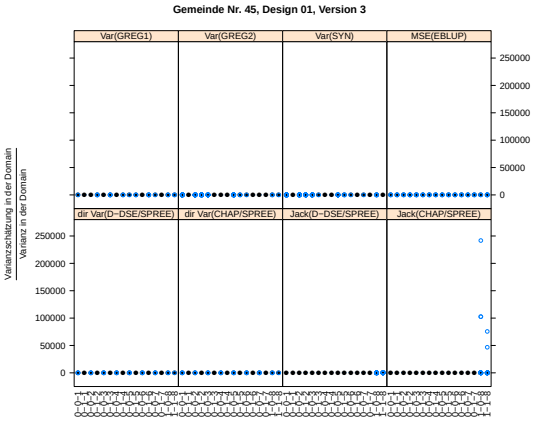


ABBILDUNG 7.22: *Relative Verzerrung der Varianzschätzungen auf Domainbasis in der Gemeinde Nr. 45. Design 01, Version 3.*

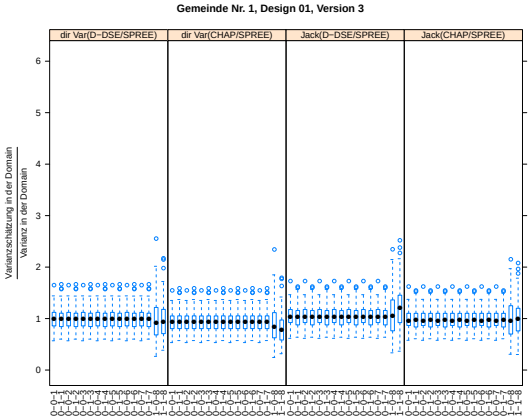


ABBILDUNG 7.23: *Relative Verzerrung der Varianzschätzungen auf Domainbasis in der Gemeinde Nr. 1. Nur direkte Varianzschätzung und Jackknife. Design 01, Version 3.*

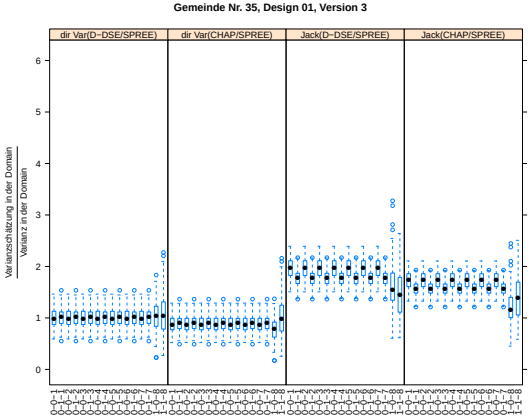


ABBILDUNG 7.24: *Relative Verzerrung der Varianzschätzungen auf Domainbasis in der Gemeinde Nr. 35. Nur direkte Varianzschätzung und Jackknife. Design 01, Version 3.*

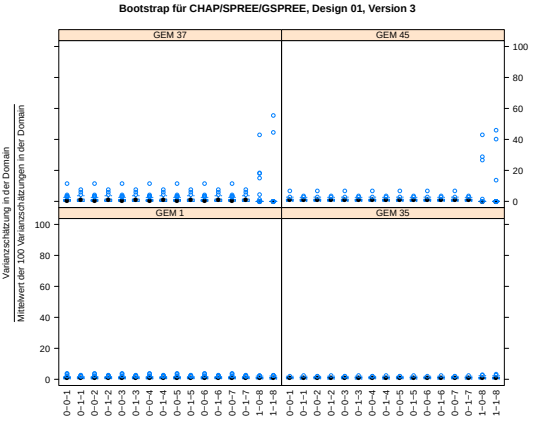


ABBILDUNG 7.25: *Verzerrung der Bootstrap–Varianzschätzungen des CHAP/SPREE/GSPREEs auf Domainbasis in Gemeinden Nr. 1, 35, 37 und 45. Design 01, Version 3.*

Wie schon gesagt hat die Zusammenlegung von Klassen sehr positiv die Bootstrap–Varianzschätzung von CHAP/SPREE/GSPREE beeinflusst. Dies ist auch nach dem Vergleich der Abbildung 7.25 mit der Abbildung 7.15 auf der Seite 121 zu erkennen.

Konfidenzintervalle

Die Abbildung 7.26 stellt die Überdeckungsrate versus Mittelwert über 100 Stichproben von relativen Konfidenzintervalllängen aller Domains d der Gemeinde $g \in \{1, 35, 37, 45\}$ dar, d.h. mathematisch

$$\frac{1}{100} \sum_{r=1}^{100} \frac{2\, z_{1-\alpha/2} \sqrt{\widehat{\mathrm{Var}}(\hat{\tau}^{\mathrm{X}}_{Z,d,g,r})}}{\tau_{Z,d,g}} \quad , \tag{7.12}$$

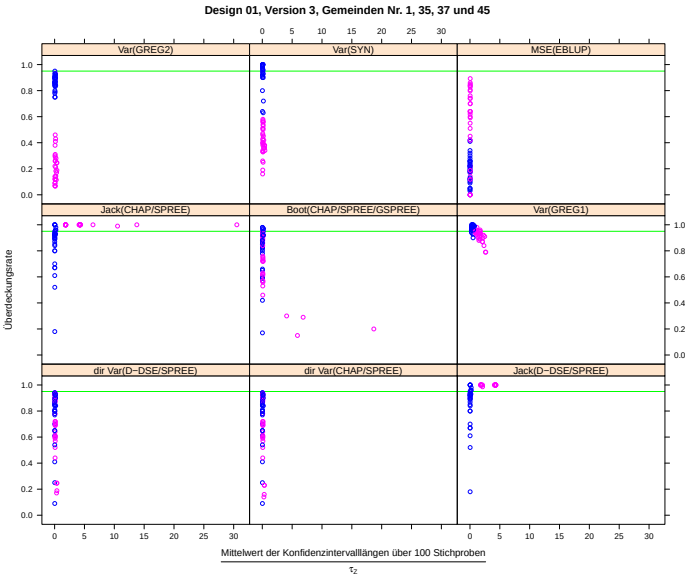


ABBILDUNG 7.26: *Überdeckungsrate versus Mittelwert der relativen Konfidenzintervalllängen über 100 Stichproben für die Domains der Gemeinden Nr. 1, 35, 37 und 45. Design 01, Version 3. Die Domains der großen Gemeinden sind blau dargestellt, Domains der kleinen Gemeinden violett. Die grüne Linie repräsentiert die Überdeckungsrate 95% (α = 0,05).*

wobei X in $\hat{\tau}^{X}_{Z,d,g,r}$ entweder D–DSE, CHAP, SPREE, GSPREE, GREG1, GREG2 oder SYN ist. Im Fall des EBLUPs ersetzt $\widehat{\text{MSE}}(\hat{\tau}^{\text{EBLUP}}_{Z,d,g,r})$ den $\widehat{\text{Var}}(\hat{\tau}^{X}_{Z,d,g,r})$. Die Domains der großen Gemeinden sind blau abgebildet, die Domains der kleinen Gemeinden violett. Wir betrachten die Überdeckungsrate 95% ($\alpha = 0{,}05$), die die grüne Linie repräsentiert. Bei diesem Konfidenzniveu erwarten wir, dass 95% der Konfidenzintervalle den wahren Wert überdecken und damit auf der grünen Linie liegen sollten. Gleichzeitig sollte man die Länge des Konfidenzintervalls beachten. Die Überdeckungsrate sollte beim möglichst kürzesten Konfidenzintervall erfüllt sein.

Es ist zu sehen, dass die Konfidenzintervalle in kleinen Gemeinen breiter als in großen Gemeinden sind, vor allem bei der Jackknife–Varianzschätzung und Bootstrap–Varianzschätzung. Dank dieser langen Konfidenzintervalle sind die Überdeckungsraten für die Jackknife–Varianzschätzungen in kleinen Gemeinden höher als 95%. Wenn wir die direkte Varianzschätzung mit Jackknife–Varianzschätzung in großen Gemeinden vegleichen, scheint die Länge des Konfidenzintervalls gleich zu sein, die Überdeckungsrate bei direkten Varianzschätzungen ist leicht niedriger.

Bei den alternativen Schätzern liefert die Varianzschätzung für GREG1–Schätzer erfreuliche Ergebnisse. Am niedrigsten liegt die Überdeckungsrate beim EBLUP.

7.2.4 Zusammenfassung bei uneingeschränkter Zufallsauswahl

Bei der uneingeschränkten Zufallsauswahl (Design 01) haben wir drei unterschiedliche Klassen untersucht, die in Version 1 bis Version 3 definiert wurden. Eine Übersicht über alle Szenarien befindet sich in Tabelle E.1 auf der Seite 196. Alle Grafiken und Tabellen für Design 01 sind im Text von Kapitel 7.2 und im Anhang auf den Seiten 203-216 zu finden.

Es hat sich gezeigt, dass es sich lohnt, D–DSE und den Chapman–Schätzer innerhalb einer Geschlecht×Nationalität Klasse zu berechnen und anschließend einen SPREE zu verwenden (Version 3), um die Schätzungen für die interessierenden Domains zu gewinnen.

In der deutschen Population verlaufen die RRMSEs für D–DSE/SPREE und CHAP/SPREE parallel und sind damit vergleichbar (s. Abbildung 7.18). Auch die ARBs für die deutsche Population sind gleich (s. Tabelle H.1). Das GSPREE–Modell kann auf die CHAP/SPREE Schätzungen der Domains angewendet werden, bringt aber keine große Verbesserung im Vergleich mit CHAP/SPREE.

In der nicht deutschen Population gibt es beim D–DSE/SPREE Probleme mit `NaN` und `Inf`. Die Anzahl der problematischen Schätzungen ist in Tabelle 7.15 Domains 1-0-8 und 1-1-8 zu sehen. Es kann CHAP/SPREE angewendet werden, dennoch kann er ein Ergebnis weit weg von der Realität liefern (s. Beispiel in (7.8)). Die unrealistischen Schätzungen in CHAP/SPREE tauchen auch in CHAP/SPREE/GSPREE auf, da dieser Schätzer CHAP/SPREE benötigt. Daher verlaufen die RRMSEs für D–DSE/SPREE, CHAP/SPREE und CHAP/SPREE/GSPREE nicht mehr parallel (s. Abbildung 7.17, (c) und (d)). Alternativ kann in der nicht deutschen Population auch der GREG1–Schätzer oder der Verhältnis–synthetische Schätzer ermittelt werden, da bei diesen Schätzern die oben genannten Probleme nicht entstehen.

Sowohl in großen als auch in kleinen Gemeinden ist die direkte Varianzschätzung und Jackknife–Varianzschätzung für D–DSE/SPREE bzw. CHAP/SPREE bei der Version 3 weniger verzerrt als bei der Version 1 (vgl. Abbildungen auf den Seiten 119-120 mit den Abbildungen auf der Seite 142). Die relative Verzerrung der direkten Varianzschätzung ist kleiner und damit besser als bei der Jackknife–Varianzschätzung . Die Überdeckungsrate bei der direkten Varianzschätzung ist sowohl in kleinen als

auch in großen Gemeinden niedriger und damit schlechter als bei der Jackknife–Varianzschätzung (s. Abbildung 7.26). Daher ist die direkte Varianzschätzung oder Jackknife–Varianzschätzung für D–DSE/SPREE bzw. CHAP/SPREE nicht zu empfehlen.

7.3 Geschichtete Zufallsauswahl

Aus der Literatur (zum Beispiel Lohr 1999, S. 95) ist bekannt, dass man in der Regel durch Schichtung mindestens so gute Ergebnisse wie bei der einfachen Zufallsauswahl erhält. In diesem Abschnitt untersuchen wir die geschichtete Zufallsauswahl mit Anschriftallokation und Mindeststichprobenumfang 1 pro Schicht, die genauer in Münnich et al. (2008) beschrieben und zu weiteren Untersuchungen vorgeschlagen ist.

Bei geschichtetem Design werden die Anschriften einer Gemeinde geschichtet. Für Saarland sind neun Schichten definiert, die durch die Anschriftgröße bestimmt sind. Die Definition der Schichten ist in Tabelle 7.17 gegeben und in der Variable ADK im Datensatz SAL gespeichert.
Die Erhebung bei geschichteter Zufallsauswahl erfolgt analog zur uneingeschränkten Zufallsauswahl. Pro Gemeinde g werden

$$n_{<g>} = \begin{cases} 550 & \text{in der Gemeinde mit } \tau_{R,<g>} \geq 10.000 \\ 550 \, \dfrac{\tau_{R,<g>}}{\tau_{R,\ll k \gg}} & \text{sonst} \end{cases} \tag{7.13}$$

Anschriften durch uneingeschränkte Zufallsauswahl ohne Zurücklegen ausgewählt. Diese Zufallsauswahl von $n_{<g>}$ Anschriften wird proportional zur Anzahl der Anschriften in den Schichten aufgeteilt, d.h. pro Schicht h wird

$$n_{h,<g>} = \frac{N_{h,<g>}}{N_{<g>}} \, n_{<g>} \tag{7.14}$$

SAL	
Personen pro Anschrift	Schicht
1	1
2	2
3	3
4	4
5	5
6	6
7-9	7
10	8
11 und mehr	9

TABELLE 7.17: *Die Definition der Anschriftenschichten in SAL.*

Anschriften ausgewählt. $N_{<g>}$ bezeichnet die Anzahl der Anschriften in Gemeinde g, $N_{h,<g>}$ die Anzahl der Anschriften in Schicht h in Gemeinde g. Die Auswahl in jeder Schicht erfolgt unabhängig von der Auswahl in einer anderen Schicht. Leere Schichten werden ignoriert. Ansonsten ist der Mindeststichprobenumfang pro Schicht bei diesem Design eine Anschrift.

Betrachten wir die Gemeinde Nr. 37 mit $\tau_{R,<37>} = 8.221$ registrierten Personen, die zum Kreis Nr. 4 mit $\tau_{R,\ll 4\gg} = 208.160$ gehört. Daher ist

$$n_{<37>} \quad = \quad 550\,\frac{\tau_{R,<37>}}{\tau_{R,\ll 4\gg}} \quad = \quad 550\,\frac{8.221}{208.160} \quad \doteq \; 22 \qquad .$$

Insgesamt $N_{<37>} = 2.232$ Anschriften der Gemeinde Nr. 37 sind in neun

Schichten aufgeteilt. Beispielsweise gehören 225 Anschriften zur Schicht Nr. 5. Nach (7.14) werden in dieser Schicht

$$n_{5,<37>} \quad = \quad \frac{N_{5,<37>}}{N_{<37>}}\, n_{<37>} \quad = \quad \frac{225}{2.232}\, 22 \quad \doteq \; 2$$

Anschriften ausgewählt.

Im Rahmen des DACSEIS Projekts wurde dieses Design als *Design 05a* bezeichnet. Der Einfachheit halber wird diese Bezeichnung auch im Folgenden benutzt. Solange es vom Kontext her klar ist, um welche Gemeinde es sich handelt, wird im Folgenden auf das g im Index verzichtet.

Für das Design 05a sind im Datensatz SAL 100 Vektoren der Länge 248.832 vorhanden, die eine 1 an der Stelle der gezogenen Anschriften haben, an der Stelle der nicht ausgewählten Anschriften eine 0. Diese 100 Vektoren (`SAL.GEM.design05a.ADR.00r.RData`, wobei `r` von 00 bis 99 läuft) zusammen mit den Vektoren aus Tabelle 7.1 werden im `R`–Programm für die Berechnungen verwendet.

Weiter ist im Datensatz SAL der Vektor `SAL.IIP.design05a.dat` der Länge 248.832 vorhanden, der die Designgewichte der Anschriften, w_a enthält. Es ist auch der Vektor `SAL.IIP.design05a.P.dat` der Länge 1.057.915 vorhanden, der die Designgewichte auf Personen–Ebene enthält. Die Designgewichte für jede Schicht in allen Gemeinden sind im Anhang B zu finden.

Ziel ist die Schätzung der Einwohnerzahl in einer Gemeinde nach Geschlecht, Nationalität und Altersgliederung, welche in Tabelle 7.3 auf der Seite 92 definiert ist.

7.3.1 Version 1 und Version 2

Wie beim Design 01, bei dem die Zusammenfassung der AGE-Klassen gewisse Vorteile brachte, werden hier die Ergebnisse der Version 1 bzw. Version 2 für Design 05a nicht präsentiert. Ein Teil der Ergebnisse ist in Dostál et al. zu finden. Im Weiteren richten wir unsere Aufmerksamkeit direkt auf die Version 3.

7.3.2 Version 3

Wie aus dem Design 01 bekannt ist, werden bei der Version 3 die AGE-Klassen sowohl für die deutsche Population als auch für die nicht deutsche Population über die ganzen Populationen definiert. D–DSE und Chapman–Schätzer werden innerhalb einer Geschlecht$\times$Nationalität Klasse berechnet. Anschließend wird ein SPREE verwendet, um die Schätzungen für die interessierenden Domains zu gewinnen.

Die Definition der AGE-Klassen der deutschen und nicht deutschen Bevölkerung in der Version 3 ergibt sich aus Tabelle 7.13 auf der Seite 131. Die Definition der Klassen findet man in Tabelle 7.14.

D–DSE/SPREE

Wegen der Teilung der Anschriften in neun Schichten kommt noch häufiger als im Design 01 das Problem mit `NaN` und `Inf` für eine Klasse bzw. für eine Domain vor. Die D–DSE/SPREE Spalte der Tabelle 7.18 gibt die Anzahl aller `NaN` und `Inf` der Gesamtpopulation pro Domain an (vgl. mit Tabelle 7.15). Die Nummerierung einer Domain dient als Schlüssel. Zum

Domain	D–DSE/SPREE		CHAP/SPREE		CHAP/SPREE/GSPREE		GREG1		GREG2		SYN		EBLUP	
	NaN	Inf	NaN	Inf	NaN	Inf	NaN	Inf	NaN	Inf	NaN	Inf	NaN	Inf
0-0-1	122	7	0	0	0	0	0	0	1.351	16	0	0	1.439	0
0-1-1	6	1	0	0	0	0	0	0	1.815	9	0	0	1.896	0
0-0-2	122	7	0	0	0	0	0	0	2.031	100	0	0	2.180	0
0-1-2	6	1	0	0	0	0	0	0	2.429	71	0	0	2.549	0
0-0-3	122	7	0	0	0	0	0	0	1.496	79	0	0	1.645	0
0-1-3	6	1	0	0	0	0	0	0	1.687	76	0	0	1.834	0
0-0-4	122	7	0	0	0	0	0	0	891	3	0	0	1.104	0
0-1-4	6	1	0	0	0	0	0	0	1.165	15	0	0	1.360	0
0-0-5	122	7	0	0	0	0	0	0	968	3	0	0	1.170	0
0-1-5	6	1	0	0	0	0	0	0	1.314	10	0	0	1.490	0
0-0-6	122	7	0	0	0	0	0	0	859	4	0	0	1.099	0
0-1-6	6	1	0	0	0	0	0	0	827	10	0	0	1.098	0
0-0-7	122	7	0	0	0	0	0	0	1.056	5	0	0	1.244	0
0-1-7	6	1	0	0	0	0	0	0	894	4	0	0	1.123	0
1-0-8	2.768	79	0	0	0	0	0	0	2.768	79	0	0	2.854	0
1-1-8	3.275	53	0	0	0	0	0	0	3.275	53	0	0	3.352	0

TABELLE 7.18: *Anzahl der* `NaN` *und* `Inf` *pro Domain in der Gesamtpopulation. Design 05a, Version 3. Die Nummerierung einer Domain dient als Schlüssel. Zum Beispiel: 0-0-2 bezeichnet eine Domain, die durch NAT=0, SEX=0, AGE-Domain=2 definiert ist.*

Schicht	Gemeinde																				Total-	
	15		17		36		37		40		45		47		49		50		anzahl	
	NaN	Inf	NaN	Inf	NaN	Inf	NaN	Inf	NaN	Inf	NaN	Inf	NaN	Inf	NaN	Inf	NaN	Inf	NaN	Inf
1	13	0	5	0	38	4	23	1	12	1	2	0	6	0	1	0	22	1	122	7
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	13	0	5	0	38	4	23	1	12	1	2	0	6	0	1	0	22	1	122	7

TABELLE 7.19: *Anzahl der* `NaN` *und* `Inf` *bei der Anwendung des D–DSEs bei den deutschen Männern im Alter von 0-95. Anzahl pro Schicht und pro Gemeinde. Design 05a, Version 3.*

Schicht	Gemeinde																				Total-	
	15		17		36		37		40		45		47		49		50		anzahl	
	NaN	Inf	NaN	Inf	NaN	Inf	NaN	Inf	NaN	Inf	NaN	Inf	NaN	Inf	NaN	Inf	NaN	Inf	NaN	Inf
1	0	0	1	0	2	1	2	0	1	0	0	0	0	0	0	0	0	1	6	1
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	1	0	2	1	2	0	1	0	0	0	0	0	0	0	0	1	6	1

TABELLE 7.20: *Anzahl der* `NaN` *und* `Inf` *bei der Anwendung des D–DSEs bei den deutschen Frauen im Alter von 0-95. Anzahl pro Schicht und pro Gemeinde. Design 05a, Version 3.*

Beispiel: 0-0-2 bezeichnet eine Domain, die durch NAT=0, SEX=0, AGE-Domain=2 definiert ist.

Die gleiche Anzahl der `NaN` bzw. `Inf` beim D–DSE/SPREE in der deutschen Population der Männer (Domains 0-0-x, x steht für 1 bis 7) bzw. Frauen (Domains 0-1-x, x steht für 1 bis 7) kommt bei Anwendung des D–DSE und SPREE vor. Zum Beispiel liefert bei der Anwendung des D–DSEs in der Klasse der deutschen Männer im Alter von 0-95 der Gemeinde Nr. 17 der Schätzer in ersten Schicht der Anschriften 5 mal `NaN` (s. Tabelle 7.19). Die anderen Schichten sind ohne Probleme. Die Anzahl der `NaN` in einzelnen Schichten wirkt sich auf die Schätzung der Klasse in der ganzen Gemeinde aus, wo die Schichten aufsummiert sind. Die anschließende Anwendung des SPREEs liefert auch 5 mal `NaN` auf alle schätzende Domains 0-0-x (x steht für 1 bis 7).

In 5.200 Simulationen (jeweils 100 Simulationen in allen 52 Gemeinden) kommt insgesamt 122 mal `NaN` für die Klasse der deutschen Männer im Alter von 0-95 vor. Daher kommt auch 122 mal `NaN` für alle 0-0-x Domains vor. Die Situation mit `Inf` ist ähnlich und wird hier nicht beschrieben. Die Anzahl der `NaN` und `Inf` bei der Anwendung des D–DSEs bei den deutschen Frauen im Alter von 0-95 ist in Tabelle 7.20 zu sehen. Die Probleme treten nur in der ersten Schicht und nur in kleinen Gemeinden.

