

Methodenbericht

Volkszählung 2000

0

Statistische Grundlagen und Übersichten
Bases statistiques et produits généraux
Basi statistiche e presentazioni generali

Statistische
Einsetzungsverfahren

Band 1

Die vom Bundesamt für Statistik (BFS) herausgegebene Reihe «Statistik der Schweiz» gliedert sich in folgende Fachbereiche:

La série «Statistique de la Suisse» publiée par l'Office fédéral de la statistique (OFS) couvre les domaines suivants:

0 Statistische Grundlagen und Übersichten

0 Bases statistiques et produits généraux

1 Bevölkerung

1 Population

2 Raum und Umwelt

2 Espace et environnement

3 Arbeit und Erwerb

3 Vie active et rémunération du travail

4 Volkswirtschaft

4 Economie nationale

5 Preise

5 Prix

6 Industrie und Dienstleistungen

6 Industrie et services

7 Land- und Forstwirtschaft

7 Agriculture et sylviculture

8 Energie

8 Energie

9 Bau- und Wohnungswesen

9 Construction et logement

10 Tourismus

10 Tourisme

11 Verkehr und Nachrichtenwesen

11 Transports et communications

12 Geld, Banken, Versicherungen

12 Monnaie, banques, assurances

13 Soziale Sicherheit

13 Protection sociale

14 Gesundheit

14 Santé

15 Bildung und Wissenschaft

15 Education et science

16 Kultur, Informationsgesellschaft, Sport

16 Culture, société de l'information, sport

17 Politik

17 Politique

18 Öffentliche Verwaltung und Finanzen

18 Administration et finances publiques

19 Kriminalität und Strafrecht

19 Criminalité et droit pénal

20 Wirtschaftliche und soziale Situation
der Bevölkerung

20 Situation économique et sociale
de la population

21 Nachhaltige Entwicklung und Disparitäten
auf regionaler und internationaler Ebene

21 Développement durable et disparités régionales
et internationales

Volkszählung 2000

Statistische Einsetzungsverfahren

Band 1

Autoren

Daniel Kilchmann, Philippe Eichenberger, Jann Potterat
Bundesamt für Statistik

Herausgeber

Bundesamt für Statistik

Vorwort

Innerhalb der Volkszählung 2000 (VZ2000) mussten fehlende Werte und widersprüchliche Daten in der Personen- und Haushaltserhebung (PHE) und der Gebäude- und Wohnungserhebung (GWE) bearbeitet werden. Diese Arbeiten wurden methodisch vom Dienst Statistische Methoden (METH) des Bundesamtes für Statistik begleitet und zu einem grossen Teil auch durchgeführt. Der Methodenbericht über die statistischen Einsetzungsverfahren ist in zwei Bände aufgeteilt und deckt die Arbeiten von METH ab.

Der vorliegende erste Band des Methodenberichts beinhaltet Erläuterungen bezüglich der Definitionen und Methoden von Einsetzungsverfahren. Zudem werden die statistischen Einsetzungsverfahren beschrieben, die in der PHE angewandt wurden.

Die methodischen Arbeiten wurden in erster Linie von Daniel Kilchmann zwischen Juli 2000 bis August 2003 durchgeführt. Methodisch unterstützt wurde er von Philippe Eichenberger und Jann Potterat (beide METH).

Die Autoren danken Thomas Alabor, Estelle Crippa, Claude Gisiger, Claude Grandjean, Katrin Imwinkelried, Pierre-Alain Pavillon, Claude Joye und Carl-Anton Schlaepfer der ehemaligen Sektion Bevölkerungs- und Haushaltsstruktur (BHS) für die fachliche Unterstützung und die gute Zusammenarbeit. José Navarro (trivadis SA), Frédéric von Kessel und Frédéric Reinhard (BFS) sei für die Erarbeitung und die Anwendung der Informatikchnittstellen gedankt.

Beat Hulliger (METH) sei für die hilfreichen Diskussionen, die fachliche Unterstützung und die Durchsicht des Berichts herzlich gedankt.

Zusammenfassung

In der Volkszählung 2000 (VZ2000) wurden fehlende Werte und widersprüchliche Daten durch Rückfragen, manuelle Bearbeitungen, deterministische Korrekturen, die Einsetzung eines nahen Nachbarn und durch modellbasierte Einsetzungsmethoden bearbeitet. Die Arbeiten des Dienstes Statistische Methoden (METH) beschränkten sich auf die statistischen Einsetzungsverfahren (Naher Nachbar Einsetzung, Modellbasierte Einsetzung).

Der vorliegende erste Band des Methodenberichts befasst sich mit den Definitionen der Lokalisierung fehlender und fehlerhafter Daten sowie von Einsetzungsverfahren. Die Erarbeitung und Anwendung statistischer Einsetzungsverfahren innerhalb der Personen- und Haushaltserhebung der VZ2000 wird ebenso im ersten Band beschrieben wie das Vorgehen und die Evaluation der Einsetzungen.

Im zweiten Band werden die statistischen Einsetzungsverfahren, die in der Gebäude- und Wohnungserhebung angewandt wurden, beschrieben. Zudem enthält der zweite Band eine Zusammenfassung der von METH durchgeführten Einsetzungsverfahren in der VZ2000 sowie die Schlussfolgerung beider Bände.

Es hat sich gezeigt, dass statistische Einsetzungsverfahren eine beträchtliche Vorbereitung voraussetzen. In der VZ2000 konnten vollständige Daten nur anhand von statistischen Einsetzungsverfahren erreicht werden. Der Einfluss der Einsetzungen hängt stark von der Qualität der vorangegangenen Prozesse der Erhebung ab. Unter Umständen können ungewollte Merkmalskombinationen durch statistische Einsetzungsverfahren vermehrt zu Tage treten.

Schlüsselwörter

Methodenbericht; Volkszählung, VZ2000, Schlussbereinigung, SB, Antwortausfall, Inkonsistenz, Fehlerlokalisierung, Einsetzungsverfahren, Fellegi-Holt, Korrekturen, Änderungen, Hot-Deck, NIM, Nearest neighbor Imputation Methodologie, Naher Nachbar Einsetzung, Empfänger, Spender, Modellbasierte Einsetzungsmethoden, Deterministische Korrekturen, CORA.

Auskunft:	Daniel Kilchmann, Tel. 032 713 60 84 Daniel.Kilchmann@bfs.admin.ch
Bearbeitung:	Dienst Statistische Methoden, BFS
Vertrieb:	Bundesamt für Statistik CH-2010 Neuchâtel Tel. 032 713 60 60 / Fax 032 713 60 61 Order@bfs.admin.ch
Internet:	http://www.statistik.admin.ch
Bestellnummer:	338-0032
Preis:	gratis
Reihe:	Statistik der Schweiz
Fachbereich:	0 Statistische Grundlagen und Übersichten
Originaltext:	Deutsch
Grafik/Layout:	BFS
Copyright:	BFS, Neuchâtel 2005 Abdruck – ausser für kommerzielle Nutzung – unter Angabe der Quelle gestattet.
ISBN:	3-303-00322-X

Inhaltsverzeichnis

1	Einführung PHE	5
2	Kontext	5
3	Methoden der Fehlerlokalisierung und Einsetzung	6
3.1	Definitionen	6
3.2	Fehlerlokalisierung	7
3.3	Einsetzungsverfahren	9
3.4	Analyse von Einsetzungsverfahren	11
4	Funktionalitäten von bestehenden Kontroll- und Einsetzungssystemen	15
4.1	EDIS	15
4.2	NIM (CANCEIS)	17
4.3	Systemvergleich	18
4.4	Das Programm NIM	19
5	Vergleich deterministischer Korrekturen mit NIM	22
5.1	Datenbasis, Anpassungen, Kontrollregeln, Fehlergenerierung	22
5.2	Evaluation	25
5.3	Fazit	29
6	Einsetzungen in der PHH, SHH und KHH der VZ2000	31
6.1	Fragebogen PHE	31
6.2	Einsetzungen mit NIM-P	32
6.2.1	Daten, Variablen und Schichtung	32
6.2.2	Anpassungen an NIM	33
6.2.3	Produktion und Analysen von NIM-P	34
6.3	KHH mit NIM-K	42
6.4	Bemerkungen zur Anwendung von NIM in der PHE	42
6.5	Modellbasierte Einsetzungen	44
Anhang		46
A	PHH/SHH/KHH der VZ2000	46
A.1	Variablenabkürzungen	46
A.2	Kontrollregeln	46
A.3	KHH-Typen	50
A.4	Distanzen und Parameter	52
A.5	Einsetzungen	54
A.5.1	Einsetzung mit NIM-P: Schichten und Variablen	54
A.5.2	Einfluss von NIM-P auf die Verteilungen	60
A.5.3	KHH	69
A.6	MBE in der PHE	70
B	Programme PHE	76

1 Einführung PHE

Die Volkszählung 2000 (VZ2000) wurde mit Stichtag 5.12.2000 vom Bundesamt für Statistik (BFS) durchgeführt. Der Dienst Statistische Methoden (METH) des BFS wurde von der ehemaligen Sektion Bevölkerungs- und Haushaltsstruktur (BHS) des BFS angefragt, die Bearbeitung von fehlenden und inkonsistenten Werten innerhalb des Teilprojekts Schlussbereinigung (SB) der Volkszählung zu evaluieren und die Plausibilisierung der Personen- und Haushaltserhebung (PHE) sowie der Gebäude- und Wohnungserhebung (GWE) methodisch zu begleiten. Die Antwortausfälle in den demografischen Variablen der PHE und den Gebäudevariablen wurden zu einem grossen Teil von METH bearbeitet. Ebenso wurden Teile der Wohnungsveränderungen von METH bearbeitet.

Der Bericht über die statistischen Einsetzungsverfahren in der Volkszählung 2000 wurde aus technischen Gründen in zwei Bände aufgeteilt. Der vorliegende Band befasst sich mit der Personen- und Haushaltserhebung, der zweite Band befasst sich mit der Gebäude- und Wohnungserhebung. Zudem befindet sich im zweiten Band eine Zusammenfassung beider Bände, die Schlussfolgerung bezüglich der statistischen Einsetzungsverfahren in der VZ2000, sowie das Literaturverzeichnis beider Bände.

In einem ersten Abschnitt des vorliegenden Bandes wird der Kontext erläutert. Darauf folgt ein Abschnitt in welchem allgemein auf Einsetzungsverfahren eingegangen wird, die in der Volkszählung zur Anwendung kamen. Danach werden die Funktionalitäten von zwei Computer-Applikationen für die Verarbeitung von Haushaltsdaten in Volkszählungen vorgestellt, welche im Vorfeld der Bearbeitung der Antwortausfälle in der VZ2000 evaluiert wurden. Im vierten Abschnitt werden deterministische Korrekturen (CORA) der Computer-Applikation NIM¹ gegenübergestellt. Diese Gegenüberstellung erstreckte sich über den Zeitraum Dezember 2000 bis Mai 2001. Im darauffolgenden Abschnitt wird die Bearbeitung der Antwortausfälle in der PHE anhand statistischer Einsetzungsverfahren dokumentiert. Dabei wurde vor allem die angepasste Computer-Applikation NIM sowie deterministische Korrekturen verwendet. In diesem Bericht wird nicht näher auf die innerhalb der VZ2000 angewandten deterministischen Korrekturen eingegangen. Die angepasste Computer-Applikation NIM wurde im Juli/August 2002 innerhalb der PHE angewandt.

2 Kontext

Die Volkszählung ist eine Vollerhebung, die alle 10 Jahre seit 1850 durchgeführt wird.

Der Prozess, der durchgeführt wird, um von den eingegangenen Fragebogen zu den Daten für Analysen zu gelangen, wird hier in zwei Teile unterteilt, die **Aufarbeitung** und die **Schlussbereinigung**² (SB).

Die Aufarbeitung der VZ2000 beinhaltete alle Arbeiten von der Eingangskontrolle der Fragebogen und dem Einscannen der Daten bis zur vollständigen Definition der Einheiten. Insbesondere wurden die Daten in mehreren Schritten auf ihre Vollständigkeit, das heisst, sämtliche Einheiten müssen vorhanden sein, die Vollständigkeit der Antworten und die Konsistenz der Antworten³ überprüft und dementsprechend bearbeitet. Ebenso wurden die verschiedenen Einheiten miteinander verknüpft. Zum Beispiel wurden die Personen mit Haushalten, die Haushalte mit Wohnungen, die Personen mit mehreren Wohnsitzen untereinander, verknüpft. Die Codierungen von Variablen (Beruf, Ausbildung, etc.) werden hier ebenfalls in die Aufarbeitung miteinbezogen.

In groben Zügen zusammengefasst, hat die Aufarbeitung zum Ziel, die Vollständigkeit der Einheiten zu sichern, und die wichtigsten Variablen so weit möglich durch Rückfragen bei den Befragten oder bei öffentlichen Ämtern und durch manuelle Korrekturen von Erfassungsfehlern zu vervollständigen. Rückfragen von fehlenden Variablenwerten können aus Kostengründen nicht für sämtliche Variablen und alle

¹NIM steht für *Nearest neighbor Imputation Methodology*

²auch Schlussbereinigungen

³im Sinne von widerspruchsfrei.

Einheiten durchgeführt werden. Deshalb wurde für die Aufarbeitung das Schwergewicht auf die Vollständigkeit der Einheiten und die Vervollständigung der demografischen Merkmale und der Fragen des Erwerbslebens gelegt.

In den Schlussbereinigungen wurden vor allem Antwortausfälle von Variablen und Inkonsistenzen bereinigt. Diese Bereinigungen wurden mehrheitlich anhand von Computer-Applikationen durchgeführt. Dies hatte vor allem den Vorteil, dass die Bearbeitungen wiederholbar sind, im Gegensatz zu manuellen Bearbeitungen. Denn manuelle Bearbeitungen basieren immer auf subjektiven Entscheiden. Hingegen hatten die manuellen Bearbeitungen in der Aufarbeitung den Anspruch, die individuell 'richtigen' Antworten zu finden, wo hingegen in der SB die statistischen Resultate und die Konsistenz im Vordergrund standen. Aus diesem Grund wurden keine Daten, die von statistischen Einsetzungen stammen, an Register geliefert.

