Validierungsprojekt zum deutschen Zensus 2011

Abschlussbericht

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

Version vom 6. März 2013

Universität Trier
Fachbereich IV, VWL
Wirtschafts- und Sozialstatistik
Universitätsring 15
54286 Trier

GESIS – Leibniz-Institut für Sozialwissenschaften
B2, 1
68159 Mannheim

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Bei der Erhebung des Zensus 2011 traten Anschriften auf, in denen laut Einwohnermelderegister keine Personen gemeldet waren. Im Zensus-Stichprobenforschungsprojekt (Münnich et al. 2012) waren von Destatis nur Anschriften im Auswahlrahmen enthalten, in denen mindestens eine Person gemeldet war. Zu untersuchen ist, inwieweit sich die Berücksichtigung unbemeldeter Anschriften auf die ursprünglich getätigten Empfehlungen sowie wie auch die Qualität der Hochrechnung auswirkt.

Dieser Bericht behandelt also den Umgang mit unbemeldeten Anschriften („Nullanschriften“) bei der Hochrechnung. Es soll untersucht werden, ob durch die Existenz von Nullanschriften eine Modifikation der zur Hochrechnung verwendeten Verfahren notwendig ist. In der Leistungsbeschreibung heißt es:

> „Es ist vom Auftragsnehmer zu untersuchen, inwieweit das *Vorhandensein von Nullanschriften* in der Stichprobe in Verbindung mit den realen Verteilungen von Melderegister-Übererfassungen (*Karteileichen*) und -Untererfassungen (*Fehlbeständen*) eine Modifikation der im Stichprobenforschungsprojekt auf Basis von simulierten Verteilungen der Registerfehler ausgesprochenen *Empfehlungen für ein Hochrechnungsverfahren* angeraten sein lässt und damit zu einer *Nachjustierung* des derzeit in der Implementation befindlichen Hochrechnungsverfahrens führen würde. “

Daraus lassen sich zwei verschiedene Fragestellungen ableiten:

- **Fragestellung 1**: Welchen Einfluss haben die Nullanschriften auf die zu erwartende Qualität der Schätzungen in der gezogenen Stichprobe?
- **Fragestellung 2**: Welchen Einfluss haben die Nullanschriften auf die zu erwartende Qualität der Schätzungen im Vergleich zur Ausgangssituation des Zensus Stichprobenforschungsprojekt?

Im Rahmen dieses Berichts wird die erste Fragestellung behandelt. Um dies zu untersuchen, werden zwei unterschiedliche Herangehensweisen gewählt. Zum einen werden die Auswirkungen auf Basis der tatsächlich gezogenen Stichprobe und zum anderen auf Basis einer Monte-Carlo-Simulation untersucht. Ausgangspunkt beider Herangehensweisen sind Dateien, die alle erfassten Anschriften enthalten.[1] Für die Hochrechnung der Zensusbevölkerung ist die Zahl der Karteileichen und Fehlbestände von enormer Bedeutung. Da diese nur für die Stichprobe bekannt sein kann, wurden auf Basis dieser Stichprobe Modelle geschätzt, die zur Erzeugung synthetischer Werte für den ganzen Datensatz genutzt wurden. Die Modelle wurden im Statistischen Bundesamt geschätzt und sind in Kapitel 2 aufgelistet.

In einem weiteren Schritt werden verschiedene Szenarien entwickelt. Diese Szenarien orientieren sich an den in der Leistungsbeschreibung genannten Möglichkeiten, wie mit dem Problem der Nullanschriften umgegangen werden soll. Insgesamt werden 24 Szenarien behandelt, wobei sich an dieser Stelle einige Unterschiede bei den beiden gewählten Ansätzen ergeben. Die Ergebnisse der Szenarioanalyse sind in Kapitel 4 zu finden, während der Aufbau und die Ergebnisse der Monte-Carlo-Simulation in Kapitel 5 beschrieben werden. Zum Schluss werden in Kapitel 6 globale Bewertungen durchgeführt und Empfehlungen für den Umgang mit den Nullanschriften gegeben.

Der Fragestellung 2 wird im Endbericht nachgegangen.

[1] Weitere Informationen über diese Dateien sind in Kapitel A zu finden.

2 Vorarbeiten

Um die Untersuchungen durchführen zu können wurden von Seiten des Auftraggebers Daten geliefert. Zur Vorbereitung der Stichprobenanalyse und der Simulationen waren allerdings noch einige Vorarbeiten zu tätigen. Die Datengrundlage und diese Vorarbeiten werden im folgenden Kapitel beschrieben.

2.1 Datengrundlage

Die wichtigste Grundlage des Projekts war ein Datensatz, der alle Anschriften in Deutschland umfasst. Für jede Anschrift sind Informationen über die Gemeinde verzeichnet, zu der die Anschrift gehört, zum Beispiel welchen 12-stelligen amtlichen Gemeindeschlüssel die Gemeinde (`AGS12`) hat und ob es sich um einen Stadtttteil einer Gemeinde handelt (`STADTTEIL`).

Anhand des Datensatzes lässt sich in Erfahrung bringen, wie viele Personen ihren Haupt- (`REG_HW`) oder Nebenwohnsitz (`REG_NW`) in der Anschrift gemeldet haben. Für die Anschriften, die in die Zensus-Stichprobe gelangt sind, ist zusätzlich die Zahl der Karteileichen (`KL_HW`) und Fehlbestände (`FB_HW`) enthalten. Diese Information steht auch für die Personen mit Nebenwohnsitz zur Verfügung (`KL_NW` und `FB_NW`).

Für Anschriften die nicht in die Stichprobe gelangt waren, mussten synthetische Karteileichen und Fehlbestände erzeugt werden. Hierfür wurden verschiedene Modelle verwendet, die in Abschnitt 2.3 näher beschrieben werden. Die Ergebnisse sind in den Variablen `KLDACH_HW_M1`, `KLDACH_NW_M1`, `KLDACH_HW_M2`, `KLDACH_NW_M2`, `KLDACH_HW_M3`, `KLDACH_NW_M3`, `KLDACH_HW_M4`, `KLDACH_NW_M4`, `FBDACH_HW_M1`, `FBDACH_NW_M1`, `FBDACH_HW_M2`, `FBDACH_NW_M2`, `FBDACH_HW_M3`, `FBDACH_NW_M3` enthalten.

Um die Allokation korrekt durchzuführen, waren die Anschriften in Schichten eingeteilt (`SCHICHTID`).

Anhand der Variable `STICHPROBENKENNUNG` lässt sich erkennen, ob eine Anschrift Teil der Stichprobe war oder nicht.

Am 24.10.2012 wurde vom Auftraggeber die Datei `KL_FB_DACH_ANSCHR.CSV` auf DVD zur Verfügung gestellt. Die Datei ist 3 173 485 KB groß und umfasste 19 802 090 Zeilen (Anschriften) und 31 Spalten.

Unter anderem waren auch folgende Variablen enthalten: Zahl der Anschriften einer Schicht in der Grundgesamtheit `GRNH`, Zahl der Anschriften einer Schicht in der Stichprobe `KLNH`, Schichtkennnung `h`, `Erhebungswelle`, SMP-Nummer `SMPNR`, `SMP_TYP`.

In der Variable `SCHICHTID` gab es 6065 Missings, deren zugehörige Anschriften gelöscht wurden. Außerdem wurden 3521 Anschriften entfernt, die nicht zur Hauptziehung gehörten. 991 Sonderanschriftennummern wurden ebenfalls gelöscht. Damit verbleiben für die weiteren Untersuchungen 19 791 513 Anschriften. Alle Vektoren beziehen sich im Folgenden nur noch auf diese Anschriften.

2.2 Karteileichen- und Fehlbestandsmodelle

Informationen darüber, ob es sich bei einer gemeldeten Person um eine Unter- oder Übererfassung handelt, können nur aus der Stichprobe gewonnen werden.[2] Allerdings werden durch die Stichprobe nur etwas weniger als 10 % aller Anschriften in Deutschland abgedeckt. Im Rahmen der Simulation resultiert das stochastische Moment allerdings aus der Stichprobenziehung. Deshalb müssen die Variablen Karteileichen und Fehlbestände künstlich aufgefüllt werden. Die künstlichen Werte resultieren aus den Vorhersagen, die aufgrund von Multilevelmodellen gemacht wurden. Da es sich bei den Einzeldaten um hochsensible Informationen handelt, durften die Schätzungen auf den Einzeldaten nur im Statistischen Bundesamt durchgeführt werden.

Es wurden vom Auftragnehmer verschiedene Karteileichen- und Fehlbestandsmodelle erstellt, deren Parameter vom Auftraggeber geschätzt und auf Anschriftenebene an den Auftragnehmer geliefert wurden. Auf Basis dieser Modelle wurden die fehlenden Werte aufgefüllt.

Die abhängigen Variablen in diesen Modellen sind also Karteileichen beziehungsweise Fehlbestände. Als erklärende Variablen wurden für das Karteileichenmodell folgende Variablen gewählt:[3]

- Geschlecht,
- Gemeindegrößenklasse,
- Wohnstatus,
- Familienstand und
- Nationalität

Beim Fehlbestandsmodell werden teilweise andere erklärende Variablen gewählt:

- Geschlecht,
- Adressgrößenklasse
- Altersklassen,
- Nationalität und
- Gemeindegrößenklasse

Bei den Multilevelmodellen handelt es sich um generalized logit mixed models. Für die ursprünglichen Modelle (KL-Modell 2 und FB-Modell 2) wurden jeweils nur Randeffekte also keine Interaktionen verwendet. Die anderen Modelle bauen jeweils auf diesem Grundmodell auf. Allerdings werden in diesen Modellen auch Interaktionseffekte berücksichtigt. Das Bundesland, die Gemeinde und die Anschrift sind zufällige Effekte, die ebenfalls in das Modell einfließen.

$$\mathrm{KL} \sim \alpha + \mathrm{SEX} + \mathrm{GGK} + \mathrm{FST} + \mathrm{WST} + \mathrm{NAT} \tag{2.1}$$

[2] Die folgenden Ausführungen gehen zurück auf Ziffer 2 des Entwurfs der Leistungsbeschreibung vom 22.6.2012.

[3] Diese Wahl der Variablen geht auf die Folien von Herrn Burgard für die Sitzung des AK für mathematische Methodik im Juni 2009 zurück. Siehe auch Burgard & Münnich (2010).

2.2.1 Karteileichenmodelle

Die vier verschiedenen Karteileichenmodelle sind in Tabelle 1 aufgeführt:

KL-Modell 1:	KL $\sim \alpha$ +SEX + AGK + GGK + FST + WST + NAT + FST:NAT + WST:NAT + AGK:WST + AGK:FST + WST:FST
KL-Modell 2:	KL $\sim \alpha$ + SEX + GGK + FST + WST + NAT
KL-Modell 3:	KL $\sim \alpha$ + SEX + AGK + GGK + FST + WST + NAT + FST:NAT + WST:NAT + AGK:WST + AGK:FST + WST:FST + GGK:FST + GGK:NAT + GGK:WST + GGK:AGK
KL-Modell 4:	KL $\sim \alpha$ + SEX + GGK + FST + WST + NAT + AGK

Tabelle 1: Die Karteileichenmodelle

2.2.2 Fehlbestandsmodelle

Die drei verschiedenen Karteileichenmodelle sind in Tabelle 2 aufgeführt:

FB-Modell 1:	FB $\sim \alpha$ + SEX + AGK + GGK + FST + WST + NAT + FST:NAT + WST:NAT + AGK:WST + AGK:FST + WST:FST
FB-Modell 2:	FB $\sim \alpha$ + SEX + AGK + GGK + WST + NAT + ADK
FB-Modell 3:	FB $\sim \alpha$ + SEX + AGK + GGK + FST + WST + NAT + GGK + FST:NAT + WST:NAT + AGK:WST + AGK:FST + WST:FST + GGK:FST + GGK:NAT + GGK:WST + GGK:AGK

Tabelle 2: Die Fehlbestandsmodelle

Aus der Kombination aller Karteileichenmodelle mit allen Fehlbestandsmodellen ergeben sich also 12 Kombinationsmöglichkeiten.

2.3 Erstellung der Karteileichen- und Fehlbestandsvektoren

Für die Stichprobenanschriften sind - bis auf einige fehlende Werte - die tatsächliche Anzahl an Karteileichen und Fehlbestände bekannt und werden in den verwendeten Vektor übernommen. Für die Stichprobenanschriften mit fehlenden Werten bei den Karteileichen und Fehlbeständen sowie bei den Nicht-Stichprobenanschriften wird auf Basis der oben beschriebenen KAL- und FEB-Modelle ein synthetischer Wert erstellt, der allerdings keine ganze Zahl ist. Daher muss eine Rundung durchgeführt werden. Um die Rundung durchzuführen, wird der Cox-Algorithmus verwendet (Cox 1987). Neben der Regel, dass Karteileichen und Fehlbestände ganzzahlig sein müssen, sind noch weitere Restriktionen zu beachten. So dürfen in einer Anschrift beispielsweise nicht mehr Karteileichen als gemeldete Personen auftauchen.

Bei der Erzeugung von Karteileichen- und Fehlbestandsvektoren werden nur Anschriften aus der Hauptziehung berücksichtigt.

Der genaue Ablauf der Erstellung der Karteileichen- und Fehlbestandsvektoren ist im Folgenden beschrieben:

2.3.1 Karteileichen

Von Destatis wurde zur Verfügung gestellt:

KL_HW	Tatsächliche Anzahl an Karteileichen in Stichprobenanschriften [ganze Zahlen]
KLDACH_HW_Mx	Durch Modell geschätzte erwartete Anzahl an Karteileichen in allen ' Anschriften [reelle nichtnegative Zahlen] (x=1,2,3,4)

Die vom Auftraggeber gelieferten tatsächlich beobachteten Karteileichenvektoren weisen nicht nur bei allen Nicht-Stichprobenelementen sondern auch in einigen Fällen bei Stichprobeneinheiten Missings auf.

Zur Unterscheidung zwischen diesen und den später durch den Auftragnehmer erzeugten Karteileichen- und Fehlbestandsvektoren werden im Folgenden die Variablenbezeichnungen KL bzw. FB verwendet, um die vom Auftraggeber gelieferten, KAL bzw. FEB, um die vom Auftragnehmer erzeugten Vektoren zu beschreiben.

Für jedes Karteileichenmodell ist ein 19 791 513 $\times$ 1 Vektor KAL_HW_M gesucht mit

a) KAL_HW_M = KL_HW für Stichprobenanschriften, die in KL_HW keinen fehlenden Wert haben (fehlend sind 10 406 Werte)

b) |KAL_HW_M - KLDACH_HW_Mx| $<$ 1 für Nichtstichprobenanschriften

c) $|\sum_{\bar{S}_{\mathrm{SMP}\times h}}$ (KAL_HW_M - KLDACH_HW_Mx)| $<$ 1

d) KAL_HW_M $\leq$ REG_HW

wobei gilt:

$S_{\mathrm{SMP}\times h}$ der Menge aller Stichprobenanschriften in SMP $\times$ h
$\bar{S}_{\mathrm{SMP}\times h}$ der Menge aller Nichtstichprobenanschriften in SMP $\times$ h

b) und c) werden nicht genau eingehalten, da es missings in den Stichprobenanschriften gibt und wegen d)

Zunächst wird KAL_HW_M = KLDACH_HW_Mx gesetzt.

1. REG_HW=0 impliziert KAL_HW_M=0
2. Binomialmodell zur Erzeugung der Karteileichen für noch 2 236 Anschriften mit missings in verbleibenden Karteileichen. Binomialmodell auf Anschrift a mit

$$\mathrm{n}_a = \mathrm{REG\_HW}_a$$

und für Anschrift $a \in S_{\mathrm{SMP}\times h}$

$$\mathrm{p}_a = \frac{\sum_{\tilde{a}\in S_{\mathrm{SMP}\times h}} \mathrm{KAL\_HW\_M}_{\tilde{a}}}{\sum_{\tilde{a}\in S_{\mathrm{SMP}\times h}} \mathrm{REG\_HW}_{\tilde{a}}},$$

d.h. p_a hängt von a nur über die Zugehörigkeit von a zum SMP und zur Schicht h ab, aus der die Anschrift a stammt.

3. Ersetzung der fehlenden 2 236 Karteileichen über Binomialmodell aus 2.
4. Ersetzung der Komponenten von KAL_HW_M in der Stichprobe durch KL_HW, falls KL_HW nicht-missing
5. KAL_HW_M > REG_HW ⇒ KAL_HW_M = REG_HW
6. Cox-Algorithmus ablaufen lassen mit n = round(sum(KAL_HW_M)) und A = KAL_HW_M aufsummiert für alle 2 365 × 8 Zellen SMPNR × h. [ftol = 10^{-10}]
7. Bei den restlichen Zellen wird $\sum_{\bar{S}_{\mathrm{SMP}\times h}}$ KAL_HW_M auf die Nichtstichprobenanschriften mittels pps-systematischer Zufallsauswahl aufgeteilt. [Cox ist hier zu langsam]

Ergebnis ist ein Vektor KAL_HW_M, der in den 1 933 337 Stichprobenanschriften mit KL_HW nicht missing übereinstimmt und für die anderen Anschriften sich in den meisten Anschriften von KLDACH_HW_Mx um weniger als 1 unterscheidet.

Hauptziehung Modell 1:	Stichproben-Anschriften	Nichtstichproben-Anschriften	Gesamt
|KAL_HW_M - KLDACH_HW_M1| < 1	1619585	16267685	17887270
|KAL_HW_M - KLDACH_HW_M1| > 1	70343	5438	75781
KLDACH_HW_M1 missing	253815	1574647	1828462
Gesamt	1943743	17847770	19791513

Hauptziehung Modell 2:	Stichproben-Anschriften	Nichtstichproben-Anschriften	Gesamt
|KAL_HW_M - KLDACH_HW_M2| < 1	1619147	16273114	17887270
|KAL_HW_M - KLDACH_HW_M2| > 1	70781	9	70790
KLDACH_HW_M2 missing	253815	1574647	1828462
Gesamt	1943743	17847770	19791513

Hauptziehung Modell 3:	Stichproben-Anschriften	Nichtstichproben-Anschriften	Gesamt
|KAL_HW_M - KLDACH_HW_M3| < 1	1619509	16273116	17892625
|KAL_HW_M - KLDACH_HW_M3| > 1	70419	7	70426
KLDACH_HW_M3 missing	253815	1574647	1828462
Gesamt	1943743	17847770	19791513

Hauptziehung Modell 4:	Stichproben-Anschriften	Nichtstichproben-Anschriften	Gesamt
|KAL_HW_M - KLDACH_HW_M4| < 1	1620030	16273119	17893149
|KAL_HW_M - KLDACH_HW_M4| > 1	69898	4	69902
KLDACH_HW_M4 missing	253815	1574647	1828462
Gesamt	1943743	17847770	19791513

Die größten Abweichungen von |KAL_HW_M - KLDACH_HW_Mx| sind in der Stichprobe zu finden und zeigen, dass in Einzelfällen die beobachteten Karteileichen und die modellierten Karteileichen sich stark unterscheiden können. Dies deutet implizit auf Ausreißer hin und verdeutlicht zudem die Schwierigkeiten, ein den Daten sehr gut entsprechendes Karteileichenmodell aufzustellen.

2.3.2 Fehlbestände

Von Destatis wurde zur Verfügung gestellt:

FB_HW	Tatsächliche Anzahl an Fehlbeständen in Stichprobenanschriften [ganze Zahlen]
FBDACH_HW_Mx	Durch Modell geschätzte erwartete Anzahl an Fehlbeständen in allen Anschriften [reelle nichtnegative Zahlen] (x=1,2,3)

Die vom Auftraggeber gelieferten tatsächlich beobachteten Fehlbestandsvektoren weisen nicht nur bei allen Nicht-Stichprobenelementen sondern auch in einigen Fällen bei Stichprobeneinheiten Missings auf.

Für jedes Modell ist ein 19 791 513 × 1 Vektor FEB_HW_M gesucht mit

a) FEB_HW_M = FB_HW für Stichprobenanschriften, die in FB_HW keinen fehlenden Wert haben (fehlend sind 10 406 Werte)

b) |FEB_HW_M - FBDACH_HW_Mx| < 1 für Nichtstichprobenanschriften

c) $|\sum_{\bar{S}_{\mathrm{SMP}\times h}}$ (FEB_HW_M - FBDACH_HW_Mx)| < 1

b) und c) werden nicht genau eingehalten.