Die Domain 1-0-8 bzw. 1-1-8 der nicht deutschen Population der Männer bzw. Frauen wird allein betrachtet. Die Situation mit `Inf` und `NaN` in den Schichten wird hier nicht beschrieben.

CHAP/SPREE, CHAP/SPREE/GSPREE

In der CHAP/SPREE Spalte der Tabelle 7.18 kann man sehen, dass die Schichtung der Anschriften keinen Einfluss auf die Schätzung durch Chap-

man–Schätzer hat, es gibt weder `NaN` noch `Inf` für Domains. Für das GSPREE–Modell werden als Pseudo–Werte wieder die durch CHAP/SPREE berechneten absoluten Häufigkeiten der Domains benutzt.

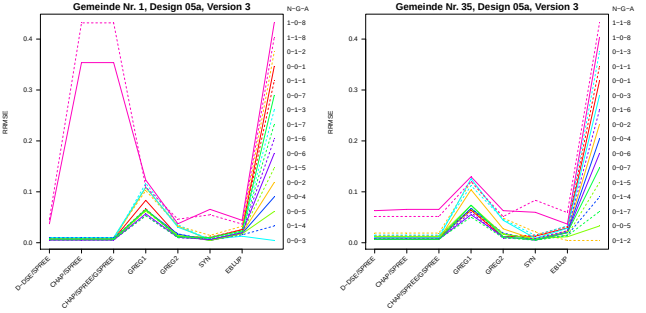
Alternative Schätzer

Die Anzahl der `NaN` und `Inf` pro Domain haben sich beim GREG2–Schätzer wegen der Anschriftenschichtung verschlechtert (siehe GREG2 Spalte in Tabelle 7.18 und vgl. mit Tabelle 7.15).

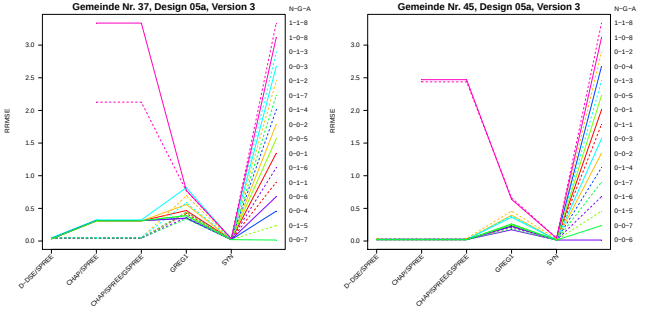
Grafische Darstellung allgemein

Für die grafische Präsentation der Ergebnisse werden die Gemeinden Nr. 1, 35, 37 und 45 genauer untersucht. Die Anzahl der im Einwohnermeldeamt registrierten Personen in diesen Gemeinden ist in Tabelle 7.8 auf der Seite 107 zu sehen.

Die Abbildung I.1 auf der Seite 217 zeigt für große Gemeinden die Mittelwerte der RRMSEs über 16 Domains einer Gemeinde auf der x–Achse versus der Varianz der 100 Stichprobenmittelwerte der RRMSEs über 16 Domains einer Gemeinde auf der y–Achse. Die Gemeinde Nr. 1 weist sowohl den höchsten Mittelwert als auch die höchste Varianz auf. Die Abbildungen I.2 bzw. I.3 auf der Seite 218 bzw. 219 zeigen für große bzw. kleine Gemeinden die Mittelwerte der RRMSEs für die 14 Domains der deutschen Population einer Gemeinde auf der x–Achse versus der Varianz der 100 Stichprobenmittelwerte der RRMSEs für 14 Domains der deutschen Population einer Gemeinde auf der y–Achse. Es ist erwähnenswert, dass die Gemeinde Nr. 1 nun einen kleinen Mittelwert und auch eine kleine Varianz aufweist, und damit die Schätzung in dieser Gemeinde eine der besten ist. Daher folgt, dass die Schätzung der nicht deutschen Population einen

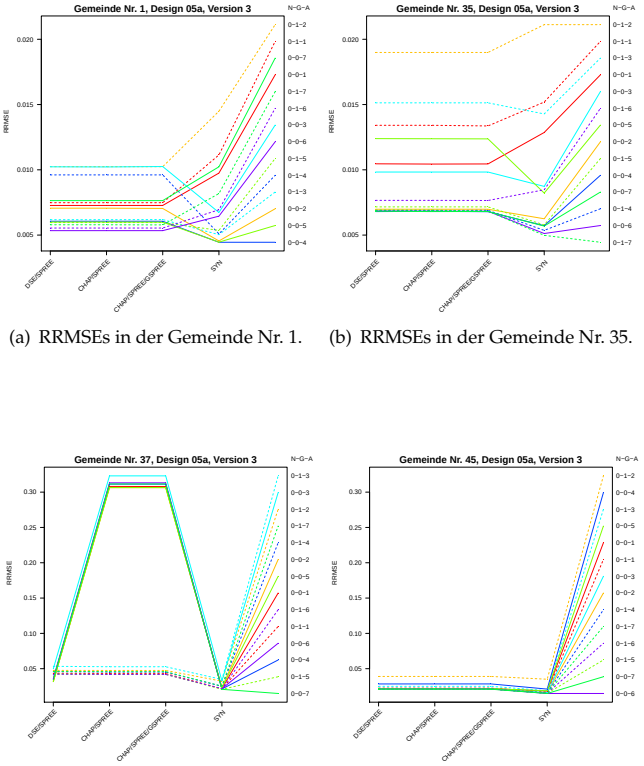


(a) RRMSEs in der Gemeinde Nr. 1. (b) RRMSEs in der Gemeinde Nr. 35.



(c) RRMSEs in der Gemeinde Nr. 37. (d) RRMSEs in der Gemeinde Nr. 45.

ABBILDUNG 7.27: *RRMSEs für die Domains in den Gemeinden Nr. 1, 35, 37 und 45. Design 05a, Version 3. 16 Linien in jeder Grafik repräsentieren 14 Domains für die deutsche Population und zwei für die nicht deutsche Population. Die Beschriftung jeder Linie definiert genau eine Domain. Die erste Ziffer in der Beschriftung ist die Nationalität, 0 für deutsch, 1 für nicht deutsch. Die zweite Ziffer ist das Geschlecht, 0 für die Männer, 1 für die Frauen. Letzte Ziffer ist die AGE-Domain, 1 bis 7 für die AGE-Domains in der deutschen Population, AGE-Domain 8 ist für die nicht deutsche Population. Zum Beispiel bedeutet die Beschriftung 0-0-2 eine Domain, die durch NAT=0, SEX=0, AGE-Domain=2 definiert ist, d.h. deutsche Männer im Alter von 18-24.*



(a) RRMSEs in der Gemeinde Nr. 1.

(b) RRMSEs in der Gemeinde Nr. 35.

(c) RRMSEs in der Gemeinde Nr. 37.

(d) RRMSEs in der Gemeinde Nr. 45.

ABBILDUNG 7.28: *RRMSEs des D–DSE/SPREEs, CHAP/SPREEs, CHAP/SPREE/GSPREEs und Verhältnis–synthetischen Schätzers für die Domains der deutschen Population (die Beschriftung beginnt mit 0) in den Gemeinden Nr. 1, 35, 37 und 45. Design 05a, Version 3.*

starken Einfluss auf die Summe der Schätzungen in der ganzen Gemeinde hat.

Grafische Darstellung der RRMSEs

Die vier Grafiken in Abbildung 7.27 zeigen die RRMSEs des Designs 05a für die untersuchten Schätzer in den Gemeinden Nr. 1, 35, 37 und 45. An der x–Achse jeder Grafik befinden sich die Schätzer, an der y–Achse links ist die Skala des RRMSEs zwischen dem gemeinsamen minimalen und maximalen Wert für die beiden großen Gemeinden oder beiden kleinen Gemeinden. Die genaue Beschreibung der Grafiken ist auf der Seite 110 zu finden.

In den zwei Grafiken der großen Gemeinden Nr. 1 und 35 sind RRMSEs für alle Schätzer in allen 16 Domains dargestellt. In den beiden Grafiken für die kleinen Gemeinden Nr. 37 und 45 fehlen RRMSEs für den GREG2–Schätzer und für EBLUP. Die RRMSEs für D–DSE/SPREE fehlen nur in zwei nicht deutschen Populationen (die Nummerierung der Domains beginnt mit 1). In diesen Fällen liefern die Schätzungen in jeder Simulation in mindestens einer Anschriftenschicht der Domain kein Ergebnis und damit werden die RRMSE nicht berechnet. Die Probleme verrät schon die höhere Anzahl der `NaN` und `Inf` in Tabelle 7.18.

Diese Probleme treten nicht im CHAP/SPREE auf, da der Chapman–Schätzer immer berechnet werden kann. Die RRMSEs für diesen Schätzer fehlen in keiner Grafik. In den Grafiken zu den Gemeinden Nr. 1, 37 und 45 sehen wir, dass die RRMSEs für CHAP/SPREE für die nicht deutsche Population (die Nummerierung der Domains beginnt mit 1) sehr hoch sind. Diese Schwankungen werden durch die Gewichtung der Anschriften und durch die Anzahl der `Inf` in diesen Domains verursacht (s. Beispiel in (7.9) auf der Seite 102).

Wie erwähnt, wird in allen Gemeinden das GSPREE–Modell angewendet, nachdem CHAP/SPREE der Domains berechnet werden. Die RRMSEs für CHAP/SPREE/GSPREE sind mit den RRMSEs für CHAP/SPREE vergleichbar.

Wegen des Ergebnisses in nicht deutschen Populationen der Gemeinde Nr. 1, 37 und 45 ist die Skala der y–Achse sehr breit und die Höhe den RRMSEs in deutschen Populationen nicht ablesbar. Für einen besseren Vergleich ist in Abbildung 7.28 ein Teil der Abbildung 7.27 dargestellt. Nur die RRMSEs für die deutsche Population (die Beschriftung der Domains beginnt nur mit 0) sind abgebildet. Die RRMSE für die Kombinationen D–DSE/SPREE, CHAP/SPREE und CHAP/SPREE/GSPREE sind nur mit RRMSEs für den Verhältnis–synthetischen Schätzer dargestellt, da die RRMSEs für diesen alternativen Schätzer am niedrigsten liegen.

Man kann sehen, dass die RRMSEs der D–DSE/SPREE, CHAP/SPREE und CHAP/SPREE/GSPREE in allen Gemeinden parallel verlaufen, mit Ausnahme in sieben Domains der deutschen Männer in der Gemeinde Nr. 37 (Domains 0-0-x, x steht für 1 bis 7). Wir können daher sagen, dass alle drei Schätzer stabil sind.

Die extrem hohen RRMSEs in CHAP/SPREE für die sieben Domains der deutschen Männer in der Gemeinde Nr. 37 kommen in der ersten Anschriftenschicht der Stichprobe Nr. 66 vor. Hier liefert der Chapman–Schätzer für die Klasse der deutschen Männer im Alter von 0-95 den Wert 11.622 und ist damit sehr weit weg von 118 tatsächlich lebenden deutschen Männern der ersten Anschriftenschicht in dieser Gemeinde. Summiert über alle neun Anschriftenschichten bekommt man 15.380 deutsche Männern in der ganzen Gemeinde Nr. 37 (vgl. mit 3.761 tatsächlich lebenden deutschen Männern in der ganzen Gemeinde Nr. 37). Diese Anzahl wird mit SPREE auf die sieben Domains 0-0-x (x steht für 1 bis 7) aufgeteilt. In Tabelle 7.21 sieht man die Anzahl der registrierten in jeder Domain benutzt in SPREE, die

Anzahl der tatsächlich lebenden in jeder Domain und die CHAP/SPREE Schätzungen der deutschen Männer in einzelnen Domains der Gemeinde Nr. 37.

Domain	τ_R	τ_Z	CHAP/ SPREE
0-0-1	625	627	2.549
0-0-2	274	276	1.117
0-0-3	212	206	864
0-0-4	648	645	2.642
0-0-5	589	593	2.402
0-0-6	826	819	3.368
0-0-7	598	595	2.438
Total	3.772	3.761	15.380

TABELLE 7.21: *Anzahl der registrierten, tatsächlich lebenden und CHAP/SPREE Schätzungen der deutschen Männer in sieben Domains der Gemeinde Nr. 37. Design 05a, Version 3.*

Die RRMSEs für den Verhältnis–synthetischen Schätzer in den großen Gemeinden Nr. 1 und Nr. 35 (s. Abbildung 7.28, Grafik (a) und (b)) sind nicht leicht einzuordnen. Für manche Domains sind sie besser als D–DSE/ SPREE, CHAP/SPREE oder CHAP/SPREE/GSPREE. Es gibt aber auch Domains, wo sie schlechter sind. In der deutschen Population der kleinen Gemeinden ist der Verhältnis–synthetische Schätzer immer besser als D–DSE/SPREE, CHAP/SPREE oder CHAP/SPREE/GSPREE.

Für einen numerischen Vergleich der RRMSEs siehe Tabelle I.1 auf der Seite 222.

Absoluter–relativer–Bias

Die Tabelle auf der Seite 223 enthält den ARB in jeder Domain der Gemeinden Nr. 1, 35, 37 und 45. `NaN` in einigen Domains der kleinen Gemeinden verraten, dass Schätzungen in keiner Stichprobe möglich sind und daher ARB nicht berechnet werden kann.

Man kann in den großen Gemeinden Nr. 1 und 35 sehen, dass die ARBs für D–DSE/SPREE der Domains für die deutsche Population (die Beschriftung beginnt mit 0) vergleichbar mit den ARBs für CHAP/SPREE und CHAP/SPREE/GSPREE sind. In der nicht deutschen Population (die Beschriftung beginnt mit 1) sind die ARBs für D–DSE/SPREE deutlich kleiner als ARBs für CHAP/SPREE oder CHAP/SPREE/GSPREE. ARBs für die alternativen Schätzer sind schwer einzuordnen. Im Durchschnitt über alle Domains sind die ARBs für GREG2–Schätzer am kleinsten (s. Tabelle 7.22).

GEM	GREG1	GREG2	SYN	EBLUP
1	0,00797	0,00295	0,01218	0,02242
35	0,00788	0,00298	0,01542	0,02346

TABELLE 7.22: *Mittelwert der ARBs über alle Domains in den Gemeinden Nr. 1 und 35. Design 05a, Version 3.*

In der deutschen Population der kleinen Gemeinden Nr. 37 und 45 sind die ARBs für D–DSE/SPREE, CHAP/SPREE und CHAP/SPREE/GSPREE vergleichbar, mit der Ausnahme der deutschen Männer (Domains 0-0-x, wo x für 1 bis 7 steht) in der Gemeinde Nr. 37. Dies spiegelt die Situation in Abbildung 7.28, (c) wieder. Am kleinsten sind die ARBs für den Verhältnis–synthetischen Schätzer.

Varianzschätzung und MSE–Schätzung

Die Abbildungen I.4 und I.5 im Anhang auf der Seite 220 geben die Varianzschätzungen für D–DSE/SPREE, CHAP/SPREE, GREG1–Schätzer, GREG2–Schätzer, Verhältnis–synthetischen Schätzer und EBLUP in jeder Domain der großen Gemeinden Nr. 1 und 37 dividiert durch ihren Varianz der Domain. Diese wird als relative Verzerrung bezeichnet und ist mathematisch in (7.11) auf der Seite 116 beschrieben.

Wegen der extrem breiten relativen Verzerrungen der Jackknife–Varianzschätzungen für D–DSE/SPREE und CHAP/SRPEE sind die Boxplots für andere Schätzer nicht gut lesbar. In den Abbildungen 7.29 bzw. 7.30 sind die relativen Verzerrungen der Varianzschätzungen für alle Schätzer in jeder Domain der großen Gemeinden Nr. 1 und 37 nochmals dargestellt, diesmal ohne Jackknife–Varianzschätzung. Es ist zu sehen, dass die relativen Verzerrungen des EBLUPs am größten sind. Am kleinsten sind die relativen Verzerrungen beim GREG1–Schätzers.

Im Nenner der Formeln für die direkte Varianzschätzung für D–DSE bzw. Chapman–Schätzer und die Varianzschätzung für GREG1–Schätzer, GREG2–Schätzer und Verhältnis–synthetischen Schätzer (siehe Kapitel 6.2.1, 6.2.2 und 6.4) steht die Anzahl der ausgewählten Anschriften einer Schicht h. Da in den kleinen Gemeinden Nr. 37 und 45 einige Schichten mit $n_h = 1$ vorkommen (s. Tabelle 7.23), ist eine direkte Varianzschätzung für diese Schicht und damit für die ganze Gemeinde nicht möglich. Daher sind in den Abbildungen 7.31 und 7.32 nur die relativen Jackknife–Varianzschätzungen für D–DSE/SPREE und CHAP/SPREE für alle Domains der kleinen Gemeinden Nr. 37 und 45 dargestellt.

Die gleiche Form der Boxplots in der Population der deutschen Männer bzw. deutschen Frauen kommt aus dem Aufbau der Klassen und der Art der Abbildung der Varianzschätzung. Die relative Schätzung von MSE für

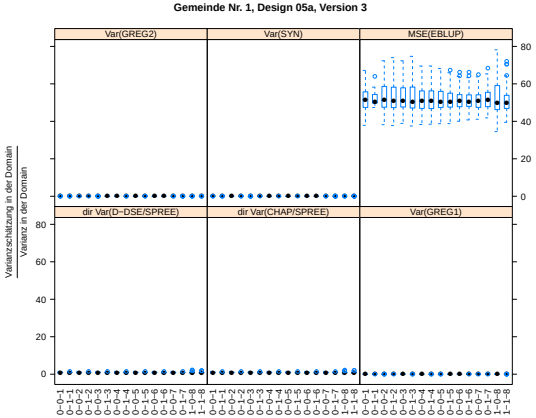


ABBILDUNG 7.29: *Relative Verzerrung der Varianzschätzungen auf Domainbasis in der Gemeinde Nr. 1. Ohne Jackknife–Varianzschätzung. Design 05a, Version 3.*

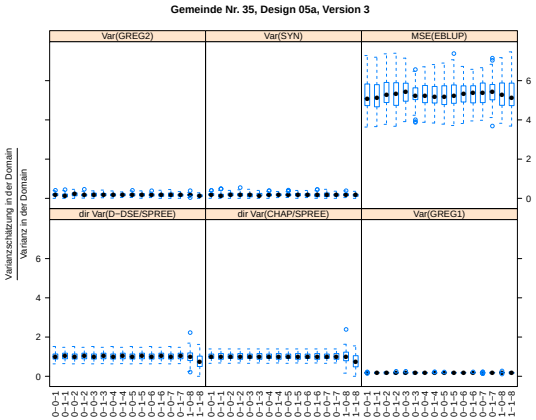


ABBILDUNG 7.30: *Relative Verzerrung der Varianzschätzungen auf Domainbasis in der Gemeinde Nr. 35. Ohne Jackknife–Varianzschätzung. Design 05a, Version 3.*

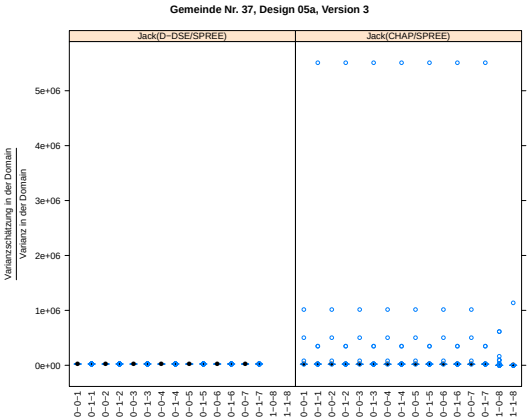


ABBILDUNG 7.31: *Relative Verzerrung der Varianzschätzungen auf Domainbasis in der Gemeinde Nr. 37. Nur Jackknife–Varianzschätzung. Design 05a, Version 3.*

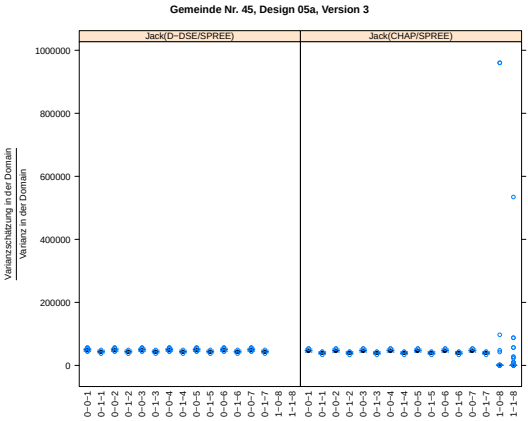


ABBILDUNG 7.32: *Relative Verzerrung der Varianzschätzungen auf Domainbasis in der Gemeinde Nr. 45. Nur Jackknife–Varianzschätzung. Design 05a, Version 3.*

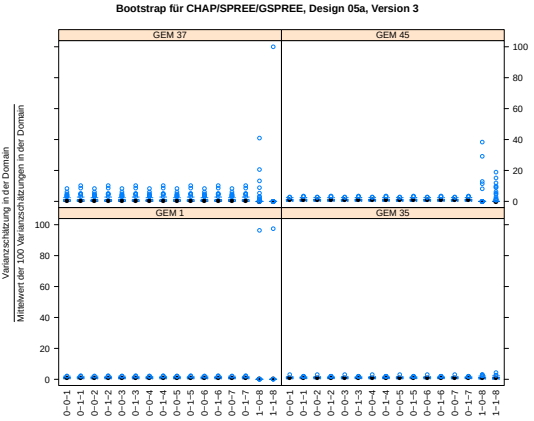


ABBILDUNG 7.33: *Relative Verzerrung der Bootstrap–Varianzschätzungen des CHAP/SPREE/GSPREEs auf Domainbasis in Gemeinden Nr. 1, 35, 37 und 45. Design 05a, Version 3.*

EBLUP ist auch nicht dargestellt, da die Definition der Domains sehr fein ist und EBLUP bei geschichteter Zufallsauswahl in den kleinen Gemeinden nicht anwendbar war.

Zu bemerken ist, dass in Abbildung 7.31 und 7.32 die Jackknife–Varianzschätzung für D–DSE/SPREE in den Domains 1-0-8 bzw. 1-1-8 nicht abgebildet ist. Es spiegelt die Situation in der Abbildung 7.27, (c) und (d) der RRMSEs für diese Domains wieder, wo eine Berechnung von RRMSE nicht möglich war. Allgemein lässt sich sagen, dass die relativen Verzerrungen der Jackknife–Varianzschätzung für D–DSE/SPREE in den kleinen Gemeinden kleiner ist als für CHAP/SPREE. Der Ursprung der Ausreißer in der Jackknife–Varianzschätzung für CHAP/SPREE ist ähnlich wie auf der Seite 122 beschrieben.