In der SB wurde von vollständigen Einheiten ausgegangen und von vollständigen und wahrheitsgetreuen Verknüpfungen der Einheiten.

Nach der SB wurden Synthesevariablen gebildet, wie zum Beispiel 'Haushaltstypologie', 'Referenzperson im Haushalt', 'Ehepaarnummer', 'Höchste abgeschlossene Ausbildung' etc. Auf die Synthesevariablen wird in diesem Bericht nicht näher eingegangen.

3 Methoden der Fehlerlokalisierung und Einsetzung

Es werden kurz die verschiedenen Methoden vorgestellt, die in der Volkszählung 2000 angewandt wurden, und die in den folgenden Abschnitten dieses Berichts wieder erwähnt werden. Zunächst werden im Folgenden Abschnitt grundlegende Begriffe definiert.

3.1 Definitionen

Eine **Einheit** ist ein Element der Population, zum Beispiel eine Person, ein Haushalt, ein Gebäude, eine Wohnung. Für die Überprüfung der Vollständigkeit und der 'Konsistenz' der Daten werden so genannte **Kontrollregeln**, auch **Plausibilitätsregeln** und **Plausibilitätstests** genannt, verwendet. Fehlen sämtliche Angaben einer Einheit, von der man weiss, dass sie existiert, so handelt es sich um einen **Antwortausfall einer Einheit**⁴. In der SB wurden keine Antwortausfälle von Einheiten bearbeitet. Im Gegenteil, es wurde davon ausgegangen, dass diese durchgeführt worden war.

Widersprechen sich zwei oder mehrere Variablenwerte oder deren Kombinationen, so wird dies **Inkonsistenz** genannt. Die betroffenen Variablen sind **inkonsistent** oder **fehlerhaft**. Ein Variablenwert **fehlt**, wenn eine Variable, die obligatorisch beantwortet werden muss, nicht beantwortet wurde. Wenn eine Variable aufgrund der Antwort anderer Variablen fehlt, so wird dieser Wert zu den Inkonsistenzen gezählt. Wenn zum Beispiel angegeben wurde, dass eine Person 9-jährig ist, aber keine Angaben zur derzeitigen Ausbildung gemacht hat, so handelt es sich um eine Inkonsistenz. Es ist durchaus möglich, dass aufgrund der Antworten zum Erwerbsleben, der Stellung im Haushalt, etc., die Wahrscheinlichkeit gross ist, dass das Alter bezüglich dieser Antworten falsch ist. Fehlt jedoch das Alter handelt es sich immer um einen fehlenden Wert, da diese Variable obligatorisch ausgefüllt sein muss. Ein **ungültiger** Variablenwert ist ein Wert der ausserhalb der erlaubten Bandbreite liegt, zum Beispiel 'Anzahl Kinder' = 99. Unter Fehlerlokalisierung wird die Lokalisierung von fehlenden und inkonsistenten Werten verstanden. Fehler in den Daten bezeichnen hier fehlende, fehlerhafte oder ungültige Werte.

Eine **Korrektur** besteht darin, dass ein Wert korrigiert wird, so dass davon ausgegangen werden kann, dass der korrigierte Wert 'richtig' ist. Wird ein Variablenwert als 'richtig' eingestuft, so heisst das, dass der Befragte, sofern die Frage bezüglich des Variablenwerts keinen Handlungsspielraum offen lässt und

⁴auch 'fehlende Einheit'.

der Befragte die Frage verstanden hat und wahrheitsgetreu antwortet, er eben diesen Variablenwert als Antwort geben würde. Lässt die Fragestellung einen Handlungsspielraum offen oder ist sie für den Befragten unverständlich, so kann im oben erwähnten Sinn nie von 'richtigen' Antworten, sprich 'richtigen' Variablenwerten, gesprochen werden.

Eine **Änderung** ist eine Abänderung eines vorhandenen Variablenwerts, welcher aufgrund von Kontrollregeln als fehlerhaft erkannt wurde. Durch die Änderung wird die Kontrollregel nicht mehr verletzt. Die Änderung hat aber keinen Anspruch auf 'Richtigkeit'.

Eine **Einsetzung** ist streng genommen die Operation, mit der ein fehlender Wert ersetzt wird. In der Fachwelt wird der Begriff der Einsetzungen oft weiter gefasst, indem auch Änderungen von vorhandenen Variablenwerten als Einsetzung betrachtet werden (siehe (Poirier, 2000)). Wenn nicht anders vermerkt, wird die weiter gefasste Definition der Einsetzung in diesem Bericht verwendet.

Eine Einheit, die mindestens eine Kontrollregel verletzt, wird **Empfänger** genannt, denn sie wird im Verlaufe der Einsetzungen mindestens einen Variablenwert 'empfangen'. Der **Spender** ist die Einheit, welche verwendet wird, um eine Variable oder mehrere Variablen des Empfängers zu ersetzen, mit dem Ziel, dass der Empfänger nach der Ersetzung die Kontrollregeln besteht. Ob eine Einsetzung auf einem Spender beruht oder nicht, hängt von der gewählten Methode ab, siehe Abschnitt 3.

Bei der **Einheit-Einsetzung** wird der Empfänger vollständig durch einen Spender ersetzt. Hingegen werden bei der **Variablen-Einsetzung** nur die Variablen, welche aufgrund der Fehlerlokalisierung als fehlerhaft erkannt werden, vom Spender oder durch ein Modell übernommen. Fehlen sämtliche Variablen so ist die Variablen-Einsetzung eine Einheit-Einsetzung. Die Einheit-Einsetzung garantiert konsistente und realistische Einheiten, da es sich bei der eingesetzten Einheit um eine Kopie einer vorhandenen Einheit handelt, welche die Plausibilitätstests bestand. Fehlt eine von zwanzig Variablen, so wird jedoch die Information der 19 vorhandenen Variablen gelöscht und durch jene des Spenders ersetzt. Die Variablen-Einsetzung hat zum Ziel, ein Maximum an Information beizubehalten. Es ist nicht immer möglich, eine Einheit zu finden, die exakt die gleiche Kombination von Variablen hat wie der Empfänger, ausser der Variablen, die eingesetzt werden müssen. In solchen Fällen können durch die Variablen-Einsetzung Kombinationen von Variablen 'kreiert' werden, die vor den Einsetzungen nicht existiert haben.

Im Folgenden wird der Einfachheit halber der Begriff Einsetzung für die Variablen-Einsetzung verwendet, da, ausser bei Ausfällen ganzer Einheiten, keine Einheit-Einsetzung in der VZ2000 durchgeführt wurde.

Unter **Einsetzungsverfahren** wird hier das Verfahren genannt, anhand dessen fehlende und inkonsistente Daten eingesetzt werden. In den Einsetzungsverfahren ist die Lokalisierung von fehlenden und inkonsistenten Werten oft beinhaltet.

3.2 Fehlerlokalisierung

Lokalisierung von fehlenden und ungültigen Werten Wie im Abschnitt 3.1 bereits erwähnt, werden hier einzig Antwortausfälle von obligatorischen Variablen als fehlende Werte definiert. Diese, sowie ungültigen Variablenwerte werden mit deterministischen Regeln lokalisiert.

Sequentielle Fehlerlokalisierung Bei der **sequentielle Fehlerlokalisierung** werden die Kontrollregeln in einer vorgegebenen Reihenfolge nacheinander angewandt. Oft werden die Variablen dadurch ebenfalls in einer bestimmten Reihenfolge verarbeitet. Dies wird zur Veranschaulichung im Folgenden angenommen. Es seien die n Variablen $y_{[1]}, \dots, y_{[n]}$ in der Reihenfolge ihrer Verarbeitung geordnet, so dass die Fehlerlokalisierung der Variablen $y_{[k]}$ vor $y_{[l]}$ durchgeführt wird, wenn $k < l$. Die Lokalisierung der fehlerhafter Variablenwerte der Variablen $y_{[k]}$ kann sich bei der sequentiellen Fehlerlokalisierung auf

die zuvor verarbeiteten Variablen abstützen:

$$y_{[k]} \begin{cases} \neq f(y_{[1]}, \dots, y_{[k-1]}), & y_{[k]} \text{ ist fehlerhaft} \\ = f(y_{[1]}, \dots, y_{[k-1]}), & y_{[k]} \text{ ist nicht fehlerhaft} \end{cases} \quad (1)$$

Vor der Fehlerlokalisierung in der Variablen $y_{[k]}$ müssen bei diesem Vorgehen die fehlerhaften Variablenwerte der Variablen $y_{[1]}, \dots, y_{[k-1]}$ eingesetzt werden. Nur so können diese Variablen uneingeschränkt in die Fehlerlokalisierung von $y_{[k]}$ miteinbezogen werden. Somit wird davon ausgegangen, dass die Variablenwerte von $y_{[1]}, \dots, y_{[k-1]}$ 'richtig' sind, das heisst, sie werden in der Folge nicht mehr abgeändert. Folglich wird der Variablen $y_{[k]}$ weniger 'Wahrheitsgehalt' zugesprochen, als den Variablen $y_{[1]}, \dots, y_{[k-1]}$.

Diese Art der Fehlerlokalisierung ist deterministisch und hat den Anspruch, Fehler eindeutig bestimmen zu können. Die Reihenfolge, in welcher die Regeln angewandt werden, hat einen entscheidenden Einfluss auf die Fehlerlokalisierung und somit auf die Resultate der Einsetzungen.

Die sequentielle Fehlerlokalisierung ist angebracht, wenn es sich um fehlende oder ungültige Werte handelt und der Kontext **eindeutig** ist; z.B. ein Gebäude mit Stockwerkeigentum-Wohnungen, die eindeutig als solche eruiert werden konnten, muss den Wert 'ja' bei der Variablen 'Stockwerkeigentum' aufweisen. Zudem ist die sequentielle Fehlerlokalisierung angebracht, wenn der multivariate Charakter der Daten vernachlässigt werden kann.

Simultane Fehlerlokalisierung Die **simultane Fehlerlokalisierung** zeichnet sich dadurch aus, dass mehrere Kontrollregeln gleichzeitig an mehrere Variablen angewandt werden. Somit können mehrere Fehler bei mehreren Variablen gleichzeitig entdeckt werden. Die Lokalisierung des Fehlers erfolgt durch Kreuzung der Regeln. Je öfter eine Variable in einer Regel vorhanden ist, die nicht erfüllt wird, desto eher wird sie als fehlerhaft erkannt.

Bei der simultanen Fehlerlokalisierung, die eine nicht-sequentielle Fehlerlokalisierung ist, wird der 'Wahrheitsgehalt' der Variablen gemäss der Gewichtung der Variablen bestimmt.

Vorausgesetzt man möchte so wenig Daten wie möglich ändern, ist das Ziel der simultanen Fehlerlokalisierung, so wenig wie möglich Variablenwerte als fehlerhaft zu lokalisieren. Dies kann erreicht werden, indem ein lineares Programm gelöst wird (*Fellegi-Holt*, c.f. (Fellegi I. P., 1976)). Dies setzt jedoch grosse Rechnerressourcen voraus. Zudem muss ein zusätzliches Kriterium für die Fehlerlokalisierung für die Fälle verwendet werden, für die mehrere optimale Lösungen existieren. Die Existenz einer optimalen Lösung garantiert nicht, dass ein Spender gefunden werden kann, der zu einer zulässigen Lösung führt. Wird die minimale Anzahl Änderungen nur durch eine einzige Kombination von Änderungen erreicht, so ist die *Fellegi-Holt* Fehlerlokalisierung deterministisch.

Bei (Bankier, 2000) wird die Vorgehensweise gegenüber *Fellegi-Holt* umgekehrt, indem zuerst Spender gesucht werden, die eine zulässige Lösung generieren können. Danach werden die Spender, die die kleinste Anzahl Variablen verändern, das heisst, die kleinste Anzahl fehlerhafter Variablen lokalisieren, ausgesucht. Dadurch wird die optimale Lösung mit machbaren, sprich vorhandenen, Lösungen angenähert. Die Fehlerlokalisierung kann dann durch den nächsten Nachbarspender definiert werden oder durch einen Spender der sich in der nächsten Umgebung befindet. Dabei hängt die Fehlerlokalisierung von der Distanzdefinition und der Umgebung, in welcher ein Nachbar oder mehrere Nachbarn gesucht werden, ab. Die Ordnung der Daten kann bei der Verwendung des nächsten Nachbarn einen entscheidenden Einfluss auf die Fehlerlokalisierung haben, da meist nicht alle potentiellen Spender zur Fehlerlokalisierung verwendet werden.

Bei der simultanen Fehlerlokalisierung kann die gleiche Konstellation der Daten bei zwei Einheiten, welche die gleichen Kontrollregeln verletzen, somit zu einer unterschiedlichen Lokalisierung des Fehlers führen, ausser bei der *Fellegi-Holt* Methode, falls nur eine einzige Lösung existiert.

Die simultane Fehlerlokalisierung ist vor allem angebracht, wenn der Kontext nicht eindeutig ist und wenn der multivariate Charakter der Daten Beibehalten werden soll.

3.3 Einsetzungsverfahren

Deterministische Korrekturen Um eine deterministische Korrektur durchführen zu können, muss der Fehler zuvor eindeutig bestimmt worden sein. Die deterministische Korrektur des lokalisierten Fehlers basiert auf Erfahrungswerten und logischen Überlegungen. Die Konstellation der Variablen, die nicht verändert werden, bestimmt dabei den Wert der eingesetzt oder abgeändert wird. Diese Art von Korrekturen können mit 'if-then-else' zusammengefasst werden. In der VZ2000 wurden die deterministischen Korrekturen 'automatische Korrekturen' genannt, da es sich im Gegensatz zu manuellen Aufarbeitungsprozessen um einen automatisierten Aufarbeitungsprozess handelt. Diese beinhalten nicht nur die deterministische Korrektur sondern auch die sequentielle Fehlerlokalisierung. In diesem Bericht werden diese Fehlerlokalisierung-Korrekturen oft **CORA**⁵ genannt.

Es hat sich gezeigt, dass bei einer grossen Anzahl Variablen den deterministischen Korrekturen Grenzen gesetzt sind, da für sämtliche Kombinationen von Variablenwerten definiert werden muss, welche Korrektur vorgenommen werden muss.