Zunächst wird FEB_HW_M = FBDACH_HW_Mx gesetzt.

1. Poissonmodell zur Erzeugung der Fehlbestände für noch 2 236 Anschriften mit missings in verbleibenden Fehlbeständen auf Anschrift $a \in S_{\mathrm{SMP}\times h}$

$$\mathrm{p}_a = \frac{\sum_{\tilde{a}\in S_{\mathrm{SMP}\times h}} \mathrm{FEB\_HW\_M}_{\tilde{a}}}{\sum_{\tilde{a}\in S_{\mathrm{SMP}\times h}} \mathrm{REG\_HW}_{\tilde{a}}},$$

   d.h. p_a hängt von a nur über die Zugehörigkeit von a zum SMP und zur Schicht h ab, aus der die Anschrift a stammt.

2. Ersetzung der fehlenden 1 828 462 Fehlbestände über Poissonmodell aus 1.

3. Ersetzung der Komponenten von FEB_HW_M in der Stichprobe durch FB_HW, falls FB_HW nicht-missing

4. Cox-Algorithmus ablaufen lassen mit n = round($\sum$(FEB_HW_M)) und A = FEB_HW_M aufsummiert für alle 2 365 × 8 Zellen SMPNR × h. [ftol = 10^{-10}]

5. Bei den restlichen Zellen wird $\sum_{\bar{S}_{\mathrm{SMP}\times h}}$ FEB_HW_M auf die Nicht-Stichprobenanschriften mittels pps-systematischer Zufallsauswahl aufgeteilt. [Cox ist hier zu langsam]

Ergebnis ist ein Vektor FEB_HW_M, der in den Stichprobenanschriften mit FB_HW für nicht missings übereinstimmt[4] und für die anderen Anschriften sich in den meisten Anschriften von FBDACH_HW_Mx um weniger als 1 unterscheidet.

[4] Im Gegensatz zu den Karteileichen bei den Hauptwohnungen gibt es bei den Fehlbeständen durch Rundungsfehler einige wenige Stichprobenanschriften bei denen sich FEB_HW_M und FB_HW um +/- 1 unterscheiden. Bei den Nebenwohnungen gibt es sowohl bei KAL_NW_M als auch bei FEB_NW_M derartige geringe Abweichungen.

Hauptziehung Modell 1:	Stichproben-Anschriften	Nichtstichproben-Anschriften	Gesamt
|FEB_HW_M - FBDACH_HW_M1| < 1	1654652	16273122	17927774
|FEB_HW_M - FBDACH_HW_M1| > 1	35276	1	35277
FBDACH_HW_M1 missing	253815	1574647	1828462
Gesamt	1943743	17847770	19791513

Hauptziehung Modell 2:	Stichproben-Anschriften	Nichtstichproben-Anschriften	Gesamt
|FEB_HW_M - FBDACH_HW_M2| < 1	1655808	16273122	17928930
|FEB_HW_M - FBDACH_HW_M2| > 1	34120	1	34121
FBDACH_HW_M2 missing	253815	1574647	1828462
Gesamt	1943743	17847770	19791513

Hauptziehung Modell 3:	Stichproben-Anschriften	Nichtstichproben-Anschriften	Gesamt
|FEB_HW_M - FBDACH_HW_M3| < 1	1655039	16273121	17928160
|FEB_HW_M - FBDACH_HW_M3| > 1	34889	2	34891
FBDACH_HW_M3 missing	253815	1574647	1828462
Gesamt	1943743	17847770	19791513

Die größten Abweichungen von |FEB_HW_M - FBDACH_HW_Mx| sind in der Stichprobe zu finden und zeigen, dass die beobachteten Fehlbestände und die modellierten Fehlbestände sich in Einzelfällen stark unterscheiden können. Dies deutet implizit auf Ausreißer hin und verdeutlicht zudem die Schwierigkeiten, ein den Daten sehr gut entsprechendes Fehlbestandsmodell aufzustellen.

Analoge Prozeduren wurden für Karteileichen- und Fehlbestandsvektoren für Nebenwohnungen angewendet.

2.4 Untersuchte Szenarien

In diesem Kapitel werden die Ergebnisse der Szenarioanalyse beschrieben. Bei dieser Herangehensweise wird die gezogene Stichprobe als gegeben angenommen. Basierend auf dieser Stichprobe wird jeweils ein GREG-Schätzer verwendet, um eine Hochrechnung für Deutschland zu erzeugen. Neben den Stichprobendaten und den oben beschriebenen Karteileichen- und Fehlbestandsvektoren werden keinerlei weitere Informationen verwendet. Neben den $4 \times 3 = 12$ Kombinationen der verschiedenen Karteileichen- und Fehlbestandsmodelle wurden noch weitere Szenarien unterschieden. Beispielsweise gibt es die Möglichkeit, keine oder nur erwünschte Nullanschriften oder auch Ausreißer zu berücksichtigen.

Die GREGs wurden für jedes der 24 Szenarien auf unterschiedliche Weise berechnet, entsprechend den Formeln in Tabelle 4. Unterschieden wird dabei zwischen folgenden Punkten, die im Anschluss genauer beschrieben werden:

- Hat eine Ausreißerbereinigung stattgefunden?
- Werden alle Nullanschriften, nur erwünschte oder keine Nullanschriften bei der Schätzung berücksichtigt?
- Werden die Regresssionsparameter β auf Kreis- oder SMP-Ebene geschätzt?

- Werden alle SMP's oder nur SMP's vom Typ 0 und 1 zur Schätzung von β herangezogen?

In einer Voruntersuchung der Anschriften aus der tatsächlich gezogenen Stichprobe stellte sich heraus, dass in den Daten Ausreißer bei der Zahl der Karteileichen und Fehlbestände vorhanden sind. Um zu beurteilen, ob es sich bei einer Anschrift um einen Ausreißer handelt oder nicht, wird Cook's Distanz gewählt. Sie liefert einen Hinweis, wie sehr eine einzelne Anschrift die $\beta-$Koeffizienten der Regressionsgleichung beeinflusst. Cook's-Distanz ist definiert durch (Cook & Weisberg 1982, S. 136):

$$D_i = \frac{\sum_{j=1}^{N}(\hat{Y}_j - \hat{Y}_{j(i)})^2}{\frac{p}{N}\sum_{i=1}^{N}(\hat{Y}_i - Y_i)^2}.$$

wobei

N	Zahl der Beobachtungen
p	Zahl der zu schätzenden Regressionsparameter ($p = 2$ oder 3)
Y_j	Beobachteter j-ter Wert
$\hat{Y}_j$	Vorhergesagter Wert für Beobachtung j unter Verwendung aller N Beobachtungen
$\hat{Y}_{j(i)}$	Vorhergesagter Wert für Beobachtung j unter Verwendung aller N Beobachtungen außer der i-ten Beobachtung

Die Definition der unerwünschten Nullanschrift ist eine vom Auftraggeber bewusst gewählte Approximation. Für die Hochrechnung sind nur die gemeldeten und existenten Personen relevant, nicht die Wohnraumgemeinschaft. Eine Anschrift ohne gemeldete und existente Personen hat immer die gleichen Auswirkungen, egal ob es sich um ein Trafohäuschen oder ein leerstehendes Wohngebäude handelt.

Alle Szenarien sind in Tabelle 3 aufgelistet.

Tabelle 3: Tabelle mit Szenarien

Szenario	Ausreißer-bereinigung	Nullanschriften	Ebene	SMP-Typ	Bezeichnung
1	keine	alle drin	SMP	nur 0 & 1	SMP-SEP.01
2	1%	alle drin	SMP	nur 0 & 1	SMP-SEP.01A
3	keine	nur erwünschte	SMP	nur 0 & 1	SMP-SEP.E01
4	1%	nur erwünschte	SMP	nur 0 & 1	SMP-SEP.E01A
5	keine	keine Nullanschriften	SMP	nur 0 & 1	SMP-SEP.N01
6	1%	keine Nullanschriften	SMP	nur 0 & 1	SMP-SEP.N01A
7	keine	alle drin	Kreis	nur 0 & 1	KRS-SEP.01
8	1%	alle drin	Kreis	nur 0 & 1	KRS-SEP.01A
9	keine	nur erwünschte	Kreis	nur 0 & 1	KRS-SEP.E01
10	1%	nur erwünschte	Kreis	nur 0 & 1	KRS-SEP.E01A
11	keine	keine Nullanschriften	Kreis	nur 0 & 1	KRS-SEP.N01
12	1%	keine Nullanschriften	Kreis	nur 0 & 1	KRS-SEP.N01A
13	keine	alle drin	SMP	alle	SMP-SEP.
14	1%	alle drin	SMP	alle	SMP-SEP.A
15	keine	nur erwünschte	SMP	alle	SMP-SEP.E
16	1%	nur erwünschte	SMP	alle	SMP-SEP.EA
17	keine	keine Nullanschriften	SMP	alle	SMP-SEP.N
18	1%	keine Nullanschriften	SMP	alle	SMP-SEP.NA
19	keine	alle drin	Kreis	alle	KRS-SEP.
20	1%	alle drin	Kreis	alle	KRS-SEP.A
21	keine	nur erwünschte	Kreis	alle	KRS-SEP.E
22	1%	nur erwünschte	Kreis	alle	KRS-SEP.EA
23	keine	keine Nullanschriften	Kreis	alle	KRS-SEP.N
24	1%	keine Nullanschriften	Kreis	alle	KRS-SEP.NA

Die Szenarien 1 bis 6 sind in Szenarien 13 bis 18 enthalten. Die Szenarien 7 bis 12 sind nicht in den Szenarien 19 bis 24 enthalten, da in die β-Schätzungen entweder nur Stichprobendaten der SMP-Typen 0 und 1 im Kreis eingehen oder alle SMP-Typen.

Die oben stehende Tabelle enthält neben den 24 Szenarien-Nummern noch weitere Informationen, welche die Szenarien definieren. Zum einen wird unterschieden, ob bei den Anschriften eine Ausreißerbereinigung stattgefunden hat. Eine Anschrift wird dann als Ausreißer definiert, wenn die zugehörige Cook's Distanz größer als der 99% Perzentilwert ist, wobei die Berechnung der Distanz auf dem Gesamtdatenbestand erfolgt.[5] In der oben stehenden Tabelle wird ein Szenario mit Ausreißerbereinigung mit dem Label "1%"gekennzeichnet.

Zudem kann die Datengrundlage für die β-Schätzungen so eingeschränkt werden, dass nur Anschriften mit mehr als 0 registrierten Personen (REG_HW + REG_NW $>$ 0) berücksichtigt werden (keine Nullanschriften). Im Fall erwünschte Nullanschriften sind 0 oder mehr Personen registriert (REG_HW + REG_NW $\geq$ 0). Ist keine Person registriert, muss dann aber mindestens ein Fehlbestand vorliegen (REG_HW + REG_NW =0 & FEB_HW + FEB_NW>0). In Spalte drei der Tabelle wird diese Unterscheidung kenntlich gemacht durch die Labels „alle drin“, „nur erwünschte“und „keine Nullanschriften“.

[5]Bei der Monte-Carlo-Simulation in Kapitel 5 wird eine andere Ausreißerbereinigung durchgeführt.

Spalte vier bezeichnet die Ebene, auf der die βs des Regressionsmodells geschätzt werden. „SMP“ bezeichnet dabei Szenerien, in denen die βs auf SMP-Ebene, „Kreis“Szenarien, in denen die βs auf Kreis-Ebene geschätzt werden.[6]

In der mit „SMP-Typ“ überschriebenen Spalte bedeutet „nur 0 & 1“, dass ausschließlich SMPs des Typs 0 und 1 betrachtet werden, das heißt, dass Anschriften aus SMPs des Typs 2 oder 3 für die Schätzungen wegfallen. Das Label „alle“ bezeichnet eine Situation, in der dieser Ausschluss nicht stattfindet.

Die Spalte „Bezeichnung“ gibt alternativ zur Szenarien-Nummer einen Marker für die einzelnen Szenarien wider, der in den folgenden Analysen etwa für eine Unterscheidung der Schätzer benutzt wird.

Wie bereits erwähnt sind sich die Szenarien 1 bis 6 beziehungsweise 13 bis 18 sehr ähnlich. Szenario 1 unterschiedet sich beispielsweise von Szenario 13 nur darin, dass bei Szenario 13 auch Ergebnisse für die SMP-Typen 2 und 3 enthalten sind. Die Matrix der Ergebnisse ist also in diesem Fall länger.

Die Schätzungen für die 24 Szenarien werden zweimal durchgeführt. Einmal werden - auf Wunsch des Auftraggebers in der Leistungsbeschreibung - die Nullanschriften in einer eigenen Dummy-Variablen definiert und einmal nicht.

3 Methodische Grundlagen

In diesem Kapitel werden die methodischen Grundlagen beschrieben. Es werden die GREG-Schätzer definiert, die zur Hochrechnung in Frage kommen und die Varianzschätzung für diese Schätzer vorgestellt.

3.1 Verallgemeinerter Regressionsschätzer

Mit jedem Element $i, i = 1, \ldots, N$ der Grundgesamtheit ist eine Ausprägung Y_i der Zielvariable und ein p-Vektor x_i von p Hilfsvariablen verbunden. Y_i ist in unserem Fall die Zahl der tatsächlich vorhandenen Personen in Anschrift i, $Y_i = \text{REG\_HW}_i - \text{KAL\_HW}_i + \text{FEB\_HW}_i$. T_x ist der p-Vektor der Summe der Hilfsvariablen in der Gesamtheit. In der $N \times p$ Matrix X sind die N Zeilen die x_i Vektoren der p Hilfsvariablen. Der Index s bezieht sich auf die Zeilen, die zur Stichprobe s gehören.

Das folgende Modell wird unterstellt:

$$E_M(Y_i) = x_i'\beta$$

$$V_M(Y_i) = v_i$$

$$V = \text{diag}\,(v_1, \ldots, v_N)$$

[6] Im Gegensatz zur Monte-Carlo-Simulation in Kapitel 5 wird weder eine Deutschland- noch Bundesland-weite einheitliche β-Schätzung durchgeführt.

$$\Pi^{-1} = \text{diag}\left(\frac{1}{\pi_1}, \ldots, \frac{1}{\pi_N}\right)$$

Die folgende Symbolik soll klarstellen, welche Anschriften der Stichprobe in die β-Schätzung eingehen und welche SMP-Typen betrachtet werden. Im Exponent steht 01, wenn bei der β-Schätzung nur Stichproben-Anschriften aus SMP-Typen 0 oder 1 berücksichtigt werden. Ein N als Subskript deutet an, dass zur β Schätzung keine Nullanschriften, E, dass keine unerwünschten Nullanschriften verwendet werden. $s(A)$ statt s bedeutet, dass zur β-Schätzung aus der Stichprobe s die Ausreißer entfernt wurden. Steht ein K im Subskript, bedeutet dies, dass die Stichprobe zur β-Schätzung aus allen Anschriften eines Kreises besteht. Diese Nomenklatur orientiert sich an den Bezeichnungen aus Tabelle **??**.

$s =$ SMP Stichprobe
$s^{01} =$ SMP Stichprobe nur vom Typ 0 und 1
$s_0 =$ SMP Stichprobe ohne jegliche Nullanschriften
$s_0^{01} =$ SMP Stichprobe ohne jegliche Nullanschriften nur vom Typ 0 und 1
$s_1 =$ SMP Stichprobe ohne unerwünschte Nullanschriften
$s_1^{01} =$ SMP Stichprobe ohne unerwünschte Nullanschriften nur vom Typ 0 und 1
$s_K =$ Stichprobe Kreis
$s_K^{01} =$ Stichprobe Kreis nur vom Typ 0 und 1

Alle Schätzungen beziehen sich auf eine Untermenge aller 19.791.513 Mio. Anschriften, die durch die jeweiligen Szenarien definiert sind.

Die Matrix der Hilfsvariablen X besteht aus 2 bzw. 3 Spalten, wobei die erste Spalte immer der 1-er Vektor ist und die zweite Spalte die Anzahl der registrierten Personen an der Hauptwohnung (REG_HW) beinhaltet. Soll das Vorhandensein einer Nullanschrift als Dummyvariable berücksichtigt werden, enthält X eine dritte Spalte, einen logischen Vektor, der den Wert 1 annimmt, wenn die Anschrift eine Nullanschrift (REG_HW + REG_NW = 0) ist und 0 sonst. Für die Szenarien mit dem Label „nur erwünschte“unter Nullanschriften, wird als Dummy-Variable der logische Vektor verwendet mit dem Wert 1, genau dann, wenn gilt:

REG_HW + REG_NW + FEB_HW + FEB_NW = 0.

Dieser Vektor ist in der Praxis allerdings nur für die Stichprobenanschriften bekannt.

$$\hat{\beta}_s = (X_s' \Pi_s^{-1} V_s^{-1} X_s)^{-1} X_s' \Pi_s^{-1} V_s^{-1} Y_s \tag{3.1}$$

Für V verwenden wir hier die Identitätsmatrix.[7] Daraus ergeben sich die Formeln der GREG-Schätzer inklusive der Regressionsparameterschätzung für die 24 Szenarien aus Tabelle 3. s bezieht sich stets auf eine konkrete Stichprobe aus einem SMP, außer im Falle s_k, wo die Stichprobe eines Kreises gemeint ist. τ_X ist zwei- oder dreidimensional, wobei die erste Komponente die Populationszahl der Anschriften im SMP, die zweite Komponente die Summe der registrierten Einwohner am Hauptwohnsitz im SMP ist. Eine dritte Komponente tritt nur auf, wenn die Nullanschriften in einer eigenen Dummy-Variable definiert sind und bezeichnet die Zahl der Nullanschriften bzw. der erwünschten Nullanschriften. $\hat{\tau}_{z,s}$ ist der Horvitz-Thompson Schätzer für die gesuchte Zensuszahl τ_z im SMP. $\hat{\tau}_{x,s}$ ist der Vektor der Horvitz-Thompson Schätzer für

[7] Definiert man $w_i = \frac{1}{\pi_1}$ erhält man gerade die Formel 5.2.

die zwei oder drei Hilfsmerkmale. In den Szenarien 1 bis 6 sind für einen SMP vom Typ 0 oder 1 bei gegebener Stichprobe $s = s^{0,1}$ die Horvitz-Thompson Schätzer berechnet.

Durch Differenzbildung von GREG-Schätzern aus verschiedenen Szenarien können weitere Punktschätzungen berechnet werden. Zum Beispiel wird durch $\hat{\tau}_{z,1}^{\text{GREG}} - \hat{\tau}_{z,3}^{\text{GREG}}$ die Zensus-Zahl in den unerwünschten Nullanschriften berechnet.

3.2 Varianzschätzung

Die für den Auftraggeber relevante Größe des relativen Standardfehlers RSE in einem SMP wird berechnet durch

$$\widehat{\text{RSE}} = \frac{\sqrt{\hat{V}}}{\tau_z} = \frac{\sqrt{\sum_{h=1}^{H} N_h^2 \frac{s_{e,h}^2}{n_h}\left(1 - \frac{n_h}{N_h}\right)}}{\tau_z} \quad ,$$

wobei $\hat{V}$ der geschichtete Residualvarianzschätzer ist und τ_z den Umfang der amtlichen Einwohnerzahl bezeichnet (ohne Sonderanschriften, etc.). $s_{e,h}^2$ ist die Stichprobenvarianz der Residuen in Schicht h des betrachteten SMP.

Die folgenden Tabellen geben einen Überblick über die SMPs von Typ 0 bzw. 1 mit den jeweils drei größten relativen Standardfehlern getrennt bei Berücksichtigung von einer bzw. zwei Hilfsvariablen.

Aus Tabelle 5 ist ersichtlich, dass innerhalb der Szenariengruppen 13-18 bzw. 19-24 die Reihenfolge der SMPs mit den drei größten RSEs sich nicht ändert. Für die Szenarien 13-18 stammen zwei der drei SMPs mit den größten RSEs aus NRW, die dritte aus Bayern. Genauer handelt es sich um Stadtteile von Köln, Dortmund sowie von Nürnberg. Für die Szenarien 19-24 stammen zwei der drei SMPs mit den größten RSEs aus Hamburg, die dritte aus NRW. Genauer handelt es sich um Stadtteile von Hamburg bzw. von Köln.