Wie im Kapitel 6.7 beschrieben ist, muss bei der geschichteten Zufallsauswahl in einer Schicht h die Bedingung $n_h \geq 2$ erfüllt sein, damit man

die Bootstrap–Methode für die Varianzschätzung des GSPREEs anwenden kann. Diese Bedingung ist in unserer Simulations–Population nicht immer erfüllt, vor allem nicht in den kleinen Gemeinden. Tabelle 7.23 gibt die Anzahl der Anschriften N_h und der ausgewählten Anschriften n_h pro Schicht h in den untersuchten Gemeinden Nr. 1, 35, 37 und 45 an. Es ist zu sehen, dass in drei bzw. zwei Schichten der Gemeinde Nr. 37 bzw. 45 nur eine Anschrift gezogen wird. An Hand der Tabelle müssen einige Schichten zusammengefasst werden.

Schicht	Gemeinde							
	1		35		37		45	
	N_h	n_h	N_h	n_h	N_h	n_h	N_h	n_h
1	5.158	96	518	94	390	4	357	8
2	5.303	98	677	122	458	4	470	11
3	4.281	80	487	88	362	4	339	8
4	3.628	67	450	81	340	3	297	7
5	2.808	52	313	57	225	2	193	4
6	2.187	41	187	34	155	1	176	4
7	3.823	71	324	58	240	2	239	5
8	577	11	40	7	26	1	26	1
9	1.833	34	51	9	36	1	37	1
Total	29.598	550	3.047	550	2.232	22	2.134	49

TABELLE 7.23: *Anzahl der Anschriften N_h und der ausgewählten Anschriften n_h pro Schicht und pro untersuchter Gemeinde Nr. 1, 35, 37 und 45.*

Ausgangspunkt in jeder Gemeinde sind die bekannten n_h. Falls in der ersten Schicht nur eine Anschrift ausgewählt ist, wird diese Schicht zusam-

Gemeinde 37			
ursprünglich		neu	
Schicht	n_h	Schicht	n_h
1	4	1	4
2	4	2	4
3	4	3	4
4	3	4	3
5	2	5	3
6	1		
7	2	6	4
8	1		
9	1		
Total	22	Total	22

Gemeinde 45			
ursprünglich		neu	
Schicht	n_h	Schicht	n_h
1	8	1	8
2	11	2	11
3	8	3	8
4	7	4	7
5	4	5	4
6	4	6	4
7	5	7	7
8	1		
9	1		
Total	49	Total	49

TABELLE 7.24: *Die Neuschichtungen in der Gemeinde Nr. 37 und 45.*

alle Gemeinden	
ursprünglich	alternativ
Schicht	Schicht
1	1
2	
3	
4	2
5	
6	
7	3
8	
9	

TABELLE 7.25: *Die alternative Schichtung in allen Gemeinden.*

men mit der zweiten Schicht der Gemeinde verbunden. Falls $n_h = 1$ für $h = 2, \ldots, 9$ wird diese Schicht h zusammen mit der vorigen Schicht $h - 1$ der Gemeinde verbunden. Auf die neuen Schichten wird die Bootstrap–Methode angewendet, um die Varianzschätzung des GSPREEs berechnen zu können. Die neue Schichtung für die Gemeinde Nr. 37 bzw. 45 gibt Tabelle 7.24, Spalten „neu“ an.

Bei der Untersuchung dieser Neuschichtung kommen beim Chapman–Schätzer gewisse Probleme vor, die eine Bootstrap–Varianzschätzung für GSPREE verhindern. Als alternative Lösung werden sowohl in großen als auch kleinen Gemeinden die Schichten nur in drei Schichten aufgeteilt (s. Tabelle 7.25).

Wegen des Vergleichs mit Design 01 (s. Abbildung 7.15 und 7.25) wird die relative Verzerrung der Bootstrap–Varianzschätzung von CHAP/SPREE/GSPREE für alle untersuchten Gemeinden separat in einer Grafik abgebildet (s. Abbildung 7.33 auf Seite 164). Allgemein können wir sagen, dass es Unterschiede bei den Varianzschätzungen von GSPREE in kleinen und großen Gemeinden gibt. Bei der deutschen Population hat die alternative Schichtung für die Bootstrap–Varianzschätzung gut funktioniert. Die relative Verzerrung beim Design 05a, Version 3 ist kleiner und damit besser als beim Design 01, Version 3 (vgl. Abbildung H.9 mit Abbildung I.6).

Konfidenzintervalle

Die Abbildung 7.34 auf der Seite 169 stellt die Überdeckungsrate versus Mittelwert über 100 Stichproben von relativen Konfidenzintervalllängen aller Domains d der Gemeinden Nr. 1, 35, 37 und 45 dar. Die mathematische Formel ist in (7.12) zu finden. Die Domains der großen Gemeinden sind blau abgebildet, die Domains der kleinen Gemeinden violett. Die grüne Linie repräsentiert die Überdeckungsrate 95% (α = 0,05).

Wie oben erwähnt können nicht alle betrachteten Varianzschätzungen in den kleinen Gemeinden berechnet werden. Daher sind in Abbildung 7.34 nur die Überdeckungsrate der Jackknife–Varianzschätzungen für D–DSE/SPREE und CHAP/SPREE und die Überdeckungsrate der Bootstrap–Varianzschätzung für CHAP/SPREE/GSRPEE für die beiden kleinen Gemeinden dargestellt.

Die relativen Konfidenzintervalllängen der direkten Varianzschätzungen für D–DSE/SPREE und CHAP/SPREE in großen Gemeinden sind kurz, die Überdeckungsraten sind aber niedrig. Hingegen sind die relativen Konfidenzintervalllängen der Jackknife–Varianzschätzungen für D–DSE/SPREE und CHAP/SPREE in großen Gemeinden lang und daher die Überdeckungsraten der Domains höher als 95%. In beiden großen Gemeinden liefert die Varianzschätzung für den GREG1–Schätzer die kürzesten Konfidenzintervalllängen und höhste Überdeckungsrate. Demgegenüber ist die Schätzung von MSE für EBLUP. Für die kleinen Gemeinden sind nur die relativen Konfidenzintervalllängen der Jackknife–Varianzschätzungen für D–DSE/SPREE und CHAP/SPREE und Bootstrap–Varianzschätzung für CHAP/SPREE/GSRPEE dargestellt.

7.3.3 Zusammenfassung bei geschichteter Zufallsauswahl

Bei der geschichteten Zufallsauswahl (Design 05a) haben wir uns auf Version 3 beschränkt (s. Seite 150). Eine Übersicht über alle Szenarien befindet sich in Tabelle E.1 auf der Seite 196. Alle Grafiken und Tabellen für Design 05a sind im Text von Kapitel 7.3 und im Anhang auf den Seiten 217-223 zu finden.

Sind die Anschriftenschichten wie auf der Seite 148 gegeben, gibt es in beiden Gemeindegrößen bei der Anwendung des D–DSE/SPREEs für die Domains Probleme mit `NaN` und `Inf` (s. Tabelle 7.18).

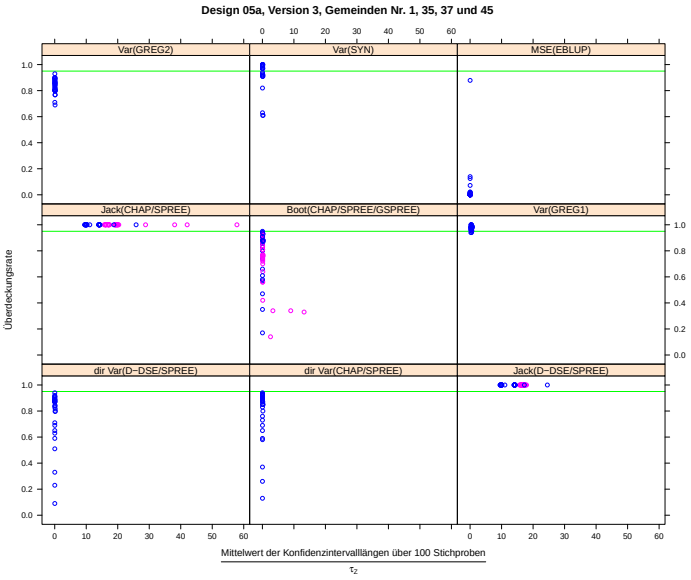


ABBILDUNG 7.34: *Überdeckungsrate versus Mittelwert der relativen Konfidenzintervalllängen über 100 Stichproben für die Domains der Gemeinden Nr. 1, 35, 37 und 45. Design 05a, Version 3. Die Domains der großen Gemeinden sind blau dargestellt, Domains der kleinen Gemeinden violett. Die grüne Linie repräsentiert die Überdeckungsrate 95% (α = 0,05).*

Für die deutsche Population der kleinen Gemeinden ist aus den Tabellen 7.19 und 7.20 die Anzahl der `NaN` und `Inf` zu sehen. Bei den deutschen Männern der untersuchten Gemeinde Nr. 37 tritt ein `Inf` auf. In der untersuchten Gemeinde Nr. 45 ist dies nicht der Fall. Dieses `Inf` erzeugt eine unrealistische CHAP/SPREE Schätzung in der Gemeinde Nr. 37, die extrem hohe RRMSEs verursacht (s. Abbildung 7.28, (c) und (d)).

Bei der deutschen Population in großen Gemeinden kommen keine Probleme mit `NaN` und `Inf` vor. Die RRMSEs für D–DSE/SPREE, CHAP/SPREE und CHAP/SPREE/GSPREE verlaufen parallel und sind damit miteinander vergleichbar (s. Abbildung 7.28, (a) und (b)).

In beiden Gemeindegrößen bei der nicht deutschen Population (Domains 1-0-8 und 1-1-8) treten die Probleme mit `NaN` und `Inf` unterschiedlich auf. Daher verlaufen die RRMSEs für D–DSE/SPREE, CHAP/SPREE und CHAP/SPREE/GSPREE in der Gemeinde Nr. 1 nicht mehr parallel (s. Abbildung 7.27, (a)). Die Gemeinde Nr. 35 ist ohne Probleme (s. Abbildung 7.27, (b)). Demgegenüber erhalten wir in beiden kleinen Gemeinden Nr. 37 und 45 für D–DSE/SPREE in jeder Simulation in mindestens einer Anschriftenschicht kein Ergebnis. RRMSEs für D–DSE/SPREE können in diesen beiden Fällen nicht berechnet werden (s. Abbildung 7.27, (c) und (d)).

Als Varianzschätzung ist in den großen Gemeinden sowohl die direkte Varianzschätzung, die Jackknife–Varianzschätzung als auch die Bootstrap–Varianzschätzung für D–DSE/SPREE, CHAP/SPREE bzw. CHAP/SPREE/GSPREE möglich (s. Abbildungen I.4, I.5 und Abbildung 7.33). Allerdings wurde im Fall der Bootstrap–Varianzschätzung die ursprüngliche Schichtung vergröbert.

In den kleinen Gemeinden kann man wegen der Definition der Varianzschätzungen nur die Jackknife–Varianzschätzung für D–DSE/SPREE und

CHAP/SPREE bzw. die Bootstrap–Varianzschätzung für CHAP/SPREE/GSPREE ermitteln. Die Verzerrungen dieser Varianzschätzungen sind deutlich höher in den Domains mit CHAP/SPREE unrealistischen Ergebnissen (s. Abbildung 7.31 und Abbildung 7.32).

8 Zusammenfassung und Ausblick

In den Jahren 2010, 2011 sollten nach der Europäischen Union (EU) alle Mitgliedstaaten einen Zensus durchführen. Auch Deutschland wird sich am Zensus 2011 beteiligen. Unter anderem wird ermittelt, wie viele Menschen in Deutschland leben. Der Zensustest 2001 (Statistisches Bundesamt 2004) hat gezeigt, dass ein neues Verfahren in Deutschland, der registergestützte Zensus, in der Lage ist, zuverlässige statistische Daten über die Bevölkerung zu liefern. Um für diese Form des Zensus optimale Designs und Schätzungen zu entwickeln, wurde von den Statistischen Ämter des Bundes und der Länder ein Forschungsprojekt aufgesetzt. Ziel dieses Projekts war die Erforschung einer optimalen Stichprobenstrategie, d.h. von Design und Schätzung.

Hilfreich für die Untersuchungen waren die im Rahmen des DACSEIS Projekts behandelten Stichprobendesigns. Die Beschreibung der Designs findet man in Münnich et al. (2008). Alle Designs beziehen sich auf eine Ziehung der Anschriften einer Gemeinde. In der vorliegenden Dissertation wird von dieser Voruntersuchung ausgegangen und zwei vorgeschlagene Stichprobendesigns im Zusammenhang mit den Schätzern genauer untersucht – die uneingeschränkte Zufallsauswahl (Design 01) und die geschichtete Zufallsauswahl mit Mindeststichprobenumfang 1 pro Schicht (Design 05a). Die Schätzverfahren werden auf der rein synthetischen Simulations–Population für das Bundesland Saarland mit 1.057.915 Einträge getestet, die im Rahmen des DACSEIS Projekts erzeugt wurde und am

Anfang des Forschungsprojekts, der sogenannten Phase 0, für den Zensus 2011 zur Verfügung stand.

Das Ziel der vorliegenden Dissertation ist es, zu untersuchen, ob auch Dual–System–Modelle beim kommenden Zensus angewendet werden können, um die amtliche Einwohnerzahl zu ermitteln.

Der Dual–System–Estimator (D–DSE) geht in Deutschland von drei Kategorien aus, im Gegensatz zu vier beim klassischen Ansatz der Dual–System–Modelle in den USA oder in der Schweiz. Man setzt voraus, dass es keine Personen gibt, die sowohl im Register fehlen als auch in der Gemeinde nicht wohnhaft sind. Bestandteil der Modelle ist die Bildung bestimmter Subpopulationen (sogenannten Klassen), innerhalb derer der D–DSE berechnet wird. In der vorliegenden Dissertation werden drei unterschiedliche Definitionen der Klassen untersucht, die in Version 1, Version 2 bzw. Version 3 definiert sind. Bei den Klassen kann es vorkommen, dass D–DSE ein `NaN` oder `Inf` liefert (s. Beispiel auf der Seite 100 bzw. 101). Um diese Probleme zu vermeiden, wird der Chapman–Schätzer verwendet. Dabei wird die Anzahl der Personen Kategorie 00 (sowohl im Register verzeichnet als auch wohnhaft in einer Gemeinde) künstlich um eins erhöht. Damit bereinigt der Chapman–Schätzer die Situationen, in denen beim D–DSE im Nenner und/oder Zähler eine „0“ steht.

In der amtlichen Statistik bildet die Schätzung von Subpopulationen (sogenannte Domains), die demografisch abgegrenzt sein können, einen besonderen Schwerpunkt. Diese Schätzung ist allgemein als Small–Area–Schätzung bekannt und wird in dieser Dissertation ebenfalls betrachtet. Um die Umfänge von Domains zu schätzen, wird innerhalb einer Gemeinde der strukturerhaltende–Schätzer (SPREE) auf D–DSE bzw. den Chapman–Schätzer angewendet. Zuallerletzt wird ein verallgemeinerter SPREE (GSPREE) untersucht, der im gesamten Bundesland Saarland verwendet wird und die Interaktionen zwischen den Gemeinden berücksichtigt.

Zwei Methoden der Varianzschätzung für D–DSE und Chapman–Schätzer werden betrachtet – die direkte Varianzschätzung und die Jackknife–Varianzschätzung. Da SPREE eine lineare Kombination des D–DSEs bzw. Chapman–Schätzers ist, lässt sich die Varianzschätzung einfach herleiten. Für die GSPREE–Varianzschätzung wird die Bootstrap–Methode untersucht.

Zu Vergleichszwecken werden andere Schätzer wie der verallgemeinerte–Regressionsschätzer (GREG1–Schätzer bzw. GREG2–Schätzer), der Verhältnis–synthetische Schätzer und EBLUP für die Schätzung der Domains herangezogen. Diese alternativen Schätzer und ihre Varianzschätzungen sind in Rao (2003) zu finden.

In die grafische Präsentation der Ergebnisse werden in der vorliegenden Dissertation nur vier Gemeinden einbezogen. Da die Grenze von 10.000 registrierten Einwohnern sehr wichtig ist, werden zwei Gemeinden über und zwei Gemeinden unter 10.000 registrierten Einwohnern genauer untersucht. Die Aussagen zur Genauigkeit aller untersuchten Schätzer werden durch unterschiedliche Abbildungen wie Boxplots der geschätzten Totalwerte, Abbildungen der „relative root mean square errors“ RRMSEs, der relativen Verzerrungen der Varianzschätzungen oder der Konfidenzintervalllängen versus Überdeckungsraten gemacht. Zusätzlich werden auch numerische Werte der RRMSE und „absolute relative bias“ ARB jeder Domain in Tabellen präsentiert, um die Güte des Schätzers zu beurteilen.

Als Fazit lassen sich daraus folgende Ergebnisse und Empfehlungen im Zusammenhang mit dem D–DSE und Chapman–Schätzer für die folgenden Bevölkerungsgruppen ableiten:

1. Deutsche Population in großen Gemeinden

Die Dual–System–Modelle sind auf die deutsche Population der großen Gemeinden anwendbar. Der Ziehungsprozess der Stichprobe spielt keine wesentliche Rolle. Die geschätzten Totalwerte der Domains sind bei beiden Stichprobendesigns vergleichbar. Wir empfehlen, den D–DSE oder den Chapman–Schätzer innerhalb einer Geschlecht×Nationalität Klasse (Version 3) mit anschließender SPREE Aufteilung auf die einzelnen Domains zu berechnen. Das GSPREE–Modell kann auf CHAP/SPREE der Domains angewendet werden, bringt aber keine große Verbesserung. Darüber hinaus darf nicht vergessen werden, dass für GSPREE zuerst der Chapman–Schätzer benötigt wird.

Innerhalb eines Ziehungsprozesses sind die relativen Verzerrungen der direkten Varianzschätzung für D–DSE/SPREE bzw. CHAP/SPREE vergleichbar. Dies gilt auch für die Jackknife–Varianzschätzung der beiden Schätzer. Allerdings empfiehlt sich die direkte Varianzschätzung, da diese Varianzschätzung innerhalb eines Designs weniger als die Jackknife–Varianzschätzung verzerrt ist.

2. Deutsche Population in kleinen Gemeinden

Beim Design 01, Version 3 können die Dual–System–Modelle auf die deutsche Population der kleinen Gemeinden angewendet werden. Wegen der Definition der Anschriftenschichten im Design 05a, Version 3 gibt es in D–DSE/SPREE Probleme mit `NaN` und `Inf`. Daher ist der Schätzer bei diesem Design nicht zu empfehlen. Wir können den Chapman–Schätzer zwar ermitteln, dennoch kann er bei manchen Stichproben ein Ergebnis weit weg von der Realität liefern (s. Seite 158). CHAP/SPREE/GSPREE bringt auch kein wertvolles Ergebnis, da er vom Chapman–Schätzer ausgegangen ist.

In solchen Situationen kann auch der GREG1–Schätzer oder der Verhältnis–synthetische Schätzer ermittelt werden.

Beim Design 01, Version 3 ist die direkte Varianzschätzung für D–DSE/SPREE und CHAP/SPREE zu empfehlen, da diese Varianzschätzung weniger als die Jackknife–Varianzschätzung verzerrt ist. Im Design 05a, Version 3 ist sinnvoll nur die Jackknife–Varianzschätzung für CHAP/SPREE oder Bootstrap–Varianzschätzung für CHAP/SPREE/GSPREE. Es ist schwer, zwischen der verzerrten Jackknife–Varianzschätzung mit breiten Konfidenzintervallen und der weniger verzerrten Bootstrap–Varianzschätzung mit niedrigen Überdeckungsraten zu entscheiden.

3. Nicht deutsche Population

Die Anwendung der Dual–System–Modelle in der nicht deutschen Population ist allgemein nicht zu empfehlen. Der Anteil dieser Population in einer Gemeinde ist sehr gering und daher schwer zu schätzen. Die Anzahl der Schätzungen, bei denen D–DSE/SPREE kein Ergebnis beim Design 01, Version 3 liefert, ist hoch. Beim Design 05a, Version 3 kommen diese Probleme bei allen Simulationen vor, dass die RRMSE, ARB und die Jackknife–Varianzschätzung für D–DSE/SPREE nicht berechnet werden können. Es gibt die Möglichkeit, CHAP/SPREE bzw. CHAP/SPREE/GSPREE zu berechnen, die aber unrealistische Ergebnisse liefern. Bei beiden Designs kann der GREG1–Schätzer oder der Verhältnis–synthetische Schätzer ermittelt werden.

Ausblick

Vorschläge für weitere Untersuchungen betreffen gewisse Bausteine der Simulationen. Eine Optimierung der Schichtanzahl und der Schichtgren-

zen könnte bedeutsam sein. Eine andere Untersuchungsvariante wäre, den Mindeststichprobenumfang pro Schicht zu erhöhen. Nicht zuletzt wäre natürlich interessant, die Dual–System–Modelle mit der Simulations–Population mit 85.790.381 Einträgen zu testen, die im Laufe des Forschungsprojekts für den Zensus 2011 mit Hilfe der Registerdaten erzeugt wurde.