Bei CORA muss bei sequentieller Verarbeitung der Variablen darauf geachtet werden, dass der Prozess nicht zirkulär wird. Dies hat zur Folge, dass der Handlungsspielraum für Korrekturen eingeschränkt wird, je mehr Variablen korrigiert, und somit als definitiv bezeichnet, wurden.

Deck-Einsetzungsmethoden Bei den *Deck*⁶-Methoden wird ein Spender verwendet um die fehlenden oder inkonsistenten Werte einzusetzen. Es werden zwei Methoden der *Deck*-Einsetzung unterschieden, einerseits die **Hot-Deck** Einsetzungsmethode und andererseits die **Cold-Deck** Einsetzungsmethode. Bei der Hot-Deck Einsetzungsmethode stammen der Spender und die fehlerhafte Einheit aus den gleichen Daten. Die Suche und Auswahl der Spender bestimmt dabei, ob es sich um eine deterministische Methode handelt oder nicht.

Bei der Cold-Deck Einsetzungsmethode wird der Spender aus einer anderen Datenquelle (Hilfsdatenquelle) entnommen. Zum Beispiel von Registern oder früheren Erhebungen. Die Cold-Deck Methode ist deterministisch, wenn die fehlerhafte Einheit in der Hilfsdatenquelle anhand eines gemeinsamen Identifikators identifiziert werden kann und die entsprechenden Werte übernommen werden (z.B. Übernahme des Wertes aus der letzten Erhebung 'Last Carry Forward').

Nachbar Einsetzungen Bei der Einsetzung eines Nachbarn wird aufgrund einer Distanz zwischen den Einheiten eine Nachbarschaft des Empfängers definiert, deren Mitglieder als Spender verwendet werden können. Die Definition der Distanz hängt normalerweise von der Anzahl Änderungen ab, welche durch den Spender hervorgerufen werden, damit der Empfänger alle Kontrollregeln erfüllt. Die Nachbarschaft wird auch **Spender-Pool** genannt. Die Definition des Spender-Pools und die Auswahl des Spenders geben den Ausschlag, ob diese Methode deterministisch oder stochastisch ist.

Die Einsetzung eines Nachbarn ist eine Hot-Deck Methode, sofern die Nachbarschaft in den gleichen Daten definiert ist, von denen der Empfänger stammt.

Im Folgenden werden zwei Nachbar Einsetzungsmethoden unterschieden, die **Nächster-Nachbar** Einsetzungsmethode und die **Naher-Nachbar** Einsetzungsmethode. Bei der Einsetzung des nächsten Nachbarn wird der nächste Nachbar bestimmt, der als Spender verwendet wird. Der Spender-Pool besteht somit aus einer einzigen Einheit. In (Bankier, 2000) sind Beispiele aufgeführt, wonach es in gewissen

⁵ Abkürzung aus dem französischen von 'corrections automatiques'.

⁶ Spielkarten, aus dem Englischen.

Fällen besser ist, nicht die minimale Anzahl Variablen abzuändern, sondern dass mehr Gewicht auf die Nähe der Variablenwerte zwischen dem Empfänger und dem Spender gelegt werden sollte.

Die Methode der Einsetzung eines nahen Nachbarn kann als abgeänderte Methode der nächster-Nachbar Einsetzungsmethode angesehen werden. Dabei wird ein Spender durch Zufallsauswahl aus der nächsten Nachbarschaft des Empfängers ausgewählt. Die nächste Nachbarschaft ist durch die Distanz zum Empfänger und der Anzahl Einheiten, die höchstens zur Nachbarschaft gehören dürfen, definiert. Die Zufallsauswahl des Spenders hängt normalerweise von der Distanz des Spenders zum Empfänger ab. Je näher der Spender ist, desto höher ist die Wahrscheinlichkeit, dass dieser für die Einsetzung ausgewählt wird.

Modellbasierte Einsetzungsmethoden Bei der **Modellbasierten Einsetzungsmethode (MBE)**, die in der VZ2000 zur Anwendung kam, muss die Fehlerlokalisierung bereits vorhanden sein. Danach wird anhand von Hilfsvariablen ein Modell postuliert, anhand dessen der plausible Wert geschätzt wird, der beim Empfänger eingesetzt wird. Die Hilfsvariablen im Modell müssen dabei vollständig vorhanden sein. Diese Methode basiert auf der Verteilung der vorhandenen Werte bedingt durch Hilfsvariablen.

Es gibt zahlreiche MBE, wie zum Beispiel Regressionsschätzer. Die in der VZ2000 verwendete Methode ist eine einfache MBE, die auf Zufallseinsetzungen von kategoriellen Variablen basiert. Wenn im Weiteren von der modellbasierten Methode gesprochen wird, so ist damit die folgend beschriebene gemeint.

In dieser MBE werden Einsetzungsklassen, resp. Schichten, gebildet, innerhalb derer die fehlenden Werte durch Zufallsauswahl eines Spenders eingesetzt werden. Die Zufallsauswahl ist dabei durch die Häufigkeit gewichtet. Allgemein kann dieses Vorgehen folgendermassen beschreiben werden:

Ω : Population (z.B. Personen, Gebäude, Wohnungen, Wohneinheiten).

R_z : Einheiten, die einen gültigen und plausiblen Wert für die Variable z angegeben haben.

y_k : beobachteter Variablenwert der Einheit k in der Variablen y , die es für bestimmte k einzusetzen gilt.

y^* : tatsächliche Variable (Variable mit den 'wahren' Werten).

$\hat{y}_k$: eingesetzter Wert von y_k^* .

$\mathbf{x}_k$: Vektor der p Hilfsvariablen $x_{k,1}, \dots, x_{k,p}$ der Einheit k .

Für die Einsetzung wird die bedingte Verteilung der Variablen y^* bedingt durch die Hilfsvariablen $\mathbf{x}$ verwendet, um die Abhängigkeiten von y^* zu $\mathbf{x}_k$ miteinzubeziehen (2).

$$y^*|\mathbf{x} \sim f(Y^*|\mathbf{X} = \mathbf{x}) \quad (2)$$

Die zufällige Einsetzung $\hat{y}_k$ beruht auf der geschätzten Verteilung $\hat{f}$ von f (3), sofern diese existiert. $\hat{y}_k$ ist eine zufällige Realisation von $\hat{f}(y|\mathbf{x}_k)$.

$$\hat{f}(Y^*|\mathbf{X} = \mathbf{x}) = f_{R_y}(Y|\mathbf{X} = \mathbf{x}) \quad (3)$$

wobei $f_{R_y}()$ die Häufigkeitsfunktion der Variablen Y aufgrund der gültigen und plausiblen Werte von Y^* ist, und die Werte der Hilfsvariablen $\mathbf{X}$ bekannt sind.

Mit der modellbasierten Methode können multivariate Abhängigkeiten miteinbezogen werden. Theoretisch ist es möglich, sämtliche Variablen, die eingesetzt werden müssen als abhängige Variablen in das Modell aufzunehmen. Dies kommt der Methode des nächsten Nachbarn gleich, bei der die Distanz zwischen Empfänger und Spender zwingend 0 sein muss, damit eine Einsetzung erfolgt. Das heisst, sämtliche Hilfsvariablen müssen beim Spender und Empfänger übereinstimmen. Fehlen beim Empfänger alle Variablen kommt die oben beschriebene modellbasierte Einsetzung einer Zufallseinsetzung gleich.

Die vorhandenen Variablen müssen ins Modell miteinbezogen werden, um die Konsistenz zwischen den vorhandenen Werten und den eingesetzten Werten zu garantieren. Die Konsistenz der eingesetzten Werte mit Variablen, die nicht als Hilfsvariablen verwendet wurden, kann nicht garantiert werden.

Für jede Kombination fehlender Werte muss ein Modell beschrieben werden. Bei n Variablen, die $C_1, \dots, C_n$ Kategorien aufweisen, müssen somit $\prod_{i=1}^n C_i$ Modelle beschrieben werden. Später wird aufgezeigt, dass im ersten Teil der Gebäudeerhebung 9 Variablen eingesetzt werden müssen. Dies führt theoretisch zu 21'726'720 Modellen. Einige Modelle wären sicherlich entfallen, da die Kombination in den Daten nicht vorhanden gewesen wäre. Wenn die Einheiten Haushalte sind, muss bei diesem Vorgehen die Anzahl Kategorien der Personenvariablen zusätzlich noch mit der Anzahl Personen im Haushalt kombiniert werden.

Aus diesen Beispielen geht hervor, dass diese Methode nur für eine kleine Anzahl Variablen angewandt werden kann, welche im Idealfall nur wenig Kategorien aufweisen.

Stepwise-backward modellbasierte Methode Die **S-bMBE** Methode wird vor allem angewandt, wenn die Hilfsvariablen des Modells fehlende Werte aufweisen, oder mit dem ganzen Modell nicht alle Variablenkombinationen der Empfänger abgedeckt werden können. In letzterem Fall heisst das, dass es Empfänger gibt, die Kombinationen von Variablenwerten aufweisen, welche kein Spender hat. Bei der *stepwise-backward* modellbasierten Methode wird im Modell (2) die Dimension von $\mathbf{X}$ iterativ verkleinert, indem bei jedem Schritt eine oder mehrere Hilfsvariablen aus dem Modell ausgeschlossen werden. In den SB der VZ2000 wurden die Variablen von den thematischen Spezialisten in Bezug zur einzusetzenden Variablen nach der Wichtigkeit ihres Einflusses geordnet. Aufgrund dieser Ordnung wurden die Variablen mit dem kleinsten Einfluss zuerst aus dem Modell entfernt. Zudem wurde jeweils untersucht, wieviele Empfänger gültige Werte in den Hilfsvariablen aufwiesen, um jene Variablen aus dem Modell zu entfernen, die bei den Empfängern nicht vorhanden waren.

Durch die Reduktion der Anzahl Variablen im Modell, besteht die Gefahr, dass Hilfsvariablen, die mit der einzusetzenden Variablen korrelieren aus dem Modell entfernt werden. Dies führt zu einem Anstieg des Bias, der hier nicht geschätzt wird. Bei dieser Methode sind die Einsetzungsklassen nicht disjunkt, da jeder Spender in mehreren Schritten verwendet werden kann.

3.4 Analyse von Einsetzungsverfahren

Die folgenden Kriterien für die Analyse der Einsetzungsverfahren basieren grundsätzlich auf jenen die in (Chambers, 2001) entwickelt wurden (siehe auch Internet Seite).

Beschreibende Analyse Die beschreibende Analyse kann oft nur einen Aufschluss über die Qualität des Einsetzungsverfahrens geben, wenn dieses mit einem oder mehreren alternativen Einsetzungsverfahren verglichen wird. Folgende drei Analyse Kriterien werden für die beschreibende Beurteilung der Einsetzungen verwendet

1. **Anzahl Änderungen**
2. **Anzahl Fehler, die nicht behandelt werden konnten**
3. **Plausibilität.**

Das erste Kriterium zielt darauf ab, die Einhaltung des *Fellegi-Holt* Prinzips zu überprüfen. Hierfür muss die minimale Anzahl Änderungen bekannt sein, die mit den vorhandenen Spendern zu fehlerfreien Daten führen. Diese Anzahl muss nicht mit der minimalen Anzahl Änderungen übereinstimmen, welche dazu führen würde, dass die Daten fehlerfrei sind, da die entsprechende Einsetzung nicht immer durchgeführt werden kann (fehlender Spender). Somit ist dieses Kriterium nicht hart, sondern kann nur als Kriterium in einem Vergleich verwendet werden, um eine Aussage über die Qualität des Einsetzungsverfahrens zu machen.

Das zweite Kriterium kann auf ein methodisches Problem hinweisen, da es zum Beispiel bei deterministischen Korrekturen nicht in jedem Fall realistisch ist, für sämtliche Kombinationen von Variablen eine Korrektur Regel zu definieren, sofern diese überhaupt existiert. Bei Einsetzungsmethoden, die auf einer Spendersuche basieren, kann dieses Kriterium auf fehlende Spender hinweisen.

Das dritte Kriterium erscheint auf den ersten Blick wenig Sinn zu machen, werden doch Einsetzungsverfahren durchgeführt, um die Plausibilität der Daten zu erhöhen. Dieses Kriterium kann jedoch auch, wie das zweite Kriterium, auf fehlende Korrekturregeln oder fehlende Spender hinweisen, wenn deterministische Korrekturen angewandt werden. Ansonsten kann dieses Kriterium auch auf widersprüchliche Kontrollregeln hinweisen.

Analyse unter Einbezug der 'Wahrheit' Unter der 'Wahrheit' werden die wahren Werte verstanden. Diese sind nie bekannt, aber oft wird diese Wahrheit konstruiert, indem ein Einsetzungsverfahren durchgeführt wird, oder aber, man geht davon aus, dass diese in Registerdaten vorhanden ist. Um die Qualität eines Einsetzungsverfahrens objektiv analysieren zu können, müssen in den 'wahren' Daten realistische Fehler generiert werden. Dies erlaubt es, die

1. **Beibehaltung der Verteilung**
2. **Wiederherstellung der wahren Werte**

zu analysieren.

Beide Kriterien hängen von der Methode ab, anhand welcher die Wahrheit konstruiert wurde.

1. Beibehaltung der Verteilung Um die Beibehaltung der Verteilung zu evaluieren, wird ein Kriterium aus (Chambers, 2001) verwendet. Es sei $\mathbf{S} \in \mathbb{N}^c$ die Randverteilung der beobachteten Variablen Y^* . $\mathbf{R} \in \mathbb{N}^c$ sei die entsprechende Randverteilung nach der Einsetzung/Korrektur, also von $\hat{Y}$. $\mathbf{T} \in \mathbb{N}^c \times \mathbb{N}^c$ sei die Kreuzmatrize der c Kategorien bezüglich $\hat{Y}$ und Y^* , $\mathbf{Q} = \text{diag}(\mathbf{R} + \mathbf{S}) - \mathbf{T} - \mathbf{T}^t$. Um die Beibehaltung der Verteilung zu testen, kann eine Wald Statistik angewandt werden:

$$W_Y = (\mathbf{R} - \mathbf{S})^t \mathbf{Q}^{-1} (\mathbf{R} - \mathbf{S}) \quad (4)$$

Die Nullhypothese lautet: H_0 : Die Randverteilung nach der Einsetzung/Korrektur ($\mathbf{R}$) ist die gleiche wie jene der Wahrheit ($\mathbf{S}$).