Darüber hinaus kann man erkennen, wenn man die RSEs zwischen den Szenariengruppen 13-18 und 19-24 miteinander vergleicht, dass eine Schätzung der βs auf Kreis-, das heißt in diesem Falle auf Großstadt-Ebene unter Einbeziehung aller SMPs des Kreises bzw. der Großstadt zu einer Verschlechterung des Varianzschätzers führt. Daher sollte die SMP-separate Schätzung vorgezogen werden, d.h. Szenarien 13-18 Berücksichtigung finden.

Die in Tabelle 6 dargestellten Ergebnisse unterscheiden sich nur marginal von den in vorigen Tabelle. Durch die Hinzunahme eines Nullanschriften-Dummys bei der β-Schätzung verringern sich einige RSEs geringfügig. Dadurch behalten auch alle weiter oben gemachten Aussagen ihre Gültigkeit.

Werden nun SMPs des Typs 1 betrachtet, so lässt sich zunächst feststellen, dass in den Szenarien 13-18 die SMP mit dem jeweils größten RSE stets aus Baden-Württemberg kommt und die SMP-Nr. 083155001006 hat. Es handelt sich dabei um Bad Krozingen (Stadt). Darüber hinaus finden sich die SMPs mit den Nummern 084360010010, 064310002002 und 055580028028. Dabei handelt es sich um Bad Wurzach (Stadt), Bensheim (Stadt), und Nordkirchen. Insgesamt liegen die drei größten RSEs für SMPs des Typs 1 höher als für SMPs des Typs 0.

Auffällig ist, dass eine Ausreißerbereinigung zu einer massiven Erhöhung der größten RSEs bei den SMP-Typen 1 führt. Eine Nichtbeachtung von Ausreißern bei der β-Schätzung scheint daher nicht empfehlenswert.

Darüber hinaus kann man erkennen, wenn man die RSEs zwischen den Szenariengruppen 13-18 und 19-24 miteinander vergleicht, dass eine Schätzung der βs auf Kreis-Ebene unter Einbeziehung aller SMPs des Kreises zu einer Verschlechterung des Varianzschätzers führt. Daher sollte die β-Schätzung stets auf der korrespondierenden Schätz-Ebene erfolgen, d.h. nur Szenarien 13-18 berücksichtigt werden.

Wie bereits bei SMPs des Typs 0 zeigt sich auch hier, dass eine Hinzunahme eines Nullanschriften-Dummys bei der β-Schätzung keinen eindeutigen Effekt auf die Größe der RSEs hat.

Aus den obigen Betrachtungen könnte abgeleitet werden, dass hinsichtlich des relativen Standardfehlers (RSE) nach wie vor eine SMP-separate Schätzung ohne Ausreißerbereinigung die beste Wahl ist.

Die folgende Abbildung illustriert beispielhaft einen möglichen Grund für den hohen RSE in der SMP mit der Nummer 083155001006. Hier ist zu erkennen, dass in den Schichten 6 und 8 starke Unterschiede zwischen τ_Z und REG_HW existieren. Insbesondere gibt es eine Anschrift aus Schicht 6, in der 900 Personen registriert sind, jedoch gleichzeitig 813 Karteileichen vorliegen. In diesem Fall ist die Regressionsgerade eine sehr schlechte Approximation für die τ_Z.

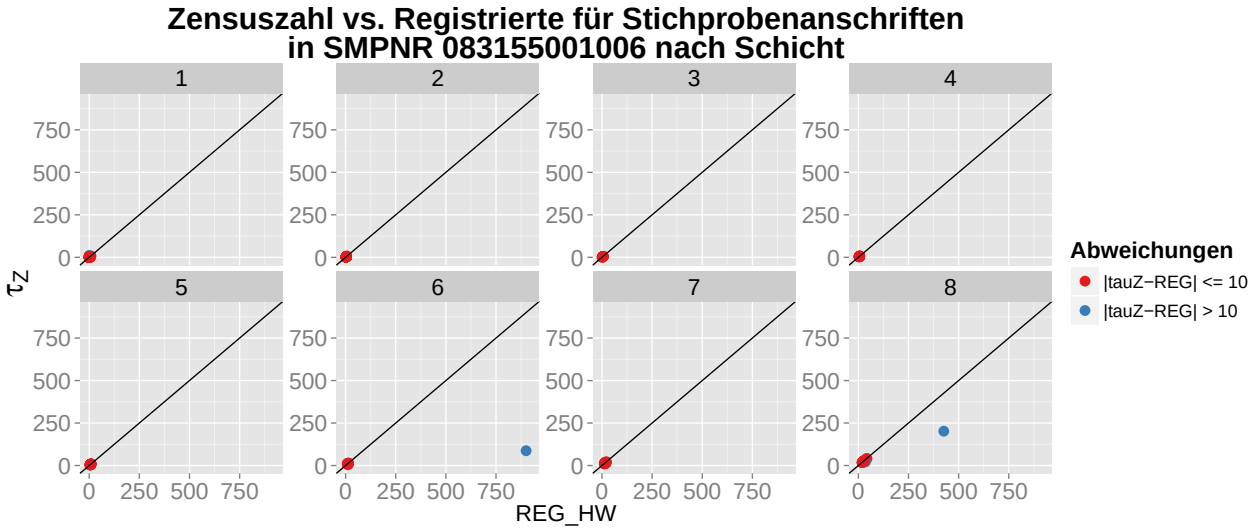

Abbildung 1: Scatterplot von REG_HW nach τ_Z für die SMPNR 083155001006.

Tabelle 4: Formeln für die GREG-Schätzer nach Szenarien

Szenario 1	$\hat{\tau}_{z,1}^{\mathrm{GREG}} = \hat{\tau}_{z,s}^{\mathrm{HT}} + \hat{\beta}'_{s^{01}} \left(\tau_X - \hat{\tau}_{x,s}^{\mathrm{HT}}\right)$
Szenario 2	$\hat{\tau}_{z,2}^{\mathrm{GREG}} = \hat{\tau}_{z,s}^{\mathrm{HT}} + \hat{\beta}'_{s(A)^{01}} \left(\tau_X - \hat{\tau}_{x,s}^{\mathrm{HT}}\right)$
Szenario 3	$\hat{\tau}_{z,3}^{\mathrm{GREG}} = \hat{\tau}_{z,s}^{\mathrm{HT}} + \hat{\beta}'_{s_E^{01}} \left(\tau_X - \hat{\tau}_{x,s}^{\mathrm{HT}}\right)$
Szenario 4	$\hat{\tau}_{z,4}^{\mathrm{GREG}} = \hat{\tau}_{z,s}^{\mathrm{HT}} + \hat{\beta}'_{s(A)_E^{01}} \left(\tau_X - \hat{\tau}_{x,s}^{\mathrm{HT}}\right)$
Szenario 5	$\hat{\tau}_{z,5}^{\mathrm{GREG}} = \hat{\tau}_{z,s}^{\mathrm{HT}} + \hat{\beta}'_{s_N^{01}} \left(\tau_X - \hat{\tau}_{x,s}^{\mathrm{HT}}\right)$
Szenario 6	$\hat{\tau}_{z,6}^{\mathrm{GREG}} = \hat{\tau}_{z,s}^{\mathrm{HT}} + \hat{\beta}'_{s(A)_N^{01}} \left(\tau_X - \hat{\tau}_{x,s}^{\mathrm{HT}}\right)$
Szenario 7	$\hat{\tau}_{z,7}^{\mathrm{GREG}} = \hat{\tau}_{z,s}^{\mathrm{HT}} + \hat{\beta}'_{s_K^{01}} \left(\tau_X - \hat{\tau}_{x,s}^{\mathrm{HT}}\right)$
Szenario 8	$\hat{\tau}_{z,8}^{\mathrm{GREG}} = \hat{\tau}_{z,s}^{\mathrm{HT}} + \hat{\beta}'_{s(A)_K^{01}} \left(\tau_X - \hat{\tau}_{x,s}^{\mathrm{HT}}\right)$
Szenario 9	$\hat{\tau}_{z,9}^{\mathrm{GREG}} = \hat{\tau}_{z,s}^{\mathrm{HT}} + \hat{\beta}'_{s_K^{01}} \left(\tau_X - \hat{\tau}_{x,s}^{\mathrm{HT}}\right)$
Szenario 10	$\hat{\tau}_{z,10}^{\mathrm{GREG}} = \hat{\tau}_{z,s}^{\mathrm{HT}} + \hat{\beta}'_{s(A)_K^{01}} \left(\tau_X - \hat{\tau}_{x,s}^{\mathrm{HT}}\right)$
Szenario 11	$\hat{\tau}_{z,11}^{\mathrm{GREG}} = \hat{\tau}_{z,s}^{\mathrm{HT}} + \hat{\beta}'_{s_K^{01}} \left(\tau_X - \hat{\tau}_{x,s}^{\mathrm{HT}}\right)$
Szenario 12	$\hat{\tau}_{z,12}^{\mathrm{GREG}} = \hat{\tau}_{z,s}^{\mathrm{HT}} + \hat{\beta}'_{s(A)_K^{01}} \left(\tau_X - \hat{\tau}_{x,s}^{\mathrm{HT}}\right)$
Szenario 13	$\hat{\tau}_{z,13}^{\mathrm{GREG}} = \hat{\tau}_{z,s}^{\mathrm{HT}} + \hat{\beta}'_{s} \left(\tau_X - \hat{\tau}_{x,s}^{\mathrm{HT}}\right)$
Szenario 14	$\hat{\tau}_{z,14}^{\mathrm{GREG}} = \hat{\tau}_{z,s}^{\mathrm{HT}} + \hat{\beta}'_{s(A)} \left(\tau_X - \hat{\tau}_{x,s}^{\mathrm{HT}}\right)$
Szenario 15	$\hat{\tau}_{z,15}^{\mathrm{GREG}} = \hat{\tau}_{z,s}^{\mathrm{HT}} + \hat{\beta}'_{s_E} \left(\tau_X - \hat{\tau}_{x,s}^{\mathrm{HT}}\right)$
Szenario 16	$\hat{\tau}_{z,16}^{\mathrm{GREG}} = \hat{\tau}_{z,s}^{\mathrm{HT}} + \hat{\beta}'_{s(A)_E} \left(\tau_X - \hat{\tau}_{x,s}^{\mathrm{HT}}\right)$
Szenario 17	$\hat{\tau}_{z,17}^{\mathrm{GREG}} = \hat{\tau}_{z,s}^{\mathrm{HT}} + \hat{\beta}'_{s_N} \left(\tau_X - \hat{\tau}_{x,s}^{\mathrm{HT}}\right)$
Szenario 18	$\hat{\tau}_{z,18}^{\mathrm{GREG}} = \hat{\tau}_{z,s}^{\mathrm{HT}} + \hat{\beta}'_{s(A)_N} \left(\tau_X - \hat{\tau}_{x,s}^{\mathrm{HT}}\right)$
Szenario 19	$\hat{\tau}_{z,19}^{\mathrm{GREG}} = \hat{\tau}_{z,s}^{\mathrm{HT}} + \hat{\beta}'_{s_K} \left(\tau_X - \hat{\tau}_{x,s}^{\mathrm{HT}}\right)$
Szenario 20	$\hat{\tau}_{z,20}^{\mathrm{GREG}} = \hat{\tau}_{z,s}^{\mathrm{HT}} + \hat{\beta}'_{s(A)_K} \left(\tau_X - \hat{\tau}_{x,s}^{\mathrm{HT}}\right)$
Szenario 21	$\hat{\tau}_{z,21}^{\mathrm{GREG}} = \hat{\tau}_{z,s}^{\mathrm{HT}} + \hat{\beta}'_{s_K} \left(\tau_X - \hat{\tau}_{x,s}^{\mathrm{HT}}\right)$
Szenario 22	$\hat{\tau}_{z,22}^{\mathrm{GREG}} = \hat{\tau}_{z,s}^{\mathrm{HT}} + \hat{\beta}'_{s(A)_K} \left(\tau_X - \hat{\tau}_{x,s}^{\mathrm{HT}}\right)$
Szenario 23	$\hat{\tau}_{z,23}^{\mathrm{GREG}} = \hat{\tau}_{z,s}^{\mathrm{HT}} + \hat{\beta}'_{s_K} \left(\tau_X - \hat{\tau}_{x,s}^{\mathrm{HT}}\right)$
Szenario 24	$\hat{\tau}_{z,24}^{\mathrm{GREG}} = \hat{\tau}_{z,s}^{\mathrm{HT}} + \hat{\beta}'_{s(A)_K} \left(\tau_X - \hat{\tau}_{x,s}^{\mathrm{HT}}\right)$

Tabelle 5: Überblick über die SMPs von Typ 0 mit den drei größten relativen Standardfehlern – Hilfsvariable: REG_HW

Szenario	SMPNR1	SMPNR2	SMPNR3	RSE1	RSE2	RSE3
13	05315000000004	05913000000001	09564000000001	0.0151	0.0122	0.0100
14	05315000000004	05913000000001	09564000000001	0.0151	0.0125	0.0103
15	05315000000004	05913000000001	09564000000001	0.0150	0.0122	0.0099
16	05315000000004	05913000000001	09564000000001	0.0151	0.0125	0.0103
17	05315000000004	05913000000001	09564000000001	0.0151	0.0122	0.0100
18	05315000000004	05913000000001	09564000000001	0.0151	0.0125	0.0103
19	02000000000006	02000000000007	05315000000004	0.0336	0.0265	0.0260
20	02000000000006	02000000000007	05315000000004	0.0341	0.0270	0.0266
21	02000000000006	02000000000007	05315000000004	0.0336	0.0265	0.0260
22	02000000000006	02000000000007	05315000000004	0.0341	0.0270	0.0266
23	02000000000006	02000000000007	05315000000004	0.0336	0.0265	0.0260
24	02000000000006	02000000000007	05315000000004	0.0342	0.0270	0.0266

Tabelle 6: Überblick über die SMPs von Typ 0 mit den drei größten relativen Standardfehlern – Hilfsvariable: REG_HW + NAD

Szenario	SMPNR1	SMPNR2	SMPNR3	RSE1	RSE2	RSE3
13	05315000000004	05913000000001	09564000000001	0.0150	0.0122	0.0100
14	05315000000004	05913000000001	09564000000001	0.0151	0.0125	0.0103
15	05315000000004	05913000000001	05111000000003	0.0183	0.0151	0.0135
16	05315000000004	05913000000001	09564000000001	0.0150	0.0125	0.0103
17	05315000000004	05913000000001	09564000000001	0.0151	0.0122	0.0100
18	05315000000004	05913000000001	09564000000001	0.0151	0.0125	0.0103
19	02000000000006	02000000000007	05315000000004	0.0335	0.0265	0.0260
20	02000000000006	02000000000007	05315000000004	0.0341	0.0269	0.0266
21	02000000000006	02000000000007	05315000000004	0.0379	0.0300	0.0286
22	02000000000006	02000000000007	05315000000004	0.0353	0.0279	0.0265
23	02000000000006	02000000000007	05315000000004	0.0336	0.0265	0.0260
24	02000000000006	02000000000007	05315000000004	0.0342	0.0270	0.0266

Tabelle 7: Überblick über die SMPs von Typ 1 mit den drei größten relativen Standardfehlern – Hilfsvariable: REG_HW

Szenario	SMPNR1	SMPNR2	SMPNR3	RSE1	RSE2	RSE3
13	083155001006	064310002002	084360010010	0.0211	0.0190	0.0188
14	083155001006	084360010010	055580028028	0.2630	0.0451	0.0311
15	083155001006	064310002002	084360010010	0.0209	0.0190	0.0180
16	083155001006	084360010010	055580028028	0.2629	0.0450	0.0311
17	083155001006	064310002002	084360010010	0.0209	0.0190	0.0180
18	083155001006	084360010010	055580028028	0.2630	0.0454	0.0314
19	083155005047	083150076076	083155017113	0.1814	0.1766	0.1759
20	083155005047	083150076076	083155017113	0.7446	0.7248	0.7218
21	083155005047	083150076076	083155017113	0.1761	0.1714	0.1707
22	083155005047	083150076076	083155017113	0.7428	0.7230	0.7201
23	083155005047	083150076076	083155017113	0.1759	0.1712	0.1705
24	083155005047	083150076076	083155017113	0.7484	0.7285	0.7255

Tabelle 8: Überblick über die SMPs von Typ 1 mit den drei größten relativen Standardfehlern – Hilfsvariable: REG_HW + NAD

Szenario	SMPNR1	SMPNR2	SMPNR3	RSE1	RSE2	RSE3
13	083155001006	064310002002	084360010010	0.0209	0.0190	0.0178
14	083155001006	084360010010	055580028028	0.2630	0.0454	0.0314
15	064310002002	083155001006	059700036036	0.0236	0.0211	0.0194
16	083155001006	084360010010	055580028028	0.2630	0.0455	0.0315
17	083155001006	064310002002	084360010010	0.0209	0.0190	0.0180
18	083155001006	084360010010	055580028028	0.2630	0.0454	0.0314
19	083155005047	083150076076	083155017113	0.1765	0.1718	0.1711
20	083155005047	083150076076	083155017113	0.7484	0.7285	0.7255
21	083155005047	083150076076	083155017113	0.1753	0.1706	0.1699
22	083155005047	083150076076	083155017113	0.7486	0.7287	0.7257
23	083155005047	083150076076	083155017113	0.1759	0.1712	0.1705
24	083155005047	083150076076	083155017113	0.7484	0.7285	0.7255

4 Aufbau und Ergebnisse der Stichprobenanalyse

In diesem Kapitel werden die Ergebnisse von Analysen der Stichprobendaten beschrieben. Alle vorgestellten Ergebnisse und Befunde basieren ausschließlich auf Schätzungen der realisierten Stichprobe. Grundsätzlich werden Analysen der Punkt- und Varianzschätzung des GREG sowie der βs aus der Regressionsparameter-Schätzung vorgestellt.

Viele der Analysen weisen Ergebnisse getrennt für die in Tabelle 3 beschriebenen Szenarien aus. Dies ermöglicht eine Bewertung verschiedener Vorgehensweisen etwa bei der Ausreißerbereinigung.

Des weiteren wird unterschieden, welche Hilfsvariablen zur Schätzung verwendet werden. Entweder wird nur die Zahl der registrierten Personen in einer Anschrift oder zusätzlich noch eine Dummyvariable als Indikator für das Vorhandensein einer Nullanschrift an einer Anschrift verwendet. Das heißt, wir betrachten[8]

$$Y_i = \beta_1 + \beta_2 \cdot \text{REG\_HW}_i + e_i \tag{4.1}$$

bzw.

$$Y_i = \beta_1 + \beta_2 \cdot \text{REG\_HW}_i + \beta_3 \cdot \delta_i + e_i \quad , \tag{4.2}$$

wobei

$$\delta_i = \begin{cases} 1 & \text{falls } i\text{-te Anschrift Nullanschrift bzw. erwünschte Nullanschrift} \\ 0 & \text{sonst} \end{cases}$$

und e_i den Fehlerterm bezeichnen. β_1 ist die additive Konstante (Intercept), das heißt, der zum 1-er Vektor zählende Regressionsparameter. β_2 bezeichnet die Steigung (Slope) des Anschriftengrößenparameters und β_3 den Regressionsparameter der Dummy-Variable zu den Nullanschriften. Da die Fehlbestände für die Stichprobenanschriften erst nach der Erhebung bekannt sind kennt man $\sum_{i=1}^{N} \delta_i$ nicht, sofern δ_i als Indikator für erwünschte Nullanschriften steht. In der Gesamtheit ist nur bekannt, ob es sich um eine Nullanschrift handelt oder nicht. Daher kann diese Information auch nur in die Schätzung von β einfließen. Wir werden im Folgenden aber unterstellen, dass bei Verwendung einer zusätzlichen Dummy-Variablen auch der δ-Vektor der erwünschten Nullanschriften in der Gesamtheit bekannt sei.

In Abbildung 2 sind die β-Schätzungen in Szenario 1 für alle 1 619 SMPs vom Typ 0 und 1 zu sehen. Die Ergebnisse der Schätzungen für $\hat{\beta}_1$ sind dunkelbraun, die Schätzungen für $\hat{\beta}_2$ hellbraun, und die Schätzungen für $\hat{\beta}_3$ türkis eingefärbt. Es ist zu sehen, dass die Unterschiede zwischen den $\hat{\beta}_1$-Schätzwerten deutlich größer als zwischen den β_2-Schätzwerten. Das bedeutet, dass die Steigungen in jeder SMP nahezu identisch groß sind und nahe bei 1 liegen. Die $\hat{\beta}_3$-Schätzungen streuen um die Null, weisen jedoch eine größere Variabilität zwischen den SMPs auf als die $\hat{\beta}_2$-Schätzungen.