A Anzahl der registrierten Personen

GEM	$\tau_{R,<g>}$	$N_{<g>}$	$n_{<g>}$	nicht Deutsche [%]	KRS
50	5.282	1.422	32	3,73	6
15	5.632	1.512	31	4,99	2
17	5.632	1.554	31	4,15	2
36	6.993	1.628	18	3,72	4
47	7.797	2.134	48	3,58	6
49	7.850	2.206	48	3,35	6
40	7.886	2.229	30	3,53	5
45	7.953	2.134	49	3,14	6
37	8.221	2.232	22	4,29	4
33	10.956	3.005	550	4,73	4
27	10.972	3.020	550	4,74	4
35	11.150	3.047	550	4,45	4
3	11.214	3.143	550	5,12	1
14	11.226	3.057	550	4,26	2
46	11.423	2.947	550	3,61	6
42	11.451	2.992	550	3,28	5
43	11.471	3.019	550	3,34	5
48	11.615	3.078	550	3,31	6
52	13.189	3.618	550	3,48	6
5	13.614	3.855	550	5,19	1
30	13.633	3.812	550	4,54	4
20	13.784	3.723	550	3,55	3
25	14.162	3.822	550	4,24	4
22	14.922	3.830	550	3,65	3
24	15.050	3.886	550	3,72	3
2	15.247	3.993	550	5,40	1
11	15.629	3.945	550	3,98	2
12	16.569	4.485	550	4,39	2
28	18.229	4.712	550	4,17	4

- Fortsetzung auf der nächsten Seite -

- Fortsetzung von der vorherigen Seite -

GEM	$\tau_{R,<g>}$	$N_{<g>}$	$n_{<g>}$	nicht Deutsche [%]	KRS
31	18.269	4.623	550	4,21	4
16	18.496	4.742	550	4,04	2
18	19.423	5.162	550	3,37	3
19	19.636	5.202	550	3,50	3
23	19.664	5.150	550	3,48	3
7	20.064	5.017	550	4,78	1
8	20.090	5.014	550	5,30	1
38	20.227	4.747	550	3,26	5
32	21.006	4.895	550	4,03	4
34	21.037	4.868	550	4,26	4
26	21.214	4.765	550	4,26	4
9	23.109	5.334	550	5,96	1
39	23.239	5.416	550	3,48	5
4	24.682	6.294	550	4,70	1
51	25.085	5.848	550	3,66	6
6	25.907	6.369	550	5,16	1
13	25.946	6.061	550	4,13	2
29	32.318	6.851	550	3,76	4
41	32.847	7.213	550	4,16	5
44	37.273	8.585	550	3,88	5
21	39.100	8.528	550	4,21	3
10	46.636	10.468	550	6,05	1
1	157.205	29.598	550	7,28	1

TABELLE A.1: *Die Gemeinden mit der Anzahl der registrierten Personen, $\tau_{R,<g>}$, Anzahl der Anschriften, $N_{<g>}$, Anzahl der ausgewählten Anschriften, $n_{<g>}$, Prozentsatz der nicht Deutschen Personen in den Gemeinden und Kreis, zu dem die Gemeinde gehört. Angeordnet nach $\tau_{R,<g>}$.*

KRS	6	2	3	5	4	1
$\tau_{R,\ll k\gg}$	90.194	99.130	141.579	144.394	208.160	357.768

TABELLE A.2: *Die Anzahl der registrierten Personen in den Kreisen, $\tau_{R,\ll k\gg}$. Angeordnet nach $\tau_{R,\ll k\gg}$.*

B Designgewichte

GEM	w_a
1	53.8145454545455
2	7.26
3	5.71454545454545
4	11.4436363636364
5	7.00909090909091
6	11.58
7	9.12181818181818
8	9.11636363636364
9	9.69818181818182
10	19.0327272727273
11	7.17272727272727
12	8.15454545454545
13	11.02
14	5.55818181818182
15	48.7741935483871
16	8.62181818181818
17	50.1290322580645
18	9.38545454545454
19	9.45818181818182
20	6.76909090909091
21	15.5054545454545
22	6.96363636363636
23	9.36363636363636
24	7.06545454545455
25	6.94909090909091
26	8.66363636363636
27	5.49090909090909
28	8.56727272727273
29	12.4563636363636
30	6.93090909090909
31	8.40545454545454
32	8.9
33	5.46363636363636
34	8.85090909090909
35	5.54
36	90.4444444444444
37	101.454545454545
38	8.63090909090091
39	9.84727272727273
40	74.3
41	13.1145454545455
42	5.44
43	5.48909090909091
44	15.6090909090909
45	43.5510204081633
46	5.35818181818182
47	45.3333333333333
48	5.59636363636364
49	45.9583333333333
50	44.4375
51	10.6327272727273
52	6.57818181818182

TABELLE B.1: *Designgewichte für Design 01*

GEM	Schicht	w_a
1	1	53.7291666666667
1	2	54.1122448979592
1	3	53.5125
1	4	54.1492537313433
1	5	54
1	6	53.3414634146341
1	7	53.8450704225352
1	8	52.4545454545455
1	9	53.9117647058824
2	1	7.22429906542056
2	2	7.28813559322034
2	3	7.3
2	4	7.22857142857143
2	5	7.23913043478261
2	6	7.22222222222222
2	7	7.25454545454545
2	8	7.375
2	9	7.3
3	1	5.67256637168142
3	2	5.71311475409836
3	3	5.72164948453608
3	4	5.72058823529412
3	5	5.71739130434783
3	6	5.71428571428571
3	7	5.68627450980392
3	8	6.2
3	9	5.92307692307692
4	1	11.421568627451
4	2	11.3809523809524
4	3	11.4285714285714
4	4	11.4927536231884
4	5	11.5357142857143
4	6	11.3488372093023
4	7	11.5223880597015
4	8	12.125
4	9	11.125
5	1	6.96491228070175
5	2	7.01587301587302
5	3	7.04494382022472
5	4	7.02816901408451
5	5	7.08333333333333
5	6	7.02857142857143
5	7	7.0204081632653
5	8	6.6
5	9	6.76923076923077
6	1	11.5346534653465
6	2	11.6442307692308
6	3	11.675
6	4	11.5
6	5	11.5344827586207
6	6	11.6744186046512
6	7	11.536231884058
6	8	11.5
6	9	11.4761904761905
7	1	9.12745098039216
7	2	9.08653846153846
7	3	9.11904761904762

- Fortsetzung auf der nächsten Seite -

- Fortsetzung von der vorherigen Seite -

GEM	Schicht	w_a
7	4	9.11764705882353
7	5	9.20408163265306
7	6	9.1304347826087
7	7	9.17391304347826
7	8	8.81818181818182
7	9	9.05882352941176
8	1	9.14563106796117
8	2	9.1047619047619
8	3	9.16666666666667
8	4	9.08695652173913
8	5	9.03846153846154
8	6	9.1304347826087
8	7	9.1159420289855
8	8	9.33333333333333
8	9	9
9	1	9.65346534653465
9	2	9.6504854368932
9	3	9.71621621621622
9	4	9.71875
9	5	9.65384615384615
9	6	9.66666666666667
9	7	9.73239436619718
9	8	10
9	9	9.83333333333333
10	1	19.0531914893617
10	2	19
10	3	19.025641025641
10	4	19.1212121212121
10	5	19.1346153846154
10	6	19.1666666666667
10	7	18.8701298701299
10	8	18.4285714285714
10	9	19.2258064516129
11	1	7.1578947368421
11	2	7.16822429906542
11	3	7.18888888888889
11	4	7.2
11	5	7.24528301886792
11	6	7.125
11	7	7.13559322033898
11	8	7.28571428571428
11	9	7.05263157894737
12	1	8.12765957446809
12	2	8.11382113821138
12	3	8.16470588235294
12	4	8.18987341772152
12	5	8.14814814814815
12	6	8.2432432432432
12	7	8.21311475409836
12	8	8.125
12	9	7.88888888888889
13	1	11.0740740740741
13	2	11.0285714285714
13	3	11.0779220779221
13	4	10.9620253164557
13	5	10.9821428571429
13	6	10.9375
13	7	10.9358974358974

- Fortsetzung auf der nächsten Seite -

- Fortsetzung von der vorherigen Seite -

GEM	Schicht	w_a
13	8	11.5
13	9	11.1875
14	1	5.56701030927835
14	2	5.52941176470588
14	3	5.57317073170732
14	4	5.51190476190476
14	5	5.58928571428571
14	6	5.53846153846154
14	7	5.59649122807018
14	8	5.57142857142857
14	9	5.77777777777778
15	1	49.8
15	2	42.875
15	3	48
15	4	48.75
15	5	55.6666666666667
15	6	53
15	7	54.6666666666667
15	8	26
15	9	0
16	1	8.62244897959184
16	2	8.65811965811966
16	3	8.63095238095238
16	4	8.58666666666667
16	5	8.61818181818182
16	6	8.55263157894737
16	7	8.61818181818182
16	8	8.22222222222222
16	9	8.8421052631579
17	1	52.6
17	2	48.5714285714286
17	3	45.5
17	4	54.25
17	5	49.3333333333333
17	6	55
17	7	43
17	8	0
17	9	0
18	1	9.3804347826087
18	2	9.46666666666667
18	3	9.35714285714286
18	4	9.37179487179487
18	5	9.3392857142857
18	6	9.425
18	7	9.31666666666667
18	8	9.375
18	9	9.33333333333333
19	1	9.5
19	2	9.41525423728813
19	3	9.52272727272727
19	4	9.41558441558442
19	5	9.41176470588235
19	6	9.5128205128205
19	7	9.48387096774193
19	8	9.33333333333333
19	9	9.33333333333333
20	1	6.77894736842105
20	2	6.7603305785124

- Fortsetzung auf der nächsten Seite -

- Fortsetzung von der vorherigen Seite -

GEM	Schicht	w_a
20	3	6.79545454545454
20	4	6.78048780487805
20	5	6.71739130434783
20	6	6.78571428571429
20	7	6.73684210526316
20	8	6.88888888888889
20	9	6.7
21	1	15.5061728395062
21	2	15.5436893203883
21	3	15.4268292682927
21	4	15.5540540540541
21	5	15.5740740740741
21	6	15.3913043478261
21	7	15.5512820512821
21	8	15.3636363636364
21	9	15.4285714285714
22	1	6.94897959183673
22	2	6.94444444444444
22	3	6.98850574712644
22	4	6.96153846153846
22	5	6.89795918367347
22	6	7.02380952380952
22	7	6.95238095238095
22	8	7
22	9	7.11764705882353
23	1	9.4065934065934
23	2	9.38392857142857
23	3	9.32608695652174
23	4	9.3076923076923
23	5	9.32075471698113
23	6	9.41463414634146
23	7	9.4375
23	8	9.42857142857143
23	9	9.08333333333333
24	1	7.07954545454545
24	2	7.075
24	3	7.05747126436782
24	4	7.06493506493507
24	5	7.12962962962963
24	6	7.025
24	7	7.0701754385965
24	8	6.81818181818182
24	9	7
25	1	6.96907216494845
25	2	6.9396551724138
25	3	6.93023255813953
25	4	6.93506493506494
25	5	7
25	6	6.925
25	7	6.96666666666667
25	8	7
25	9	6.75
26	1	8.6829268292683
26	2	8.63829787234043
26	3	8.61038961038961
26	4	8.61333333333333
26	5	8.73214285714286
26	6	8.56521739130435

- Fortsetzung auf der nächsten Seite -

- Fortsetzung von der vorherigen Seite -

GEM	Schicht	w_a
26	7	8.7283950617284
26	8	8.76923076923077
26	9	8.76923076923077
27	1	5.47169811320755
27	2	5.5045045045045
27	3	5.47777777777778
27	4	5.46666666666667
27	5	5.47272727272727
27	6	5.53658536585366
27	7	5.5
27	8	5.5
27	9	5.75
28	1	8.61052631578947
28	2	8.55084745762712
28	3	8.57471264367816
28	4	8.5625
28	5	8.56
28	6	8.54054054054054
28	7	8.5
28	8	8.6
28	9	8.70588235294118
29	1	12.4204545454545
29	2	12.4903846153846
29	3	12.5333333333333
29	4	12.3684210526316
29	5	12.4385964912281
29	6	12.5128205128205
29	7	12.4487179487179
29	8	12.7692307692308
29	9	12.25
30	1	6.98969072164948
30	2	6.92682926829268
30	3	6.91489361702128
30	4	6.96052631578947
30	5	6.88461538461538
30	6	6.86486486486486
30	7	6.93103448275862
30	8	6.57142857142857
30	9	7.16666666666667
31	1	8.39784946236559
31	2	8.38938053097345
31	3	8.44318181818182
31	4	8.46052631578947
31	5	8.41818181818182
31	6	8.33333333333333
31	7	8.4406779661017
31	8	8.22222222222222
31	9	8.22222222222222
32	1	8.9186046511628
32	2	8.90291262135922
32	3	8.85526315789474
32	4	8.91891891891892
32	5	8.94230769230769
32	6	8.88372093023256
32	7	8.86585365853658
32	8	8.75
32	9	9.04545454545454
33	1	5.46464646464646

- Fortsetzung auf der nächsten Seite -

- Fortsetzung von der vorherigen Seite -

GEM	Schicht	w_a
33	2	5.44166666666667
33	3	5.45882352941176
33	4	5.42682926829268
33	5	5.47058823529412
33	6	5.47368421052632
33	7	5.47457627118644
33	8	5.75
33	9	5.75
34	1	8.84883720930232
34	2	8.82
34	3	8.83783783783784
34	4	8.85714285714286
34	5	8.84905660377358
34	6	8.88636363636364
34	7	8.82716049382716
34	8	9.16666666666667
34	9	8.8695652173913
35	1	5.51063829787234
35	2	5.54918032786885
35	3	5.53409090909091
35	4	5.55555555555556
35	5	5.49122807017544
35	6	5.5
35	7	5.58620689655172
35	8	5.71428571428571
35	9	5.66666666666667
36	1	89.6666666666667
36	2	84.75
36	3	83.3333333333333
36	4	101
36	5	81.5
36	6	112
36	7	84
36	8	0
36	9	98
37	1	97.5
37	2	114.5
37	3	90.5
37	4	85
37	5	112.5
37	6	77.5
37	7	120
37	8	0
37	9	0
38	1	8.64285714285714
38	2	8.66
38	3	8.64102564102564
38	4	8.58108108108108
38	5	8.64814814814815
38	6	8.61538461538461
38	7	8.59550561797753
38	8	8.75
38	9	8.65
39	1	9.7948717948718
39	2	9.8173076923077
39	3	9.8961038961039
39	4	9.8421052631579
39	5	9.76923076923077

- Fortsetzung auf der nächsten Seite -

- Fortsetzung von der vorherigen Seite -

GEM	Schicht	w_a
39	6	9.88888888888889
39	7	9.91666666666667
39	8	10.1538461538462
39	9	9.66666666666667
40	1	81.4
40	2	71.7142857142857
40	3	74.2
40	4	84
40	5	65
40	6	73.5
40	7	67
40	8	39
40	9	0
41	1	13.0987654320988
41	2	13.1057692307692
41	3	13.1625
41	4	13.0657894736842
41	5	13.1296296296296
41	6	13.1304347826087
41	7	13.1168831168831
41	8	12.5833333333333
41	9	13.45
42	1	5.45555555555556
42	2	5.43103448275862
42	3	5.40217391304348
42	4	5.42105263157895
42	5	5.47169811320755
42	6	5.48717948717949
42	7	5.44444444444444
42	8	5.375
42	9	5.53846153846154
43	1	5.50537634408602
43	2	5.48305084745763
43	3	5.47191011235955
43	4	5.475
43	5	5.51020408163265
43	6	5.51428571428571
43	7	5.484375
43	8	5.44444444444444
43	9	5.53846153846154
44	1	15.578313253012
44	2	15.6504854368932
44	3	15.5853658536585
44	4	15.5696202531646
44	5	15.6111111111111
44	6	15.6511627906977
44	7	15.5949367088608
44	8	16
44	9	15.5
45	1	44.625
45	2	42.7272727272727
45	3	42.375
45	4	42.4285714285714
45	5	48.25
45	6	44
45	7	47.8
45	8	26
45	9	37

- Fortsetzung auf der nächsten Seite -

- Fortsetzung von der vorherigen Seite -

GEM	Schicht	w_a
46	1	5.38461538461538
46	2	5.35398230088496
46	3	5.36904761904762
46	4	5.34146341463415
46	5	5.33333333333333
46	6	5.39024390243902
46	7	5.35
46	8	5.18181818181818
46	9	5.42857142857143
47	1	43.8888888888889
47	2	45.4
47	3	47.375
47	4	44.4285714285714
47	5	48.5
47	6	41.5
47	7	46.8
47	8	0
47	9	26
48	1	5.59090909090909
48	2	5.58333333333333
48	3	5.63333333333333
48	4	5.58536585365854
48	5	5.60416666666667
48	6	5.58536585365854
48	7	5.62295081967213
48	8	5.25
48	9	5.66666666666667
49	1	47.3333333333333
49	2	45.8
49	3	44.375
49	4	48.2857142857143
49	5	49
49	6	52
49	7	44.8
49	8	28
49	9	25
50	1	48.6
50	2	43.4285714285714
50	3	43.2
50	4	43.8
50	5	44.6666666666667
50	6	51.5
50	7	39.25
50	8	29
50	9	0
51	1	10.6202531645571
51	2	10.7047619047619
51	3	10.6883116883117
51	4	10.6025641025641
51	5	10.6296296296296
51	6	10.5918367346939
51	7	10.6282051282051
51	8	10.2857142857143
51	9	10.5625
52	1	6.6
52	2	6.55
52	3	6.55056179775281
52	4	6.59493670886076

- Fortsetzung auf der nächsten Seite -

- Fortsetzung von der vorherigen Seite -

GEM	Schicht	w_a
52	5	6.63461538461538
52	6	6.5
52	7	6.55172413793103
52	8	6.85714285714286
52	9	7

TABELLE B.2: *Designgewicht für Design 05a.*

C Übersicht über Schätzer

Name des Schätzers	Beschreibung	Benutzt
D–DSE	DSE in einer Domain einer Gemeinde	Version 1
CHAP	Chapman–Schätzer in einer Domain einer Gemeinde	Version 1
CHAP/GSPREE	Chapman–Schätzer im GSPREE–Modell angewendet	Version 1
D–DSE/SPREE	D–DSE mit anschließender Durchführung des SPREEs	Version 2, 3
CHAP/SPREE	Chapman–Schätzer mit anschließender Durchführung des SPREEs	Version 2, 3
CHAP/SPREE/GSPREE	Chapman–Schätzer mit anschließender Durchführung des SPREEs, dann im GSPREE–Modell angewendet	Version 2, 3
GREG1	verallgemeinerter Regressionsschätzer unter Berücksichtigung der Populationsanzahl des Registers für die Gemeinde	Version 1, 2, 3
GREG2	verallgemeinerter Regressionsschätzer unter Berücksichtigung der Populationsanzahl des Registers für die Domain in der Gemeinde	Version 1, 2, 3
SYN	Verhältnis–synthetischer Schätzer	Version 1, 2, 3
EBLUP	Schätzer basiert auf Unit–level Modell	Version 1, 2, 3

TABELLE C.1: *Bezeichnung der verschiedenen verwendeten Schätzer.*

D Übersicht über Hilfsinformationen

Name des Schätzers	Hilfsinformation
D–DSE	Anzahl der registrierten Personen einer Domain
CHAP	Anzahl der registrierten Personen einer Domain
CHAP/GSPREE	Anzahl der registrierten Personen einer Domain Anzahl der tatsächlich lebenden Personen einer Domain
D–DSE/SPREE	Anzahl der registrierten Personen einer Domain Anzahl der registrierten Personen einer Klasse Anzahl der tatsächlich lebenden Personen einer Klasse
CHAP/SPREE	Anzahl der registrierten Personen einer Domain Anzahl der registrierten Personen einer Klasse Anzahl der tatsächlich lebenden Personen einer Klasse
CHAP/SPREE/GSPREE	Anzahl der registrierten Personen einer Domain Anzahl der registrierten Personen einer Klasse Anzahl der tatsächlich lebenden Personen einer Klasse Anzahl der tatsächlich lebenden Personen einer Domain
GREG1	Anzahl der registrierten Personen einer Gemeinde
GREG2	Anzahl der registrierten Personen einer Domain
SYN	Anzahl der registrierten Personen einer Domain
EBLUP	Anzahl der registrierten Personen einer Domain

TABELLE D.1: *Verwendete Hilfsinformationen in verschiedenen Schätzer.*

E Schematische Übersicht

	Population	Stichprobendesign	Def. der Klassen	Schätzung der Totalwerte einer Klasse	Schätzung der Totalwerte einer Domain	Varianzschätzung
1	SAL	01	Version 1	DSE	-	Direkte
1	SAL	01	Version 1	DSE	-	Jackknife
1	SAL	01	Version 1	Chapman	-	Direkte
1	SAL	01	Version 1	Chapman	-	Jackknife
1	SAL	01	Version 1	Chapman	GSPREE	Bootstrap
1	SAL	01	Version 2	DSE	SPREE	Direkte
1	SAL	01	Version 2	DSE	SPREE	Jackknife
1	SAL	01	Version 2	Chapman	SPREE	Direkte
1	SAL	01	Version 2	Chapman	SPREE	Jackknife
1	SAL	01	Version 2	Chapman	GSPREE	Bootstrap
1	SAL	01	Version 3	DSE	SPREE	Direkte
1	SAL	01	Version 3	DSE	SPREE	Jackknife
1	SAL	01	Version 3	Chapman	SPREE	Direkte
1	SAL	01	Version 3	Chapman	SPREE	Jackknife
1	SAL	01	Version 3	Chapman	GSPREE	Bootstrap
1	SAL	01	-	-	GREG1	Direkte
1	SAL	01	-	-	GREG2	Direkte
1	SAL	01	-	-	SYN	Direkte
1	SAL	01	-	-	EBLUP	Direkte
1	SAL	05a	Version 1	DSE	-	Direkte
1	SAL	05a	Version 1	DSE	-	Jackknife
1	SAL	05a	Version 1	Chapman	-	Direkte
1	SAL	05a	Version 1	Chapman	-	Jackknife
1	SAL	05a	Version 1	Chapman	GSPREE	Bootstrap
1	SAL	05a	Version 2	DSE	SPREE	Direkte
1	SAL	05a	Version 2	DSE	SPREE	Jackknife
1	SAL	05a	Version 2	Chapman	SPREE	Direkte
1	SAL	05a	Version 2	Chapman	SPREE	Jackknife
1	SAL	05a	Version 2	Chapman	GSPREE	Bootstrap
1	SAL	05a	Version 3	DSE	SPREE	Direkte
1	SAL	05a	Version 3	DSE	SPREE	Jackknife

- Fortsetzung auf der nächsten Seite -

- Fortsetzung von der vorherigen Seite -

	Popula-tion	Stichproben-design	Def. der Klassen	Schätzung der Totalwerte einer Klasse	Schätzung der Totalwerte einer Domain	Varianz-schätzung
1	SAL	05a	Version 3	Chapman	SPREE	Direkte
1	SAL	05a	Version 3	Chapman	SPREE	Jackknife
1	SAL	05a	Version 3	Chapman	GSPREE	Bootstrap
1	SAL	05a	-	-	GREG1	Direkte
1	SAL	05a	-	-	GREG2	Direkte
1	SAL	05a	-	-	SYN	Direkte
1	SAL	05a	-	-	EBLUP	Direkte

TABELLE E.1: *Die Szenarien der Simulationen.*

F Übersicht über Aufbau

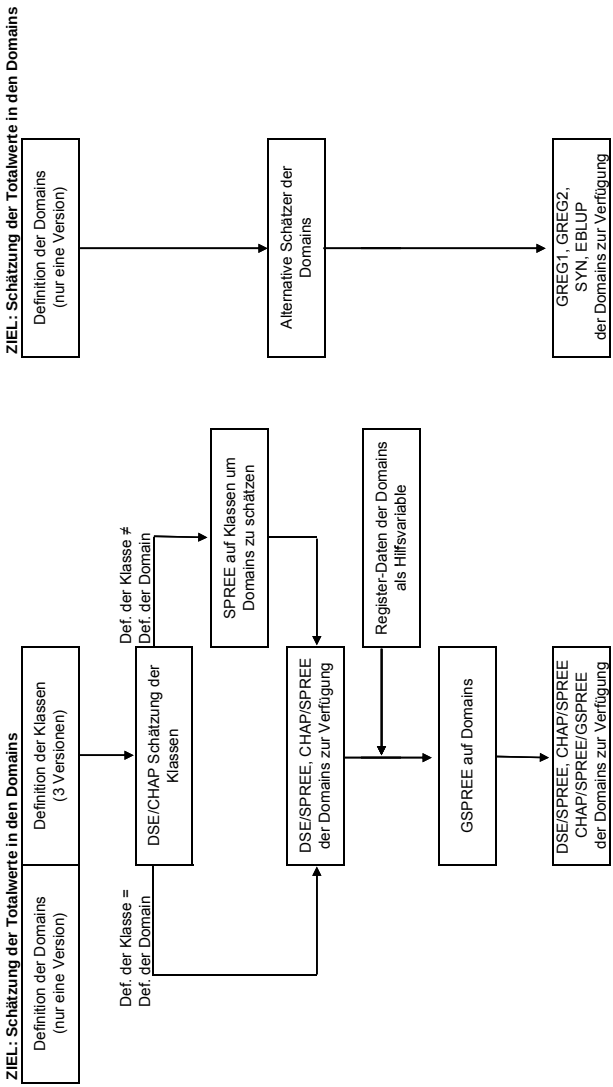
Einen besonderen Schwerpunkt in der amtlichen Statistik bildet die Schätzung von Subpopulationen, die demografisch abgegrenzt sein können (sogenannten Domains). Bezüglich der Homogenität (s. Seite 12) müssen für die Dual–System–Modelle bestimmte Klassen definiert werden.