Bei einer grossen Population ist: $W_Y \sim \chi_c^2$. Bei grossen Werten von W_Y wird H_0 verworfen.

2. Wiederherstellung der wahren Werte Die Notationen und das Vorgehen für die Evaluierung der Wiederherstellung der wahren Werte basieren auf (Chambers, 2001). Es sei

$RI_Y = \sum I(\hat{Y}_i = Y_i^*)$ die Anzahl der Einsetzungen/Korrekturen, welche den wahren Werten entsprechen.

$FI_Y = \sum I(\hat{Y}_i \neq Y_i^*)$ die Anzahl der Einsetzungen/Korrekturen, welche nicht den wahren Werten entsprechen. Der Anteil an richtigen Einsetzungen/Korrekturen an der gesamten Anzahl Einsetzungen/Korrekturen ist $A_Y = \frac{RI_Y}{RI_Y + FI_Y}$. Der Anteil der falsch eingesetzten/korrigierten Werte wird als Qualitätsmass der Wiederherstellung der wahren Werte verwendet:

$$D_Y = \frac{FI_Y}{n} \quad (5)$$

Wenn die Einsetzung/Korrektur die 'wahren Werte' wiederherstellt, dann ist $D_Y = 0$. Unter der Voraussetzung, dass die Randverteilungen beibehalten werden, kann die Varianz von D_Y folgendermassen geschätzt werden, $\hat{V}(D_Y) = \frac{1-D_Y}{n}$.

Vergleich mit anderen Erhebungen Bei dieser Art Analysen wird die 'Wahrheit' durch Daten ersetzt, von denen man weiss, dass sie nicht der Wahrheit entsprechen. Somit kann die Qualität des Einsetzungsverfahrens anhand anderer Erhebungen nicht quantifiziert werden. Die Auswahl der Vergleichserhebungen spielt hierbei eine zentrale Rolle, da der Vergleich durch eine schlechte Auswahl verzerrt wird. Oft stehen jedoch nicht beliebig viele Erhebungen zur Verfügung, um sie für einen Vergleich heranzuziehen. Zudem können oft nur Aggregate verglichen werden.

Aus diesen Gründen wird darauf verzichtet, Qualitätsmasse wie sie im Abschnitt 3.4 beschrieben wurden, zu verwenden, um den Vergleich zu quantifizieren. Hingegen sollten, sofern möglich, Randverteilungen mit anderen Erhebungen analysiert werden, um sehr grobe Fehlverhalten des Einsetzungsverfahrens zu eruieren.

Varianzschätzer Durch die Einsetzungsmethode bedingt, kann die Varianz geschätzt werden, welche durch den Einsetzungsprozess hinzugefügt wurde. Dies ist nur bei Methoden möglich, denen ein stochastisches Modell zugrunde liegt. Bei deterministischen Methoden ist es nicht möglich die Varianz zu schätzen, da sie den Schätzungen einen Bias und keine Varianz hinzufügen. Bei der *stepwise-backward* modellbasierten Methode wurde die Varianz nicht geschätzt, da dies aus Zeitgründen nicht möglich war, und da der Aufwand bei mehreren Schritten zu gross war.

Im Folgenden wird ein Varianzschätzer definiert, für dessen Schätzung davon ausgegangen wird, dass die Einsetzungen durch NIM gemacht wurden. Dieser Varianzschätzer kam aus Zeitgründen erst bei der Verarbeitung der Gebäudevariablen zur Anwendung.

Es sei Ω : die Population.

S_h : der Spenderpool in der Schicht h .

E_h : der Pool der Empfänger in der Schicht h .

d_{k,k_1} : die Distanz $k \in E_h$ und $k_1 \in S_h$. Anhand dieser Distanz werden die Spender geordnet: $d_{k,[1]} \leq d_{k,[2]} \leq \dots \leq d_{k,[|S_h|]}$

$\mathcal{N}_n(k)$: die nächste Nachbarschaft der Einheit k von der Grösse n , wobei $0 < n \leq n_{\max}$ = wählbarer Parameter. $\mathcal{N}_n(k) = \{r \in S_h | r \leq n, d_{k,[1]} \leq d_{k,r} \leq d_{k,[n]}\}$

y_k : der Variablenwert von y der Einheit k . Es wird davon ausgegangen, dass die Variable y eine kategoriale Variable mit C_y Kategorien ist, $y(1), \dots, y(C_y)$. Es sei die Indikatorvariable

$$i_{y_r}(y(c)) = \begin{cases} 1 & \text{wenn } y_r = y(c), \quad c \in \{1, \dots, C_y\} \\ 0 & \text{andernfalls.} \end{cases} \quad (6)$$

Die Wahrscheinlichkeit, dass die Einheit r aus $\mathcal{N}_n(k)$ als Spender dient, ist umgekehrt proportional zur Distanz zum Empfänger $d_{k,r}$:

$$p_k(r) = d_{k,r}^{-1} \left(\sum_{k_1=1}^n d_{k,[k_1]}^{-1} \right)^{-1}, \quad k \in E_h, r \in \mathcal{N}_n(k).$$

Für jede Variable y und ihre zugehörigen Kategorien wird eine Varianz geschätzt, welche durch die Einsetzung hervorgerufen worden ist. Die Wahrscheinlichkeit, dass dem Empfänger k die Kategorie c in der Variable y eingesetzt wird, ist $p_{y,c,k} = \sum_{r \in \mathcal{N}_n(k)} p_k(r) i_{y_r}(y(c))$. Das Ereignis 'dem Empfänger k wird

die Kategorie c in der Variablen y eingesetzt', ist eine Realisierung einer Bernoulli-Zufallszahl mit der Wahrscheinlichkeit $p_{y,c,k}$, dass das Ereignis eintritt. Unter der Voraussetzung, dass die Einsetzungen unabhängig voneinander sind, ist die Varianz für die Anzahl Einsetzungen in der Kategorie c der Variablen y

$$Var(y(c)) = \sum_h \sum_{k \in E_h} p_{y,c,k} (1 - p_{y,c,k}) \quad (7)$$

$Var(y(c))$ bezieht sich nur auf die Einsetzungen, die Mehrheit der Daten erfahren aber keine Änderungen oder Einsetzungen. Um die Genauigkeit der Endresultate zu schätzen, müssen diese auch in die Schätzung miteinbezogen werden. Der Variationskoeffizient des Schätzers der Anzahl Ausprägungen von $y(c)$ ist somit

$$cv(y(c)) \doteq \frac{\sqrt{Var(y(c))}}{\sum_{r \in \Omega} i_{y_r}(y(c))} \quad , \text{ wobei } c \in \{1, \dots, C_y\}, \forall y. \quad (8)$$

Einfache Ausreisser-Lokalisierung der Entwicklung Um überdurchschnittliche Veränderungen einzelner Merkmale im Vergleich zur VZ90 zu eruieren, wurde auf Gemeindeebene eine einfache Ausreisser-Lokalisierung durchgeführt. Diese Analyse wurde vor der Einsetzung und nach der Einsetzung durchgeführt. Zuerst wurde die relative und die absolute Veränderung untersucht. Da die absolute Entwicklung heterogener war als die relative, wurde schlussendlich nur die relative Veränderung verwendet. Im Folgenden wird die Methode vorgestellt, welche für die Lokalisierung von Ausreissern der relativen Veränderung verwendet wurde. Es sei

$N_{i,00}$: die Anzahl Einheiten der Gemeinde i in der VZ2000.

$N_{i,90}$: die Anzahl Einheiten der Gemeinde i in der VZ90.

$f_{y,i,j,k}$: die relative Häufigkeit der Kategorie k der Variablen y , der Gemeinde i in der VZ j , $j \in \{90, 00\}$.

C_y : die Anzahl Kategorien der Variablen y , die in der VZ90 und in der VZ2000 vorhanden sind.

$h_{y,i,j,k}$: die absolute Häufigkeit der Kategorie k der Variablen y , der Gemeinde i in der VZ j , $j \in \{90, 00\}$.

Die Gemeinden wurden in 5 Grössenklassen aufgrund von $f_{y,i,90,k}$ anhand eines Klassifikationsbaums aufgeteilt. Diese Klassifikation wurde anhand der beiden SAS Prozeduren `proc cluster` (Median Methode) und `proc tree` gebildet. $Cl_1, \dots, Cl_5$ bezeichnen die 5 Grössenklassen der Gemeinden. Somit wird die Hypothese vorausgesetzt, dass Gemeinden mit dem gleichen Anteil einer Variablenkategorie in der VZ90 sich ungefähr gleich verändern und somit in der VZ2000 wiederum einen ähnlichen Anteil aufweisen.

Um grosse relative Veränderungen zu glätten, wurde der Logarithmus der relativen Veränderung innerhalb jeder Kategorie k von y verwendet $dr_{y,i,k} = \log \left(\frac{f_{y,i,00,k}}{f_{y,i,90,k}} \right)$.

$dr_{y,i,k}$ wurde innerhalb jeder Kategorie von y robust zentriert und standardisiert:

$$\tau_{y,i,k} = \frac{0.6745 \left(dr_{y,i,k} - \underset{c \in Cl_h}{\text{med}}(dr_{y,c,k}) \right)}{\underset{c}{\text{MAD}}(dr_{y,c,k})} \quad (9)$$

Der Multiplikator '0.6745' dient dazu, dass $\frac{\underset{c \in Cl_h}{\text{MAD}}(dr_{y,c,k})}{0.6745}$ ein konsistenter Schätzer der Standardabweichung einer Normalverteilung ist.

Es sei N_h die Anzahl Gemeinden in der Grössenklasse h und $\kappa_{y,h,k} > 0$ die obere Grenze, so dass die Veränderung der relativen Häufigkeit der Kategorie k der Variablen y der Gemeinde $i \in Cl_h$ ein Ausreisser ist, wenn $|\tau_{y,i,k}| > \kappa_{y,h,k}$.

$\tau_{y,i,k}$ ist ein Ereignis der Zufallsvariablen $\tau_{y,k}$. Vorausgesetzt dass $\tau_{y,k} \sim \mathcal{N}(0, 1)$ und dass die Daten keine Ausreisser enthalten, so ist die Grenze folgendermassen definiert

$$\kappa_{y,h,k} \doteq \max 2N_h \Phi(\kappa_{y,h,k}) < 1$$

Somit ist

$$\kappa_{y,h,k} = \Phi^{-1} \left(\frac{1}{2N_h} \right) \quad (10)$$

Die Gemeinden i mit den extremsten Werten wurden durch die folgenden beiden Kriterien bestimmt:

1. $|\tau_{y,i,k}| > \kappa_{y,h,k}, \quad \forall i \in Cl_h, h \in \{1, \dots, 5\}$ und
2. (a) $(h_{y,i,00,k} - h_{y,i,90,k} > 20) \cup (h_{y,i,90,k} - h_{y,i,00,k} > 50)$ und
 (b) $\left(\frac{h_{y,i,00,k}}{h_{y,i,90,k}} < 0.9\right) \cup \left(\frac{h_{y,i,00,k}}{h_{y,i,90,k}} > 1.05\right)$

Bei Kategorien mit wenigen Observationen hat die relative Veränderung der absoluten Häufigkeit keinen Einfluss, da eine untere Grenze für das Ausmass der absoluten Veränderung definiert wurde (20, respektive 50). Das Kriterium 2.(b) hat zur Folge, dass das absolute Ausmass der Veränderung für Kategorien mit vielen Observationen grösser werden kann, je mehr Observationen vorhanden sind, bis die Veränderung als Ausreisser bezeichnet wird. Die Grenzen im zweiten Kriterium wurden subjektiv mit Unterstützung der Fachspezialisten ausgewählt.

4 Funktionalitäten von bestehenden Kontroll- und Einsetzungssystemen

Folgend werden die wichtigsten Funktionalitäten von bestehenden Kontroll- und Einsetzungssystemen für Einsetzungen von Haushaltsdaten von Volkszählungen beschrieben. Dieser Untersuchungen wurden Mitte 2000 durchgeführt, mit der Absicht eine oder zwei Applikationen noch vor der VZ2000 zu testen, um eine Anwendung bei befriedigenden Tests in den Schlussbereinigungen der VZ2000 zu verwenden. Die Basis für die Vergleiche bildete (Poirier, 1999), aufgrund dessen zwei Applikationen ausgewählt wurden, die hier näher beschrieben werden und die den Ansprüchen des BFS am ehesten genügten. Die folgenden beiden Berichte über die ausgewählten Systeme bildeten sodann für diesen Abschnitt die Basisdokumentation: EDIS (Vickers, 1999) und NIM (Bankier, 2000). Weiterreichende Informationen finden sich auch unter (Poirier, 2000).

4.1 EDIS

EDIS ist die Kontroll- und Einsetzungs-Software, welche bei der Volkszählung 2001 in Grossbritannien verwendet wurde und vom *Office of National Statistics* ONS entwickelt wurde. Es ist ein Kontroll- und Einsetzungsprogramm, das einerseits Fehler lokalisiert und andererseits eine Nächster Nachbar Methode verwendet, um die Fehler zu beheben.

Variablen und Datencharakteristiken EDIS kann qualitative und quantitative Variablen behandeln, wobei die Kontrolle quantitativer Variablen nach Intervallen von statten geht.

In der britischen Volkszählung werden Plausibilitätstests in vier Gruppen aufgeteilt:

1. Regeln bezüglich von Haushaltsvariablen
2. Regeln für Personenvariablen, die unabhängig von Variablenwerten anderer Haushaltsmitglieder durchgeführt werden
3. Regeln für Personenvariablen, die Personenvariablen anderer Haushaltsmitglieder miteinbeziehen
4. Regeln für Personen in Kollektivhaushalten

In der britischen Volkszählung wird die 4. Kategorie mit mehr als 30 Personen im Haushalt speziell ausserhalb des EDIS Programms behandelt. Falls mehr als 30 Personen im gleichen Haushalt leben, wird jede Person so bearbeitet, als wohne sie alleine und hat somit auch keine Variable 'Stellung im Haushalt'.

Prozesse Die im Abschnitt 'Variablen und Datencharakteristiken' aufgeführten Gruppen werden in sämtlichen Prozessen unterschieden, und differenziert behandelt.