[8] In unserem speziellen Fall entsprechen den Y_i den Zensuszahlen Z_i in Anschrift i.

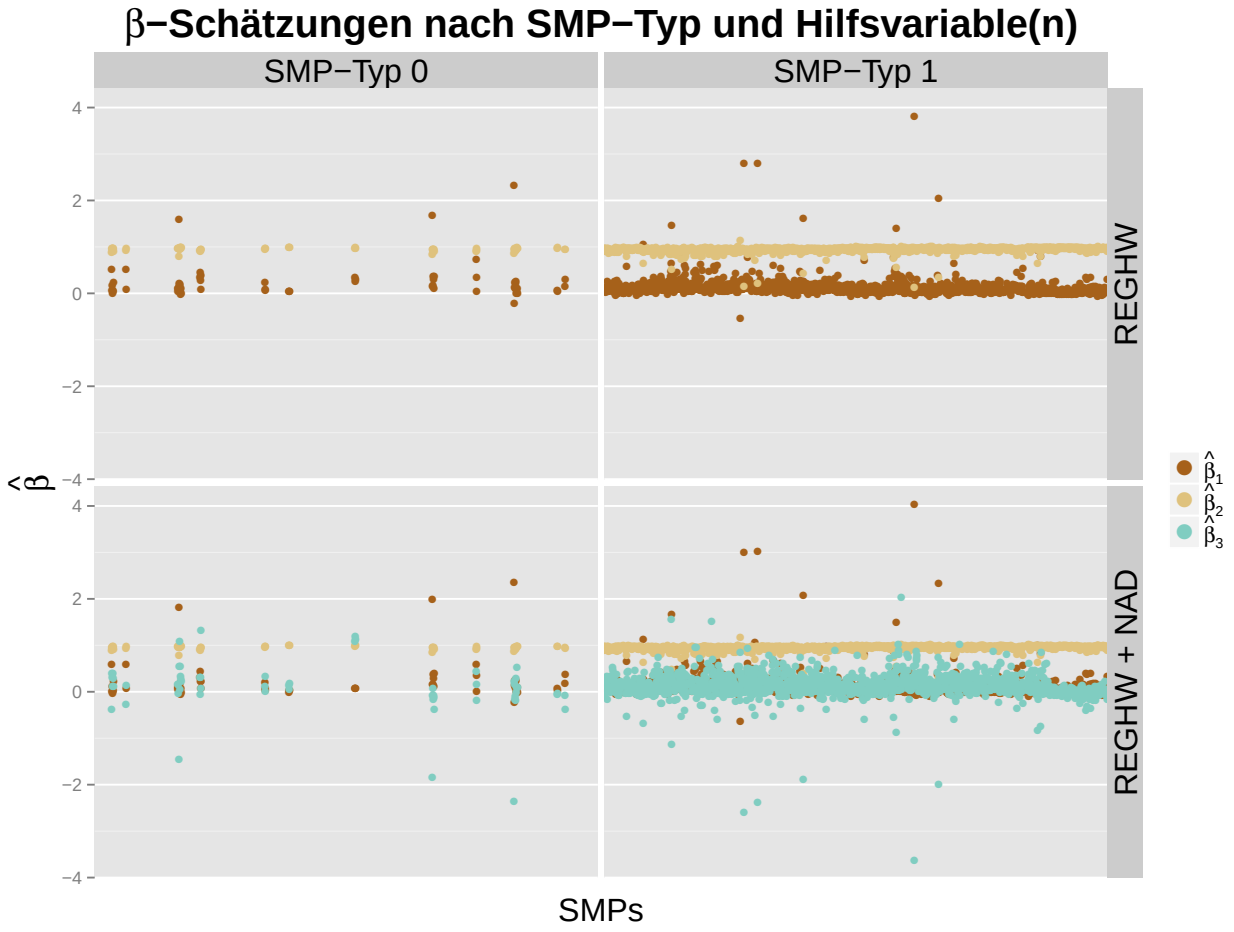

Abbildung 2: $\hat{\beta}_1$, $\hat{\beta}_2$ und $\hat{\beta}_3$; KAL/FEB Modell 3/3

In Abbildung 3 sind die über alle SMPs eines Bundeslandes gemittelten $\hat{\beta}_1$ (Intercept) zu sehen. Diese sind in Berlin (11) und Hamburg (02) am größten. Die Kennungen für die Bundesländer sind in Tabelle 11 gegeben.

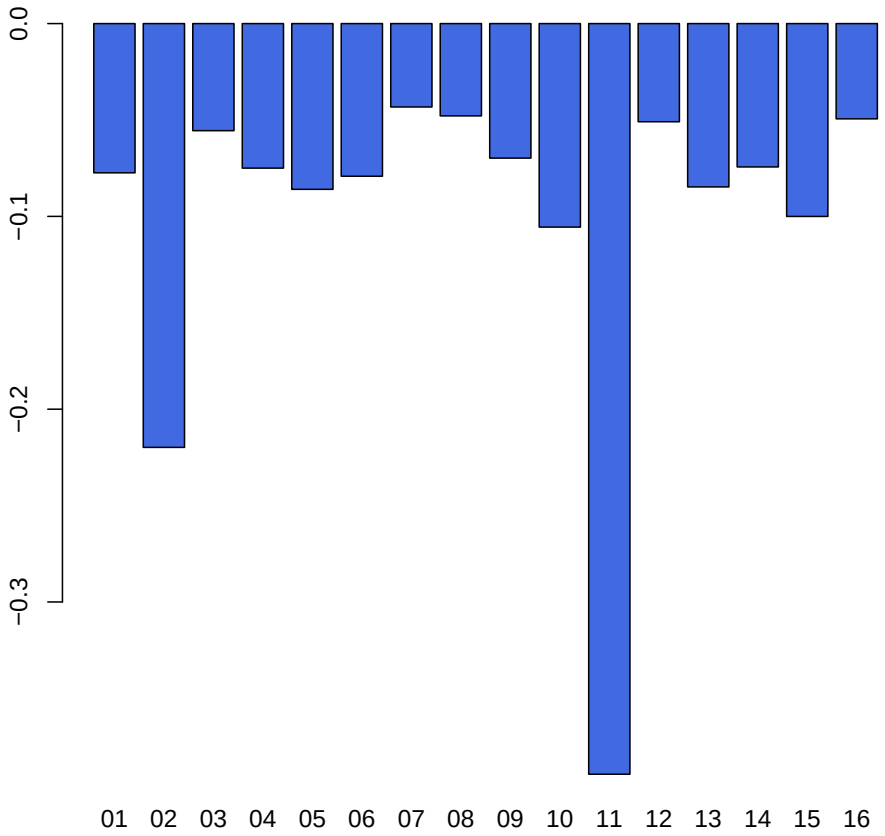

Abbildung 3: Mittelwerte von $\hat{\beta}_1$ nach Bundesländern; Szenario 1; KAL/FEB Modell 1/1

4.1 Berücksichtigung der Nullanschriften

Eines der Hauptziele des Projekts ist es, den Einfluss der Nullanschriften zu untersuchen. Dabei kommen verschiedene Szenarien infrage, die in Abschnitt 2.4 vorgestellt wurden.

In Abbildung 4 ist ein Vergleich zwischen zwei Szenarien realisiert. In beiden Szenarien fand keine Ausreißerbereinigung statt, es wurde auf Ebene der SMP geschätzt und nur die SMP-Typen 0 und 1 berücksichtigt. Der Unterschied besteht darin, dass in Szenario 5 die Nullanschriften unberücksichtigt bleiben, während sie bei Szenario 1 in die Berechnungen einfließen. Szenario 5 beschreibt die bisher im Zensus-Forschungsprojekt betrachtete Zensus-Situation mit der realisierten Allokation.

Dargestellt wird jeweils der relative Bias, also die Differenz zwischen τ_Z und dem GREG-Schätzwert geteilt durch τ_Z. Der jeweilige GREG-Schätzer kann Tabelle 4 entnommen werden. Der Wert für die Abszisse ξ, für den SMP k und das Szenario 1 wird beispielsweise folgendermaßen berechnet:

$$\frac{\tau_{Z,k} - \hat{\tau}^{\text{GREG}}_{z,1,k}}{\tau_{Z,k}}$$

Den entsprechende Wert der Ordinate, für SMP k und Szenario 5 berechnet man aus:

$$\frac{\tau_{Z,k} - \hat{\tau}^{\text{GREG}}_{z,5,k}}{\tau_{Z,k}}$$

Wenn sich die Schätzergebnisse bei den beiden Szenarien nicht unterscheiden, liegen alle Punkte auf der Winkelhalbierenden. Der Punkt ist rot eingefärbt, wenn es sich um einen SMP vom Typ 1 handelt und ist blau eingefärbt, wenn es ein SMP von Typ 0 ist. Die Größe des Punkts richtet sich nach dem mittleren relativen Bias $= \frac{1}{2}$(Abszissenwert + Ordinatenwert).

Im Panel mit der Überschrift REGHW + NAD wird die Schätzung unter Einbeziehung der zusätzlichen Nullanschriften-Dummy-Variablen (NAD) verstanden.

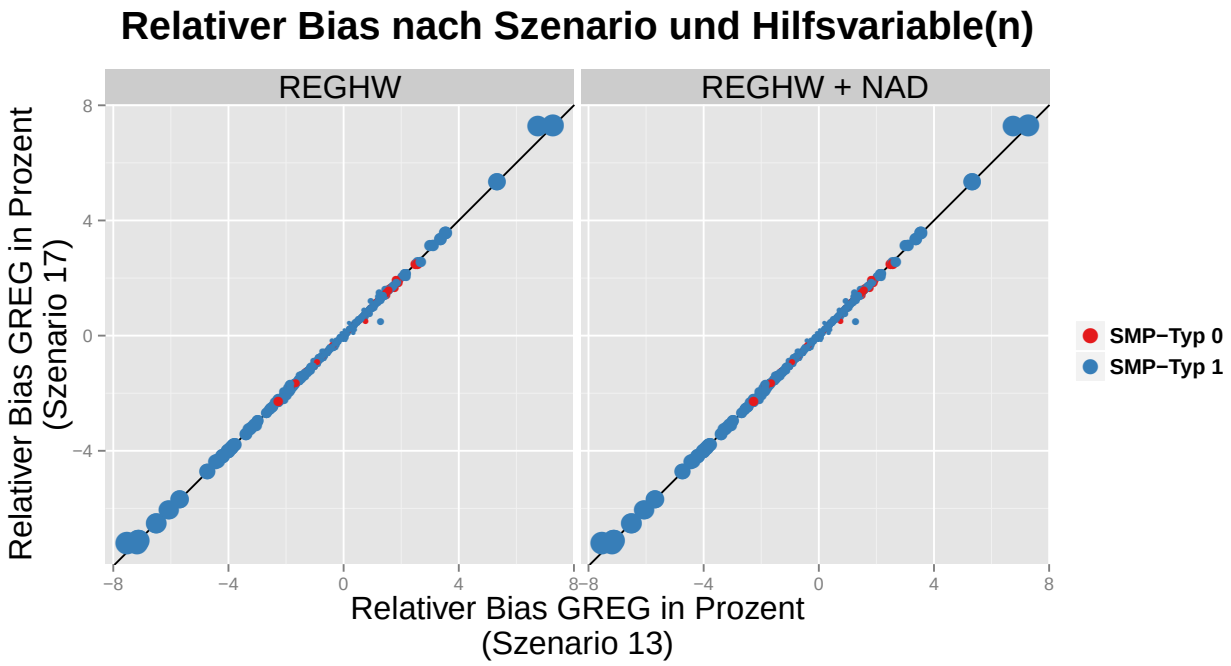

Abbildung 4: Bias; Vergleich von Szenario 13 mit Szenario 17; KAL/FEB Modell 3/3

Nicht alle Punkte liegen auf der Winkelhalbierenden. Die Schätzergebnisse der beiden Szenarien unterscheiden sich also. Allerdings sind keine systematischen Muster zu erkennen.

Man kann eruieren, dass ein besonders hoher absoluter Bias (τ_Z ist deutlich größer als der GREG-Schätzwert) vor allem in großen SMPs auftritt.

4.2 Ausreißerbereinigung

In Abbildung 5 sind die $\hat{\beta}_1$ für zwei Szenarien und die 54 Sampling-Points vom SMP-Typ 1 (SMP-Typ 0 nicht vorhanden) aus Schleswig-Holstein abgebildet. Die rote Linie bildet die geschätzten $\hat{\beta}_1$ für alle Sampling-Points im Szenario 1 ab. Dabei wurde nach der Größe des geschätzten Parameters sortiert. Der kleinste Schätzwert ist auf der linken Seite zu finden, während der größte Schätzwert auf der rechten Seite abgebildet ist. Die türkisfarbene Linie bildet die geschätzten Parameter für $\hat{\beta}_1$ im Szenario 2 ab. Dieses Szenario unterscheidet sich von dem ersten Szenario darin, dass eine robuste Schätzung durchgeführt wurde, das heißt, dass eine Ausreißerbereinigung stattgefunden hat.

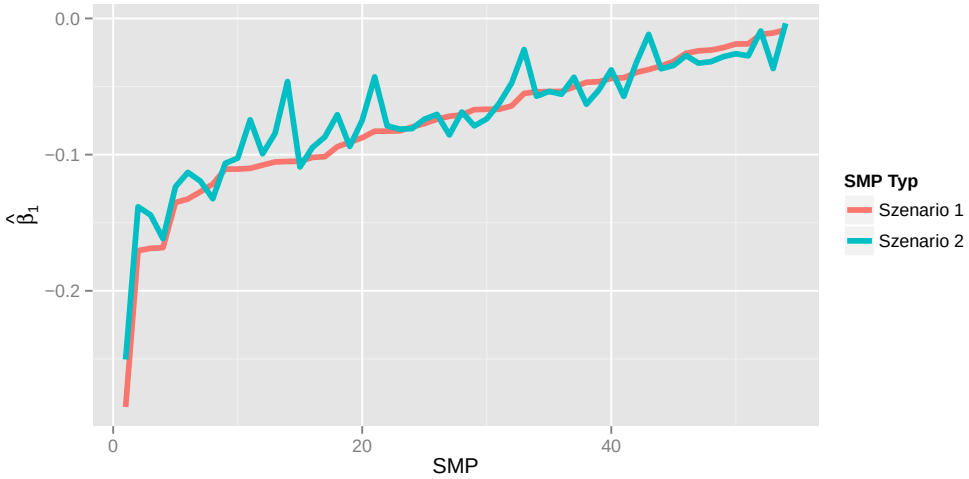

Abbildung 5: Verteilung der $\hat{\beta}_1$ in den 54 SMPs aus Schleswig-Holstein für Szenarien 1 und 2; KAL/FEB Modell 1/1

Abbildung 5 legt einerseits nahe, dass die Unterschiede in Bezug auf die $\hat{\beta}_1$-Schätzungen (Intercept) zwischen den SMPs deutlich bedeutsamer sind als die Unterschiede zwischen den Szenarien 1 und 2. Das drückt sich darin aus, dass die beiden Kurven über die x-Achse relativ nahe beieinander liegen. Dies ist ein Hinweis darauf, dass der Unterschied zwischen den Szenarios von relativ kleiner Dimension ist. Andererseits kann man sehen, dass durch eine Ausreißerbereinigung die $\hat{\beta}_1$-Schätzungen in der Tendenz eher größer ausfallen als ohne Ausreißerbereinigung.

Bei der Abbildung 6 handelt es sich um eine Visualisierung des Vergleichs von zwei Szenarien. Das Szenario 1 wird mit dem Szenario 2 verglichen. In beiden Szenarien werden alle Nullanschriften berücksichtigt, geschätzt wird jeweils auf SMP Ebene und es werden bei der Schätzung der β nur die SMP-Typen 0 und 1 berücksichtigt. Der einzige Unterschied besteht darin, dass in Szenario 1 keine Ausreißer bereinigt werden, während dies in Szenario 2 der Fall ist. Verglichen wird jeweils die Differenz zwischen den τ_Z und den GREG Schätzwerten. Wenn aus der Differenz eine negative Zahl resultiert, dann ist für diesen SMP τ_Z größer, wenn eine positive Zahl resultiert, dann ist der GREG-Schätzer größer. Auf der x-Achse ist diese Differenz für Szenario 2 abgetragen, während auf der y-Achse die Differenz für Szenario 1 abgetragen ist. Da die Punkte alle auf der

Winkelhalbierenden liegen, ist, was die Differenz zwischen GREG-Schätzwert und τ_Z anbelangt kein großer Unterschied zwischen den beiden Szenarien zu erkennen. Vermutlich wurden also zu wenige Ausreißer entfernt um die Schätzwerte zu verbessern.

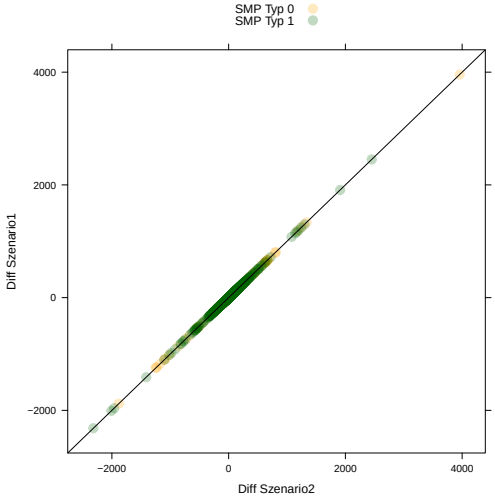

Abbildung 6: xy-Plot der Differenzen zwischen tau.Z und GREG – KAL-Modell 3, FEB Modell 1 – Vergleich von Szenario 1 mit Szenario 2

4.3 Varianzschätzung

Im Folgenden werden die Schätzergebnisse für Szenarien/Schätzer-Kombinationen verglichen. Die einzelnen Panels eines Plots ergeben sich aus der Kombination der fünf Varianzschätzer und jeweils sechs Szenarien. In jedem Panel sind auf der x-Achse die relativen Standardfehler abgetragen, auf der y-Achse τ_z. Ein Punkt bzw. ein Dreieck symbolisiert den $\widehat{\mathrm{SE}}_R$ Schätzwert und den Wert für τ_z in einer SMP. Punkte symbolisieren den GREG-Schätzer mit nur einer Hilfsvariable (`REG_HW`), Dreiecke stehen für den GREG-Schätzer mit zwei Hilfsvariablen (`REG_HW` und `NAD`). Eine Unterscheidung zwischen SMP Typen erfolgt durch farbliche Kodierung (• SMP Typ 0, • SMP Typ 1, • SMP Typ 2, • SMP Typ 3). Eine gestrichelte rote Linie markiert die von Destatis geforderte obere Schranke für den relativen Standardfehler von 0,5%.

In Abbildung 7 werden die Szenarien 13-18 dargestellt (vgl. Tabelle 3), eingeschränkt auf SMPs des Typs 0 und 1 und daher äquivalent zu Szenarien 1-6.

In diesen Szenarien werden die βs auf SMP-Ebene geschätzt und nur SMPs des Typs 0 und 1 betrachtet. Die Szenarien unterscheiden sich darin, ob keine Ausreißerbereinigung stattfindet (Szenarien 13, 15, 17) oder ob 1% der Ausreißer entfernt werden (Szenarien 14, 16, 18). Außerdem wird unterschieden, ob Nullanschriften in die Schätzung mit aufgenommen (Szenarien 13, 14), nur erwünschte Nullanschriften aufgenommen (Szenarien 15, 16) oder generell von der Schätzung ausgeschlossen werden (Szenarien 17, 18).