Innerhalb der Klassen werden D–DSE oder der Chapman–Schätzer berechnet. Wenn die Definition der Klassen mit der Definition der Domains übereinstimmt, hat man automatisch die gewünschten Ergebnisse auch für die Domains. Wenn die Definition der Klassen mit der Definition der Domains nicht übereinstimmt, benötigt man SPREE, um die Schätzung der Totalwerte in den Domains zu gewinnen. Es wird die Schätzung für die Domains nach diesem Schritt als D–DSE/SPREE bzw. CHAP/SPREE bezeichnet.

Weiter werden die Schätzungen CHAP/SPREE der Domains zusammen mit den Register–Daten in den Domains im GSPREE–Modell angewendet, um weitere Schätzungen der Domains zu gewinnen. Diese werden als CHAP/SPREE/GSPREE bezeichnet.

Zu Vergleichszwecken werden in den Domains alternative Modelle angewendet, die zum GREG1–Schätzer, GREG2–Schätzer, Verhältnis–synthetischen Schätzer bzw. EBLUP für die Domains führen.

Für die grafische Darstellung siehe nächste Seite.


ZIEL: Schätzung der Totalwerte in den Domains
Definition der Domains (nur eine Version)
Definition der Klassen (3 Versionen)
Def. der Klasse = Def. der Domain
DSE/CHAP Schätzung der Klassen
Def. der Klasse ≠ Def. der Domain
SPREE auf Klassen um Domains zu schätzen
DSE/SPREE, CHAP/SPREE der Domains zur Verfügung
Register-Daten der Domains als Hilfsvariable
GSPREE auf Domains
DSE/SPREE, CHAP/SPREE CHAP/SPREE/GSPREE der Domains zur Verfügung
ZIEL: Schätzung der Totalwerte in den Domains
Definition der Domains (nur eine Version)
Alternative Schätzer der Domains
GREG1, GREG2, SYN, EBLUP der Domains zur Verfügung

G Übersicht über Versionen

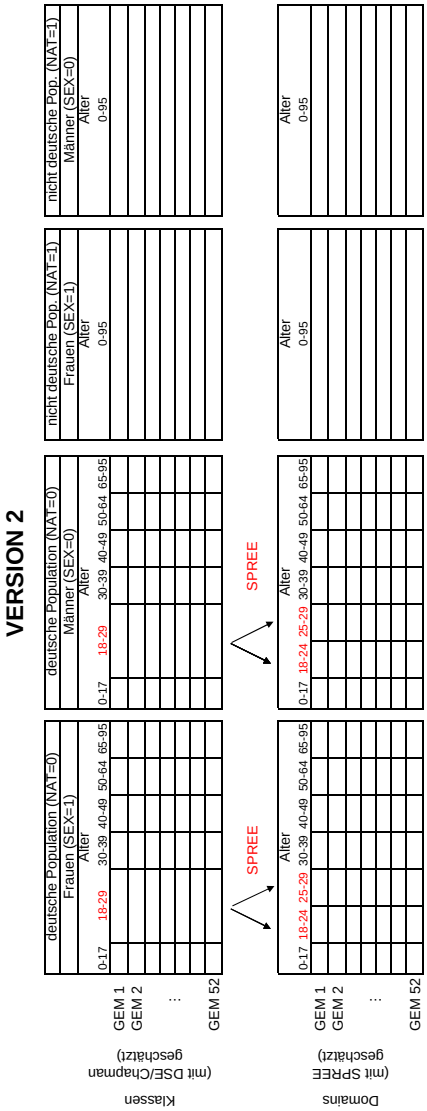
Es werden drei unterschiedliche Versionen der Klassen in einer Gemeinde untersucht.

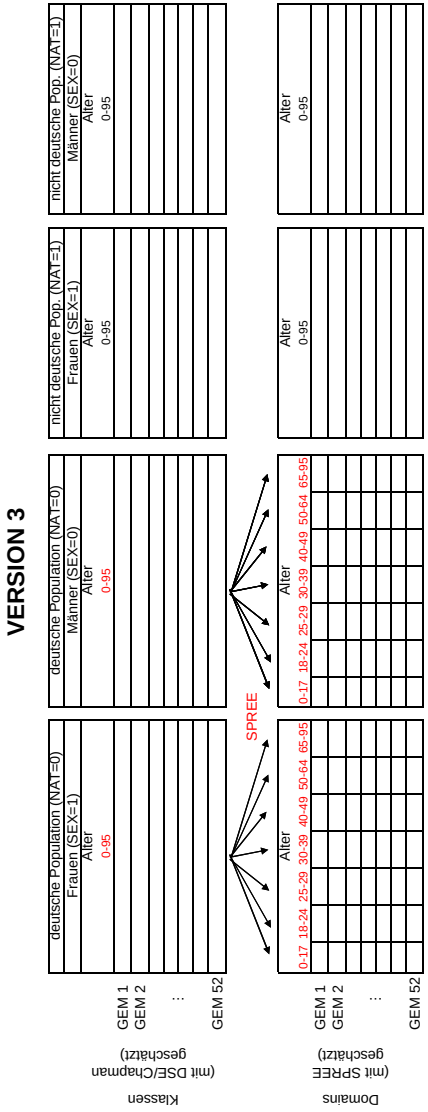
In der Version 1 stimmen die Klassen mit den Domains überein. In der deutschen Population sind sie durch die zwei Geschlechter und sieben Altersgruppen definiert, in der nicht deutschen Population nur durch die zwei Geschlechter. Nach der Anwendung D–DSE bzw. Chapman–Schätzer in den Klassen hat man automatisch die gewünschten Ergebnisse für die Domains. In dieser Version ist also keine anschließende Durchführung des SPREEs für die Domains nötig (s. grafische Abbildung auf der Seite 200).

In der Version 2 werden die Klassen der deutschen Frauen/Männer im Alter von 18-24/25-29 zusammengefasst. Diese Klassen sind breiter definiert als in der Version 1. Um die Ergebnisse der Domains zu erhalten, wird SPREE verwendet (s. grafische Abbildung auf der Seite 201). Die Klassen in der nicht deutschen Population bleiben unverändert.

In der Version 3 werden die Klassen weiter vergröbert. Die Definition der Klassen ist wie bei der nicht deutschen Population durch das Geschlecht bestimmt. Um die Ergebnisse der Domains zu erhalten wird SPREE verwendet (s. grafische Abbildung auf der Seite 202).

VERSION 1
Klassen=Domains
(mit DSE/Chapman
geschätzt)
GEM 1
GEM 2
...
GEM 52
deutsche Population (NAT=0)
Frauen (SEX=1)
Alter
0-17 18-24 25-29 30-39 40-49 50-64 65-95
deutsche Population (NAT=0)
Männer (SEX=0)
Alter
0-17 18-24 25-29 30-39 40-49 50-64 65-95
nicht deutsche Pop. (NAT=1)
Frauen (SEX=1)
Alter
0-95
nicht deutsche Pop. (NAT=1)
Männer (SEX=0)
Alter
0-95


VERSION 2
deutsche Population (NAT=0)
Frauen (SEX=1)
Alter
0-17 18-29 30-39 40-49 50-64 65-95
deutsche Population (NAT=0)
Männer (SEX=0)
Alter
0-17 18-29 30-39 40-49 50-64 65-95
nicht deutsche Pop. (NAT=1)
Frauen (SEX=1)
Alter
0-95
nicht deutsche Pop. (NAT=1)
Männer (SEX=0)
Alter
0-95
SPREE
SPREE
GEM 1
GEM 2
...
GEM 52
Klassen
(mit DSE/Chapman
geschätzt)
Alter
0-17 18-24 25-29 30-39 40-49 50-64 65-95
Alter
0-95
Domains
(mit SPREE
geschätzt)


VERSION 3
deutsche Population (NAT=0)
Frauen (SEX=1)
Alter
0-95
deutsche Population (NAT=0)
Männer (SEX=0)
Alter
0-95
nicht deutsche Pop. (NAT=1)
Frauen (SEX=1)
Alter
0-95
nicht deutsche Pop. (NAT=1)
Männer (SEX=0)
Alter
0-95
SPREE
Alter
0-17 18-24 25-29 30-39 40-49 50-64 65-95
Alter
0-95
GEM 1
GEM 2
...
GEM 52
Klassen
(mit DSE/Chapman
geschätzt)
Domains
(mit SPREE
geschätzt)

H Anhang zur uneingeschränkten Zufallsauswahl

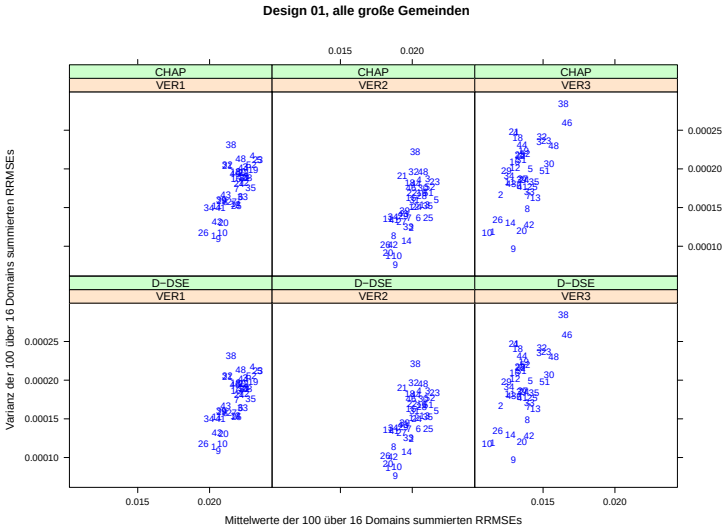


ABBILDUNG H.1: *Die Summen der RRMSEs über alle 16 Domains in jeder großen Gemeinde werden über die 100 Stichproben gemittelt und versus ihrer Varianzen abgebildet. Design 01, alle Versionen.*

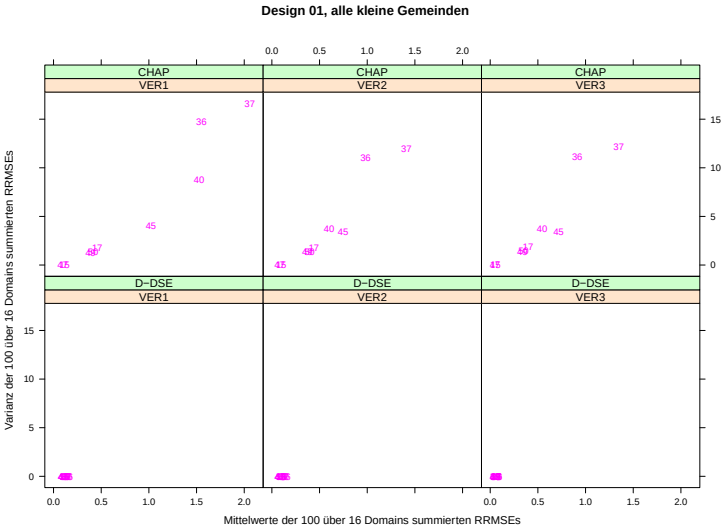


ABBILDUNG H.2: *Die Summen der RRMSEs über alle 16 Domains in jeder kleinen Gemeinde werden über die 100 Stichproben gemittelt und versus ihrer Varianzen abgebildet. Design 01, alle Versionen.*

GEM	D–DSE/ SPREE	CHAP/ SPREE	CHAP/ SPREE/ GSPREE
1	0,00423	0,00423	0,00423
35	0,00660	0,00660	0,00659
37	0,01093	0,01093	0,01096
45	0,00695	0,00695	0,00694

TABELLE H.1: *Mittelwerte der ARBs für die 14 Domains der deutschen Population in den Gemeinden Nr. 1, 35, 37 und 45. Design 01, Version 3.*

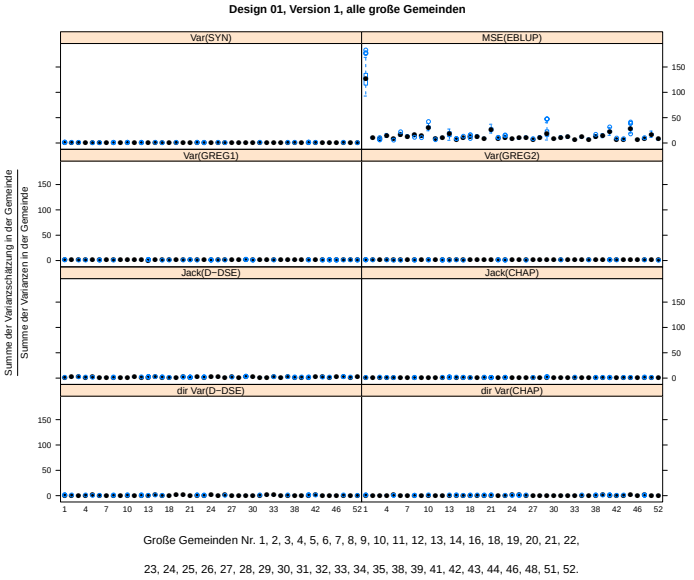


ABBILDUNG H.3: *Relative Verzerrung der Varianzschätzungen auf Gemeindebasis, alle große Gemeinden. Design 01, Version 1.*

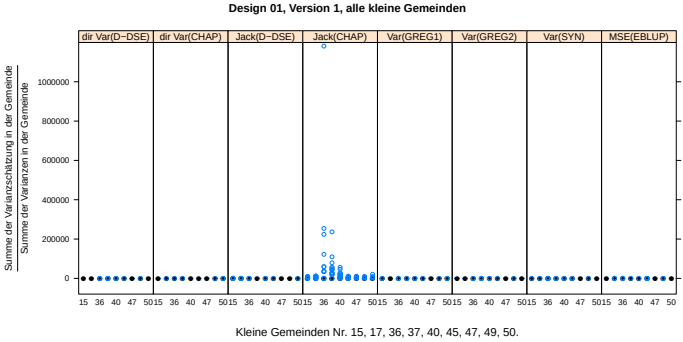


ABBILDUNG H.4: *Relative Verzerrung der Varianzschätzungen auf Gemeindebasis, alle kleine Gemeinden. Design 01, Version 1.*

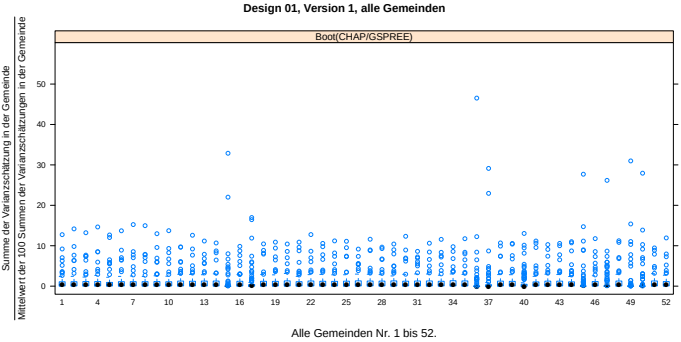


ABBILDUNG H.5: *Verzerrung der Bootstrap–Varianzschätzungen des CHAP/GSPREEs auf Gemeindebasis, alle Gemeinden. Design 01, Version 1.*

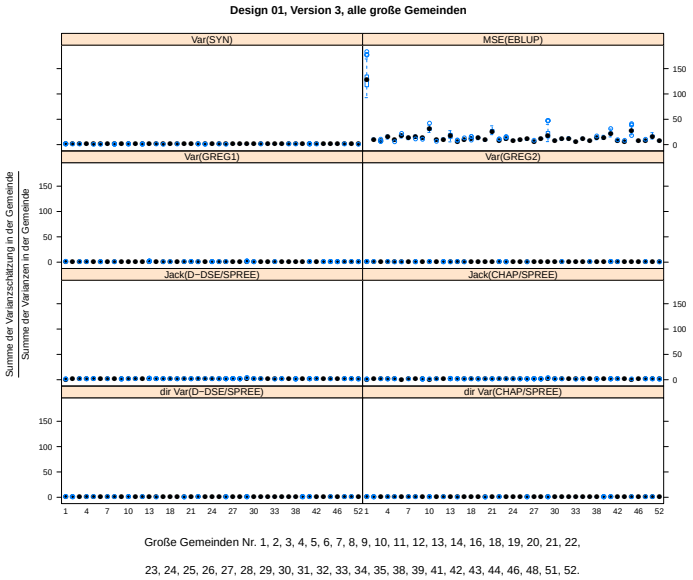


ABBILDUNG H.6: *Relative Verzerrung der Varianzschätzungen auf Gemeindebasis, alle große Gemeinden. Design 01, Version 3.*

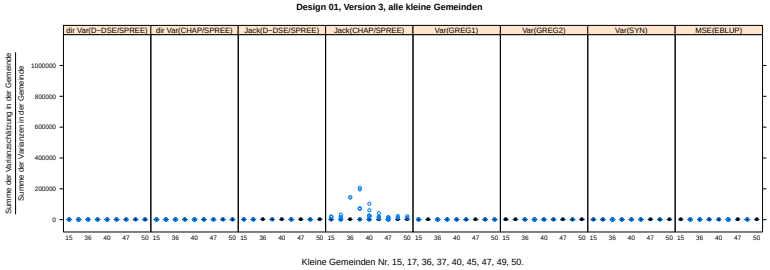


ABBILDUNG H.7: *Relative Verzerrung der Varianzschätzungen auf Gemeindebasis, alle kleine Gemeinden. Design 01, Version 3.*

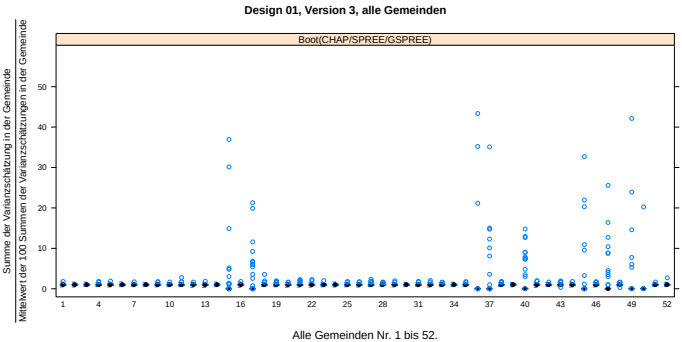


ABBILDUNG H.8: *Verzerrung der Bootstrap–Varianzschätzungen des CHAP/SPREE/GSPREEs auf Gemeindebasis, alle Gemeinden. Design 01, Version 3.*

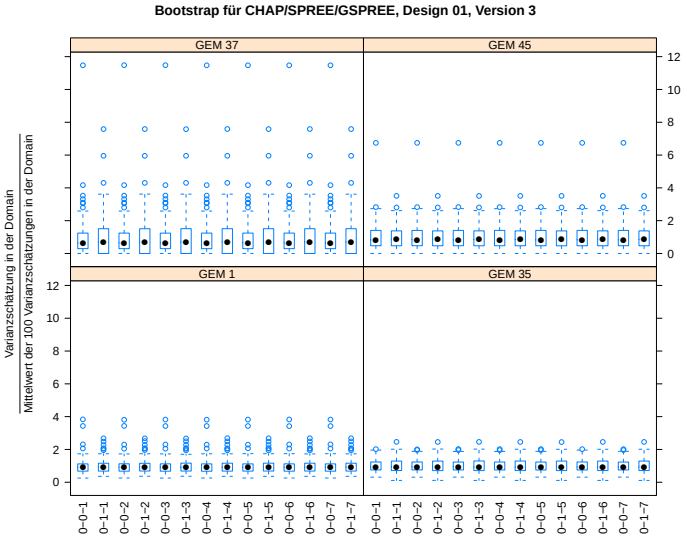


ABBILDUNG H.9: *Verzerrung der Bootstrap–Varianzschätzungen des CHAP/SPREE/GSPREEs auf Domainbasis in Gemeinden Nr. 1, 35, 37 und 45 in der deutschen Population. Design 01, Version 3.*