Kontrolle

Die Fehlerlokalisierung verläuft sequentiell und deterministisch. Es werden inkonsistente Datenwerte einzelner Personen und zwischen den Personen lokalisiert. Im weiteren bestehen auch einige Kontrollen der Gebäude- und Wohnungsvariablen. Die Fehlerlokalisierung ist in der Programmiersprache C programmiert. Anpassungen der Regeln verlangen einen Eingriff in den Quellcode.

Einsetzung

Die Kontrollregeln beinhalten deterministische Korrekturen, sowie Regeln, welche keine Aktion nach sich ziehen, und somit zu statistischen Einsetzung führen.

Die potentiellen Spender werden anhand von Schlüsselvariablen, einer statistischen und einer geographischen Distanz ermittelt. Dabei hängen die Schlüsselvariablen jeweils von der/den Variablenwerten ab, die es einzusetzen gilt. Somit werden jeder Variablen ihre zugehörigen Schlüsselvariablen zugeteilt. Diese haben gegenüber den Variablen die eingesetzt werden einen höheren Stellenwert und müssen vorgängig bereinigt werden. Die statistische Distanz wird anhand der Variablen 'Alter', 'Zivilstand', 'Geschlecht' und 'Stellung im Haushalt' berechnet. Dies gilt unabhängig davon, welche Variablen eingesetzt werden. Die statistische Distanz zwischen zwei Haushalten setzt sich aus den einzelnen statistischen Distanzen aller Haushaltsmitglieder zusammen.

Die geographische Distanz verlangt Geokoordinaten und entspricht einer euklidischen Distanz. Die Einsetzung wird in der folgenden Reihenfolge durchgeführt:

1. Haushalte mit nur einem fehlenden oder inkonsistenten Variablenwert
2. Haushalte mit nur einer Person, die fehlende oder inkonsistente Werte aufweist
3. Haushalte mit zwei oder mehr Personen, welche fehlende oder inkonsistente Daten aufweisen.

Für die Eruiierung eines Spenders wird ein iteratives Verfahren angewandt, welches stoppt, sobald ein Spender gefunden wurde, der zu einem vollständigen und konsistenten Haushalt führt.

1. Einsetzung anhand eines Haushalt-Spenders unter Berücksichtigung des ganzen Haushalts.
2. Personen Einsetzung, unter Berücksichtigung der Grösse des Haushalts. Personen mehrerer Spender-Haushalte können zur Einsetzung im gleichen Empfänger-Haushalt verwendet werden. Die Grösse des Haushalts muss bei Empfänger und Spender übereinstimmen.
3. Personen Einsetzung, ohne die Grösse des Haushalts zu berücksichtigen. Jede Person wird als eine einzelne Einheit betrachtet. Verknüpfungsvariablen im Haushalt werden nicht eingesetzt und bleiben somit leer oder inkonsistent.
4. Personen Einsetzung aufgrund einer iterativ reduzierten Menge von Schlüsselvariablen, ohne die Grösse des Haushalts zu berücksichtigen. Verknüpfungsvariablen im Haushalt werden nicht eingesetzt und bleiben somit leer oder inkonsistent.
5. Manuelle Überprüfung anhand eines 'Outputs'.

Der Fall, dass mehr als 6 Felder eines Records eingesetzt werden sollten, war im Frühjahr 2000 bei ONS noch in Bearbeitung. Ungeklärt war auch noch das Verhalten, wenn kein Spender gefunden werden kann. Es musste also noch ein Werkzeug entwickelt werden, welches die Records mit dem jeweiligen verletzten Regeln auflistet und allenfalls bearbeitet.

Dokumentation der Verarbeitung

Die Dokumentation der Prozesse ist für jeden Prozessdurchlauf detailliert nach den aufgelisteten Punkten unter 'Einsetzung' und nach der Art und Weise der Einsetzung aufgegliedert. Die Anzahl oder die prozentuale Rate der geänderten Daten wird aufgelistet.

4.2 NIM (CANCEIS)

NIM steht für *Nearest Neighbor Imputation Methodology*, vormals *New Imputation Methodology*. Die Software, in welcher diese Methodologie zur Anwendung kommt, wird ebenfalls mit NIM bezeichnet. NIM wurde von Statistics Canada für die kanadische Volkszählung von 1996 entwickelt und kam in dieser auch erfolgreich zum Einsatz. CANCEIS (*Canadian Census Edit and Imputation System*) ist der Nachfolger und die Weiterentwicklung von NIM und kam bei der Volkszählung 2001 in Kanada zum Einsatz.

Prozesse Die Fehler werden anhand einer simultanen Fehlerlokalisierung eruiert. Die Einsetzung erfolgt durch eine Naher-Nachbar Einsetzung, obwohl der Name an eine Nächster-Nachbar Einsetzung denken lässt.

Die multivariate Verteilung der Variablen wird bei NIM durch die Einsetzung (Imputation) der Werte der Variablen in nächster Umgebung, so weit dies möglich ist, gewährleistet. Um fehlerhafte Daten einer Person, respektive Haushalts, zu ändern wird folgendermassen vorgegangen:

1. Suche innerhalb einer zu definierenden Umgebung nach n nächsten Nachbarn, welche eine möglichst gute Übereinstimmung der qualitativen Variablenwerte mit dem Empfänger aufweisen und deren quantitative Variablenwerte nahe den quantitativen Variablenwerten des Empfängers sind. n ist ein vom Anwender zu bestimmender Parameter.
2. Für jeden der n unter 1. bestimmten Nachbarn wird die minimale Teilmenge an Variablenwerten bestimmt, welche die fehlerhaften Daten bereinigen kann.
3. Zufällige Auswahl eines Nachbarn, welcher sich in einer minimalen Anzahl Variablenwerten, und so wenig wie möglich von den fehlerhaften Daten unterscheidet (im Sinne einer benutzerdefinierten Distanz).

Für jeden Empfänger wird die Nachbarschaft aufgrund der Charakteristiken des Empfängers implizit neu definiert. Die Anzahl potentieller Spender, die untersucht wird, ist durch einen frei wählbaren Parameter N begrenzt. So wird die ASCII-Datei sequentiell gelesen, bis höchstens N Haushalte die Kontrollregeln verletzt haben oder bis höchstens N Haushalte alle Kontrollregeln erfüllen. Nimmt man an, dass N_v Haushalte die Regeln verletzt haben, und dass N_s alle Regeln erfüllt haben. Dann wird für jeden der N_v Haushalte eine Nachbarschaft aus den N_s potentiellen Spenderhaushalten eruiert, wobei diese aus maximal n Haushalten besteht. Hiermit ist gesagt, dass N nicht zu klein sein darf, um Abhängigkeiten zwischen der Nachbarschaft zweier aufeinander folgender Empfängerhaushalte so klein wie möglich zu halten.

Die Variablen werden innerhalb von NIM in einer vorgegebenen Reihenfolge behandelt. Da fehlende und ungültige Variablenwerte deterministisch Lokalisiert werden können und diese immer eingesetzt werden müssen werden diese zuerst eingesetzt.

Für den zweiten Schritt wird davon ausgegangen, dass die Distanz am kleinsten ist, eine Variable einzusetzen, als mehrere. Deshalb werden in diesem Schritt Variablen eingesetzt, die in mehreren Kontrollregeln vorkommen, welche nicht erfüllt werden.

In einem letzten Schritt wird die minimale Anzahl Variablen eruiert, die eingesetzt werden müssen, damit keine Kontrollregel mehr verletzt wird.

1. fehlende und ungültige Variablenwerte
2. Variablen, die in mehreren Kontrollregeln vorkommen, welche nach 1 nicht erfüllt werden
3. Minimale Anzahl Variablen, die in den Kontrollregeln vorkommen, welche nach 2 nicht erfüllt werden.

Die Komplexität von NIM wird auf $O(m)$ geschätzt, wobei m die Anzahl expliziter Regeln darstellt.

Kontrollregeln Die Kontrollregeln werden anhand von DLT (*decision logic tables*) gehandhabt, welche als Tabellen mit logischen Aussagen in der ersten Spalten und Kontrollregeln in den weiteren Spalten betrachtet werden können. Die Elemente der Tabellen bestehen aus Y/N oder '-' als Elemente betrachtet werden können, siehe Tabelle 2. Ist eine Regel wahr, so liegt ein Fehler im entsprechenden Record vor. Beispiel: Es sei ein 3 Personenhaushalt, der sich laut Fragebogen aus einem Kind, einer Mutter und einer Grossmutter zusammensetzt. Die Kontrollregeln sind in der Tabelle 1 aufgeführt, welche in der Tabelle 2 in der DLT Form wiedergegeben sind.

Tabelle 1: Beispiel für Kontrollregeln im Haushaltszusammenhang eines 3-Personen Haushalts. Die Zahlen in Klammern beziehen sich auf die Personen 1,2,3 im Haushalt. 'Relat(X)' steht für die Stellung im Haushalt der Person X bezüglich der Referenzperson. In unserem Beispiel ist die Person 1 die Referenzperson.

Regel	logische Regel
1	$(\text{Relat}(1) = \text{Referenz}) \cap (\text{Relat}(2) = \text{Mutter}) \cap (\text{Alter}(2) - \text{Alter}(1) < 15)$
2	$(\text{Relat}(2) = \text{Mutter}) \cap (\text{Alter}(2) < 15)$
3	$(\text{Relat}(1) = \text{Referenz}) \cap (\text{Relat}(3) = \text{Grossmutter}) \cap (\text{Alter}(3) - \text{Alter}(1) < 30)$
4	$(\text{Relat}(3) = \text{Grossmutter}) \cap (\text{Alter}(2) < 45)$

Tabelle 2: Beispiel einer DLT (decision logic table) eines Haushalts mit drei Personen, 1 ist die Referenzperson im Haushalt, 2 die Mutter, 3 die Grossmutter. Es werden die Stellung im Haushalt (Relat(X)) und die Altersunterschiede untersucht.

Aussage	Regel 1	Regel 2	Regel 3	Regel 4
Relat(1) = Referenz	Y	-	Y	-
Relat(2) = Mutter	Y	Y	-	-
Alter(2) - Alter(1) < 15	Y	-	-	-
Alter(2) < 15	-	Y	-	-
Relat(3) = Grossmutter	-	-	Y	Y
Alter(3) - Alter(1) < 30	-	-	Y	-
Alter(3) < 45	-	-	-	Y

4.3 Systemvergleich

Die beiden untersuchten Applikationen verwenden eine Nachbar Einsetzung (siehe auch Tabelle 3). Für die Fehlerlokalisierung ist der Aufwand bei EDIS grösser als bei NIM, da diese deterministisch durchge-

führt wird. Somit muss aufgrund von Analysen festgelegt werden, welche Variablen der grössere Wahrheitsgehalt zugesprochen werden soll.

EDIS ist in seiner Entwicklung weit fortgeschritten. Für eine allfällige Verwendung in der VZ2000 fehlen dennoch entscheidende Teile wie die Verarbeitung der Daten, für welche kein Spender gefunden werden konnte. Ebenso müsste im BFS ein spezieller Prozess durchgeführt werden, um die demographischen Variablen vorgängig zu bereinigen. Diese Bedingungen sind zu restriktiv, als dass EDIS für die Verarbeitung der Personen- und Haushaltsvariablen der VZ2000 in Frage kommen könnte. Im Gegensatz zu EDIS existiert von NIM eine lauffähige Version, die bereits erfolgreich in der kanadischen Volkszählung 1996 eingesetzt worden ist. Es gilt jedoch abzuklären, inwiefern diese an die schweizerische Volkszählung angepasst werden kann. Zudem wurde in keinem der vorhandenen Berichte ((Bankier, 2000), (Poirier, 1999), (Poirier, 2000)) eingehend auf die Qualität der Einsetzungen eingegangen.

Aufgrund der obigen Überlegungen musste vom Einsatz von EDIS für die VZ2000 abgesehen werden. Für einen definitiven Entscheid, ob NIM für die VZ2000 verwendet werden kann, wurde ein eingehender Test mit Daten der Schweiz durchgeführt, um abzuklären, welche Anpassungen nötig sind und mit welchem Aufwand dies verbunden ist.

Tabelle 3: Vergleich der Eigenschaften von NIM und EDIS. 'Spender (N-N)' bezieht sich auf die 'nächster Nachbar Spendersuche'. '(Ja)' bedeutet, dass die entsprechende Charakteristik nicht vollständig getestet wurde.

		NIM	EDIS
Autor		Statistics Canada	ONS, UK
Variablen	Quantitative	(Ja)	(Ja)
	Qualitative	Ja	Ja
Kontrolle	Daten Überprüfung	Ja	Ja
	minimale Änderungen	Ja	Ja
	Fehler Lokalisierung	simultan	deterministisch
Einsetzung	Spender (Zufall)	(Ja)	(Ja)
	Spender (N-N)	Ja	Ja
	Spender Suche vor	Zu ändernde Variablen	
Spender Suche		Bestimmung, der zu	werden vor der Spender
	ändernden Variablen	Suche aufgrund einer	
		Hierarchie bestimmt	
Erprobt		<i>Canadian Census 1996</i>	nein

4.4 Das Programm NIM

Der in diesem Abschnitt aufgeführte Beschrieb basiert einerseits auf den Erklärungen in (Bankier, 2000) und andererseits auf den Erfahrungen des Autors mit dem Quellcode vom Prototypen von NIM, der dem BFS im Sommer 2000 von STATCAN zur Verfügung gestellt wurde.

Im folgenden wird nicht das gesamte Programm beschrieben, sondern der Ablauf geschildert, wie mit einem fehlerhaften Haushalt umgegangen wird.