In Abbildung 7 liegen die meisten Punkte rechts von der gestrichelten roten Linie. Das bedeutet, dass unabhängig von dem betrachteten Szenario und dem verwendeten Varianzschätzer in den meisten Fällen die Präzisionsanforderungen nicht erfüllt werden. Es fällt auf, dass einige Punkte in Szenarien 14, 16 und 18 deutlich weiter rechts liegen als in den anderen Szenarien. Dies hat mit dem Einfluss der Ausreißerbereinigung auf die β-Schätzung zu tun, die sich negativ auf die

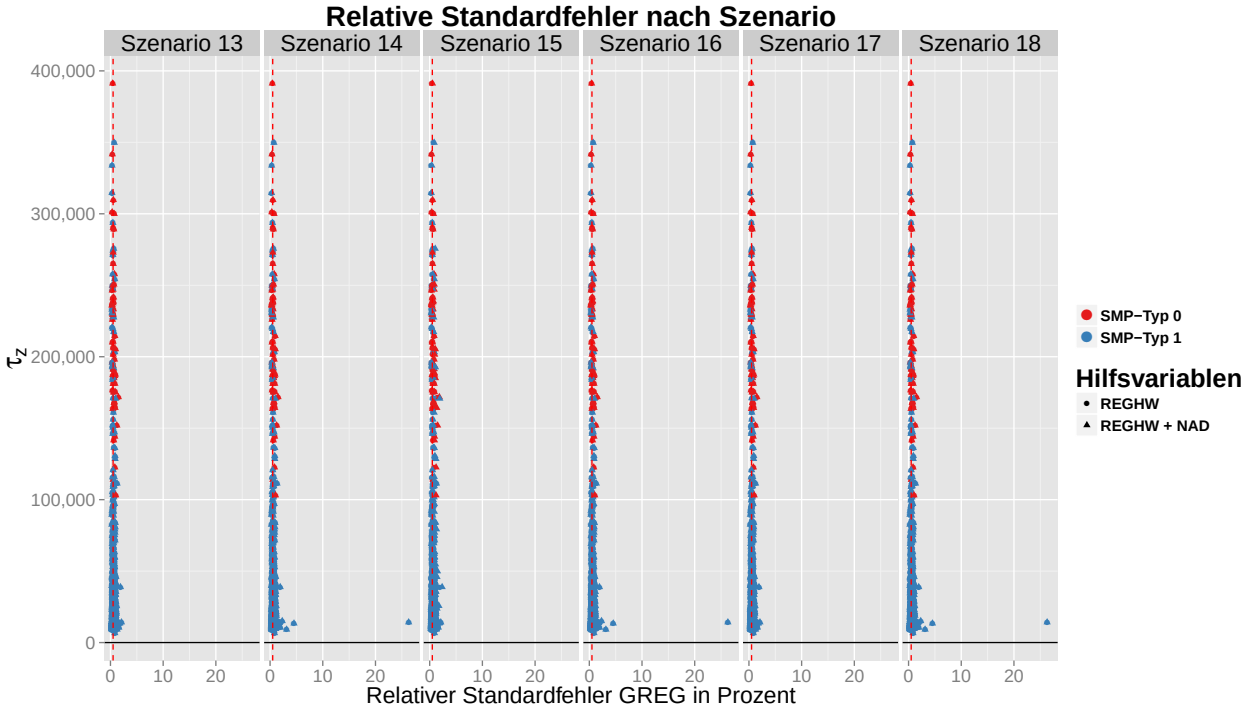

Abbildung 7: Relativer Standardfehler des GREG Schätzers; KAL/FEB Modell 3/3, Szenarien 13-18

Qualität der Schätzung auswirkt. Außerdem ist, wie zu erwarten war, erkenntlich, dass größere SMPs eher kleine Varianzschätzungen aufweisen als kleinere SMPs.

Insgesamt lässt sich festhalten, dass die vom Auftraggeber gestellten Präzisionsanforderungen in allen Szenarien von der Mehrheit der SMPs erfüllt werden. Eine β-Schätzung auf SMP-Ebene erweist sich als besser im Sinne kleiner Standardfehler. Ebenso ist eine Ausreißerbereinigung unter der Maßgabe hoher Präzision zu empfehlen.

5 Aufbau und Ergebnisse der Simulationen

5.1 Allokation

Bisher wurde zu Fragestellung 1 eingehend die realisierte Stichprobe analysiert. Dabei wurde die von DESTATIS errechnete und verwendete Allokation für die Stichprobenziehung verwendet. Es wurden die Umfänge n_h und N_h aus dem im Frühjahr 2012 gelieferten Datensatz *SMPUND-SCHICHT0* übernommen. Dabei musste in einigen Schichten noch eine Anpassung gemacht werden, da es in dem Simulationsdatensatz in zwei Schichten die Anzahl der Anschriften kleiner ist als die Anzahl zu ziehender Elemente. In diesen Schichten wird eine Vollerhebung durchgeführt und somit n_h auf das ausgezählte N_h aus den Daten gesetzt. Innerhalb der Schichten wird eine uneingeschränkte Zufallsauswahl ohne Zurücklegen durchgeführt.

In diesem Kapitel soll die Untersuchung mit Hilfe von Simulationen ergänzt werden, um eine zum Zensus-Stichprobenforschungsprojekt äquivalente Vorgehensweise zu erreichen. Ziel ist es, die im Projekt getätigten Empfehlungen auf mögliche Anpassungen auf Grund der Existenz von

Nullanschriften zu überprüfen.

5.2 Stichprobenziehung

Es werden für alle Designs 1000 Stichproben gezogen. In diesem Teil der Untersuchung handelt es sich jedoch lediglich um das in der Realität verwendete Design.

Design 1: voller Datensatz
Es wird mit den n_h aus der Allokation eine uneingeschränkte Zufallsauswahl ohne Zurücklegen innerhalb der Schichten durchgeführt.

5.3 Schätzer

Für Fragestellung 1 wurden gruppierte und separate GREG Schätzer verwendet, um zu sehen, welcher der GREG-Schätzer mit den Nullen am Besten zurechtkommt.

Beim GREG wird ein statistisches Modell verwendet. Ausgehend vom linearen Regressionsmodell

$$y = X'\beta + \varepsilon \tag{5.1}$$

erhält man unter den üblichen Annahmen für das Regressionsmodell als GREG-Schätzer

$$\widehat{\tau}_{Y,d,\mathrm{GREG}} = \sum_{i=1}^{n_d} w_i y_i + \Big(\sum_{i=1}^{N_d} X_i - \sum_{i=1}^{n_d} w_i x_i \Big)^T \widehat{\beta} \tag{5.2}$$

mit

$$\widehat{\beta} = \left(\sum_{i=1}^{n_d} w_i x_i x_i^T \right)^{-1} \sum_{i=1}^{n_d} w_i x_i y_i \qquad , \tag{5.3}$$

bei gewichteter Schätzung von β und

$$\widehat{\beta} = \left(\sum_{i=1}^{n_d} x_i x_i^T \right)^{-1} \sum_{i=1}^{n_d} x_i y_i \qquad . \tag{5.4}$$

bei ungewichteter Schätzung von β, wobei β die Lösung des linearen Regressionsmodells (5.1) ist. x_i bzw. X_i sind die Hilfsinformationen der i-ten Anschrift in der Stichprobe bzw. in der Grundgesamtheit, welche auch als Vektor vorliegen können. Die Matrix der Hilfsinformationen enthält in der ersten Spalte den 1-Vektor, welcher den Wert des Koeffizienten für den Achsenabschnitt des Regressionsmodells β_1 liefert.

Der Regressionskoeffizient β kann auch basierend auf Gruppen g geschätzt werden (β_g), wobei sich diese Gruppen aus Areas und/oder Schichten zusammensetzen können. Bei der Schätzung auf Gruppen unterscheiden wir folgende Schätzer

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

BLA-SEP Bundesland-separate Schätzung. Es wird über alle Schichten hinweg separat in jedem Bundesland ein eigenes Regressionsmodell betrachtet.

KRS-SEP Kreis-separate Schätzung. Es wird über alle Schichten hinweg separat in jedem Kreis ein eigenes Regressionsmodell betrachtet.

SMP-SEP SMP-separate Schätzung. Es wird über alle Schichten hinweg separat in jedem SMP ein eigenes Regressionsmodell betrachtet.

Der GREG-Schätzer (5.2) kann im Allgemeinen auch durch

$$\widehat{\tau}_{Y,d,\mathrm{GREG}} = \sum_{i=1}^{N_d} x_{i,d}^T \widehat{\beta}_g + \sum_{i=1}^{n_d} w_i \cdot \big(\underbrace{y_{i,d} - x_{i,d}^T \widehat{\beta}_g}_{e_{i,d}} \big) \tag{5.5}$$

beschrieben werden. In Analogie zum Varianzschätzer vom HT erhält man damit für den GREG

$$\widehat{\mathrm{V}}\big(\widehat{\tau}_{d,\mathrm{HT}}\big) = \sum_{h=1}^{H} N_{h,d}^2 \cdot \frac{s_{e,h,d}^2}{n_{h,d}} \cdot \Big(1 - \frac{n_{h,d}}{N_{h,d}}\Big) \quad . \tag{5.6}$$

Hierbei handelt es sich um einen Residualvarianzschätzer, der sowohl für den Fall eines gewichteten β und eines ungewichteten β anwendbar ist, da die resultierenden Residuen verwendet werden.

Die unterschiedlichen Variationen der Schätzung der β und der Hochrechnung können wie folgt strukturiert werden. Aufgrund der Eigenschaften des GREG-Schätzers, kann das β, welches zur Hochrechnung verwendet wird, relativ frei gewählt werden. Es sollte möglichst nah an dem *wahren* β der Grundgesamtheit liegen. Bei Störungen in den Daten, wie zum Beispiel durch Clusterungen in den Ausprägungspaaren (y, x) oder Ausreißern im Allgemeinen, muss überlegt werden, ob das β sich in der einzelnen Stichprobe nicht geeigneter schätzen ließe, als wenn die Störfaktoren unterdrückt werden. Das Vorgehen zur Ermittlung des GREG-Schätzwertes kann also zweigeteilt werden:

Schätzung der β Zunächst muss das Vorgehen zur Schätzung der β präzisiert werden. Die verschiedenen Faktoren, die hierbei berücksichtigt werden müssen sind:

Wahl der Kovariablen Es wurden zwei Kovariablen Modelle untersucht. Zum einen das Modell in dem $X = (1, REG\_HW)$. Dabei ist REG_HW mit der hohen Korrelation zur Zensusanzahl als Kovariable gesetzt. Das andere Modell ist $X = (1, REG\_HW, D)$ wobei D die Dummy-Variable für Nullanschriften ist, d.h. $D_i = \begin{cases} 1 & \mathrm{REGHW}_i + \mathrm{REGNW}_i = 0 \\ 0 & sonst \end{cases}$

Behandlung von Ausreißern Als Ausreißer wurden in der Simulation diejenigen Anschriften identifiziert die eine Cooks-Distanz von über $0,5$ aufweisen. Die Schätzer die bei der Schätzung des β die Ausreißer auslassen, sind mit dem Zusatz **A** im Namen versehen.

Behandlung von Nullanschriften Es wurden einerseits alle Nullanschriften REG_HW + REG_NW=0 in der Schätzung der β verwendet, andererseits ausgelassen. IM Falle der Nichtberücksichtigung ist der Schätzer mit dem Zusatz **N** versehen.

Verwendung von SMP-Typ 0 und 1 Da der Fokus auf der Zensusbevölkerung für die SMP-Typen 0 und 1 liegt, kann argumentiert werden, dass bei der Berechnung von gruppierten GREGs nur Informationen aus den SMP-Typen 0 und 1 einfließen sollen. Falls also bei der Schätzung des β nur Beobachtungen aus den SMP-Typen 0 und 1 verwendet wurden, dann wird der Schätzer mit dem Zusatz **01** gekennzeichnet.

Gewichtung Das β kann entweder gewichtet oder ungewichtet geschätzt werden. Dies hat keinen Einfluss auf die weitere Verwendung der Gewichte im GREG-Schätzer.

Hochrechnung Bei der Hochrechnung mit dem geschätzten β muss darauf geachtet werden, dass die Residuen korrekt berechnet wurden. Die Residuen e_i sind gegeben durch $e_i = y_i - x_i\widehat{\beta}$. Hierbei spielt es keine Rolle auf welcher Basis das β berechnet wurde, es sind stets Residuen für alle in der Stichprobe beobachteten Werte zu berechnen gemäß Formel (5.5). Gleiches gilt für die Residuen, die in Formel (5.6) für die Berechnung von $s^2_{e,h,d}$ verwendet werden.

5.4 Simulationen zur Fragestellung 1

Die Monte-Carlo-Simulation ist im Vorgehen analog zu denjenigen aus dem Zensus-Stichprobenforschungsprojekt. Es wurden 1000 Stichproben gezogen und für die Monte-Carlo-Simulation verwendet. Mit dem Bericht wird ein *RData*-File geliefert, in dem die Maße für die verschiedenen Kombinationsmöglichkeiten herausgelesen werden können. Da die Monte-Carlo-Simulation sehr umfassend ist, werden in diesem Bericht lediglich die aus Sicht des Auftragnehmers wichtigsten Eckpunkte aufgeführt. Dabei liegt der Fokus sehr stark auf den nach Absprache mit dem Auftraggeber am meisten interessierenden Aspekten. Diese umfassen insbesondere die beiden folgenden Fragen:

- Kann weiterhin davon ausgegangen werden, dass SMP-separate GREG-Schätzungen den Genauigkeitsanforderungen genügen?
- Welches der vier oben genannten Szenarien ist am besten geeignet, um die Hochrechnung durchzuführen?

Die Genauigkeitsanforderungen für SMP-Typ 0 und 1 lauten, dass im Mittel die Hälfte der SMPs vom Typ 0 und 1 einen RRMSE von unter $0,5\%$ erreicht. Diese $0,5\%$-Grenze ist in den folgenden Abbildungen durch die rote Linie gekennzeichnet. An die SMP-Typen 2 und 3 ist diese Genauigkeitsanforderung nicht gestellt, die rote Linie wurde hier nur als optische Hilfestellung gezeichnet. In Abbildungen 8 und 9 ist der RRMSE der GREG-Schätzer unter verschiedenen Szenarien in Form von Boxplots dargestellt. Hierbei entsprechen die Namen der Schätzer der oben gegebenen Nomenklatur.

Für die Simulationen der Fragestellung 1 wurden die Karteileichen- Fehlbestandsmodellkombinationen 1/1, 2/2, und 3/3 verwendet. Die Ergebnisse der Schätzung, gegeben der drei verschiedenen Karteileichen- und Fehlbestandsmodelle, sind durchaus sichtbar unterschiedlich. Dabei schneiden die Kombinationen 1/1 und 3/3 ähnlich ab. Beim Modell 2/2 werden niedrigere RRMSEs erzielt, als bei den beiden anderen Modellen. Im Vergleich mit der Stichprobe ist davon auszugehen, dass die Modelle 1/1 und 3/3 näher an der Realität liegen als das Modell 2/2. Aus diesem Grund werden die Ergebnisse im Folgenden nur für das Modell 3/3 betrachtet,

da die anderen Modelle fast gleiche oder bessere Ergebnisse erzielen, und somit eine gewisse Abschätzung nach unten gegeben ist.

5.4.1 Punkt-Schätzung

Es wurden sowohl gewichtete als auch ungewichtete β-Schätzungen durchgeführt. Die gewichteten Schätzungen sind in den Abbildungen 8 und 9 in den oberen Zeilen, die ungewichteten in den unteren Zeilen der Abbildungen dargestellt. In den Spalten sind die verschiedenen SMP-Typen getrennt aufgeführt. Zur besseren Einschätzung der Verteilung der RRMSE in den Szenarien sind die mittleren RRMSE in Prozent über die SMP eines Schätzer angegeben. In Klammern steht der Anteil der SMPs, die einen RRMSE von über $0,5\%$ haben. Weiterhin ist derjenige Schätzer orange markiert, welcher den geringsten mittleren RRMSE je Kombination aufweist. Erfreulicherweise ist dies durchgängig der SMP-separate GREG-Schätzer der das β ohne Ausreißerbehandlung und ohne das Herausnehmen der Nullanschriften schätzt.

Für die gesetzlich Relevanten SMP-Typen 0 und 1 stellt sich der simulierte RRMSE wie folgt dar. Für den SMP-Typ 0 kann im Prinzip jeder der hier vorgestellten GREG-Schätzer verwendet werden, da bei allen das Qualitätserfordernis erfüllt ist. So erreichen alle SMPs des Typs 0 einen simulierten RRMSE von unter 0,5%. Die Unterschiede zwischen den Schätzern sind so klein, dass sie nicht von der Monte-Carlo-Variation unterscheidbar sind. Für den SMP-Typ 1 unterscheiden sich die vorgestellten GREG-Schätzer zwar etwas mehr als beim SMP-Typ 0, jedoch auch nicht erheblich. So erreichen alle Schätzer einen mittleren simulierten RRMSE von 0,49%, und bei allen Schätzern liegen über 58% der SMPs des Typs 1 unter einem simulierten RRMSE von 0,5%. Hierbei zeigt es sich jedoch, dass bei der Verwendung des Schätzers SMP-SEP, also demjenigen ohne Behandlung von Ausreißer und ohne gesonderter Berücksichtigung der Nullanschriften, mit die höchste Rate der SMPs vom Typ 1, die unter 0,5% simuliertem RRMSE liegen, erreicht wird.

Auch wenn keine gesetzliche Anforderung an die Qualität der SMP-Typen 2 und 3 gestellt sind, ist es nicht nur für die Forschungsgemeinschaft interessant wie gut dort die Schätzungen verlaufen. Auch hier liegt der SMP-SEP GREG-Schätzer vorne, bei den simulierten RRMSE, und erreicht, vor allem bei SMP-Typ 3, in vielen SMPs einen erfreulich kleinen simulierten RRMSE.

Der Unterschied zwischen der gewichteten Schätzung und der ungewichteten Schätzung des β ist in diesem Zusammenhang nicht von der reinen Monte-Carlo-Variation zu unterscheiden, sodass davon ausgegangen werden kann, dass für die Punktschätzung bei Ziel 1 die Gewichtung des linearen Regressionsmodells keine größere Rolle spielt. Dies ist nicht weiter erstaunlich, da die lineare Regression durch sehr viele Beobachtungen gestützt ist, und somit sehr stabil ist, weiterhin wird der überwiegende Teil der Variation der abhängigen Variable durch die Registervariable erklärt, so dass das Modell an sich schon sehr viel Erklärungskraft bietet.

Es stellt sich nun die Frage, ob es nicht vielleicht besser wäre, für die Nullanschriften, anstelle einer Herausnahme aus der Schätzung des β eine Dummy-Variable zu verwenden. Diese Dummy-Variable liefert den Wert 1, falls eine Anschrift eine Nullanschrift ist und 0, falls in der Anschrift Personen registriert sind. Diese Betrachtung ist in Abbildung 9 dargestellt. Aus den mittleren RRMSE und den Anteilen an SMPs, die einen RRMSE über $0,5\%$ aufweisen, ist zu sehen, dass das Einführen einer Dummy-Variablen nicht zu einer sichtbaren Verbesserung der Punkt-Schätzung in der Monte-Carlo-Simulation führt (Vergleich mit Abbildung 8).

5.4.2 Varianz-Schätzung

Ausgehend von den Punktschätzergebnissen wird die Analyse der Varianzschätzung auf die interessierenden Schätzer reduziert. Die Varianzschätzung wird hier auf zweierlei Weise analysiert. Zum einen wird der relative Bias der Varianzschätzung betrachtet. Zum Anderen werden Konfidenzintervallüberdeckungsraten untersucht. Der relative Bias der Varianzschätzung für die Area d und den Schätzer „est“ ist definiert als

$$\mathrm{VBias}_d^{\mathrm{est}} = \frac{\mathrm{E}\left(\widehat{\mathrm{V}}(\widehat{\tau}_d^{\mathrm{est}})\right) - \mathrm{V}(\widehat{\tau}_d^{\mathrm{est}})}{\mathrm{V}(\widehat{\tau}_d^{\mathrm{est}})} \tag{5.7}$$

wobei hier für $\mathrm{V}(\widehat{\tau}_d^{\mathrm{est}})$ die Monte-Carlo Varianz der Punktschätzer und für $\mathrm{E}\left(\widehat{\mathrm{V}}(\widehat{\tau}_d^{\mathrm{est}})\right)$ der Monte-Carlo-Erwartungswert der Varianzschätzer verwendet wurde. Dieses Maß kann Werte im Intervall $(-\infty, 1]$ annehmen, wobei Werte kleiner Null eine Unterschätzung und Werte über Null eine Überschätzung der Varianz der Punktschätzer signalisieren. Erwünscht ist somit ein VBias von 0.