GEM Nr. 1 Version 1	D–DSE	CHAP	CHAP/ GSPREE	GREG1	GREG2	SYN	EBLUP
0-0-1	0,01116	0,01116	0,02647	0,07449	0,01116	0,00965	0,02743
0-1-1	0,01261	0,01261	0,00660	0,06254	0,01261	0,01105	0,02884
0-0-2	0,03067	0,03067	0,05570	0,10072	0,03067	0,00429	0,01851
0-1-2	0,02668	0,02668	0,02319	0,09498	0,02668	0,01441	0,03231
0-0-3	0,03153	0,03153	0,08289	0,10114	0,03152	0,00656	0,01500
0-1-3	0,03010	0,03010	0,05189	0,09850	0,03010	0,00485	0,02171
0-0-4	0,01637	0,01637	0,01695	0,06205	0,01638	0,00421	0,01994
0-1-4	0,01608	0,01608	0,00800	0,05862	0,01608	0,00490	0,01709
0-0-5	0,01438	0,01438	0,01062	0,06390	0,01439	0,00423	0,02009
0-1-5	0,01364	0,01364	0,00817	0,05122	0,01364	0,00516	0,02216
0-0-6	0,01015	0,01015	0,01118	0,05944	0,01015	0,00629	0,02367
0-1-6	0,01099	0,01099	0,00699	0,05835	0,01099	0,00678	0,02423
0-0-7	0,01296	0,01296	0,01232	0,06738	0,01296	0,01011	0,02790
0-1-7	0,00782	0,00782	0,00623	0,05998	0,00782	0,00805	0,02565
1-0-8	0,03533	0,03533	0,03533	0,13177	0,03533	0,06564	0,04663
1-1-8	0,04417	0,04417	0,04417	0,10623	0,04417	0,05493	0,03641
GEM Nr. 35 Version 1	D–DSE	CHAP	CHAP/ GSPREE	GREG1	GREG2	SYN	EBLUP
0-0-1	0,01228	0,01228	0,01014	0,06977	0,01223	0,01246	0,02702
0-1-1	0,01064	0,01064	0,00998	0,07055	0,01062	0,01467	0,02922
0-0-2	0,02900	0,02900	0,05526	0,09594	0,02904	0,00631	0,02091
0-1-2	0,03791	0,03791	0,04821	0,11944	0,03805	0,01669	0,00395
0-0-3	0,03009	0,03008	0,10098	0,13201	0,03015	0,01146	0,02657
0-1-3	0,03629	0,03629	0,05276	0,11553	0,03645	0,01636	0,03168
0-0-4	0,01630	0,01630	0,01099	0,06950	0,01631	0,00567	0,01959
0-1-4	0,01730	0,01730	0,00854	0,06481	0,01727	0,00505	0,01380
0-0-5	0,01604	0,01604	0,01309	0,06048	0,01602	0,00776	0,01024
0-1-5	0,01807	0,01807	0,00891	0,06139	0,01803	0,00521	0,01335
0-0-6	0,01110	0,01110	0,01070	0,05951	0,01114	0,00521	0,01866
0-1-6	0,00937	0,00937	0,00828	0,05828	0,00940	0,00847	0,02270
0-0-7	0,01306	0,01306	0,00794	0,07399	0,01306	0,00543	0,01938
0-1-7	0,01074	0,01074	0,00867	0,05643	0,01077	0,00496	0,01356
1-0-8	0,04776	0,04775	0,04775	0,14060	0,04774	0,05544	0,03799
1-1-8	0,05001	0,05001	0,05001	0,14922	0,04981	0,07852	0,06013
GEM Nr. 37 Version 1	D–DSE	CHAP	CHAP/ GSPREE	GREG1	GREG2	SYN	EBLUP
0-0-1	0,12792	0,12726	0,60573	0,44599	0,12783	0,02628	0,03433
0-1-1	0,05617	0,05614	0,03312	0,36848	0,05615	0,02478	0,02649
0-0-2	0,16965	0,16325	0,55451	0,61266	0,16937	0,02794	0,03797
0-1-2	0,13332	0,12703	0,05126	0,65863	0,13309	0,03789	0,05018
0-0-3	0,31953	10,49727	1,25505	0,63976	0,31954	0,03355	0,02442
0-1-3	0,17928	0,16993	0,06661	0,62849	0,17915	0,03502	0,02394
0-0-4	0,10066	0,10061	0,48986	0,37066	0,10077	0,02469	0,02909
0-1-4	0,07203	0,07202	0,03303	0,33494	0,07201	0,02661	0,02379
0-0-5	0,11078	0,11067	0,52304	0,33950	0,11071	0,02777	0,03696
0-1-5	0,13523	0,13491	0,03373	0,35516	0,13519	0,02472	0,02978
0-0-6	0,07150	0,07144	0,54040	0,33937	0,07150	0,02473	0,02691
0-1-6	0,04971	0,04969	0,03361	0,32263	0,04971	0,02568	0,03291
0-0-7	0,04914	0,04913	0,52494	0,38840	0,04924	0,02461	0,02886
0-1-7	0,05898	0,05897	0,03379	0,27358	0,05899	0,02716	0,02356
1-0-8	0,29141	10,56280	10,56280	0,81864	0,29122	0,03726	0,02539
1-1-8	0,23120	10,57381	10,57381	0,76264	0,23125	0,03788	0,02643
GEM Nr. 45 Version 1	D–DSE	CHAP	CHAP/ GSPREE	GREG1	GREG2	SYN	EBLUP
0-0-1	0,03324	0,03322	0,03027	0,26135	0,03318	0,01676	0,02792
0-1-1	0,05555	0,05553	0,22881	0,25104	0,05551	0,01560	0,02549
0-0-2	0,12032	0,11923	0,06034	0,41827	0,12038	0,01629	0,02692
0-1-2	0,15837	4,27260	0,33152	0,47262	0,15836	0,03382	0,04815
0-0-3	0,13180	0,12965	0,08368	0,39639	0,13160	0,01624	0,02671
0-1-3	0,16745	0,16614	0,23045	0,43547	0,16743	0,01835	0,03040
0-0-4	0,05695	0,05694	0,03030	0,22433	0,05690	0,02274	0,01460
0-1-4	0,05800	0,05797	0,22420	0,23829	0,05783	0,01484	0,02143
0-0-5	0,04466	0,04465	0,02645	0,25197	0,04466	0,01695	0,02830
0-1-5	0,07954	0,07951	0,23135	0,28182	0,07945	0,01499	0,02384
0-0-6	0,04169	0,04169	0,02632	0,22202	0,04168	0,01503	0,01967
0-1-6	0,03508	0,03508	0,22057	0,20639	0,03514	0,01482	0,02089
0-0-7	0,05885	0,05884	0,02701	0,24699	0,05885	0,01478	0,02109
0-1-7	0,03296	0,03295	0,24018	0,22053	0,03291	0,01476	0,02216
1-0-8	0,35207	6,51934	6,51934	0,58593	0,35174	0,04348	0,02849
1-1-8	0,27859	4,63779	4,63779	0,56964	0,27831	0,04852	0,03267

TABELLE H.2: *RRMSEs aller Domains in den Gemeinden Nr. 1, 35, 37 und 45, Design 01, Version 1.*

GEM Nr. 1 Version 2	D–DSE/ SPREE	CHAP/ SPREE	CHAP/ SPREE/ GSPREE	GREG1	GREG2	SYN	EBLUP
0-0-1	0,01116	0,01116	0,00737	0,07449	0,01116	0,00965	0,02743
0-1-1	0,01261	0,01261	0,00590	0,06254	0,01261	0,01105	0,02884
0-0-2	0,02209	0,02209	0,00881	0,10072	0,03067	0,00429	0,01851
0-1-2	0,02136	0,02136	0,01310	0,09498	0,02668	0,01441	0,03231
0-0-3	0,02223	0,02223	0,01085	0,10114	0,03152	0,00656	0,01500
0-1-3	0,02195	0,02195	0,00836	0,09850	0,03010	0,00485	0,02171
0-0-4	0,01637	0,01637	0,00771	0,06205	0,01638	0,00421	0,01994
0-1-4	0,01608	0,01608	0,00696	0,05862	0,01608	0,00490	0,01709
0-0-5	0,01438	0,01438	0,00707	0,06390	0,01439	0,00423	0,02009
0-1-5	0,01364	0,01364	0,00686	0,05122	0,01364	0,00516	0,02216
0-0-6	0,01015	0,01015	0,00647	0,05944	0,01015	0,00629	0,02367
0-1-6	0,01099	0,01099	0,00596	0,05835	0,01099	0,00678	0,02423
0-0-7	0,01296	0,01296	0,00835	0,06738	0,01296	0,01011	0,02790
0-1-7	0,00782	0,00782	0,00526	0,05998	0,00782	0,00805	0,02565
1-0-8	0,03533	0,03533	0,03533	0,13177	0,03533	0,06564	0,04663
1-1-8	0,04417	0,04417	0,04417	0,10623	0,04417	0,05493	0,03641
GEM Nr. 35 Version 2	D–DSE/ SPREE	CHAP/ SPREE	CHAP/ SPREE/ GSPREE	GREG1	GREG2	SYN	EBLUP
0-0-1	0,01228	0,01228	0,00678	0,06977	0,01223	0,01246	0,02702
0-1-1	0,01064	0,01064	0,00844	0,07055	0,01062	0,01467	0,02922
0-0-2	0,02140	0,02140	0,00825	0,09594	0,02904	0,00631	0,02091
0-1-2	0,03421	0,03421	0,01997	0,11944	0,03805	0,01669	0,00395
0-0-3	0,02087	0,02087	0,01182	0,13201	0,03015	0,01146	0,02657
0-1-3	0,02832	0,02832	0,01604	0,11553	0,03645	0,01636	0,03168
0-0-4	0,01630	0,01630	0,00767	0,06950	0,01631	0,00567	0,01959
0-1-4	0,01730	0,01730	0,00722	0,06481	0,01727	0,00505	0,01380
0-0-5	0,01604	0,01604	0,01145	0,06048	0,01602	0,00776	0,01024
0-1-5	0,01807	0,01807	0,00814	0,06139	0,01803	0,00521	0,01335
0-0-6	0,01110	0,01110	0,00610	0,05951	0,01114	0,00521	0,01866
0-1-6	0,00937	0,00937	0,00713	0,05828	0,00940	0,00847	0,02270
0-0-7	0,01306	0,01306	0,00624	0,07399	0,01306	0,00543	0,01938
0-1-7	0,01074	0,01074	0,00826	0,05643	0,01077	0,00496	0,01356
1-0-8	0,04776	0,04775	0,04775	0,14060	0,04774	0,05544	0,03799
1-1-8	0,05001	0,05001	0,05001	0,14922	0,04981	0,07852	0,06013
GEM Nr. 37 Version 2	D–DSE/ SPREE	CHAP/ SPREE	CHAP/ SPREE/ GSPREE	GREG1	GREG2	SYN	EBLUP
0-0-1	0,12792	0,12726	0,04294	0,44599	0,12783	0,02628	0,03433
0-1-1	0,05617	0,05614	0,03416	0,36848	0,05615	0,02478	0,02649
0-0-2	0,16730	0,16620	0,04163	0,61266	0,16937	0,02794	0,03797
0-1-2	0,13531	0,13340	0,04480	0,65863	0,13309	0,03789	0,05018
0-0-3	0,17719	0,17605	0,05560	0,63976	0,31954	0,03355	0,02442
0-1-3	0,14876	0,14670	0,04125	0,62849	0,17915	0,03502	0,02394
0-0-4	0,10066	0,10061	0,04219	0,37066	0,10077	0,02469	0,02909
0-1-4	0,07203	0,07202	0,03373	0,33494	0,07201	0,02661	0,02379
0-0-5	0,11078	0,11067	0,04112	0,33950	0,11071	0,02777	0,03696
0-1-5	0,13523	0,13491	0,03424	0,35516	0,13519	0,02472	0,02978
0-0-6	0,07150	0,07144	0,04462	0,33937	0,07150	0,02473	0,02691
0-1-6	0,04971	0,04969	0,03421	0,32263	0,04971	0,02568	0,03291
0-0-7	0,04914	0,04913	0,04292	0,38840	0,04924	0,02461	0,02886
0-1-7	0,05898	0,05897	0,03506	0,27358	0,05899	0,02716	0,02356
1-0-8	0,29141	10,56280	10,56280	0,81864	0,29122	0,03726	0,02539
1-1-8	0,23120	10,57381	10,57381	0,76264	0,23125	0,03788	0,02643
GEM Nr. 45 Version 2	D–DSE/ SPREE	CHAP/ SPREE	CHAP/ SPREE/ GSPREE	GREG1	GREG2	SYN	EBLUP
0-0-1	0,03324	0,03322	0,02073	0,26135	0,03318	0,01676	0,02792
0-1-1	0,05555	0,05553	0,02253	0,25104	0,05551	0,01560	0,02549
0-0-2	0,06732	0,06729	0,02187	0,41827	0,12038	0,01629	0,02692
0-1-2	0,08898	0,08892	0,03483	0,47262	0,15836	0,03382	0,04815
0-0-3	0,06732	0,06729	0,02189	0,39639	0,13160	0,01624	0,02671
0-1-3	0,09047	0,09041	0,02266	0,43547	0,16743	0,01835	0,03040
0-0-4	0,05695	0,05694	0,02609	0,22433	0,05690	0,02274	0,01460
0-1-4	0,05800	0,05797	0,02225	0,23829	0,05783	0,01484	0,02143
0-0-5	0,04466	0,04465	0,02168	0,25197	0,04466	0,01695	0,02830
0-1-5	0,07954	0,07951	0,02257	0,28182	0,07945	0,01499	0,02384
0-0-6	0,04169	0,04169	0,02151	0,22202	0,04168	0,01503	0,01967
0-1-6	0,03508	0,03508	0,02252	0,20639	0,03514	0,01482	0,02089
0-0-7	0,05885	0,05884	0,02137	0,24699	0,05885	0,01478	0,02109
0-1-7	0,03296	0,03295	0,02216	0,22053	0,03291	0,01476	0,02216
1-0-8	0,35207	6,51934	6,51934	0,58593	0,35174	0,04348	0,02849
1-1-8	0,27859	4,63779	4,63779	0,56964	0,27831	0,04852	0,03267

TABELLE H.3: *RRMSEs aller Domains in den Gemeinden Nr. 1, 35, 37 und 45, Design 01, Version 2.*

GEM Nr. 1 Version 3	D–DSE/ SPREE	CHAP/ SPREE	CHAP/ SPREE/ GSPREE	GREG1	GREG2	SYN	EBLUP
0-0-1	0,00836	0,00836	0,00835	0,07449	0,01116	0,00965	0,02743
0-1-1	0,00690	0,00690	0,00689	0,06254	0,01261	0,01105	0,02884
0-0-2	0,00734	0,00734	0,00734	0,10072	0,03067	0,00429	0,01851
0-1-2	0,00953	0,00953	0,00952	0,09498	0,02668	0,01441	0,03231
0-0-3	0,01016	0,01016	0,01017	0,10114	0,03152	0,00656	0,01500
0-1-3	0,00650	0,00650	0,00649	0,09850	0,03010	0,00485	0,02171
0-0-4	0,00656	0,00656	0,00656	0,06205	0,01638	0,00421	0,01994
0-1-4	0,01023	0,01023	0,01023	0,05862	0,01608	0,00490	0,01709
0-0-5	0,00651	0,00651	0,00651	0,06390	0,01439	0,00423	0,02009
0-1-5	0,00621	0,00621	0,00621	0,05122	0,01364	0,00516	0,02216
0-0-6	0,00634	0,00634	0,00634	0,05944	0,01015	0,00629	0,02367
0-1-6	0,00549	0,00549	0,00549	0,05835	0,01099	0,00678	0,02423
0-0-7	0,00872	0,00872	0,00872	0,06738	0,01296	0,01011	0,02790
0-1-7	0,00551	0,00551	0,00551	0,05998	0,00782	0,00805	0,02565
1-0-8	0,03533	0,03533	0,03533	0,13177	0,03533	0,06564	0,04663
1-1-8	0,04417	0,04417	0,04417	0,10623	0,04417	0,05493	0,03641
GEM Nr. 35 Version 3	D–DSE/ SPREE	CHAP/ SPREE	CHAP/ SPREE/ GSPREE	GREG1	GREG2	SYN	EBLUP
0-0-1	0,00978	0,00978	0,00984	0,06977	0,01223	0,01246	0,02702
0-1-1	0,01285	0,01285	0,01283	0,07055	0,01062	0,01467	0,02922
0-0-2	0,00570	0,00570	0,00582	0,09594	0,02904	0,00631	0,02091
0-1-2	0,02000	0,02000	0,01977	0,11944	0,03805	0,01669	0,00395
0-0-3	0,00890	0,00890	0,00905	0,13201	0,03015	0,01146	0,02657
0-1-3	0,01461	0,01461	0,01462	0,11553	0,03645	0,01636	0,03168
0-0-4	0,00559	0,00559	0,00549	0,06950	0,01631	0,00567	0,01959
0-1-4	0,00804	0,00804	0,00799	0,06481	0,01727	0,00505	0,01380
0-0-5	0,01166	0,01166	0,01163	0,06048	0,01602	0,00776	0,01024
0-1-5	0,00830	0,00830	0,00830	0,06139	0,01803	0,00521	0,01335
0-0-6	0,00564	0,00564	0,00565	0,05951	0,01114	0,00521	0,01866
0-1-6	0,00753	0,00753	0,00751	0,05828	0,00940	0,00847	0,02270
0-0-7	0,00559	0,00559	0,00559	0,07399	0,01306	0,00543	0,01938
0-1-7	0,00800	0,00800	0,00797	0,05643	0,01077	0,00496	0,01356
1-0-8	0,04776	0,04775	0,04775	0,14060	0,04774	0,05544	0,03799
1-1-8	0,05001	0,05001	0,05001	0,14922	0,04981	0,07852	0,06013
GEM Nr. 37 Version 3	D–DSE/ SPREE	CHAP/ SPREE	CHAP/ SPREE/ GSPREE	GREG1	GREG2	SYN	EBLUP
0-0-1	0,03150	0,03149	0,03157	0,44599	0,12783	0,02628	0,03433
0-1-1	0,02629	0,02629	0,02617	0,36848	0,05615	0,02478	0,02649
0-0-2	0,03187	0,03187	0,03199	0,61266	0,16937	0,02794	0,03797
0-1-2	0,04196	0,04196	0,04211	0,65863	0,13309	0,03789	0,05018
0-0-3	0,04454	0,04454	0,04456	0,63976	0,31954	0,03355	0,02442
0-1-3	0,03389	0,03389	0,03376	0,62849	0,17915	0,03502	0,02394
0-0-4	0,03225	0,03225	0,03221	0,37066	0,10077	0,02469	0,02909
0-1-4	0,02713	0,02713	0,02708	0,33494	0,07201	0,02661	0,02379
0-0-5	0,03179	0,03179	0,03185	0,33950	0,11071	0,02777	0,03696
0-1-5	0,02703	0,02703	0,02703	0,35516	0,13519	0,02472	0,02978
0-0-6	0,03331	0,03331	0,03331	0,33937	0,07150	0,02473	0,02691
0-1-6	0,02849	0,02849	0,02849	0,32263	0,04971	0,02568	0,03291
0-0-7	0,03233	0,03233	0,03228	0,38840	0,04924	0,02461	0,02886
0-1-7	0,02741	0,02741	0,02737	0,27358	0,05899	0,02716	0,02356
1-0-8	0,29141	10,56280	10,56280	0,81864	0,29122	0,03726	0,02539
1-1-8	0,23120	10,57381	10,57381	0,76264	0,23125	0,03788	0,02643
GEM Nr. 45 Version 3	D–DSE/ SPREE	CHAP/ SPREE	CHAP/ SPREE/ GSPREE	GREG1	GREG2	SYN	EBLUP
0-0-1	0,02162	0,02162	0,02168	0,26135	0,03318	0,01676	0,02792
0-1-1	0,02002	0,02002	0,02010	0,25104	0,05551	0,01560	0,02549
0-0-2	0,02128	0,02127	0,02128	0,41827	0,12038	0,01629	0,02692
0-1-2	0,03431	0,03431	0,03403	0,47262	0,15836	0,03382	0,04815
0-0-3	0,02128	0,02127	0,02132	0,39639	0,13160	0,01624	0,02671
0-1-3	0,02159	0,02159	0,02186	0,43547	0,16743	0,01835	0,03040
0-0-4	0,02821	0,02821	0,02816	0,22433	0,05690	0,02274	0,01460
0-1-4	0,02003	0,02003	0,02002	0,23829	0,05783	0,01484	0,02143
0-0-5	0,02176	0,02176	0,02177	0,25197	0,04466	0,01695	0,02830
0-1-5	0,01986	0,01986	0,01988	0,28182	0,07945	0,01499	0,02384
0-0-6	0,02114	0,02114	0,02119	0,22202	0,04168	0,01503	0,01967
0-1-6	0,02014	0,02014	0,02017	0,20639	0,03514	0,01482	0,02089
0-0-7	0,02082	0,02081	0,02074	0,24699	0,05885	0,01478	0,02109
0-1-7	0,01992	0,01992	0,01989	0,22053	0,03291	0,01476	0,02216
1-0-8	0,35207	6,51934	6,51934	0,58593	0,35174	0,04348	0,02849
1-1-8	0,27859	4,63779	4,63779	0,56964	0,27831	0,04852	0,03267

TABELLE H.4: *RRMSEs aller Domains in den Gemeinden Nr. 1, 35, 37 und 45, Design 01, Version 3.*