NIM besteht aus drei Hauptprogrammen, die in vorgegebener Reihenfolge nacheinander kompiliert und ausgeführt werden müssen. Im ersten Hauptprogramm ('Dict.c') werden die Definitionen der Variablen, wie sie in den entsprechenden ASCII-Dateien gegeben werden müssen, in C-Code umgewandelt. Im zweiten Hauptprogramm ('Nimprot1.c') werden die Regeln in C-Code umgewandelt, welche in der Datei 'DLT.txt' in der 'DLT'-Form (siehe Tabelle 2) vom Benutzer abgespeichert worden sind. Im dritten Hauptprogramm ('Nimprot2.c') werden die Haushalte aus der Datei 'hhld.txt' gelesen. Pro Zeile muss der Eintrag einer Person stehen. Die Reihenfolge der Variablen muss der Reihenfolge entsprechen, die in der Funktion zum Einlesen 'read_hhld' (Datei 'Variable.h') verwendet wird, und in der Datei 'Var-

name.h' definiert ist. In dieser Datei sind ebenfalls die Gewichte der einzelnen Variablen definiert. Die Parameter des verwendeten Algorithmus und die Gewichte der Distanzfunktionen sind in der Datei 'Nimprdef.h' definiert. Die Distanzen selbst sind in der Datei 'Variable.h' programmiert, die Zuordnungen zu den einzelnen Variablen in der Datei 'Varname.h'.

Vorgehen NIM liest sequentiell die Daten der Personen aus der ASCII-Datei, speichert sie in einem Vektor⁷ der Haushalte und testet, ob Kontrollregeln verletzt werden. Fehlerhafte Haushalte werden in eine Datei geschrieben. Dieses Vorgehen wird wiederholt, bis entweder die Anzahl Spender oder Empfänger eine maximale Anzahl erreicht. Diese obere Grenze (1'000) kann in der Datei 'Nimprdef.h' festgelegt werden. Wird die maximale Anzahl erreicht, so werden die Empfänger sequentiell einzeln bearbeitet. Zuerst wird, wie bereits gesagt, für jeden Empfänger die Nachbarschaft bestimmt. Die maximal 200 nächsten Nachbarn, ein Parameter, der vom Benutzer gewählt werden kann, werden einer nach dem anderen verwendet, um zu testen, ob sie als Spender in Frage kommen, d.h. dass diese so verwendet werden können, dass der Empfänger nach der Einsetzung keine Regel mehr verletzt. Zuerst werden fehlenden und ungültigen Werte ersetzt. Danach wird getestet, welche Kontrollregeln nun verletzt werden. Falls Variablen in mehreren verletzten Kontrollregeln vorkommen, werden diese eingesetzt, da auf diese Weise die Anzahl eingesetzter Variablen minimiert wird. Nach einem neuerlichen Test anhand der Kontrollregeln, wird für die restlichen Inkonsistenzen eingesetzt. Dabei werden mehrere mögliche Einsetzungen getestet, indem einerseits die Variablen, die eingesetzt werden, miteinander kombiniert werden⁸. Andererseits werden die Personen des Spenderhaushalts permutiert, um unterschiedlichen Ordnungen in den Haushalten Rechnung zu tragen⁹. Aus den potentiellen Spendern wird einer aufgrund der totalen Einsetzungsdistanz zufällig ausgewählt. Die Einsetzungsdistanz berechnet sich aus der Haushaltsdistanz zwischen Spender und Empfänger vor der Einsetzung, der Haushaltsdistanz zwischen Spender und Empfänger nach der Einsetzung, der Haushaltsdistanz zwischen dem Empfänger vor der Einsetzung und dem Empfänger nach der Einsetzung und der Anzahl eingesetzter Variablen. Anhand eines Parameters kann der Benutzer wählen, ob er es vorzieht, dass der eingesetzte Empfänger eher dem ursprünglichen gleicht, oder ob er eher dem Spender gleichen soll (Parameter ' α ' in der Tabelle 4). Im ersten Fall werden im Vergleich zum zweiten Fall eher weniger Variablen eingesetzt, was aber zu speziellen Haushalten führen kann. Im zweiten Fall gleicht der eingesetzte Haushalt eher einem bereits existierenden, der alls solcher nicht in Frage gestellt wird.

Ein weiterer Parameter (' γ ' in der Tabelle 4) kann vom Benutzer festgelegt werden, anhand dessen festgelegt werden kann, ob die Einsetzung eher eine kleine Distanz haben soll, oder ob sie so wenig wie möglich Variablen einsetzen soll. Im ersten Fall werden Variablen, die nahe beieinander liegen oder die ein kleines Gewicht haben öfters eingesetzt als im zweiten Fall. Im zweiten Fall kann es sein, dass dadurch, dass so wenig wie möglich Variablen eingesetzt werden, aufgrund von Variablen 'sekundärer' Bedeutung, Variablen 'primärer' Bedeutung einen Wert annehmen, der im Vergleich zum restlichen Haushalt akzeptabel ist aber nicht unbedingt erwünscht ist.

In der Version von STATCAN wurde immer der nächste Nachbar ausgewählt. Für die Anwendung in der VZ2000 wurde die Einsetzung umgekehrt proportional zur totalen Distanz zufällig ausgewählt. Die maximale Anzahl Einsetzungen, die hierfür verwendet wird kann auch vom Benutzer mittels eines Parameters festgelegt werden (Parameter 'NBRNMCIAS' in der Tabelle 4).

Parameter Abgesehen von Parametern, welche es dem Benutzer erlauben, mehr oder weniger Information über die Kontrollen und Einsetzungen zu generieren - diese sind in der Tabelle 4 nicht aufgeführt

⁷Jedes Element des Vektors ist eine C-Struktur.

⁸Es werden nicht alle möglichen Kombinationen getestet.

⁹Es werden nicht alle möglichen Permutationen verwendet

- kann vor allem die Spenderauswahl und die Leistung des Programms beeinflusst werden. Die beiden Parameter 'MAX_FAILED_IMPUTE' und 'NUM_DONORS_DIST_CAL' begrenzen die Anzahl Empfänger und die Anzahl Spender. STATCAN empfiehlt dringend, die beiden Werte gleich zu setzen.

In der Version von STATCAN wurde unter den Spendern nur solange ein Spender zur Auswahl gesucht, solange weniger als 'NBRNMCIAS' Einsetzungen gefunden wurden. Für die Produktion in der VZ2000 wurde dieser Parameter nicht mehr verwendet. Statt dessen wurde ein neuer kreiert ('NBRNMCIAS-KEPT'), der die Anzahl möglicher Einsetzungen beschränkt, aus denen zufällig eine ausgewählt wird, nachdem sämtliche Spender ('NUM_DONORS_DIST_CAL') verwendet wurden, um Einsetzungen zu generieren. Dies hat zur Folge, dass ein beträchtlicher Verlust an Leistung in Abhängigkeit der Werte der Parameter zu bemerken ist. Hingegen sind sämtliche NUM_DONORS_DIST_CAL Haushalte der Spenderliste potentielle Spender. Die restlichen aufgeführten Parameter dienen vor allem der Leistungs-optimierung des Programms.

Tabelle 4: Parameter von NIM, ihre Abkürzung und ihr 'üblicher' Wert. 'Max.' steht für maximal, 'Min.' für minimal und '#' steht für Anzahl.

Beschreibung	Abkürzung	Wert
Max. # potentieller Spender	NUM_DONORS_DIST_CAL	1'000
Max. # Empfänger	MAX_FAILED_IMPUTE	1'000
Max. # Spender aus denen die Einsetzung zufällig ausgewählt wird	NBRNMCIAS	10
Max. # DLTs	MAX_DLTs	20
Max. # Behauptungen in den DLTs	MAXPROPOSITIONS	100
Max. # Regeln	MAX_NUMBER_EDITS	60
α	ALPHA	0.95
γ	GAMMA	1.1

5 Vergleich deterministischer Korrekturen mit NIM

Für diesen Vergleich wurde ein Prototyp von NIM von STATCAN verwendet, der eine verbesserte Version des in der kanadischen Volkszählung 1996 verwendeten Version war. Die deterministischen Korrekturen für diesen Vergleich wurden von einem Informatiker in C++ programmiert.

Der Vergleich der deterministischen Korrekturen mit NIM soll Aufschluss über die Möglichkeiten und Grenzen dieser Software geben, dies in technischer sowie in qualitativer Hinsicht.

Der Vergleich basiert auf den Daten der Kantone Freiburg, Wallis und der Stadt Zürich der Volkszählung 1990 (VZ90). In der Aufarbeitung der VZ90 wurden deterministische Korrekturen als Einsetzungsverfahren verwendet. Die Daten weisen **Flags** auf, die auf Korrekturen hinweisen. Anhand der Häufigkeiten dieser Flags wurden für diesen Test in den vollständigen Daten der VZ90 Fehler zufällig eingefügt. Somit wurden nicht die deterministischen Korrekturen der VZ90 wieder rückgängig gemacht, da sonst die zum Vergleich herangezogenen deterministischen Korrekturen (CORA) im Vorteil gegenüber NIM gewesen wären.

Die Vergleichstests erfolgten unter Einbezug der Variablen in der Tabelle 5.

Aus Zeitgründen konnten keine weiteren Variablen in den Vergleich miteinbezogen werden.

5.1 Datenbasis, Anpassungen, Kontrollregeln, Fehlergenerierung

Für die Tests wurden die Daten der Kantone Freiburg, Wallis und der Stadt Zürich aus der VZ90 verwendet.

Da NIM auf ganze Regionen angewandt werden muss, um die lokalen Eigenheiten beizubehalten, wurde darauf verzichtet, eine Stichprobe von Haushalten in Regionen zu ziehen. Denn die Ausarbeitung eines Klumpen-Stichprobenplans hätte zu viel Zeit in Anspruch genommen.

Anpassungen an NIM Der NIM Prototyp von *Statistics Canada* (STATCAN) ist auf die Volkszählung in Kanada abgestimmt. Die Kanadier haben einen Fragebogen pro Haushalt, wobei die erste Person auf dem Fragebogen sich als Haushaltschef bezeichnen muss. Die Stellung im Haushalt der anderen Personen muss im Verhältnis zur ersten Person angegeben werden. Hiermit entstehen innerhalb des Haushalts genauer definierte Zusammenhänge als in der Schweiz, so wird zum Beispiel bei den Kanadiern unter Schwiegermutter und Mutter unterschieden. Dies ist in der Schweiz nicht der Fall. Somit müssen in der Schweiz weniger Regeln bezüglich der Haushaltszusammensetzung geschrieben werden, hingegen besteht immer eine gewisse Unsicherheit in Bezug auf den Haushaltszusammenhang. So kann zum Beispiel bei einer Angabe 'Sohn, Tochter' nicht gesagt werden, ob beide Eltern im selben Haushalt leben oder nur der Vater oder die Mutter. Dies wiederum führt zu aufwendigeren Alters Tests innerhalb des Haushalts, die jedoch nie zu einer Eindeutigkeit führen, wie dies in Kanada der Fall ist.

Programmierkenntnisse in der Programmiersprache C sind die Voraussetzung, um neue Distanzen zu definieren, neue Variablen einzubeziehen und den Programmablauf zu verstehen. Die Logik des Programms gleicht in manchen Teilen jedoch eher der Programmiersprache Fortran. Es wurden im ganzen Programm vor allem globale Variablen verwendet, was aufgrund der vorhandenen Komplexität und des Umfangs des Programms ein gewisses Risiko darstellt.

Datenaufarbeitung und Variablen Die Variable Ehepaarnummer (EPNRW) wurde nicht wie von STATCAN verwendet, da sie in der Schweiz nicht direkt im Fragebogen erhoben wird. Hingegen wurden nicht nur die demographischen Variablen verarbeitet, sondern auch noch andere Variablen mit einer hohen Priorität in der Volkszählung und zusätzlich die Variable Verkehrsmittel, die eine weniger hohe Priorität hat. Im weiteren wurden Hilfsvariablen miteinbezogen, die vor allem zur Beibehaltung lokaler Charakteristiken dienen. In der Tabelle 5 sind die hauptsächlichsten Variablen aufgelistet, die von NIM

und CORA verwendet wurden. CORA verwendete noch mehrere Variablen innerhalb der Kontrollregeln, auf deren Auflistung verzichtet wurde. Die Variable NOHH ist ein Haushalts Identifikator der sich von der VZ90 Variablen 'Haushaltsnummer' (HHNR) grundsätzlich unterscheidet. Die Variable ABGA bezeichnet die Kombination aller angekreuzten Werte der Variablen 'Abgeschlossene Ausbildung'. Die Variablen BESIT und VEMI wurden ebenso mit Mehrfachangaben verarbeitet.

Da es von Vorteil ist, wenn der Haushalt bei NIM geordnet ist, wurden die Daten nach der Einführung des Haushaltsidentifikators nach folgenden Variablen geordnet: Zählgemeinde (ZGDE), Gebäude- nummer (GBNR), Haushaltsnummer (HHNR), Geokoordinaten (XACH, YACH), Haushaltsidentifikator (NOHH), wirtschaftliche Referenzperson (RPHHW), absteigende wirtschaftliche Ehepaarnummer (EP- NRW) und absteigendes Alter (ALTJ).

Hiermit sollen die regionalen und lokalen Strukturen der Haushalte beibehalten werden.

Tabelle 5: Variablen, die von NIM und CORA verwendet wurden. Hilfsvariablen wurden zur Spen- dersuche (NIM) oder zur Spezifizierung der Kontrollregel und Korrektur (CORA) verwendet. Weitere Variablen, welche CORA verwendet hat, wurden unter 'weitere' zusammengefasst.

Variable	Abkürzung	NIM		CORA	
		Änderung möglich	Hilfs- var.	Änderung möglich	Hilfs- var.
Haushaltsnummer	NOHH				
Anzahl Personen ¹⁰	APERW		x		
Laufnummer	LAUF				
Zählgemeinde	ZGDE		x		x
Geburtsort	GORT		x		x
Wohnort vor 5 Jahren	WOR5		x		x
Nationalität	NATI	x			
Alter	ALTJ	x			x
Geschlecht	GESL	x			x
Zivilstand	ZIVL	x		x	x
Stellung im Haushalt	STHHW	x		x	x
Muttersprache	SPRA	x		x	x
Abgeschlossene Aus- bildung	ABGA	x			
Beschäftigungssituation	BESIT	x			x
Berufliche Stellung	BSKO	x			x
Verkehrsmittel	VEMI	x		x	
<i>weitere</i>					x

Bit-Codierung Variablen mit Mehrfachangaben (mehrere Werte gleichzeitig), ABGA, BESIT und VE- MI, wurden aus Gründen der Effizienz umcodiert. Im Datensatz der VZ90 existiert für jeden Wert eines Merkmales eine Variable, die mit der Bit-Codierung allesamt zu einer einzigen Variablen zusammenge- fasst wurden.