In Abbildung 10 ist der simulierte RRMSE der Punktschätzer gegen den simulierten relativen Bias der Varianzschätzung aufgetragen. Zunächst ist festzustellen, dass das Einfügen eines Dummys für die Nullanschriften (rechte Grafik in Abbildung 10) keinen deutlichen Unterschied macht, weder beim simulierten RRMSE noch beim simulierten relativen Bias der Varianzschätzung. Ein deutlicher Unterschied der Varianzschätzungen ist jedoch zwischen den Modellen SMP-SEP und SMP-SEP.A zu verzeichnen. Beim SMP-SEP.A, der Ausreißer aus der Schätzung des β herausnimmt, ist zu sehen, dass der simulierte relative Bias der Varianzschätzung zum Teil wesentlich geringer ist, als ohne Ausreißerbehandlung. Dies geschieht jedoch auf Kosten eines höheren RRMSEs. Dieser höhere RRMSE ist durch eine stark erhöhte Variabilität des Punktschätzers getrieben, falls Ausreißer aus dem Regressionsmodell rausgeworfen werden. Es ist auch zu sehen, dass die allermeisten SMPs einen geringen simulierten relativen Bias der Varianzschätzung gepaart mit einem geringen simulierten RRMSE aufweisen und somit unproblematisch erscheinen.

Ein ähnliche Bild liefert die Betrachtung der Konfidenzintervallüberdeckungsraten in Abbildung 11. Die Konfidenzintervallüberdeckungsrate ist definiert als der Anteil der in der Simulation erzeugten Konfidenzintervalle, die den wahren Wert überdecken. Zu sehen ist, dass die Ausreißerbehandlung zwar dazu führt, dass geringfügig mehr Konfidenzintervalle den wahren Wert überdecken, eine KI-Rate von 24% ist jedoch auch unbefriedigend. Hierbei wird durch die Verwendung der Ausreißerbehandlung eine etwas höhere KI-Rate auf Kosten einer wesentlich größeren KI-Länge erkauft. Wiederum sind keine bedeutsamen Unterschiede zwischen dem Regressionsmodell ohne (links) und dem Regressionsmodell mit Dummy Variable (rechts) in Abbildung 11 zu erkennen. Bei den allermeisten SMPs funktioniert die Kombination von Punkt und Varianzschätzer sehr gut, hier liegen die simulierten KI-Raten sehr nahe der nominalen KI-Rate von 95%.

An Abbildung 12 ist gut zu erkennen, dass im Falle von SMP-SEP der Hauptgrund für die niedrige KI-Rate in einer leichten Verzerrung des Punktschätzers zu verzeichnen ist. Je höher die Verzerrung, desto geringer ist hier die KI-Rate. Im Gegensatz dazu ist bei SMP-SEP.A für die niedrigen KI-Raten nicht eine Verzerrung verantwortlich, sondern eine hohe Variabilität der Punktschätzung. In Anbetracht der zum Teil viel höheren RRMSE die der SMP-SEP.A

gegenüber dem SMP-SEP in diesen problematischen SMPs hat, ist der Vorzug der etwas niedrigeren relativen Verzerrung jedoch zu vernachlässigen.

In Abbildungen 13 und 14 ist die Puntk- und Varianzschätzverteilung für den SMP-SEP und den SMP-SEP.A für SMP 084360010010 und 083155001006 dargestellt. Diese SMPs stellen sich sowohl was den simulierten RRMSE wie auch bei der Varianzschätzung in der Simulation als problematisch heraus. Anhand der Punktschätzverteilung wird die Problematik klar. Die Ausreißer, die in diesen SMPs vorliegen, führen zu einer Separation der Punktschätzverteilung in zwei Massebereiche. So entsteht eine zweigipflige Verteilung der Punktschätzer. Das selbe passiert bei der simulierten Verteilung der Varianzschätzer. Daraus kann geschlossen werden, dass die vorhandenen Ausreißer einen ausgeprägt starken negativen Einfluss auf die simulierten RRMSEs dieser SMPs haben.

5.4.3 Vergleich der Realisierten Stichprobe mit den Ergebnissen der Simulation

Über die Simulation hinaus, ist es von Interesse zu sehen, ob Simulation und realisierte Stichproben die gleichen SMPs problematisieren. Hierzu liefert die Abbildung 15 eine Übersicht. Zunächst fällt auf, dass die geschätzten RRMSEs aus der Stichprobe meistens höher liegen, als die simulierte RRMSEs. Es ist davon auszugehen, dass die KAL-FEB-Modelle nicht alle Effekte, die im Zensus-Stichprobenforschungsprojekt berücksichtigt werden konnten, verwenden. Dies resultiert auch daraus, dass im aktuellen Datensatz einige sehr auffällige Ausreißer enthalten sind, die im Zensus-Test nicht beobachtet wurden.

Da nicht alle SMPs separat betrachtet werden können, wurden drei Auffällige eingehender dargestellt, die auch in Abbildung 15 gekennzeichnet sind. Bei diesen wird im Folgenden genauer darauf eingegangen, warum diese einen höheren RRMSE aufweisen. In den folgenden Abbildungen kennzeichnen rote Kreise die Stichprobenbeobachtungen und blaue Kreise die Beobachtungen in der Grundgesamtheit. Die Größe der Kreise ist proportional zur Wurzel der Häufigkeit des Auftretens der Beobachtungen. Jede Beobachtung entspricht einer Anschrift mit der Anzahl der registrierten beziehungsweise tatsächlich vorhanden Personen, wobei das KAL/FEB-Modell 3/3 zugrunde gelegt wurde.

SMP 05315000000004 Der SMP 05315000000004 weist in der Simulation einen deutlich niedrigeren RRMSE auf (0,453%) als in der Stichprobe geschätzt wird (1,458%). Dabei scheint zunächst das lineare Regressionsmodell recht gut auf die Daten zu passen. Sehr auffällig ist dabei jedoch der eine Punkt bei etwa (0,80). Diese Beobachtung ist in der Stichprobe enthalten. Das absolute Residuum für diese Beobachtung ist extrem hoch. Dies erhöht die geschätzte Varianz erheblich, zumal die Inklusionswahrscheinlichkeit für Anschriften ohne registrierte Personen verhältnismäßig gering ist, und somit das relativ hohe Gewicht diese Abweichung noch betont. In diesem SMP liegt der hohe geschätzte RRMSE an einem Ausreißer. In diesem Falle, wenn auch nicht gesondert dargestellt, liefern gewichtete und ungewichtete Hochrechnung unterschiedliche Ergebnisse, insbesondere für die Varianzschätzung. In den meisten anderen Fällen erkennt man sonst höchstens vernachlässigbare Unterschiede.

SMP 083155001006 In SMP 083155001006 sieht das Bild ähnlich aus. In diesem Fall liegen jedoch hohe Karteileichen-Anzahlen in zwei Anschriften vor. Auch wenn fast alle Punkte

genau auf der Winkelhalbierenden liegen und so das lineare Regressionmodell im Grunde eine sehr genaue Anpassung erreichen sollte, reichen diese wenigen Ausreißer aus, um sowohl den simulierten RRMSE wie auch den geschätzten RRMSE dieses SMPs extrem zu erhöhen. Es liegt somit wiederum ein Ausreißerproblem vor. Ausreißer rechts unten in der Graphik erweisen sich auch als Hebelpunkte für die Regression.

SMP 084360010010 Hier resultiert in etwa die gleiche Aussage wie für SMP 083155001006.

SMP 084160041041 Im SMP 084160041041 sind besonders viele Nullanschriften vorhanden. Damit soll der Einfluss der Nullanschriften auf die Schätzung untersucht werden. Das Ergebnis ist jedoch sehr überzeugend, sowohl der simulierte RRMSE als auch der geschätzte RRMSE liegen beide unter der Qualitätsanforderung von 0,5% RRMSE.

5.4.4 Zusammenfassung der Ergebnisse aus den Simulationen zu Ziel 1

Auf Grundlage der Simulationen ergibt sich die Situation, dass die Empfehlungen aus dem Stichproben-Forschungsprojekt nicht revidiert werden müssen. Als geeigneter Schätzer für die Anwendung auf der realisierten Stichprobe wird weiterhin der SMP-SEP GREG-Schätzer ohne Ausreißerbehandlung und ohne gesonderte Behandlung der Nullanschriften empfohlen.

Es ist weiterhin zu erkennen, dass die Qualitätseinbußen größer ausfallen, als im Stichprobenforschungsprojekt skizziert wurde. Dieser Umstand folgt durch das Auftreten teilweise sehr auffälliger Ausreißer, was in der Form nicht zu vermuten war. Die großen Anschriften weisen hier extreme Klumpungs-Effekte auf, die sich sehr negativ auf die Schätzungen und deren Qualität auswirken.

Weiterhin ist auffällig, dass fast alle Ausreißer die im Simulationsdatensatz enthalten sind, auch in der Stichprobe beobachtete Ausreißer sind. Das heißt, dass die verwendeten Karteileichen- und Fehlbestandvektoren möglicherweise glatter sind, als in der tatsächlichen Population. Aus mangelnder Information ist dies allerdings nur schwerlich überprüfbar. Dennoch sind die Aussagen der Simulation in Bezug auf die Empfehlungen der Methodik belastbar. Die Qualitätserfordernisse können aber nicht in der Form gehalten werden. Nicht berücksichtigt werden konnten in diesen Untersuchungen die Nachziehung sowie die Sonderanschriften.

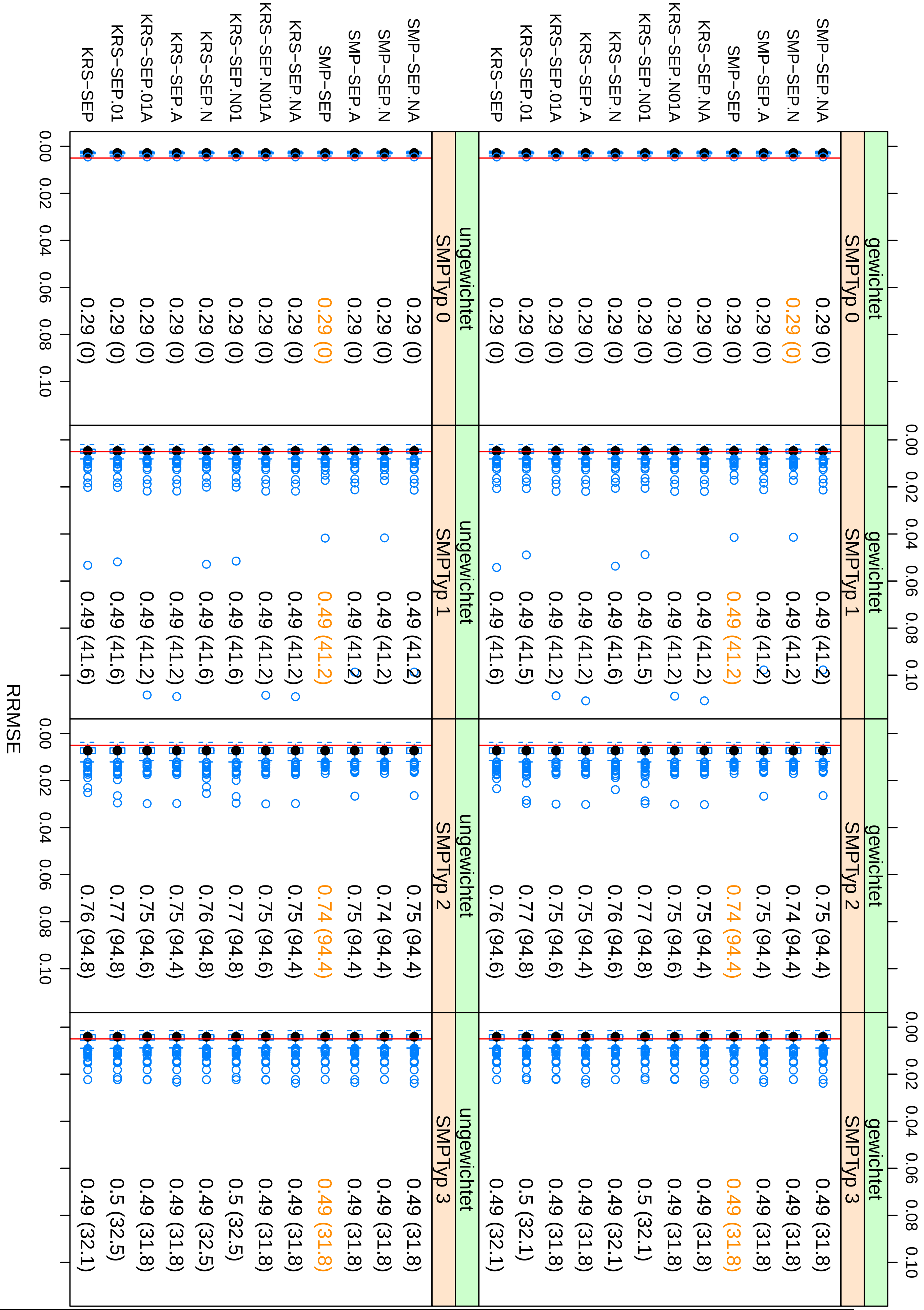

Abbildung 8: RRMSE bei KAL/FEB-Modell 3/3 für das Regressionsmodell $\tau_Z = \beta_1 + \text{REGHW}\beta_2 + e$

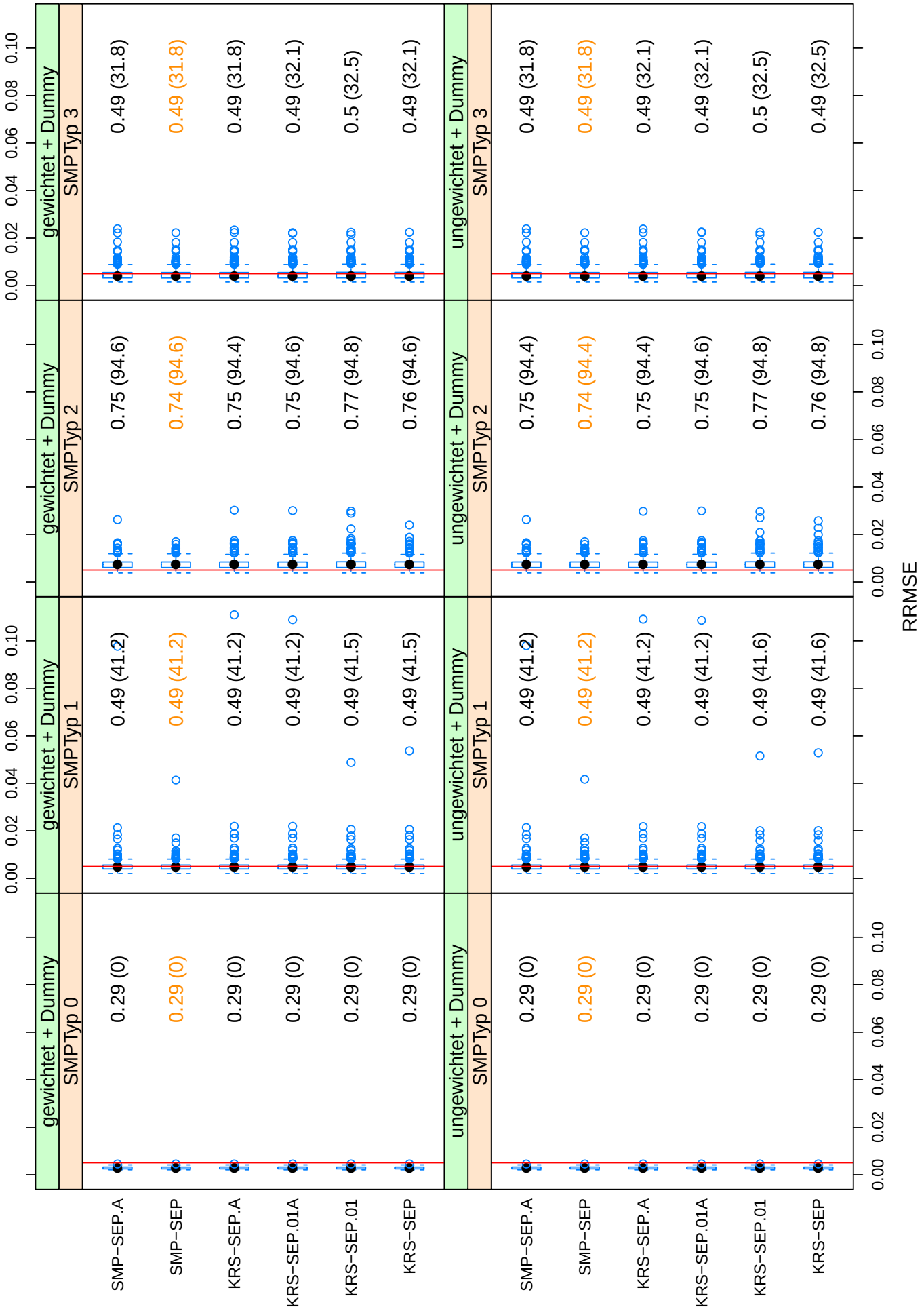

Abbildung 9: RRMSE bei KAL/FEB-Modell 3/3 für das Regressionsmodell $\tau_Z = \beta_1 +$ REGHW$\beta_2 +$ Nullanschrift$\beta_3 + e$

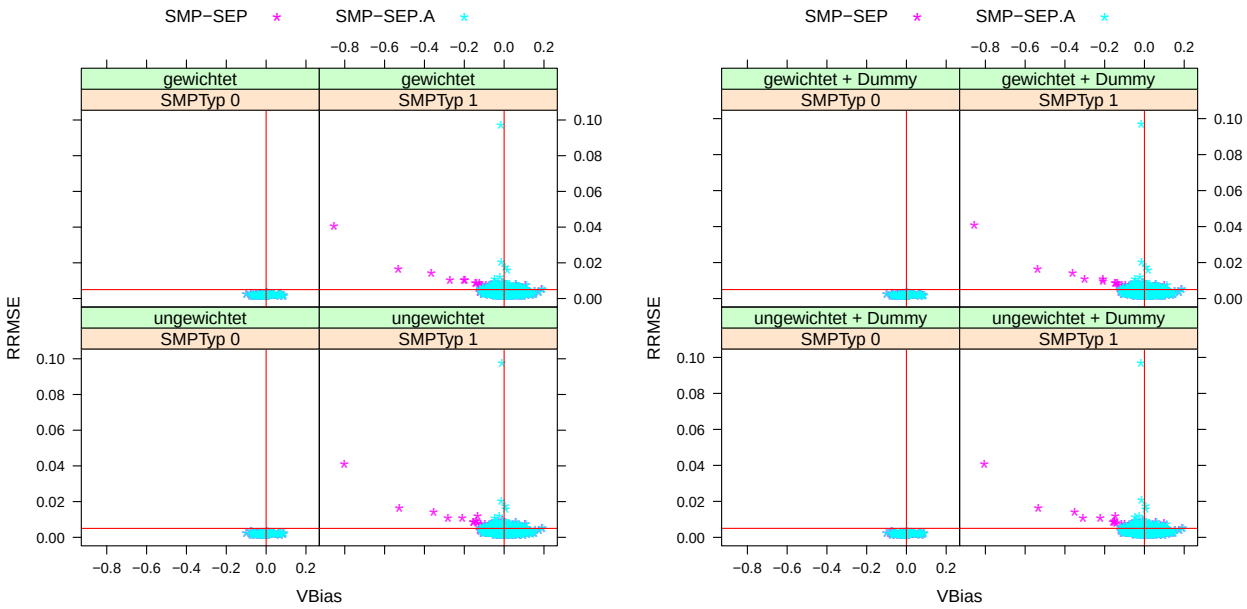

Abbildung 10: RRMSE versus Bias der Varianzschätzung bei KAL/FEB-Modell 3/3 beim Regressionsmodell SMP-SEP und SMP-SEP-A. Links für das Modell ohne und rechts für das Modell mit Dummy für die Nullanschriften.