GEM Nr. 1 Version 1	D–DSE	CHAP	CHAP/ GSPREE	GREG1	GREG2	SYN	EBLUP
0-0-1	0,00165	0,00165	0,00080	0,00200	0,00165	0,00872	0,02646
0-1-1	0,00188	0,00188	0,00172	0,01516	0,00188	0,01025	0,02793
0-0-2	0,00330	0,00330	0,01070	0,01975	0,00330	0,00102	0,01699
0-1-2	0,00184	0,00184	0,00192	0,00870	0,00184	0,01382	0,03150
0-0-3	0,00259	0,00259	0,01446	0,00692	0,00259	0,00507	0,01306
0-1-3	0,00319	0,00319	0,00331	0,00047	0,00317	0,00251	0,02044
0-0-4	0,00031	0,00031	0,00340	0,00132	0,00031	0,00065	0,01857
0-1-4	0,00121	0,00121	0,00090	0,00817	0,00121	0,00257	0,01544
0-0-5	0,00066	0,00066	0,00008	0,00649	0,00066	0,00079	0,01872
0-1-5	0,00062	0,00062	0,00111	0,00052	0,00062	0,00307	0,02093
0-0-6	0,00036	0,00036	0,00239	0,00200	0,00036	0,00473	0,02253
0-1-6	0,00193	0,00193	0,00180	0,00554	0,00193	0,00537	0,02313
0-0-7	0,00045	0,00045	0,00653	0,00626	0,00045	0,00923	0,02696
0-1-7	0,00045	0,00045	0,00097	0,00568	0,00045	0,00690	0,02462
1-0-8	0,00225	0,00225	0,00225	0,00661	0,00225	0,06549	0,04593
1-1-8	0,00219	0,00219	0,00219	0,00524	0,00220	0,05476	0,03556
GEM Nr. 35 Version 1	D–DSE	CHAP	CHAP/ GSPREE	GREG1	GREG2	SYN	EBLUP
0-0-1	0,00001	0,00001	0,00397	0,00454	0,00003	0,01172	0,02677
0-1-1	0,00092	0,00092	0,00607	0,00490	0,00092	0,01406	0,02900
0-0-2	0,00264	0,00264	0,00527	0,01598	0,00277	0,00473	0,02057
0-1-2	0,00549	0,00549	0,02893	0,01508	0,00536	0,01611	0,00013
0-0-3	0,00137	0,00137	0,00070	0,00234	0,00130	0,01067	0,02630
0-1-3	0,00164	0,00163	0,01313	0,00620	0,00173	0,01585	0,03147
0-0-4	0,00139	0,00140	0,00568	0,00126	0,00139	0,00375	0,01922
0-1-4	0,00396	0,00396	0,00225	0,00130	0,00399	0,00265	0,01330
0-0-5	0,00050	0,00050	0,00774	0,00119	0,00054	0,00645	0,00955
0-1-5	0,00108	0,00108	0,00428	0,00463	0,00115	0,00305	0,01280
0-0-6	0,00067	0,00067	0,00006	0,00544	0,00067	0,00296	0,01830
0-1-6	0,00118	0,00118	0,00383	0,00545	0,00120	0,00732	0,02241
0-0-7	0,00059	0,00059	0,00003	0,00848	0,00057	0,00345	0,01903
0-1-7	0,00043	0,00043	0,00518	0,00430	0,00043	0,00260	0,01304
1-0-8	0,00616	0,00616	0,00616	0,00953	0,00613	0,05526	0,03775
1-1-8	0,00579	0,00578	0,00578	0,00139	0,00560	0,07837	0,05997
GEM Nr. 37 Version 1	D–DSE	CHAP	CHAP/ GSPREE	GREG1	GREG2	SYN	EBLUP
0-0-1	0,01429	0,01419	0,06807	0,03161	0,01429	0,00992	0,02598
0-1-1	0,00283	0,00283	0,00633	0,02028	0,00283	0,00261	0,01367
0-0-2	0,00833	0,00720	0,06152	0,03717	0,00834	0,01399	0,03058
0-1-2	0,00172	0,00056	0,02255	0,01750	0,00168	0,02958	0,04529
0-0-3	0,02534	1,07470	0,16010	0,06578	0,02534	0,02233	0,00470
0-1-3	0,01709	0,01847	0,02264	0,09165	0,01710	0,02441	0,00789
0-0-4	0,01971	0,01969	0,05591	0,06158	0,01971	0,00219	0,01833
0-1-4	0,00679	0,00678	0,00109	0,04544	0,00679	0,00974	0,00657
0-0-5	0,00762	0,00761	0,05069	0,00057	0,00762	0,01347	0,02945
0-1-5	0,01516	0,01511	0,00632	0,01713	0,01516	0,00327	0,01943
0-0-6	0,00615	0,00615	0,06923	0,01071	0,00612	0,00171	0,01452
0-1-6	0,00378	0,00379	0,00861	0,01569	0,00378	0,00796	0,02406
0-0-7	0,00053	0,00053	0,06479	0,01711	0,00052	0,00180	0,01795
0-1-7	0,00817	0,00817	0,00865	0,00525	0,00816	0,01085	0,00555
1-0-8	0,00389	1,06441	1,06441	0,00191	0,00385	0,02751	0,00949
1-1-8	0,04591	1,02642	1,02642	0,02095	0,04593	0,02863	0,01122
GEM Nr. 45 Version 1	D–DSE	CHAP	CHAP/ GSPREE	GREG1	GREG2	SYN	EBLUP
0-0-1	0,00175	0,00175	0,00280	0,00624	0,00176	0,00829	0,02404
0-1-1	0,00087	0,00087	0,02730	0,00857	0,00088	0,00533	0,02114
0-0-2	0,00947	0,00933	0,00194	0,10589	0,00953	0,00706	0,02282
0-1-2	0,00146	0,42798	0,01195	0,08095	0,00131	0,03071	0,04608
0-0-3	0,01593	0,01561	0,00120	0,03547	0,01584	0,00692	0,02262
0-1-3	0,01642	0,01615	0,01572	0,03987	0,01644	0,01114	0,02691
0-0-4	0,00157	0,00156	0,01337	0,00127	0,00162	0,01715	0,00100
0-1-4	0,00153	0,00153	0,01887	0,02761	0,00147	0,00009	0,01595
0-0-5	0,00160	0,00160	0,00922	0,00947	0,00160	0,00866	0,02449
0-1-5	0,00088	0,00087	0,02274	0,01393	0,00087	0,00318	0,01909
0-0-6	0,00728	0,00728	0,00335	0,03374	0,00725	0,00243	0,01347
0-1-6	0,00028	0,00028	0,02718	0,00324	0,00032	0,00066	0,01522
0-0-7	0,00025	0,00026	0,00221	0,01820	0,00030	0,00045	0,01546
0-1-7	0,00039	0,00039	0,02734	0,02695	0,00046	0,00112	0,01694
1-0-8	0,00250	0,92133	0,92133	0,09560	0,00252	0,04048	0,02409
1-1-8	0,04077	0,42972	0,42972	0,07018	0,04107	0,04602	0,02900

TABELLE H.5: *ARBs aller Domains in den Gemeinden Nr. 1, 35, 37 und 45, Design 01, Version 1.*

GEM Nr. 1 Version 2	D–DSE/ SPREE	CHAP/ SPREE	CHAP/ SPREE/ GSPREE	GREG1	GREG2	SYN	EBLUP
0-0-1	0,00165	0,00165	0,00077	0,00200	0,00165	0,00872	0,02646
0-1-1	0,00188	0,00188	0,00182	0,01516	0,00188	0,01025	0,02793
0-0-2	0,00176	0,00176	0,00282	0,01975	0,00330	0,00102	0,01699
0-1-2	0,00501	0,00501	0,00748	0,00870	0,00184	0,01382	0,03150
0-0-3	0,00227	0,00227	0,00680	0,00692	0,00259	0,00507	0,01306
0-1-3	0,00639	0,00639	0,00306	0,00047	0,00317	0,00251	0,02044
0-0-4	0,00031	0,00031	0,00152	0,00132	0,00031	0,00065	0,01857
0-1-4	0,00121	0,00121	0,00173	0,00817	0,00121	0,00257	0,01544
0-0-5	0,00066	0,00066	0,00127	0,00649	0,00066	0,00079	0,01872
0-1-5	0,00062	0,00062	0,00224	0,00052	0,00062	0,00307	0,02093
0-0-6	0,00036	0,00036	0,00095	0,00200	0,00036	0,00473	0,02253
0-1-6	0,00193	0,00193	0,00177	0,00554	0,00193	0,00537	0,02313
0-0-7	0,00045	0,00045	0,00520	0,00626	0,00045	0,00923	0,02696
0-1-7	0,00045	0,00045	0,00047	0,00568	0,00045	0,00690	0,02462
1-0-8	0,00225	0,00225	0,00225	0,00661	0,00225	0,06549	0,04593
1-1-8	0,00219	0,00219	0,00219	0,00524	0,00220	0,05476	0,03556
GEM Nr. 35 Version 2	D–DSE/ SPREE	CHAP/ SPREE	CHAP/ SPREE/ GSPREE	GREG1	GREG2	SYN	EBLUP
0-0-1	0,00001	0,00001	0,00270	0,00454	0,00003	0,01172	0,02677
0-1-1	0,00092	0,00092	0,00526	0,00490	0,00092	0,01406	0,02900
0-0-2	0,00437	0,00437	0,00245	0,01598	0,00277	0,00473	0,02057
0-1-2	0,02122	0,02122	0,01825	0,01508	0,00536	0,01611	0,00013
0-0-3	0,00143	0,00143	0,00896	0,00234	0,00130	0,01067	0,02630
0-1-3	0,01128	0,01128	0,01406	0,00620	0,00173	0,01585	0,03147
0-0-4	0,00139	0,00140	0,00436	0,00126	0,00139	0,00375	0,01922
0-1-4	0,00396	0,00396	0,00164	0,00130	0,00399	0,00265	0,01330
0-0-5	0,00050	0,00050	0,00919	0,00119	0,00054	0,00645	0,00955
0-1-5	0,00108	0,00108	0,00470	0,00463	0,00115	0,00305	0,01280
0-0-6	0,00067	0,00067	0,00142	0,00544	0,00067	0,00296	0,01830
0-1-6	0,00118	0,00118	0,00353	0,00545	0,00120	0,00732	0,02241
0-0-7	0,00059	0,00059	0,00112	0,00848	0,00057	0,00345	0,01903
0-1-7	0,00043	0,00043	0,00563	0,00430	0,00043	0,00260	0,01304
1-0-8	0,00616	0,00616	0,00616	0,00953	0,00613	0,05526	0,03775
1-1-8	0,00579	0,00578	0,00578	0,00139	0,00560	0,07837	0,05997
GEM Nr. 37 Version 2	D–DSE/ SPREE	CHAP/ SPREE	CHAP/ SPREE/ GSPREE	GREG1	GREG2	SYN	EBLUP
0-0-1	0,01429	0,01419	0,00912	0,03161	0,01429	0,00992	0,02598
0-1-1	0,00283	0,00283	0,00810	0,02028	0,00283	0,00261	0,01367
0-0-2	0,00033	0,00045	0,00159	0,03717	0,00834	0,01399	0,03058
0-1-2	0,01197	0,01228	0,03083	0,01750	0,00168	0,02958	0,04529
0-0-3	0,03630	0,03618	0,03451	0,06578	0,02534	0,02233	0,00470
0-1-3	0,04330	0,04297	0,02311	0,09165	0,01710	0,02441	0,00789
0-0-4	0,01971	0,01969	0,00760	0,06158	0,01971	0,00219	0,01833
0-1-4	0,00679	0,00678	0,00195	0,04544	0,00679	0,00974	0,00657
0-0-5	0,00762	0,00761	0,00067	0,00057	0,00762	0,01347	0,02945
0-1-5	0,01516	0,01511	0,00481	0,01713	0,01516	0,00327	0,01943
0-0-6	0,00615	0,00615	0,01637	0,01071	0,00612	0,00171	0,01452
0-1-6	0,00378	0,00379	0,00754	0,01569	0,00378	0,00796	0,02406
0-0-7	0,00053	0,00053	0,01313	0,01711	0,00052	0,00180	0,01795
0-1-7	0,00817	0,00817	0,01062	0,00525	0,00816	0,01085	0,00555
1-0-8	0,00389	1,06441	1,06441	0,00191	0,00385	0,02751	0,00949
1-1-8	0,04591	1,02642	1,02642	0,02095	0,04593	0,02863	0,01122
GEM Nr. 45 Version 2	D–DSE/ SPREE	CHAP/ SPREE	CHAP/ SPREE/ GSPREE	GREG1	GREG2	SYN	EBLUP
0-0-1	0,00175	0,00175	0,00049	0,00624	0,00176	0,00829	0,02404
0-1-1	0,00087	0,00087	0,00423	0,00857	0,00088	0,00533	0,02114
0-0-2	0,00112	0,00113	0,00681	0,10589	0,00953	0,00706	0,02282
0-1-2	0,01113	0,01115	0,02795	0,08095	0,00131	0,03071	0,04608
0-0-3	0,00112	0,00113	0,00688	0,03547	0,01584	0,00692	0,02262
0-1-3	0,00870	0,00868	0,00873	0,03987	0,01644	0,01114	0,02691
0-0-4	0,00157	0,00156	0,01450	0,00127	0,00162	0,01715	0,00100
0-1-4	0,00153	0,00153	0,00363	0,02761	0,00147	0,00009	0,01595
0-0-5	0,00160	0,00160	0,00813	0,00947	0,00160	0,00866	0,02449
0-1-5	0,00088	0,00087	0,00081	0,01393	0,00087	0,00318	0,01909
0-0-6	0,00728	0,00728	0,00498	0,03374	0,00725	0,00243	0,01347
0-1-6	0,00028	0,00028	0,00550	0,00324	0,00032	0,00066	0,01522
0-0-7	0,00025	0,00026	0,00300	0,01820	0,00030	0,00045	0,01546
0-1-7	0,00039	0,00039	0,00246	0,02695	0,00046	0,00112	0,01694
1-0-8	0,00250	0,92133	0,92133	0,09560	0,00252	0,04048	0,02409
1-1-8	0,04077	0,42972	0,42972	0,07018	0,04107	0,04602	0,02900

TABELLE H.6: *ARBs aller Domains in den Gemeinden Nr. 1, 35, 37 und 45, Design 01, Version 2.*

GEM Nr. 1 Version 3	D–DSE/ SPREE	CHAP/ SPREE	CHAP/ SPREE/ GSPREE	GREG1	GREG2	SYN	EBLUP
0-0-1	0,00574	0,00574	0,00574	0,00200	0,00165	0,00872	0,02646
0-1-1	0,00427	0,00427	0,00426	0,01516	0,00188	0,01025	0,02793
0-0-2	0,00403	0,00403	0,00402	0,01975	0,00330	0,00102	0,01699
0-1-2	0,00785	0,00785	0,00784	0,00870	0,00184	0,01382	0,03150
0-0-3	0,00808	0,00808	0,00809	0,00692	0,00259	0,00507	0,01306
0-1-3	0,00352	0,00352	0,00351	0,00047	0,00317	0,00251	0,02044
0-0-4	0,00235	0,00235	0,00236	0,00132	0,00031	0,00065	0,01857
0-1-4	0,00863	0,00863	0,00863	0,00817	0,00121	0,00257	0,01544
0-0-5	0,00221	0,00221	0,00222	0,00649	0,00066	0,00079	0,01872
0-1-5	0,00296	0,00296	0,00296	0,00052	0,00062	0,00307	0,02093
0-0-6	0,00174	0,00174	0,00174	0,00200	0,00036	0,00473	0,02253
0-1-6	0,00065	0,00065	0,00065	0,00554	0,00193	0,00537	0,02313
0-0-7	0,00626	0,00626	0,00626	0,00626	0,00045	0,00923	0,02696
0-1-7	0,00090	0,00090	0,00090	0,00568	0,00045	0,00690	0,02462
1-0-8	0,00225	0,00225	0,00225	0,00661	0,00225	0,06549	0,04593
1-1-8	0,00219	0,00219	0,00219	0,00524	0,00220	0,05476	0,03556
GEM Nr. 35 Version 3	D–DSE/ SPREE	CHAP/ SPREE	CHAP/ SPREE/ GSPREE	GREG1	GREG2	SYN	EBLUP
0-0-1	0,00806	0,00806	0,00809	0,00454	0,00003	0,01172	0,02677
0-1-1	0,01143	0,01143	0,01140	0,00490	0,00092	0,01406	0,02900
0-0-2	0,00119	0,00119	0,00127	0,01598	0,00277	0,00473	0,02057
0-1-2	0,01906	0,01906	0,01881	0,01508	0,00536	0,01611	0,00013
0-0-3	0,00697	0,00697	0,00709	0,00234	0,00130	0,01067	0,02630
0-1-3	0,01337	0,01337	0,01342	0,00620	0,00173	0,01585	0,03147
0-0-4	0,00002	0,00002	0,00004	0,00126	0,00139	0,00375	0,01922
0-1-4	0,00537	0,00537	0,00533	0,00130	0,00399	0,00265	0,01330
0-0-5	0,01020	0,01020	0,01018	0,00119	0,00054	0,00645	0,00955
0-1-5	0,00575	0,00575	0,00576	0,00463	0,00115	0,00305	0,01280
0-0-6	0,00074	0,00074	0,00075	0,00544	0,00067	0,00296	0,01830
0-1-6	0,00464	0,00464	0,00468	0,00545	0,00120	0,00732	0,02241
0-0-7	0,00024	0,00024	0,00021	0,00848	0,00057	0,00345	0,01903
0-1-7	0,00531	0,00531	0,00528	0,00430	0,00043	0,00260	0,01304
1-0-8	0,00616	0,00616	0,00616	0,00953	0,00613	0,05526	0,03775
1-1-8	0,00579	0,00578	0,00578	0,00139	0,00560	0,07837	0,05997
GEM Nr. 37 Version 3	D–DSE/ SPREE	CHAP/ SPREE	CHAP/ SPREE/ GSPREE	GREG1	GREG2	SYN	EBLUP
0-0-1	0,00186	0,00186	0,00185	0,03161	0,01429	0,00992	0,02598
0-1-1	0,00116	0,00116	0,00111	0,02028	0,00283	0,00261	0,01367
0-0-2	0,00593	0,00593	0,00587	0,03717	0,00834	0,01399	0,03058
0-1-2	0,03338	0,03338	0,03352	0,01750	0,00168	0,02958	0,04529
0-0-3	0,03050	0,03050	0,03063	0,06578	0,02534	0,02233	0,00470
0-1-3	0,02069	0,02069	0,02079	0,09165	0,01710	0,02441	0,00789
0-0-4	0,00599	0,00599	0,00605	0,06158	0,01971	0,00219	0,01833
0-1-4	0,00605	0,00605	0,00605	0,04544	0,00679	0,00974	0,00657
0-0-5	0,00542	0,00542	0,00541	0,00057	0,00762	0,01347	0,02945
0-1-5	0,00699	0,00699	0,00694	0,01713	0,01516	0,00327	0,01943
0-0-6	0,00989	0,00989	0,00990	0,01071	0,00612	0,00171	0,01452
0-1-6	0,01167	0,01167	0,01169	0,01569	0,00378	0,00796	0,02406
0-0-7	0,00638	0,00638	0,00645	0,01711	0,00052	0,00180	0,01795
0-1-7	0,00707	0,00707	0,00711	0,00525	0,00816	0,01085	0,00555
1-0-8	0,00389	1,06441	1,06441	0,00191	0,00385	0,02751	0,00949
1-1-8	0,04591	1,02642	1,02642	0,02095	0,04593	0,02863	0,01122
GEM Nr. 45 Version 3	D–DSE/ SPREE	CHAP/ SPREE	CHAP/ SPREE/ GSPREE	GREG1	GREG2	SYN	EBLUP
0-0-1	0,00674	0,00674	0,00680	0,00624	0,00176	0,00829	0,02404
0-1-1	0,00301	0,00301	0,00312	0,00857	0,00088	0,00533	0,02114
0-0-2	0,00545	0,00545	0,00540	0,10589	0,00953	0,00706	0,02282
0-1-2	0,02837	0,02837	0,02814	0,08095	0,00131	0,03071	0,04608
0-0-3	0,00545	0,00545	0,00551	0,03547	0,01584	0,00692	0,02262
0-1-3	0,00889	0,00889	0,00881	0,03987	0,01644	0,01114	0,02691
0-0-4	0,01876	0,01876	0,01877	0,00127	0,00162	0,01715	0,00100
0-1-4	0,00228	0,00228	0,00225	0,02761	0,00147	0,00009	0,01595
0-0-5	0,00721	0,00721	0,00723	0,00947	0,00160	0,00866	0,02449
0-1-5	0,00091	0,00091	0,00087	0,01393	0,00087	0,00318	0,01909
0-0-6	0,00399	0,00399	0,00399	0,03374	0,00725	0,00243	0,01347
0-1-6	0,00301	0,00301	0,00299	0,00324	0,00032	0,00066	0,01522
0-0-7	0,00200	0,00200	0,00199	0,01820	0,00030	0,00045	0,01546
0-1-7	0,00126	0,00126	0,00124	0,02695	0,00046	0,00112	0,01694
1-0-8	0,00250	0,92133	0,92133	0,09560	0,00252	0,04048	0,02409
1-1-8	0,04077	0,42972	0,42972	0,07018	0,04107	0,04602	0,02900

TABELLE H.7: *ARBs aller Domains in den Gemeinden Nr. 1, 35, 37 und 45, Design 01, Version 3.*

I Anhang zur geschichteten Zufallsauswahl

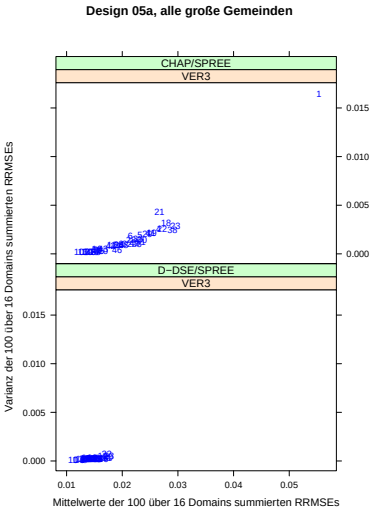


ABBILDUNG I.1: *Die Summen der RRMSEs über alle 16 Domains in jeder großen Gemeinde werden über die 100 Stichproben gemittelt und versus ihrer Varianzen abgebildet. Design 05a, Version 3.*

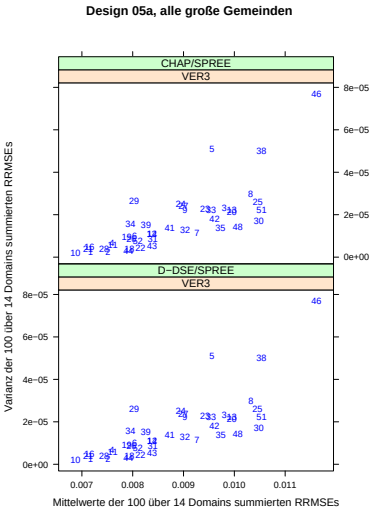


ABBILDUNG I.2: *Die Summen der RRMSEs über alle 14 Domains in jeder kleinen Gemeinde werden über die 100 Stichproben gemittelt und versus ihrer Varianzen abgebildet. Design 05a, Version 3.*

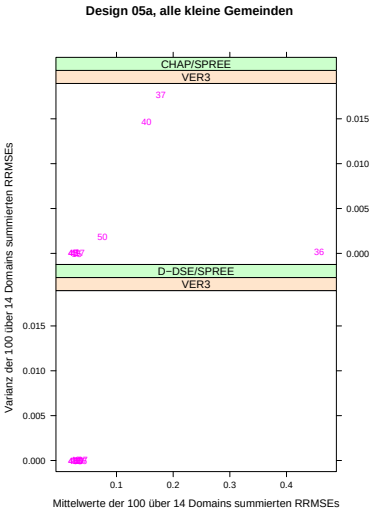


ABBILDUNG I.3: *Mittelwerte der RRMSEs für die 14 Domains der deutschen Population versus Varianz der RRMSEs für alle kleine Gemeinden. Design 05a, Version 3.*

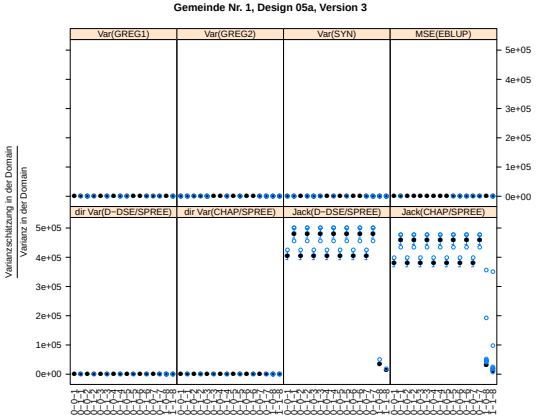


ABBILDUNG I.4: *Relative Verzerrung der Varianzschätzungen auf Domainbasis in der Gemeinde Nr. 1. Design 05a, Version 3.*

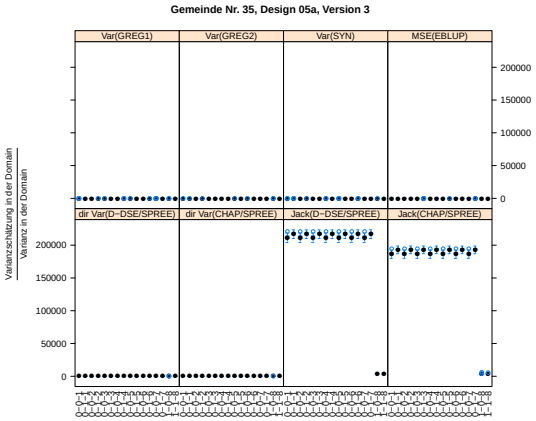


ABBILDUNG I.5: *Relative Verzerrung der Varianzschätzungen auf Domainbasis in der Gemeinde Nr. 35. Design 05a, Version 3.*