Es sei v_i , die i -te Antwortmöglichkeit, $v_i = 1$ wenn eine Antwort an i -ter Stelle vorhanden ist, $v_i = 0$ ansonsten; $v_0 = 1$ wenn die Antwort ungültig ist und $v_0 = 0$ ansonsten.

¹⁰wirtschaftlicher Wohnsitz

Beispiel: $ABGA = \sum_i v_i 2^i = 2112 = 2^6 + 2^{11}$, bedeutet, dass die 6. und die 11. Antwort angekreuzt (Gymnasium und Universität) sind.

Berechnung der Personen-Haushaltsdistanz Die Distanz zwischen einer Empfängerperson x an i -ter Stelle im 'Empfängerhaushaltsvektor' und einer Spenderperson $\tilde{x}$ an j -ter Position im 'Spenderhaushaltsvektor' ist

$$d_{fp,i,j} = \sum_v \omega_v d_v(x, \tilde{x}). \quad (11)$$

fp steht für *failed-passed* und bedeutet, dass die Distanz zwischen dem Empfänger (*failed*) und dem Spender (*passed*) berechnet wird. ω_v ist die vom Benutzer festgelegte Gewichtung der Variablen v . $\omega_v = 1$, ausser $\omega_{VEMI} = 0.01$, $\omega_{NATI} = 0.5$ und $\omega_{WORS} = 0.5$ (siehe (Bankier, 2000) für nähere Informationen bezüglich der Berechnung der Distanzen).

Die Haushaltsdistanz d_{fp} ist die Summe der Distanzen $d_{fp,i,j}$. NIM geht mehrere Ordnungen der Personen im 'Spenderhaushaltsvektor' durch, um die d_{fp} zu minimieren. Der Prototyp von STATCAN ging nur einen kleinen Teil der möglichen Kombinationen durch. Für die Produktion in der VZ2000 wurden alle Kombinationen verwendet: $d_{fp} = \min_{\sigma} \sum_i d_{fp,i,\sigma_i}$. σ bezeichnet die Permutation der Ordnung im Spenderhaushalt.

Kontrollregeln Als Kontrollregeln von CORA wurden die Regeln der VZ90 verwendet, siehe (Boruvka, 1994), sofern diese automatisiert werden konnten. Die zugehörigen Aktionen änderten in manchen Fällen mehrere Variablen gleichzeitig.

Die Kontrollregeln für NIM basierten auf den gleichen Regeln, wie jene für CORA. Wobei die Regeln für die zusätzlich verarbeiteten Variablen in NIM mit den entsprechenden Regeln der VZ90 komplettiert wurden.

Fehlergenerierung Die Häufigkeit der Fehler, die in die Daten der VZ90 eingefügt wurden, sind anhand der entsprechenden Flags auf Ebene Kanton respektive der Stadt Zürich berechnet worden. Insgesamt wurden vier Flags in den Daten der VZ90 geführt, siehe Tabelle 6. Es wurde auf die Generierung von Fehlern aufgrund des Flags 4 verzichtet, da dieses Flag nicht von einer deterministischen Korrektur herrührt und nur bei der Erwerbstätigkeit vorhanden ist.

Tabelle 6: Flags der VZ90.

Flagnr.	Bedeutung	Grössenordnung
1	Ein vorhandener Wert wurde gelöscht	< 1%
2	Ein fehlender Wert wurde eingesetzt	≈ 3% bis > 10%
3	Ein vorhandener Wert wurde durch einen anderen ersetzt	≈ 2% bis > 6% ¹¹
4	Hot-Deck Einsetzung	≈ 1%

Für die demographischen Variablen Alter (ALTJ), Geschlecht (GESL), Zivilstand (ZIVL) und Nationalität (NATI) standen keine Flags zur Verfügung, um Fehler zu generieren. Deshalb wurden die Flag 1 und Flag 2 Fehler für diese Variablen frei festgelegt, siehe Tabellen 7. Auf Flag 3 Fehler wurde verzichtet. Die Generierung der Flag 3 Fehler von STHHW und SPRA betraf nur Doppelangaben und für SPRA wurden im Speziellen nur Doppelangaben der vier offiziellen Sprachen der Schweiz (deutsch, französisch, italienisch, räto-romanisch) generiert. Für die Variable Verkehrsmittel (VEMI) bestanden keine

¹¹STHHW.

Flags, da diese Variable bei der VZ90 nicht aufgearbeitet wurde. Es bestehen jedoch noch rund 4% fehlende Angaben bei erwerbstätigen Personen, für welche versucht wird Einsetzungen vorzunehmen. Die Variablen BSKO und ABGA wiesen ebenso noch fehlende Werte auf, die ebenfalls eingesetzt wurden. Die Fehler wurden zufällig und unabhängig vom Kontext eingesetzt. Deshalb kann nicht ausgeschlossen werden, dass die eingefügten Fehler sich nicht als Fehler bemerkbar machen. Die zufällig generierte Konstellation kann somit durchaus 'plausibel' sein.

Tabelle 7: Flag 1 und Flag 2 Fehleranteil der demographischen Variablen, das heisst, die Wahrscheinlichkeit mit der ein richtigerweise fehlender Wert zufällig bei der Fehlergenerierung ersetzt wurde.

Variable	Flag 1 [%]	Flag 2 [%]
ALTJ	0.5	0.5
GESL	0.5	0.5
ZIVL	0.5	0.5
NATI	0.0	1.0

5.2 Evaluation

Evaluationskriterien Die Evaluation wird anhand der in Tabelle 8 aufgeführten Kriterien durchgeführt.

Bei der Variablen 'VEMI' wurden keine Fehler eingefügt, womit CORA Korrekturen einzig aufgrund des verfälschten Alters machen konnte. Dies zeigt einzig auf, dass NIM fähig ist, fehlende Werte einzusetzen, ohne komplizierte Aktionen zu definieren, was bei CORA nicht der Fall ist. Aus diesem Grund wurde die Variable 'VEMI' aus der Vergleichsanalyse ausgeschlossen.

Anzahl Änderungen Zur Vereinfachung der Diskussion wird eine Änderung als eine Abänderung eines bestehenden Wertes oder eine Einsetzung eines fehlenden Wertes definiert.

Da NIM und CORA nicht die gleichen Variablen behandelt haben, muss dieser Vergleich mit Vorsicht genossen werden. NIM strebt die kleinste Anzahl Änderungen aufgrund der vorhandenen Spender an. Die Anzahl Änderungen hängt von der Anzahl Variablen ab, die zu Verletzungen der Kontrollregeln geführt haben, sowie von den Charakteristiken der Spender.

Bei CORA kann die Anzahl Änderungen auch nicht kontrolliert werden. Da die deterministischen Korrekturen sequentiell durchgeführt werden, ist nicht auszuschliessen, dass eine Korrektur zu einer vermeidbaren weiteren Korrektur führt, wobei davon ausgegangen wird, dass dies eher seltene Fälle sind.

Ein Vergleich der Anzahl Änderungen, die NIM gemacht hat, mit der Anzahl Änderungen die CORA vorgenommen hat, macht streng genommen keinen Sinn. Denn einige Regeln die bei den deterministischen Korrekturen in der VZ90 manuell gelöst wurden, konnten in CORA nicht aufgenommen werden. Hingegen wurden diese Regeln, vor allem STHHW betreffend, in NIM miteinbezogen. Andererseits hat

Tabelle 8: Kriterienliste zur Evaluation des Vergleichs CORA - NIM

Nr.	Kriterium
1	Anzahl Änderungen
2	Vollständigkeit und Konsistenz
3	Fehler, die nicht behandelt werden konnten
4	Beibehaltung der Verteilung
5	Wiederherstellung der wahren Werte

NIM mehr Variablen verarbeitet als CORA, womit die zusätzlichen Kontrollregeln in NIM zu mehr Fehlern führen konnte. Aus diesen Gründen wird darauf verzichtet, die Anzahl Änderungen zu vergleichen.

Vollständigkeit und Konsistenz Die Vollständigkeit und Konsistenz hängt direkt von den Kontrollregeln ab. Durch diese wird 'bestimmt', welche Werte und Kombinationen von Werten zulässig, also plausibel, sind. In CORA konnten nicht alle Regeln der Fehlererkennung implementiert werden, da die zugehörige Korrektur manuell vorgenommen wurde. Die Records und Haushalte, in welchen NIM Einsetzungen vorgenommen hat, sind vollständig und konsistent in Bezug auf die verwendeten Regeln. NIM kann aber nicht immer einen Spender finden. Die entsprechenden Records bleiben somit weiterhin fehlerhaft.

Bei CORA kann keine Garantie abgegeben werden, ob die verarbeiteten Records plausibel sind, da eine Korrektur eine weitere nach sich ziehen kann. Zumindest theoretisch ist es möglich, dass dieser Prozess zirkulär wird, womit nie eine vollständige Plausibilität erreicht werden kann. In dieser Testphase wurde keine zirkuläre Korrekturkette beobachtet.

Fehler, die nicht behandelt werden konnten Da die Generierung der Fehler zufällig durchgeführt wurde, kann nicht garantiert werden, dass ein generierter 'Fehler' sich als solcher bemerkbar macht. Deshalb, und weil bei CORA nur die Fehler entdeckt werden, die korrigiert werden, ist für CORA kein Kriterium vorhanden, um das Ausmass der nicht behandelten Inkonsistenzen zu beziffern. Es kann nur die Anzahl fehlender Werte, die nicht eingesetzt wurden, eruiert werden. Aus der Tabelle 9 geht hervor, dass nach CORA vor allem bei den Variablen Zivilstand und der Stellung im Haushalt noch fehlende Werte existieren. Diese 'Restfälle' wurden in der VZ90 manuell bearbeitet. Fehler die von CORA nicht

Tabelle 9: Anzahl Personen, die nach der Verarbeitung mit NIM respektive CORA noch fehlende Werte aufweisen. Die 'Restfälle' von CORA wurden in der VZ90 manuell bearbeitet.

Variable	FR		VS		ZH		Total	
	NIM	CORA	NIM	CORA	NIM	CORA	NIM	CORA
ZIVL	25	349	27	480	38	909	90	1738
SPRA	31	10	14	9	39	99	84	118
STHHW		4	11	281	10	146	21	431
Total	56	363	52	770	87	1154	195	2287

behandelt wurden, basieren auf fehlenden Kontrollregeln oder aber weil nicht das gesamte Spektrum möglicher Antwort Kombinationen in die Aktionen eingebaut werden konnte. Fehlende Kontrollregeln haben bei CORA und NIM die gleichen Auswirkungen. Fehlende Aktionen bei CORA haben jedoch keinerlei Auswirkungen auf NIM, da die Einsetzung nur von den Kontrollregeln und den zur Verfügung stehenden Spendern beeinflusst wird.

Aus der Tabelle 10 ist ersichtlich, dass NIM bei genügend grossen Schichten, bei mehr als 85% der fehlerhaften Haushalte einen Spender findet. Um die restlichen fehlerhaften Haushalte einzusetzen, sollte in Betracht gezogen werden, die geographischen Grenzen etwas lockern, wie dies in der englischen Volkszählung in EDIS gemacht wird.

Beibehaltung der Verteilung Die verarbeiteten Daten wurden mit den Daten der VZ90 verglichen, um die Effizienz von NIM und CORA bezüglich der 'wahren Verteilung' der Variablen, zu beurteilen. Nach der Verarbeitung werden die Daten, je nach Verarbeitung mit CORA-Daten respektive NIM-Daten bezeichnet. Die Daten der VZ90 vor der Fehlergenerierung, werden VZ-90 Daten genannt.

Zuerst wurde eine einfache Statistik erstellt, in welcher für jede Variable und jede Kategorie ermittelt wird, ob die Randverteilungen der NIM-Daten oder die Randverteilungen der CORA-Daten näher bei

Tabelle 10: Anzahl Haushalte, die als fehlerhaft erkannt wurden, für die aber NIM keinen Spender finden konnte (n. einges. = nicht eingesetzt). Die Prozentzahlen beziehen sich auf die Anzahl Haushalte, welche als fehlerhaft erkannt wurden. 'APERW' steht für die Anzahl Personen des Haushalts mit wirtschaftlichem Wohnsitz im entsprechenden Wohnsitz.

	APERW		FR		VS		ZH	
			n. einges.	[%]	n. einges.	[%]	n. einges.	[%]
1			0	0.00	0	0.00	0	0.00
2			344	9.34	335	2.71	963	5.38
3			389	13.20	471	4.98	658	7.81
4			639	15.75	644	5.75	697	9.45
5			279	14.48	346	7.42	226	10.32
6			116	18.41	86	8.57	66	11.91
7			37	19.47	22	9.65	30	19.87
8			12	16.22	5	7.35	11	18.03
9			9	33.33	1	5.26	3	14.20
10			1	14.28	5	100.00	2	18.18
11			0	0.00	1	100.00	3	33.33
12			1	100.00	2	100.00	1	33.33
Total			1827	12.03	1948	4.30	2660	4.88

den Randverteilungen der VZ90-Daten sind. Aus der Tabelle 11 ist ersichtlich, dass CORA einen Vorteil hat gegenüber NIM. Die Randverteilungen von NIM und CORA lagen im Kanton Freiburg jedoch meist sehr nahe beieinander; so hatten sie in einer Sprachkategorie die gleiche Frequenz. Zudem basiert dieser Test ausschliesslich auf drei Variablen, wobei die Verteilung der Sprache bei CORA in sämtlichen Schichten näher an jener der VZ90 war als NIM.

Tabelle 11: Frequenzen Vergleich VZ90 - NIM, VZ90 - CORA, nach Variablen und Schichten aufgeteilt. N_{NIM} steht für die Anzahl Kategorien, in welchen NIM näher an der Frequenz der VZ90 ist, N_{CORA} steht für die Anzahl Kategorien, in welchen CORA näher an der Frequenz der VZ90 ist. In den letzten beiden Spalten steht das kummulierte Total von N_{NIM} und N_{CORA} nach Variablen aufgeteilt. In der letzten Zeile stehen die Spalten Totale.