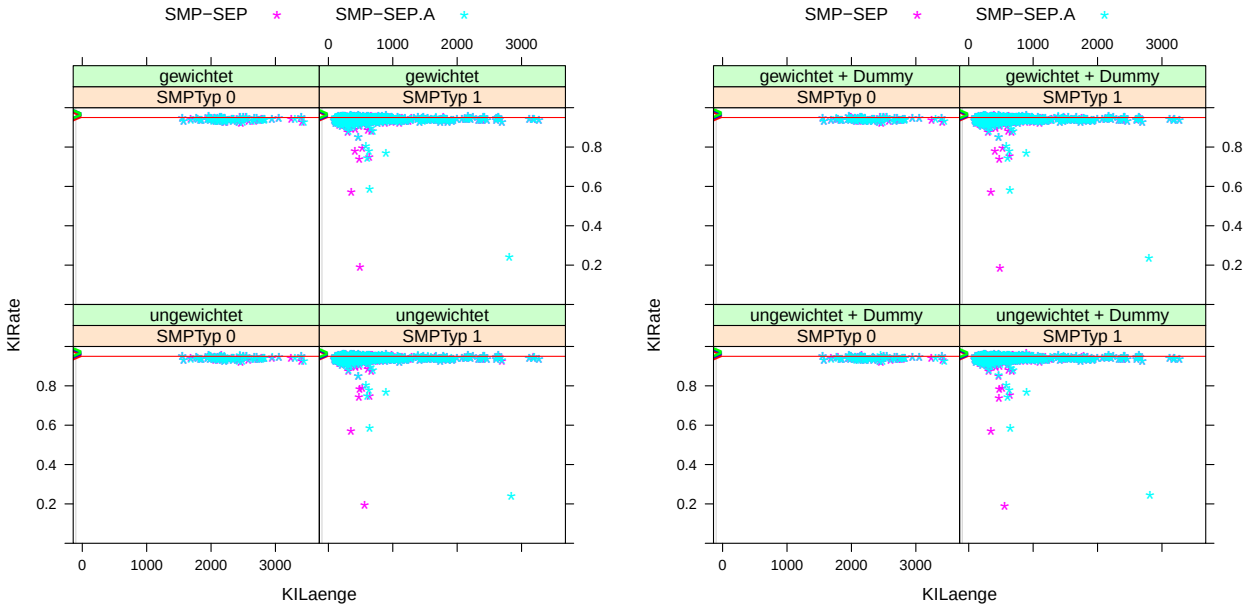

Abbildung 11: KI-Rate versus KI-Länge bei KAL/FEB-Modell 3/3 beim Regressionsmodell SMP-SEP und SMP-SEP-A. Links für das Modell ohne und rechts für das Modell mit Dummy für die Nullanschriften.

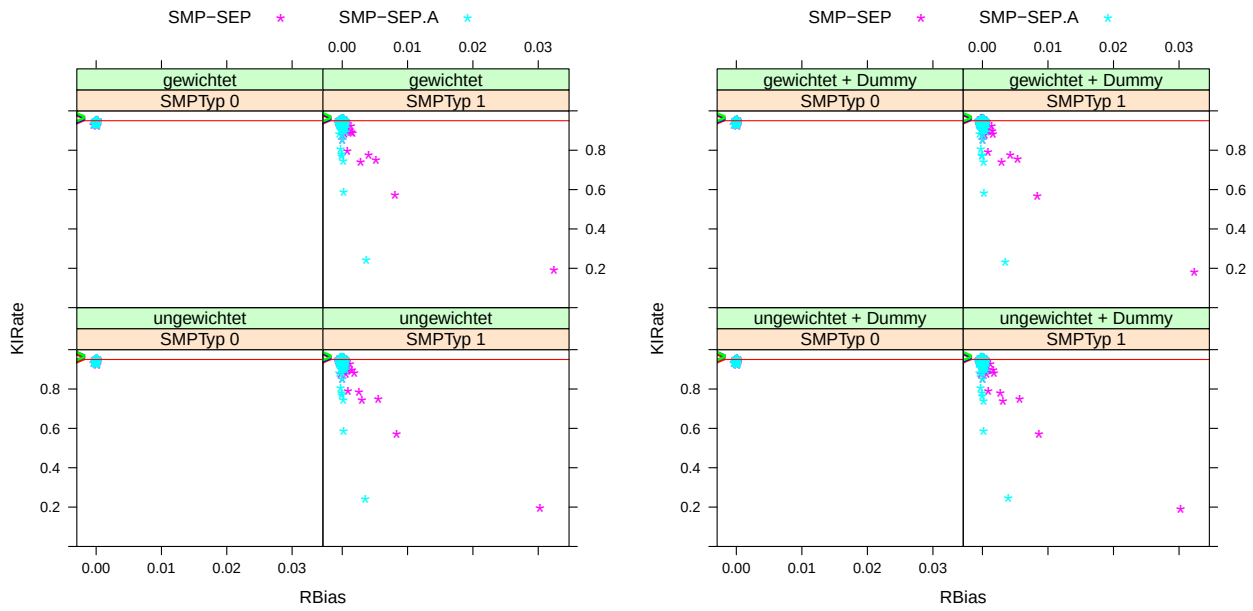

Abbildung 12: KI-Rate versus Bias der Punktschätzung bei KAL/FEB-Modell 3/3 beim Regressionsmodell SMP-SEP und SMP-SEP-A. Links für das Modell ohne und rechts für das Modell mit Dummy für die Nullanschriften.

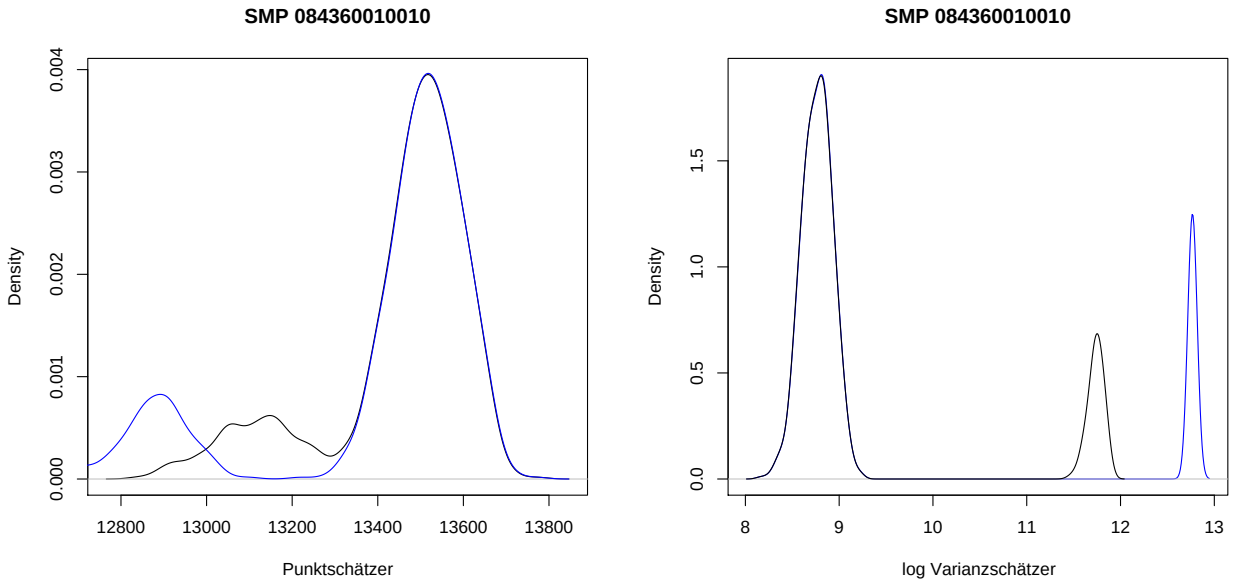

Abbildung 13: Punktschätzverteilung (links) und Varianzschätzverteilung (rechts) für die Schätzer SMP-SEP (schwarz) und SMP-SEP.A (blau) in SMP 084360010010

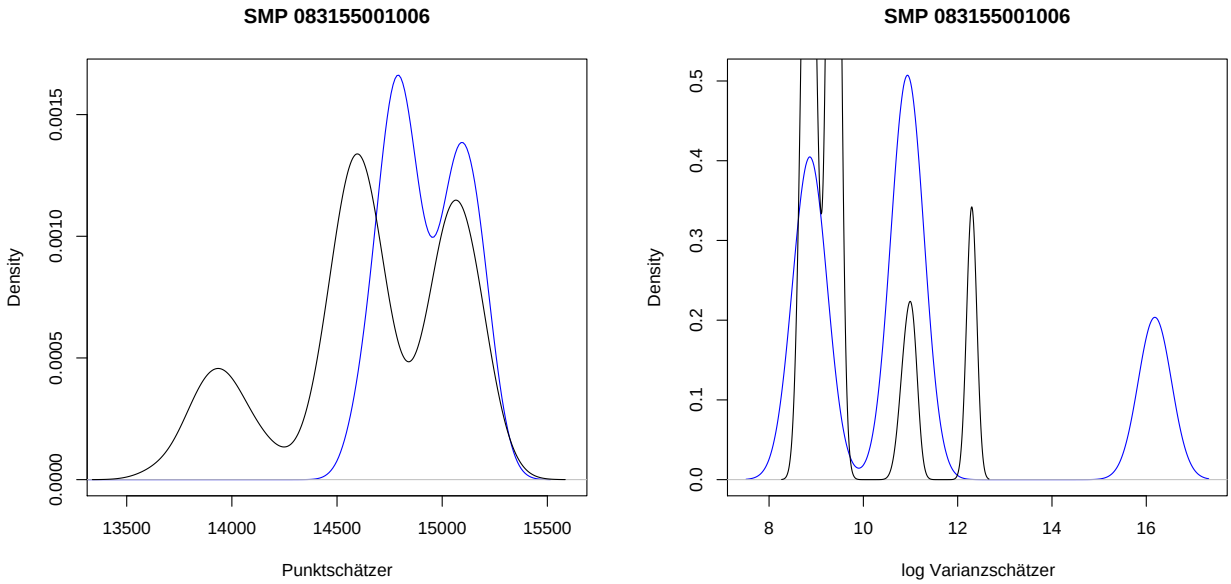

Abbildung 14: Punktschätzverteilung (links) und Varianzschätzverteilung (rechts) für die Schätzer SMP-SEP (schwarz) und SMP-SEP.A (blau) in SMP 083155001006

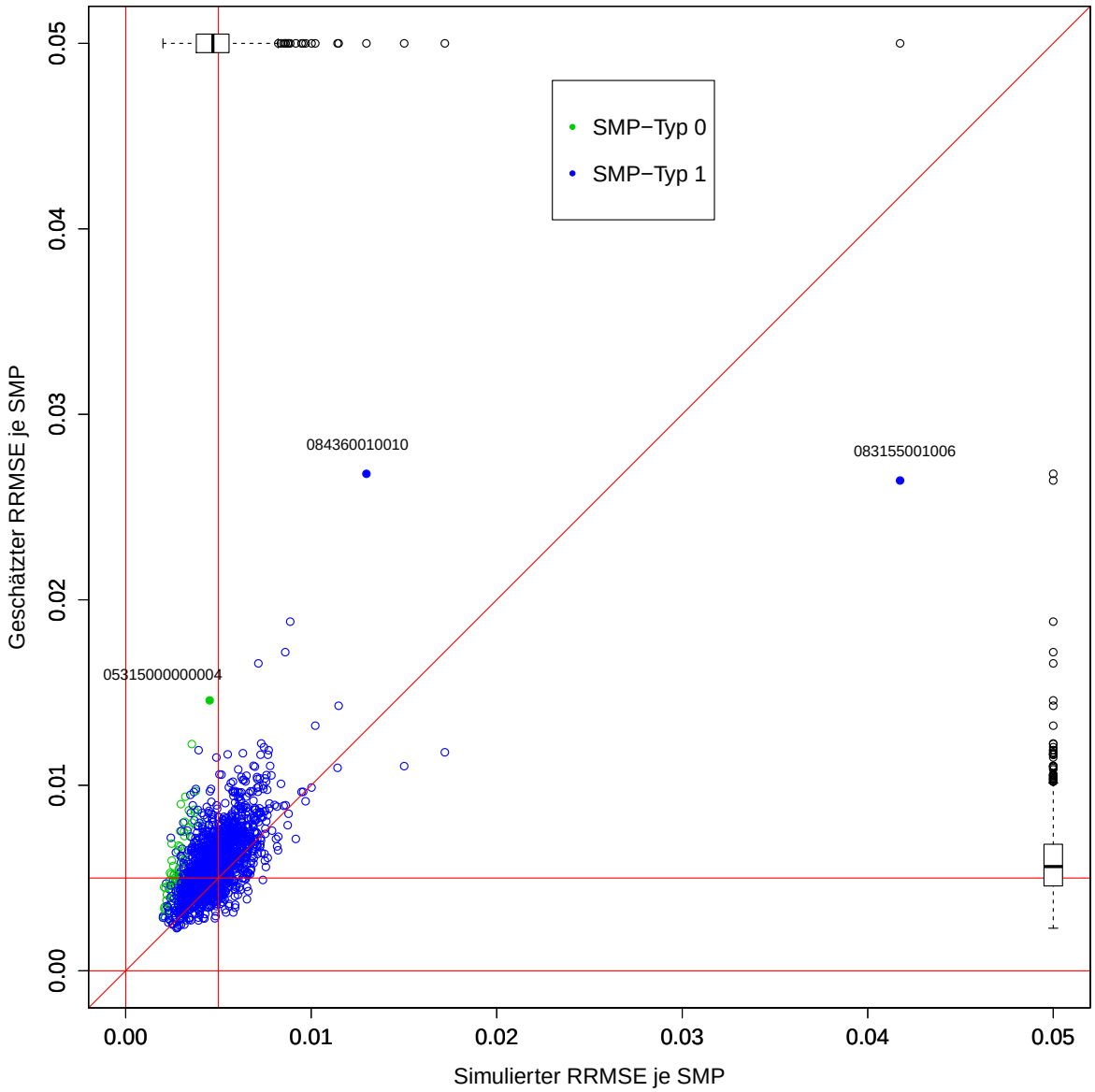

Abbildung 15: Geschätzer RRMSE in der Stichprobe versus Simulierter RRMSE je SMP vom Typ 0 oder 1.

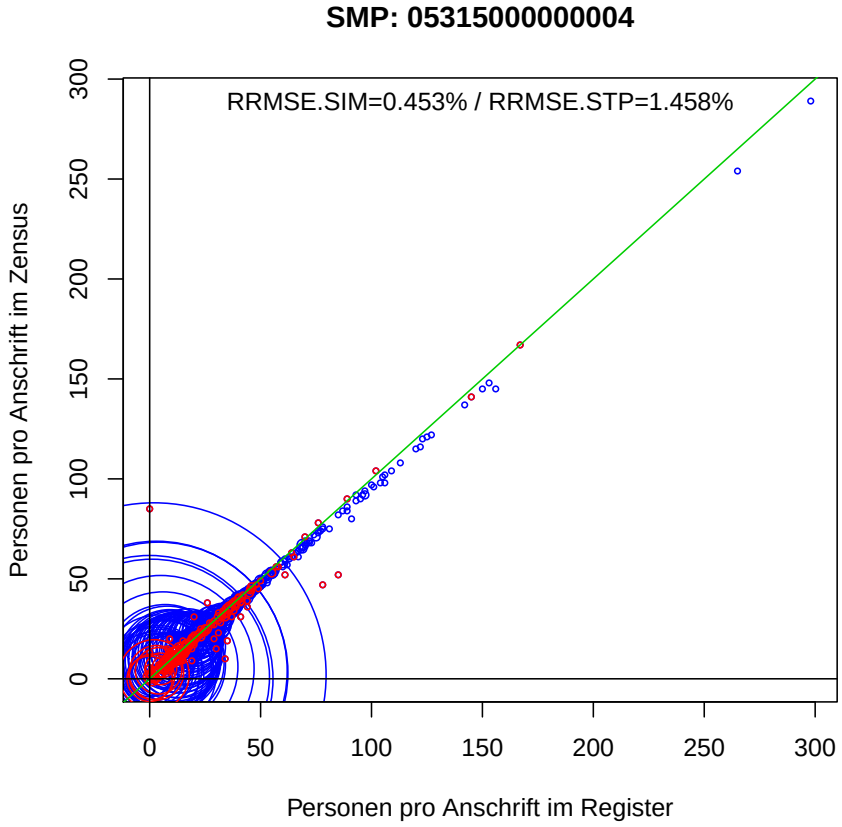

Abbildung 16: Registerbevölkerung versus Zensusbevölkerung in der Simulation für SMP 05315000000004.

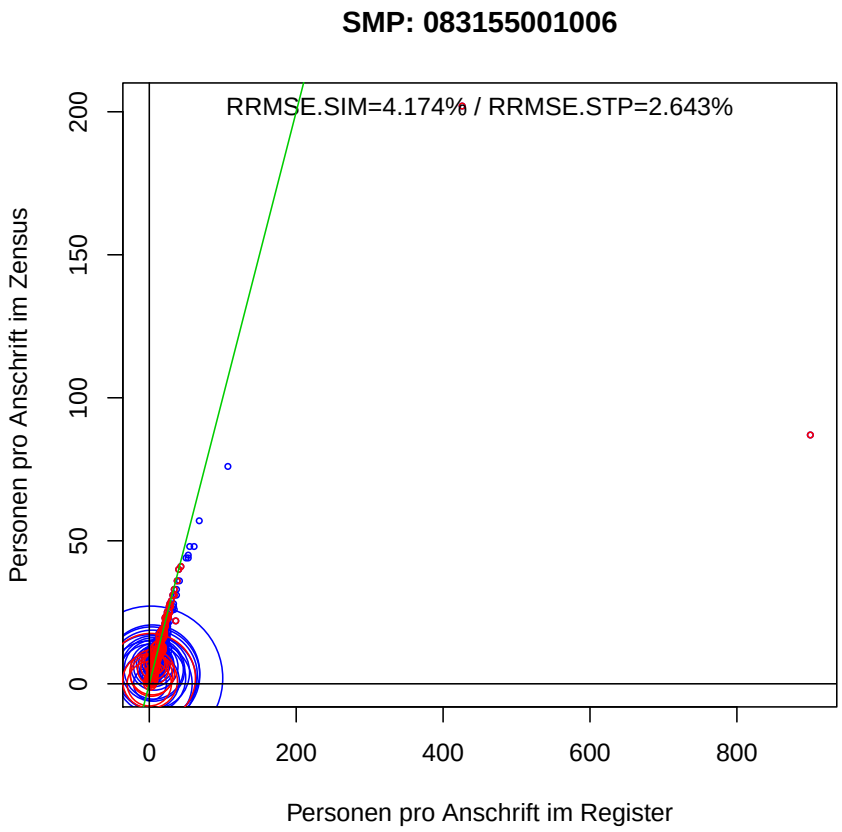

Abbildung 17: Registerbevölkerung versus Zensusbevölkerung in der Simulation für SMP 083155001006.

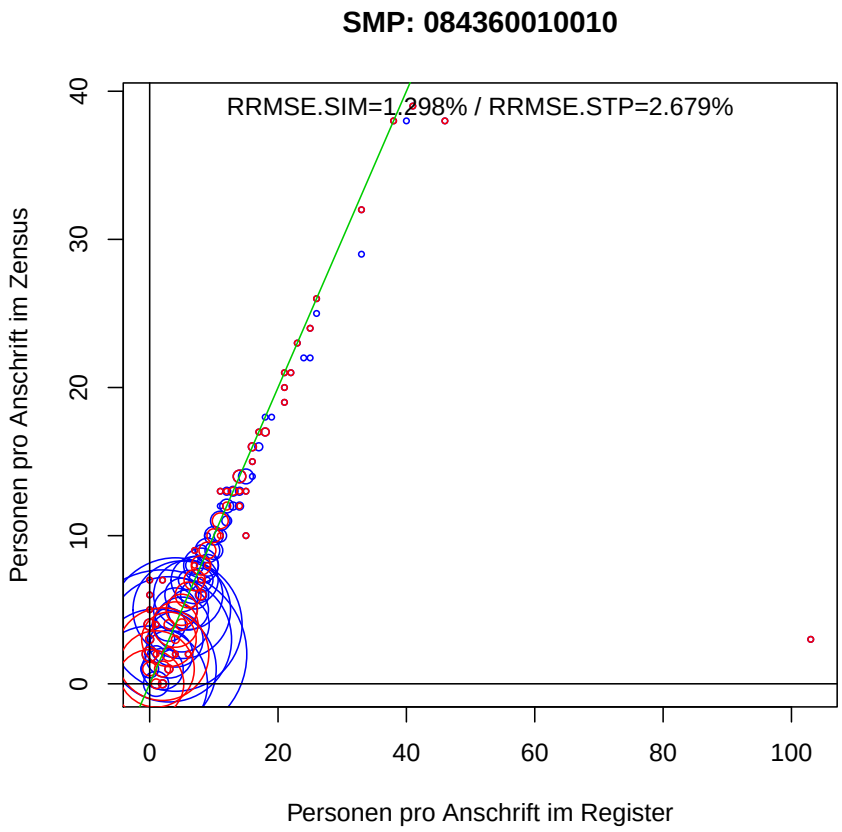

Abbildung 18: Registerbevölkerung versus Zensusbevölkerung in der Simulation für SMP 084360010010.