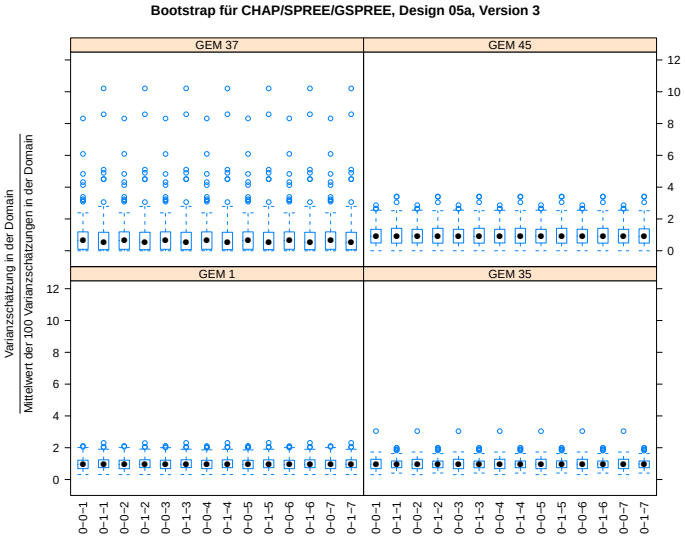


ABBILDUNG I.6: *Verzerrung der Bootstrap–Varianzschätzungen des CHAP/SPREE/GSPREEs auf Domainbasis in Gemeinden Nr. 1, 35, 37 und 45 in der deutschen Population. Design 05a, Version 3.*

GEM Nr. 1 Version 3	D–DSE/ SPREE	CHAP/ SPREE	CHAP/ SPREE/ GSPREE	GREG1	GREG2	SYN	EBLUP
0-0-1	0,00727	0,00727	0,00727	0,08332	0,01317	0,00975	0,02600
0-1-1	0,00748	0,00748	0,00748	0,06678	0,01015	0,01114	0,02544
0-0-2	0,00706	0,00706	0,00704	0,10824	0,03284	0,00455	0,01864
0-1-2	0,01024	0,01024	0,01024	0,10256	0,03310	0,01449	0,03234
0-0-3	0,01023	0,01022	0,01025	0,10941	0,03054	0,00673	0,01276
0-1-3	0,00618	0,00618	0,00619	0,11679	0,03550	0,00503	0,02216
0-0-4	0,00606	0,00606	0,00606	0,06426	0,01697	0,00445	0,01807
0-1-4	0,00961	0,00961	0,00961	0,06349	0,01746	0,00509	0,01557
0-0-5	0,00600	0,00600	0,00599	0,06349	0,01468	0,00446	0,01751
0-1-5	0,00596	0,00596	0,00596	0,06189	0,01430	0,00536	0,01920
0-0-6	0,00534	0,00534	0,00534	0,05678	0,01068	0,00645	0,01986
0-1-6	0,00554	0,00554	0,00554	0,05340	0,00993	0,00694	0,02018
0-0-7	0,00764	0,00764	0,00764	0,06562	0,01198	0,01023	0,02312
0-1-7	0,00579	0,00579	0,00578	0,05288	0,01027	0,00818	0,02155
1-0-8	0,03731	0,35380	0,35380	0,12316	0,03731	0,06563	0,04398
1-1-8	0,04622	0,43184	0,43184	0,10945	0,04622	0,05487	0,03664
GEM Nr. 35 Version 3	D–DSE/ SPREE	CHAP/ SPREE	CHAP/ SPREE/ GSPREE	GREG1	GREG2	SYN	EBLUP
0-0-1	0,01046	0,01044	0,01045	0,06381	0,01153	0,01286	0,02905
0-1-1	0,01341	0,01341	0,01336	0,06608	0,01181	0,01519	0,03167
0-0-2	0,00698	0,00698	0,00698	0,10416	0,02731	0,00625	0,02298
0-1-2	0,01899	0,01899	0,01899	0,11998	0,04837	0,02112	0,00445
0-0-3	0,00983	0,00983	0,00983	0,12572	0,04415	0,00873	0,02847
0-1-3	0,01513	0,01513	0,01513	0,11238	0,04595	0,01428	0,03339
0-0-4	0,00680	0,00682	0,00679	0,06745	0,01750	0,00572	0,02075
0-1-4	0,00688	0,00688	0,00690	0,06763	0,01905	0,00536	0,01494
0-0-5	0,01239	0,01238	0,01237	0,07357	0,01655	0,00822	0,01171
0-1-5	0,00716	0,00716	0,00716	0,06539	0,01809	0,00578	0,01496
0-0-6	0,00685	0,00683	0,00683	0,05515	0,01197	0,00512	0,01977
0-1-6	0,00766	0,00766	0,00765	0,05968	0,00899	0,00847	0,02361
0-0-7	0,00683	0,00680	0,00681	0,07380	0,01193	0,00569	0,01970
0-1-7	0,00691	0,00691	0,00684	0,05046	0,00979	0,00498	0,01366
1-0-8	0,06300	0,06551	0,06551	0,12964	0,06300	0,05982	0,03665
1-1-8	0,05158	0,05162	0,05162	0,12040	0,05142	0,08359	0,05919
GEM Nr. 37 Version 3	D–DSE/ SPREE	CHAP/ SPREE	CHAP/ SPREE/ GSPREE	GREG1	GREG2	SYN	EBLUP
0-0-1	0,03221	0,30806	0,30790	0,46686	NaN	0,02204	NaN
0-1-1	0,04349	0,04310	0,04307	0,41960	NaN	0,02165	NaN
0-0-2	0,03170	0,30622	0,30622	0,55889	NaN	0,02304	NaN
0-1-2	0,04759	0,04751	0,04751	0,69042	NaN	0,03229	NaN
0-0-3	0,04990	0,32302	0,32302	0,81748	NaN	0,03278	NaN
0-1-3	0,05336	0,05275	0,05275	0,58431	NaN	0,03473	NaN
0-0-4	0,03456	0,31135	0,31134	0,35131	NaN	0,02127	NaN
0-1-4	0,04577	0,04530	0,04529	0,44052	NaN	0,02498	NaN
0-0-5	0,03165	0,30654	0,30654	0,39041	NaN	0,02302	NaN
0-1-5	0,04240	0,04197	0,04197	0,43580	0,02641	0,02083	NaN
0-0-6	0,03631	0,31315	0,31316	0,34658	NaN	0,02162	NaN
0-1-6	0,04223	0,04189	0,04191	0,36327	0,08726	0,02170	NaN
0-0-7	0,03481	0,31151	0,31151	0,38796	0,00504	0,02078	NaN
0-1-7	0,04619	0,04579	0,04580	0,33835	0,01773	0,02533	NaN
1-0-8	NaN	3,33744	3,33744	0,77603	NaN	0,03676	NaN
1-1-8	NaN	2,12668	2,12668	0,76896	NaN	0,03807	NaN
GEM Nr. 45 Version 3	D–DSE/ SPREE	CHAP/ SPREE	CHAP/ SPREE/ GSPREE	GREG1	GREG2	SYN	EBLUP
0-0-1	0,02172	0,02169	0,02169	0,22394	0,04128	0,01841	NaN
0-1-1	0,02250	0,02250	0,02250	0,23687	0,02687	0,01676	NaN
0-0-2	0,02146	0,02144	0,02147	0,39555	NaN	0,01564	NaN
0-1-2	0,03902	0,03902	0,03902	0,45695	NaN	0,03508	NaN
0-0-3	0,02154	0,02150	0,02150	0,36398	NaN	0,01610	NaN
0-1-3	0,02502	0,02502	0,02502	0,38248	NaN	0,01883	NaN
0-0-4	0,02840	0,02843	0,02844	0,22969	0,06108	0,02111	NaN
0-1-4	0,02123	0,02122	0,02125	0,24467	0,11099	0,01544	NaN
0-0-5	0,02189	0,02182	0,02182	0,24684	0,01507	0,01867	NaN
0-1-5	0,02191	0,02190	0,02190	0,26217	0,08137	0,01518	NaN
0-0-6	0,02123	0,02127	0,02129	0,17184	0,02896	0,01491	NaN
0-1-6	0,02123	0,02123	0,02119	0,21334	0,05439	0,01519	NaN
0-0-7	0,02097	0,02099	0,02096	0,26077	0,13296	0,01491	NaN
0-1-7	0,02147	0,02147	0,02145	0,18616	0,02375	0,01528	0,02235
1-0-8	NaN	2,47087	2,47087	0,63808	NaN	0,04333	NaN
1-1-8	NaN	2,43736	2,43736	0,65949	NaN	0,04805	NaN

TABELLE I.1: *RRMSEs aller Domains in den Gemeinden Nr. 1, 35, 37 und 45, Design 05a, Version 3.*

GEM Nr. 1 Version 3	D–DSE/ SPREE	CHAP/ SPREE	CHAP/ SPREE/ GSPREE	GREG1	GREG2	SYN	EBLUP
0-0-1	0,00506	0,00506	0,00507	0,00483	0,00278	0,00873	0,02532
0-1-1	0,00506	0,00506	0,00506	0,00977	0,00493	0,01025	0,02486
0-0-2	0,00470	0,00470	0,00469	0,00732	0,00122	0,00101	0,01739
0-1-2	0,00865	0,00865	0,00864	0,00545	0,00527	0,01383	0,03150
0-0-3	0,00876	0,00876	0,00878	0,00428	0,00982	0,00506	0,01109
0-1-3	0,00272	0,00272	0,00273	0,01723	0,00086	0,00247	0,02100
0-0-4	0,00303	0,00303	0,00303	0,01011	0,00237	0,00065	0,01705
0-1-4	0,00782	0,00782	0,00782	0,01383	0,00121	0,00257	0,01429
0-0-5	0,00289	0,00289	0,00289	0,00317	0,00275	0,00079	0,01658
0-1-5	0,00216	0,00216	0,00216	0,00121	0,00158	0,00308	0,01844
0-0-6	0,00107	0,00107	0,00107	0,01018	0,00266	0,00474	0,01922
0-1-6	0,00015	0,00015	0,00015	0,00865	0,00135	0,00538	0,01957
0-0-7	0,00558	0,00558	0,00558	0,00749	0,00031	0,00925	0,02263
0-1-7	0,00170	0,00170	0,00170	0,00206	0,00075	0,00691	0,02094
1-0-8	0,00019	0,07067	0,07067	0,01280	0,00019	0,06547	0,04289
1-1-8	0,00921	0,07280	0,07280	0,00914	0,00921	0,05467	0,03593
GEM Nr. 35 Version 3	D–DSE/ SPREE	CHAP/ SPREE	CHAP/ SPREE/ GSPREE	GREG1	GREG2	SYN	EBLUP
0-0-1	0,00799	0,00798	0,00798	0,00650	0,00092	0,01202	0,02871
0-1-1	0,01234	0,01234	0,01230	0,00215	0,00033	0,01447	0,03133
0-0-2	0,00125	0,00125	0,00125	0,01507	0,00277	0,00346	0,02256
0-1-2	0,01821	0,01821	0,01821	0,00079	0,00625	0,02056	0,00098
0-0-3	0,00702	0,00702	0,00702	0,00916	0,00455	0,00719	0,02816
0-1-3	0,01418	0,01418	0,01418	0,00225	0,01406	0,01354	0,03316
0-0-4	0,00012	0,00011	0,00011	0,01009	0,00306	0,00343	0,02036
0-1-4	0,00436	0,00436	0,00442	0,01054	0,00099	0,00288	0,01438
0-0-5	0,01034	0,01033	0,01032	0,00343	0,00000	0,00674	0,01095
0-1-5	0,00483	0,00483	0,00483	0,00698	0,00273	0,00325	0,01435
0-0-6	0,00084	0,00085	0,00083	0,00244	0,00219	0,00275	0,01939
0-1-6	0,00552	0,00552	0,00555	0,00362	0,00138	0,00723	0,02331
0-0-7	0,00036	0,00038	0,00036	0,00953	0,00010	0,00345	0,01931
0-1-7	0,00439	0,00439	0,00436	0,00008	0,00090	0,00254	0,01302
1-0-8	0,00443	0,01051	0,01051	0,02838	0,00422	0,05960	0,03632
1-1-8	0,00287	0,00005	0,00005	0,01507	0,00327	0,08354	0,05901
GEM Nr. 37 Version 3	D–DSE/ SPREE	CHAP/ SPREE	CHAP/ SPREE/ GSPREE	GREG1	GREG2	SYN	EBLUP
0-0-1	0,00516	0,03506	0,03502	0,03089	NaN	0,00726	NaN
0-1-1	0,00867	0,00846	0,00846	0,01741	NaN	0,00529	NaN
0-0-2	0,00110	0,03076	0,03080	0,15141	NaN	0,01105	NaN
0-1-2	0,02372	0,02394	0,02394	0,02991	NaN	0,02639	NaN
0-0-3	0,03762	0,06840	0,06840	0,02049	NaN	0,02568	NaN
0-1-3	0,03077	0,03051	0,03051	0,01409	NaN	0,02760	NaN
0-0-4	0,01306	0,04307	0,04304	0,03612	NaN	0,00071	NaN
0-1-4	0,01596	0,01571	0,01575	0,01655	NaN	0,01276	NaN
0-0-5	0,00144	0,03123	0,03126	0,06445	NaN	0,01047	NaN
0-1-5	0,00282	0,00259	0,00259	0,01875	0,02641	0,00048	NaN
0-0-6	0,01693	0,04711	0,04712	0,06654	NaN	0,00441	NaN
0-1-6	0,00193	0,00216	0,00216	0,01224	0,07108	0,00516	NaN
0-0-7	0,01347	0,04355	0,04353	0,01555	0,00504	0,00113	NaN
0-1-7	0,01702	0,01680	0,01680	0,02232	0,01773	0,01360	NaN
1-0-8	NaN	0,88220	0,88220	0,06295	NaN	0,03150	NaN
1-1-8	NaN	0,22387	0,22387	0,01673	NaN	0,03214	NaN
GEM Nr. 45 Version 3	D–DSE/ SPREE	CHAP/ SPREE	CHAP/ SPREE/ GSPREE	GREG1	GREG2	SYN	EBLUP
0-0-1	0,00676	0,00664	0,00659	0,02584	0,02853	0,01113	NaN
0-1-1	0,00792	0,00792	0,00793	0,00248	0,01067	0,00808	NaN
0-0-2	0,00547	0,00536	0,00536	0,06815	NaN	0,00794	NaN
0-1-2	0,03319	0,03319	0,03319	0,04343	NaN	0,03210	NaN
0-0-3	0,00541	0,00530	0,00530	0,03402	NaN	0,00833	NaN
0-1-3	0,01356	0,01356	0,01356	0,02775	NaN	0,01233	NaN
0-0-4	0,01883	0,01892	0,01894	0,00732	0,02145	0,01414	NaN
0-1-4	0,00267	0,00265	0,00265	0,01675	0,01702	0,00307	NaN
0-0-5	0,00709	0,00698	0,00698	0,02845	0,00882	0,01129	NaN
0-1-5	0,00586	0,00586	0,00586	0,00446	0,06994	0,00541	NaN
0-0-6	0,00402	0,00415	0,00420	0,04599	0,00153	0,00039	NaN
0-1-6	0,00194	0,00192	0,00194	0,03241	0,00052	0,00236	NaN
0-0-7	0,00210	0,00221	0,00221	0,00017	0,13296	0,00206	NaN
0-1-7	0,00369	0,00372	0,00368	0,01616	0,00271	0,00383	0,02235
1-0-8	NaN	0,48768	0,48768	0,00904	NaN	0,04120	NaN
1-1-8	NaN	0,72354	0,72354	0,10708	NaN	0,04566	NaN

TABELLE I.2: *ARBs aller Domains in den Gemeinden Nr. 1, 35, 37 und 45, Design 05a, Version 3.*

Literaturverzeichnis

Agresti, A. (2002). *Categorical Data Analysis*. Wiley.

Andress, H.-J., Hagenaars, J. A., and Kühnel, S. (1997). *Analyse von Tabellen und kategorialen Daten*. Springer-Verlag.

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey ans satellite data. *Journal of the American Statistical Association*, 83:28–36.

Bell, W. R. (1993). Using information from demographic analysis in post–enumeration survey estimation. *Journal of the American Statistical Association*, 88:1106–1118.

Chapman, D. (1951). *Some properties of the hypergeometric distribution with applications to zoological censuses.* Univ. Calif. Publ. Stat. 1.

Datta, G. S. and Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area extimation problems. *Statistica Sinica*, 10:613–627.

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology*, Vol. 25, No. 2:193–203.

Dostál, L., Gabler, S., Ganninger, M., and Münnich, R. (In submission). Du-

al system estimation in the German register-based census 2011.

Fahrmeir, L. and Hamerle, A. (1984). *Multivariate statistische Verfahren*. Walter de Gruyter.

Faraway, J. J. (2006). *Extending the Linear Model with R*. Chapman & Hall/-CRC.

Fienberg, S. E. (1992). Bibliography on capture-recapture modelling with application to census undercount adjustment. *Survey Methodology*, 18:143–154.

Goldstein, H. (2003). *Multilevel Statistical Models*. Oxford University Press, New York.

Hintze, J. L. and Nelson, R. D. (1998). Statistical computing and graphics, violin plots: A box plot-density trace synergism. *The American Statistician*, 52:181–184.

Hogan, H. (1993). The 1990 post–enumeration survey: Operations and results. *Journal of the American Statistical Association*, 88:1047–1060.

Lohr, S. L. (1999). *Sampling: Design and Analysis*. Duxbury Press.

Münnich, R. (2008). Varianzschätzung in komplexen Erhebungen. *Austrian Journal of Statistics*, 37:319–334.

Münnich, R., Bihler, W., Bjørnstad, J., Davison, A., Sardy, S., Haslinger, A., Knottnerus, P., Laaksonen, S., Ohly, D., Schürle, J., Wiegert, R., Oetliker, U., Renfer, J.-P., Quatember, A., Skinner, C., and Berger, Y. (2003). Data quality in complex surveys. *DACSEIS Deliverable 1.1*. www.dacseis.de.

Münnich, R., Gabler, S., Ganninger, M., Burgard, J. P., and Kolb, J.-P. (2012). *Stichprobenoptimierung und Schätzung im Zensus 2011*. Wiesbaden: Statis-

tisches Bundesamt.

Münnich, R., Gabler, S., Ganninger, M., and Thees, N. (2008). *Zensus Stichproben-Projekt, Technical Annex*. Version: 5. Februar 2008.

Petersen, C. G. J. (1896). The yearly immigration of young plaice into the Limfjord from the German Sea. *Report of the Danish Biological Station*, 6:1–48.

Prasad, N. G. N. and Rao, J. N. K. (1990). The estimation of the mean squared error of small-area-estimators. *Journal of the American Statistical Association*, Vol. 85, No. 409:163–171.

Purcell, N. J. and Kish, L. (1980). Postcensal estimates for local areas (or domains). *International Statistical Review*, 48:3–18.

Rao, J. N. K. (2003). *Small Area Estimation*. Wiley.

Renaud, A. (2004). Coverage estimation for the Swiss population census 2000, estimation, methodology and results. Technical report, Swiss Federal Statistical Office.

Robinson, J. G., Ahmed, B., Gupta, P. D., and Woodrow, K. A. (1993). Estimation of population coverage in the 1990 United States census based on demographic analysis. *Journal of the American Statistical Association*, 88:1061–1071.

Saei, A., Zhang, L.-C., and Chambers, R. (2005). Generalised structure preserving estimation for small areas. *Statistics in Transition*, 7(3):685–696.

Sekar, G. A. F. and Deming, W. E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*, 44:101–115.

Shao, J. (1996). Resampling methods in sample surveys (with discussion). *Statistics*, 27:203–254.

Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap*. Springer-Verlag.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag.

Statistisches Bundesamt (2004). Ergebnisse des Zensustests. *Wirtschaft und Statistik*, Vol. 8:813–833.

The European Parliament and the Council (2008). Regulation (EC) No 763/2008 of the European Parliament and of the Council of 9 July 2008 on the population and housing censuses. *Official Journal of the European Union*, L 218:14–20.

Thompson, S. K. (1992). *Sampling*. John Wiley & Sons, Inc.

Wolter, K. M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*, 81:338–346.

Zhang, L.-C. and Chambers, R. L. (2004). Small area estimates for cross-classifications. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 66(Part 2):479–496.

Studienverlauf

Lucie Dostál
Geburtsjahr/–ort: 1980 in Brno, Tschechische Republik

Ausbildung

Mai 08 – Juli 11	**Universität Trier** Promotion am Lehrstuhl für Wirtschafts– und Sozialstatistik
Sept. 02 – Mai 05	**Palackého Universität in Olomouc** Studium der Wirtschaftsmathematik
Sept. 98 – Juni 02	**Masarykova Universität in Brno** Studium der Mathematik
Sept. 94 – Juni 98	**Gymnasium Brno**

Auszeichnungen

2009 Best Poster Award for the most inspiring Small Area research at the RRC09 - the international conference on small area statistics in Germany

2005 Auszeichnung der Naturwissenschaftlichen Fakultät der Palackého Universität in Olomouc für die beste wissenschaftliche Arbeit 2005

Berufliche Tätigkeiten

Seit Okt. 09	**Deutsches Krebsforschungszentrum, Heidelberg** Wissenschaftliche Mitarbeiterin in Abteilung Epidemiologie von Krebserkrankungen
Mai 08 – Sept. 09	**Universität Trier** Wissenschaftliche Mitarbeiterin am Zensus 2011 Stichprobenforschungsprojekt

Die Auswirkungen von Rahmenfehlern in Zensen werden bereits seit vielen Jahren untersucht. Eine Methode, um aktuelle Bevölkerungszahlen zu gewinnen, basiert auf Fortschreibung. Wegen Ungenauigkeiten in der Fortschreibung wurden aber auch andere Modelle entwickelt – die capture–recapture–Modelle. Am 29. August 2006 hat die Bundesregierung beschlossen, dass in Deutschland 2011 ein registergestützter Zensus durchgeführt wird. Der Schwerpunkt dieser Dissertation liegt in der Anwendung des capture–recapture–Modells im deutschen Zensus 2011. Die Dissertation vergleicht den dual system estimator (DSE) und alternative Schätzer (Verallgemeinerter–Regressionsschätzer, Verhältnis–synthetischer Schätzer, Schätzer basierend auf dem Unit–level Modell) für die Schätzung der Anzahl der tatsächlich vorhandenen Personen. Die empirische Untersuchung der Güte der Schätzer basiert auf Monte Carlo Simulationen synthetischer Populationen des Bundeslandes Saarland.

The effects of coverage errors in censuses have been studied for several years now. One of the methods to estimate these errors uses demographic analysis. Because of some limitation in this method, alternative models were developed. These models are related to the capture–recapture models or to the dual system models. On August 29, 2006, the German Cabinet decided to conduct the census in Germany using a register–based procedure. One of the possible methods to estimate the true population is considered to be dual system estimator (DSE). This thesis compares DSE, which could be used for the register–based census 2011 in Germany, and other well–known estimators (ratio estimator, ratio synthetic estimator, estimated best linear unbiased predictor). The empirical results are based on a Monte Carlo simulation study using a synthetic population of the federal state of Saarland.

Universität Trier