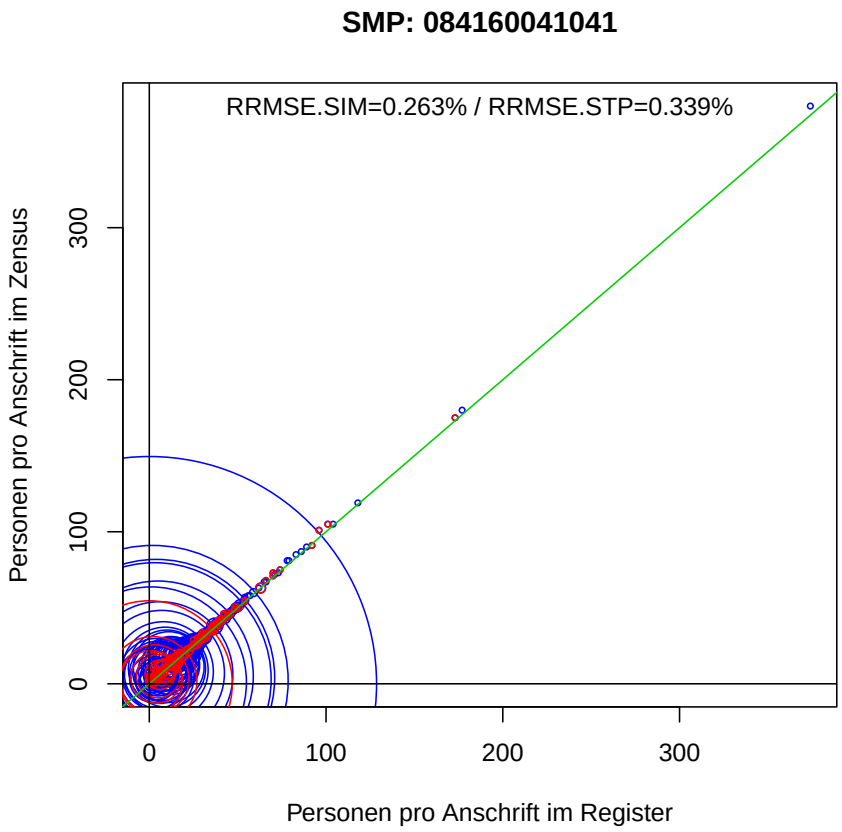

Abbildung 19: Registerbevölkerung versus Zensusbevölkerung in der Simulation für SMP 084160041041.

5.5 Simulationen zur Fragestellung 2

Um den Einfluss der Nullanschriften auf die Ergebnisse des Zensusforschungsprojektes zu untersuchen, wurde eine weiterführende Simulationsstudie durchgeführt. Mit dieser soll geklärt werden, inwieweit ungerechtfertigte Nullanschriften die optimale Allokation und damit die Qualität der Schätzung der amtlichen Einwohnerzahl auf SMPs beeinflussen können. Hierzu wurden vier Szenarien entwickelt:

Tabelle 9: Szenarien der Simulationsstudie II

Szenario	Grundgesamtheit	Schichtung
M1(wie in Teil 1)	vollständiges Register	Schichtung wie im Zensus
M1b	vollständiges Register	Schichtung neu durchgeführt
M2	um die unerwünschten Nullanschriften verkürztes Register	Schichtung neu durchgeführt
M3	alle Nullanschriften aus dem Register entfernt	Schichtung neu durchgeführt

Da die Simulationen sehr rechenintensiv sind, wurde ausschließlich das KAL/FEB-Modell 3/3 verwendet. Die zu beantwortende Frage ist, wie stark sich die Nullanschriften im Register auf die Qualität der Schätzung auswirken. In Abbildung 20 sind die relativen Veränderungen der RRMSEs aufgrund der verschiedenen Nullanschriften dargestellt. Als Referenz-Szenario wurde hierbei das Szenario M1b verwendet. Die relativen Veränderungen r^*_{**} der RRMSEs für Szenario $*$ und Schätzer $**$ sind wie folgt berechnet

$$r^*_{**} = \frac{RRMSE^*_{**} - RRMSE^{\mathrm{M1b}}_{**}}{RRMSE^{\mathrm{M1b}_{**}}} \quad .$$

Eine relative Veränderung r^* von größer Null für Szenario $*$ bedeutet, dass das Szenario $*$ einen um $r^* \cdot 100$ Prozent höheren RRMSE aufweist als das Szenario M1b.

Wie an Abbildung 20 zu sehen ist, verändert sich die Qualität der Punktschätzung zwischen dem tatsächlich realisierten Szenario M1 und dem Szenario M1b nur unwesentlich. Das tatsächlich realisierte Szenario schneidet teilweise sogar leicht besser ab. Durch die Schichtwechsler und Ausreißer in der realisierten Stichprobe und die damit verbundenen unterschiedlich wirkenden Effekte auf die RRMSEs sind präzise Interpretationen in allgemeiner Form nicht möglich.

Im Szenario M2 liegt der RRMSE der Punktschätzung im Schnitt um 1,2 Prozent niedriger bei SMPs vom Typ 0 und um 1,25 Prozent niedriger bei SMPs vom Typ 1 als im Referenz-Szenario. Das Herausnehmen einiger Nullanschriften aus dem Register führt also zu einer Verbesserung der Punktschätzung. Noch deutlicher wird dies am Beispiel des Szenario M3, bei dem sich die Punktschätzung im Schnitt bei allen vier SMP-Typen verbessert, speziell um 2,3 Prozent bei SMPs vom Typ 0 und 4,5 Prozent bei SMPs vom Typ 1. Wenn also keine Nullanschriften im Register gewesen wären, wie es im Zensusstichproben-Forschungsprojekt der Fall war, so wären im Durchschnitt der SMPs vom Typ 1 Verbesserungen um 4,5 Prozent präzisere Schätzergebnisse zu erwarten gewesen.

Die Ergebnisse zwischen ungewichteter und gewichteter β-Schätzung unterscheiden sich überwiegend erst nach der zweiten Nachkommastelle und liefern daher eine identische Beurteilung.

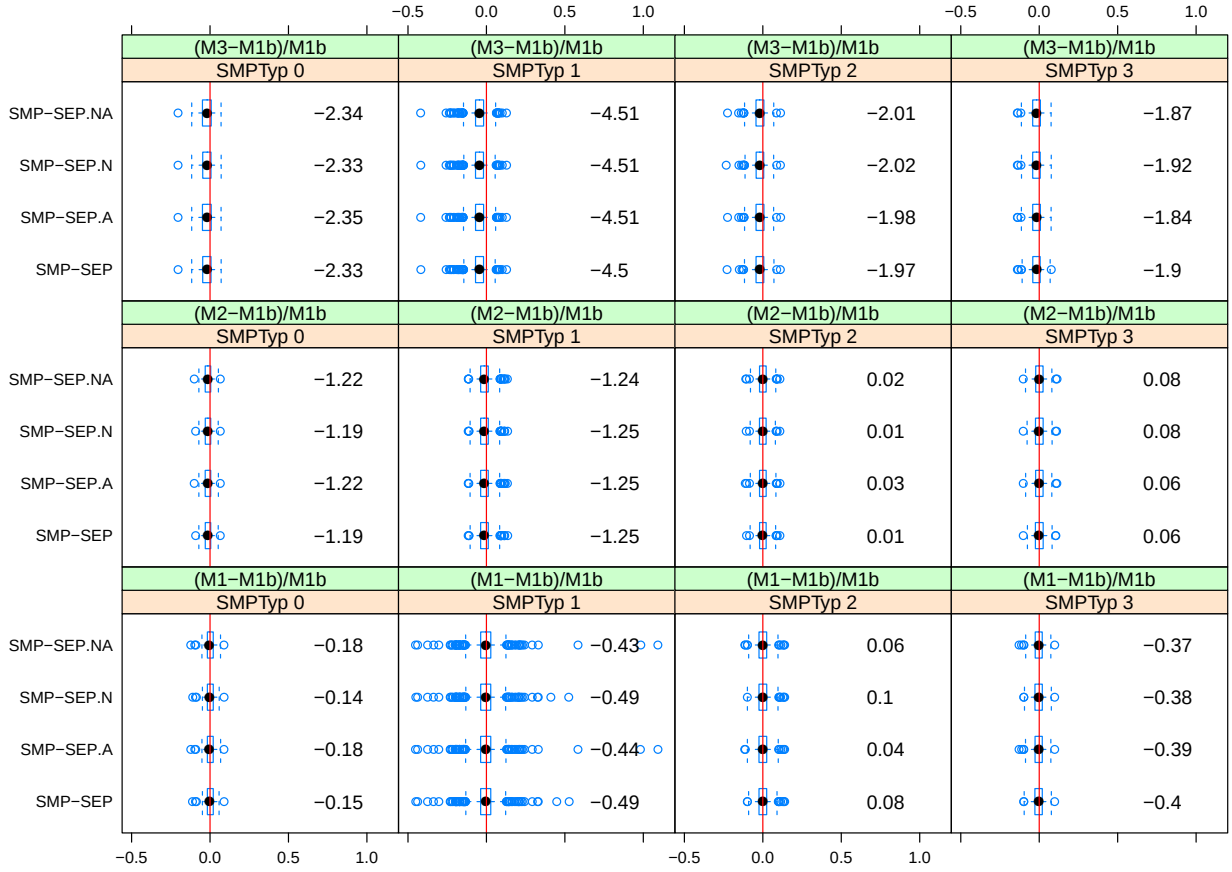

Abbildung 20: Relative Veränderung der RRMSE der Punktschätzung der Szenarien M1, M2 und M3 zum Referenz-Szenario M1b bei ungewichteter β-Schätzung.

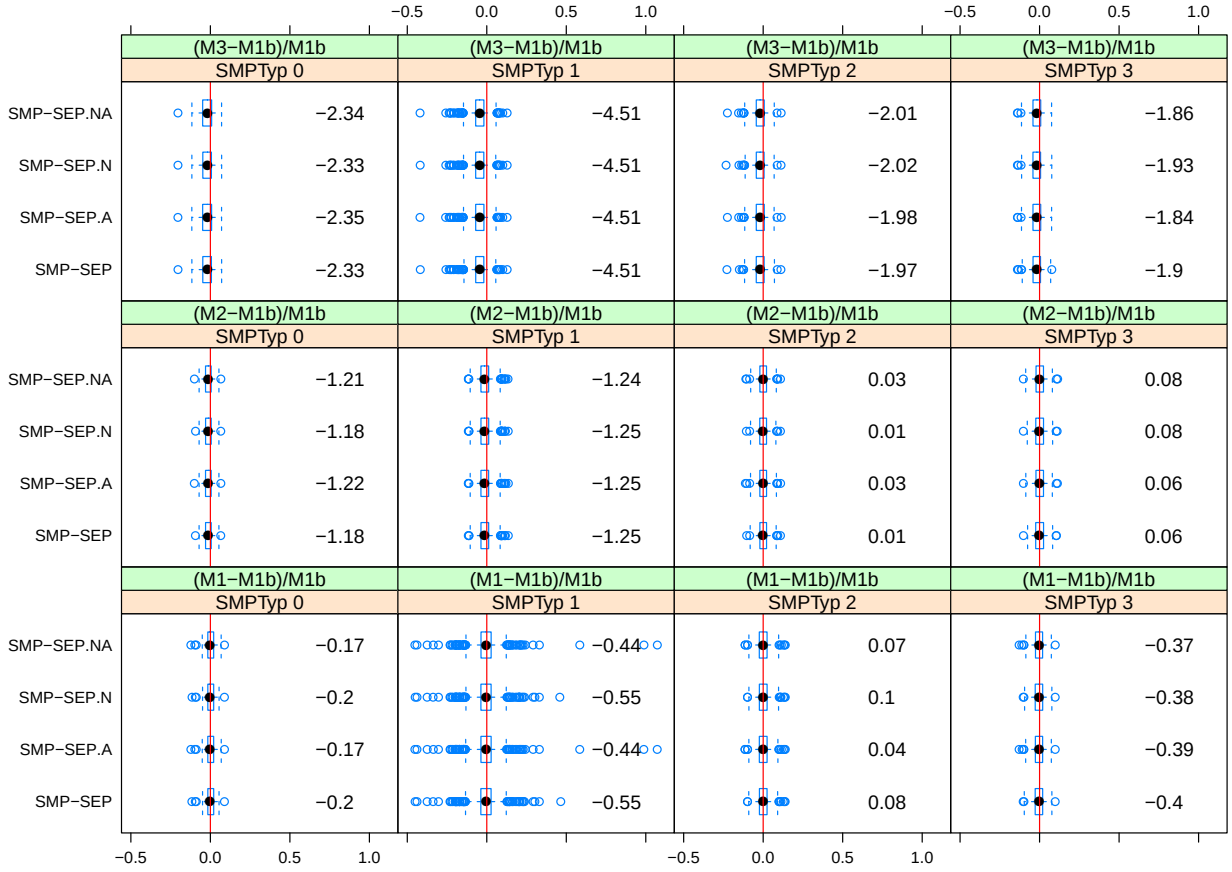

Abbildung 21: Relative Veränderung der RRMSE der Punktschätzung der Szenarien M1, M2 und M3 zum Referenz-Szenario M1b bei gewichteter β-Schätzung.

6 Bewertungen und Empfehlungen

6.1 Ziele des Auftrags

Ziel des Validierungsprojektes ist es herauszufinden, welchen Einfluss die Nullanschriften auf die im Zensus-Stichprobenforschungsprojekt getätigten Empfehlungen sowie die Qualität der Hochrechnung haben.

> „Es ist vom Auftragsnehmer zu untersuchen, inwieweit das *Vorhandensein von Nullanschriften* in der Stichprobe in Verbindung mit den realen Verteilungen von Melderegister-übererfassungen (*Karteileichen*) und -untererfassungen (*Fehlbeständen*) eine Modifikation der im Stichprobenforschungsprojekt auf Basis von simulierten Verteilungen der Registerfehler ausgesprochenen *Empfehlungen für ein Hochrechnungsverfahren* angeraten sein lässt und damit zu einer *Nachjustierung* des derzeit in der Implementation befindlichen Hochrechnungsverfahren führen würde“

6.2 Zusammenfassung

Die Untersuchungen in den Stichproben und den Simulationen haben gezeigt, dass für die gegebene Stichprobe trotz des massiven Auftretens von Nullen die Empfehlungen aus dem Stichprobenforschungsprojekt nicht revidiert werden müssen. Dies war a priori in der Form nicht vorauszusehen, da der Umfang der Nullanschriften erheblich war.

Wie die Simulationen zu Fragestellung 2 gezeigt haben, ist jedoch eine Verschlechterung der Qualität durch die unerwarteten Nullanschriften im Register erkennbar.

Wenig erfreulich ist zudem die Tatsache, dass die Qualitätsschätzung aus den vorliegenden Daten die Erfordernisse nicht in dem gewünschten Maße erfüllen. Diese Aussage gilt verstärkt in SMPs mit zum Teil gravierenden Ausreißern (hohe Registerzahl bei nur wenigen tatsächlich vorhandenen Einwohnern), aber auch in SMPs mit unerwartet vielen Nullanschriften im Register.

Literatur

Burgard, J. P. & Münnich, R. (2010), ‘Modelling over- and undercounts for design-based monte carlo studies in small area estimation: an application to the german reister-assisted census’, *Computational Statistics and Data Analysis* .

Cook, R. & Weisberg, S. (1982), *Residuals and influence in regression*, Vol. 5, Chapman and Hall New York.

Cox, L. H. (1987), ‘A constructive procedure for unbiased controlled rounding’, *JASA* **82**, 520–524.

Münnich, R., Gabler, S., Ganninger, M., Burgard, J. P. & Kolb, J.-P. (2012), *Stichprobenoptimierung und Schätzung im Zensus 2011*, number 21 *in* ‘Statistik und Wissenschaft’, DESTATIS.

A Datensätze

Variablename	Inhalt	SAS-Variablentyp
SMP_NR	Nummer des Samplingpoints	alphanumerisch, 14 Stellen
AN_GRKL	Anschriftengrößenklasse	alphanumerisch, eine Stelle
SMP_TYP	SMP-Typ	alphanumerisch, eine Stelle
OPT	Kennzeichnung, ob Stichprobenumfang gesetzt oder berechnet werden soll (1=zu berechnen, 0=gesetzt)	alphanumerisch, eine Stelle
F_U	Untergrenze für den Auswahlsatz in % (=leer, falls OPT=0) numerisch	
F_O	Obergrenze für den Auswahlsatz in % (=leer, falls OPT=0) n	umerisch
SCHICHTUMF	Schichtumfang: Zahl der Anschriften in der Auswahlgesamtheit in der Schicht	numerisch
BEV_SCHICHT	Anzahl der gemeldeten Personen (HW+NW) in der Schicht insgesamt	numerisch
EW_SCHICHT	Anzahl der gemeldeten Personen (HW) in der Schicht insgesamt	numerisch
SHQ_BEV	Varianz von BEV in der Schicht	numerisch
SHQ_EW	Varianz von EW in der Schicht	numerisch
NSIZE	Anzahl der Stichprobenanschriften	numerisch

Tabelle 10: Der Ausgangsdatensatz

A.1 Anmerkungen

Die Auswahlgrundlage (Anschriften- und Gebäuderegister) hatte den Stand 1.9.2010. Wenn man von den Nullanschriften einmal absieht, können sich allein durch den unterschiedlichen Stand schon deutliche Unterschiede gegenüber den Ergebnissen der Phase-1-Daten ergeben. Auch die SMPs haben sich wegen eines aktuelleren Gebietsstand und neu berechneter 10000er Grenze geändert.

A.2 Aufbau der Auswertungsdatensätze

Bei den Auswertungsdatensätzen gibt es zwei verschiedene Arten von Daten-Dateien.

Modell_11_X1_REGHW.RData bzw.
Modell_11_X1_REGHW_NAD.RData

Hier sind die Ergebnisse gespeichert für KAL-Modell 1 und FEB-Modell 1 sowie für das Simulationszenario 1. In diesem Szenario werden keine Ausreißer bereinigt, alle Nullanschriften sind enthalten, die β's werden auf Ebene der SMP berechnet und es werden nur die SMP-Typen 0 und 1 berücksichtigt. Für das Beispiel Modell_11_X1.RData wäre das eine Daten-Datei mit 1619 Zeilen und 13 bzw. 14 Spalten, je nachdem, ob nur REG_HW oder auch der Nullanschriften-Dummy als Hilfsvariable dient. Die SMP Typen 2 und 3 sind hier also nicht enthalten. Schaut

man sich die tau.Z an, dann sind die meisten Points größer als 10000 Einwohner. Einzelne Points gibt es, die weniger als 10000 Einwohner haben, die meisten liegen aber über 9000 Einwohner. Der Sampling Point 031520009009 hat nur 6743 Einwohner. Dies ist aber auch schon so, wenn man in der Ausgangsdatei von Destatis die Zahl der REG_HW aufsummiert.

Die zweite Art von Daten sind:

X11_REGHW.RData bzw.
X11_REGHW_NAD.RData

Hier sind die Ergebnisse aller Szenarien für eine Kombination aus KAL- und FEB-Modell zusammengefasst, in diesem Beispiel jeweils für das erste Modell. Dieser Data Frame hat immer 47808 Zeilen und 14 bzw. 15 Spalten. Es sind hier Daten für alle Szenarien zusammengefasst. Wenn man sich aber davon den Teil für das Szenario 1 anschaut, dann kommt man wieder auf die Zahl von 1619 Beobachtungen.

Tabelle 11: Codes zu den Bundesländern

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

A.3 Verallgemeinerter Regressionsschätzer bei Destatis

In Tabelle 12 sind die von Destatis verwendeten Hilfsvariablen enthalten.

=1 (Konstante)	
Bemeldete Anschrift (1=ja, 0=nein)	
Gemeldete Personen insgesamt	
Geschlecht / Staatsangehörigkeit	Deutsch, männlich Deutsch, männlich Nicht-deutsch, männlich
Familienstand	Ledig und unbekannt Verheiratet und Lebenspartnerschaft Verwitwet und Lebenspartner verstorben
Alter	Unter 6 6 bis unter 18 18 bis unter 25 25 bis unter 30 30 bis unter 40 40 bis unter 50 50 bis unter 60 60 bis unter 65
Erwerbstätigkeit	SV-pflichtig und geringfügig Beschäftigte (ERWERB=1,4) Beamte, Richter und Soldaten (ERWERB=2) Arbeitslose und Personen in Umschulungsmaßnahmen (ERWERB=3)

Tabelle 12: Hilfsmerkmale des GREG-Schätzers für die Einwohnerzahl