Variable	FR		VS		ZH		Total	
	N_{NIM}	N_{CORA}	N_{NIM}	N_{CORA}	N_{NIM}	N_{CORA}	N_{NIM}	N_{CORA}
ZIVL	2	2	2	2	1	3	5	7
SPRA	1	6	1	6	3	4	5	16
STHHW	1	6	6	1	6	1	13	8
Total	4	14	9	9	10	8	23	31

Um die Beibehaltung der Verteilung zu evaluieren, wird ein Kriterium aus (Chambers, 2001) verwendet, welches im Abschnitt 3.4 beschrieben wurde.

Bei einigen Variablen wurden Kategorien zusammengelegt: 'r<Variablen Name>' steht für die Daten der VZ90 (z.B. rSPRS), 'ri<Variablen Name>' steht für die Daten von NIM, 'rcc<Variablen Name>' steht für die Daten von CORA.

Aufgrund der Komplexität der Variablen ABGA wurde die Variable HABG nach der Einsetzung respektive der Korrektur aufgrund der Variablen ABGA kreiert, wie dies in der VZ90 gemacht worden war.

Aus den Tabellen 12 und 13 ist ersichtlich, dass weder NIM noch CORA die Randverteilungen wieder-

Tabelle 12: Wald Statistik für die Variablen, die mit NIM und CORA eingesetzt, resp. korrigiert, wurden. 'NA' bedeutet, dass die Statistik nicht berechnet werden konnte, da Q der Gleichung (4) singulär ist.

Variable	FR W_Y			VS W_Y			ZH W_Y		
	df	NIM	CORA	df	NIM	CORA	df	NIM	CORA
ZIVL	4	294	1110	4	362	1430	4	278	1732
rSPRA	6	666	53	7	661	NA	7	1270	698
rSTHHW	6	26	NA	7	844	20139	7	737	17271

Tabelle 13: Wald Statistik für die Variablen, die nur von NIM eingesetzt wurden. 'NA' bedeutet, dass die Statistik nicht berechnet werden konnte, da Q der Gleichung (4) singulär ist.

Variable	FR-NIM		VS-NIM		ZH-NIM	
	df	W_Y	df	W_Y	df	W_Y
AGRP	7	840	7	1104	7	2702
GESL	2	31	2	37	2	41
rNATI	7	NA	7	NA	7	1223
rBESIT	10	2384	10	4042	10	4897
BSKO	6	581	6	3099	6	5416
HABG	8	187	8	4628	8	1347

herstellen konnten. Die verwendete Statistik ist sehr sensibel auf geringfügige Abweichungen. Daher ist sie bei einer grossen Anzahl Beobachtungen meist signifikant.

CORA war bei der Korrektur der Sprache effizienter als NIM, ansonsten hatte NIM einen klaren Vorteil gegenüber CORA. Im Vergleich zur Tabelle 11 wurde anhand der Statistik W_Y eine statistische Distanz der Abweichung zur wahren Verteilung berechnet, womit die Resultate der obig genannten Tabelle relativiert werden.

NIM konnte die Randverteilung von AGRP wegen der beiden höchsten Altersklassen nicht beibehalten. Werden diese ausgeschlossen, so wird diese Randverteilung beibehalten. Dem liegt die Fehlergenerierung zugrunde. Da bei 0.5% der Daten das Alter zufällig zwischen 0 und 120 generiert wurde. Somit ist beinahe die Hälfte dieser generierten Alter über 65. In NIM wurde jedoch erschwert, dass eine Person mit mehr als 65 Jahren verjüngt wird, um die Struktur Rentner/Nicht-Rentner beizubehalten. Da bei einigen Variablen nach der Einsetzung/Korrektur immer noch ein grosser Teil an fehlenden Werten besteht sind die entsprechenden Werte von W_Y sehr gross.

Bemerkung: Da im Kanton Fribourg bei den Variablen rSPRA und rSTHHW keine zusätzlichen Kategorien für fehlende Werte kreiert werden mussten, sind die entsprechenden Freiheitsgrade (df) kleiner als jene bei VS und ZH.

Wiederherstellung der 'wahren' Werte Aufgrund eines Kriteriums in (Chambers, 2001), welches im Abschnitt 3.4 beschrieben wurde, kann die Wiederherstellung der 'wahren Werte' quantifiziert werden. Da die Resultate im Abschnitt weiter oben gezeigt haben, dass die Randverteilung keiner Variablen, weder von NIM noch von CORA, wiederhergestellt werden konnte, wird $\hat{V}(D_Y)$ aus 3.4 in den folgenden Resultaten nicht näher erläutert, da diese Schätzung die Beibehaltung der Verteilung voraussetzt.

Aus der Tabelle 14 ist ersichtlich, dass anhand von NIM das genaue Alter von Personen nicht eingesetzt

werden kann. Dies liegt einerseits an der Altersdistanz, die aufgrund von Altersgruppen definiert wurde. Andererseits ist das Spektrum möglicher Antworten zu gross, als dass NIM dessen Verteilung reproduzieren könnte. Hinzu kam, wie bereits erwähnt, das Problem mit der Fehlergenerierung des Alters. Dies war ebenso der Grund dafür, dass die Altersgruppen nicht wieder hergestellt werden konnten.

Ebenfalls fragwürdig scheint die Einsetzung bei den Variablen BESIT, BSKO und HABG. Die Variablen BESIT und HABG lassen Kombinationen von Werten zu, was zu einer grossen Anzahl erlaubter Antworten führt. Hingegen ist der Anteil der Werte, die wiederhergestellt werden konnte bei den anderen Variablen ansprechend. Bei der Stellung im Haushalt im Kanton Fribourg hat NIM die Variable STHH überproportional eingesetzt, siehe Tabelle 15.

CORA zeigt gegenüber NIM einen eindeutigen Vorteil in der Sprache auf. Bei den beiden anderen Variablen die verglichen wurden, gelang es NIM öfters, die wahren Werte wiederherzustellen.

Der Anteil der falsch eingesetzten Daten hängt von der Variablen und der Schicht ab. Im Allgemeinen kann gesagt werden, dass CORA und NIM im Wallis am meisten Probleme bekundeten. Bei NIM wird die 1% Grenze in den Variablen ALTJ, BESIT und HABG (nur VS) überschritten. CORA überschreitet diese Grenze bei der Variablen STHHW in den Schichten VS und ZH.

Tabelle 14: Effizienz von NIM: Gegenüberstellung von Variablen der VZ90 mit den eingesetzten Werten. 'RI_Y' steht für gleiche Werte, 'FI_Y' steht für unterschiedliche Werte.

Variable	Kanton	RI _Y	FI _Y	A _Y [%]	D _Y [%]
ALTJ	FR	127	2204	5.45	1.07
	VS	139	3271	4.08	1.36
	ZH	136	3636	3.61	1.05
AGRP	FR	1294	1037	55.51	0.50
	VS	1544	1866	45.28	0.77
	ZH	1625	2147	43.08	0.62
NATI	FR	673	92	87.97	0.04
	VS	819	105	88.64	0.04
	ZH	1483	350	80.91	0.10
GESL	FR	1296	259	83.34	0.13
	VS	1567	306	83.66	0.13
	ZH	1921	486	79.81	0.14
BESIT	FR	3279	2396	57.78	1.16
	VS	7608	7238	51.25	3.00
	ZH	7400	12396	37.38	3.59
BSKO	FR	288	300	48.98	0.15
	VS	1122	751	59.90	0.31
	ZH	1136	708	61.61	0.20
HABG	FR	1076	692	60.86	0.34
	VS	5894	6362	48.09	2.64
	ZH	2103	1591	56.93	0.46

5.3 Fazit

Bei Gegenüberstellung von deterministischen Korrekturen (CORA) und der Methode NIM zeigt, dass die Anzahl Änderungen und die Vollständigkeit und Konsistenz nur bei NIM beeinflussbar sind. Weder mit NIM noch mit CORA konnten sämtliche fehlerhaften Haushalte/Records behandelt werden. Bei NIM

Tabelle 15: Effizienz von NIM und CORA: Gegenüberstellung der Werte von Variablen der VZ90 mit den eingesetzten Werten von NIM und den korrigierten Werten von CORA. 'RI_Y' steht für gleiche Werte, 'FI_Y' steht für unterschiedliche Werte.

Variable	Kt	NIM				CORA			
		RI _Y	FI _Y	A _Y [%]	D _Y [%]	RI _Y	FI _Y	A _Y [%]	D _Y [%]
ZIVL	FR	1119	178	86.28	0.09	871	1019	46.08	0.49
	VS	969	239	80.22	0.10	983	1324	42.61	0.55
	ZH	1513	495	75.35	0.14	1024	2069	33.11	1.02
SPRA	FR	1558	359	81.27	0.17	1616	198	89.08	0.10
	VS	3939	483	89.08	0.20	5507	266	95.39	0.11
	ZH	7393	1352	84.54	0.39	7500	740	91.02	0.21
STHHW	FR	5	69	6.76	0.03	1	4	20.00	0.00
	VS	10747	637	94.40	0.26	1668	7062	19.11	2.93
	ZH	14727	914	94.16	0.26	4654	6403	42.09	1.85

ist dies eine Frage von fehlenden Spendern und/oder der Anteil fehlerhafter Haushalte war zu gross. Die Spendersuche kann nach unseren Erfahrungen bei 9 Personen-Haushalten bei rund einem drittel erfolglos sein und ab 10 Personen-Haushalten kann sie gänzlich ohne Erfolg bleiben. Die Daten der VZ90, die als Wahrheit betrachtet wurden, sind anhand von deterministischen Korrekturen verarbeitet worden, bevor sie die gegenwärtige Form hatten. Der Test hat gezeigt, dass CORA trotz seinem deterministischen Charakter nicht wiederholbar ist. Der Grund dafür ist einerseits, dass nicht alle Korrekturen automatisierbar waren, andererseits wohl auch wegen des sequentiellen Ablaufs der deterministischen Korrekturen, der einen entscheidenden Einfluss auf die Fehlerlokalisierung hat. Somit blieben auch nach der Verarbeitung der Daten mit CORA noch fehlerhafte Records übrig.

Weder NIM noch CORA konnten die Randverteilungen wiederherstellen. Hingegen war die von NIM kreierte Randverteilung der Randverteilung der VZ90 bei den vergleichbaren Variablen näher als CORA. Einzig in der Variablen SPRA war CORA effizienter. Der Test hat auch gezeigt, dass die Altersdistanz, welche in NIM definiert worden war, zu restriktiv war, zumindest für diesen Test.

Die Wiederherstellung der wahren Werte konnte mit NIM, abgesehen von den Variablen AGRP, BESIT, BSKO und HABG, gewährleistet werden. Bei den vergleichbaren Variablen ist aufgefallen, dass NIM ganz klar öfters die wahren Werte eingesetzt hat, als CORA. Dies liegt wohl daran, dass die Aktionen von CORA nicht alle möglichen Kombinationen abdecken. Der Anteil an fehlerhaften Korrekturen an der Anzahl gemachter Korrekturen ist bei CORA für die Variable SPRA besser als der gleiche Anteil bei NIM.

Es konnte nicht garantiert werden, dass die generierten Fehler Inkonsistenzen nach sich gezogen haben. Dies hat die Wiederherstellung der wahren Werte und der Verteilung für NIM und CORA erschwert. Deshalb konnten nur relative Vergleiche angestellt werden.

Unter den gegebenen Voraussetzungen war CORA bei der Variablen SPRA eindeutig effizienter als NIM. Wobei NIM 34 verschiedene Sprachen und zusätzlich noch Doppelangaben der offiziellen Sprachen der Schweiz behandelt hat.

Von den Evaluationskriterien ausgehend, hat NIM einen kleinen Vorteil gegen über CORA, da die Kriterien öfters erfüllt werden. Ist der Anteil fehlerhafter Haushalte zu gross, führt dies in NIM zu einem Spenderproblem. Bei einer hohen Anzahl Antwortmöglichkeiten ist die Effizienz von NIM nicht mehr gewährleistet. Der Aufwand, um Variablen mit einer hohen Anzahl Antwortmöglichkeiten in CORA einzubauen, ist jedoch enorm, da für beinahe jede Kombination eine Aktion definiert werden müsste.

Analysen auf kleineren geographischen Niveaus konnten aus Zeitgründen nicht durchgeführt werden.

Aufgrund der durchgeführten Tests kann keine eindeutige Aussage gemacht werden, ob NIM oder CORA für eine bestimmte Variable, oder Variablengruppe, verwendet werden soll. Deshalb ist es angebracht,

fallweise zu entscheiden, welche Methode angewandt wird. Hierfür müssen jedoch aufwendige Tests durchgeführt werden.

6 Einsetzungen in der PHH, SHH und KHH der VZ2000

Die Anwendung von NIM in der Volkszählung 2000 ist die erste Anwendung einer derartigen Einsetzungssoftware in dieser Erhebung. Der Entscheid, für welche Variablen NIM verwendet wurde, basierte auf den Tests im Abschnitt 5. Aus Ressourcenmangel wurde auf weitere Vergleichstests zwischen NIM und CORA, wie dies in 5.3 vorgeschlagen worden war, verzichtet.

In einer ersten Phase wurden die Privathaushalte und Kollektivhaushalte der ganzen Schweiz und des Fürstentum Liechtensteins verarbeitet. Wenn im Folgenden von den Kantonen gesprochen wird, so sind der Einfachheit halber damit die 26 Kantone der Schweiz und das Fürstentum Liechtenstein gemeint.

Es wurden die Variablen 'Geburtsjahr', 'Geschlecht', 'Zivilstand', 'Jahr der letzten Änderung des Zivilstands', 'Elternschaft', 'Anzahl Kinder', 'Geburtsjahre der Kinder' sowie 'Stellung im Haushalt' verarbeitet. Die restlichen demographischen Variablen wurden vor allem durch deterministische Korrekturen (CORA) behandelt. In den folgenden Abschnitten wird die Anwendung von NIM an die Daten der Privathaushalte der Personen (PHH) und der Sammelhaushalte (SHH) der VZ2000 beschrieben. Im Abschnitt 6.3 wird dann auf die Kollektivhaushalte (KHH) eingegangen. Die Version von NIM, die für die PHH verwendet wurde, wird **NIM-P** genannt, jene die für die Sammelhaushalte verwendet wurde, **NIM-S** und jene die für die KHH verwendet wurde wird mit **NIM-K** bezeichnet. Wie aus dem Abschnitt 6.2.2 hervorgeht, ist diese Bezeichnungsweise nicht korrekt, da für die PHH mehrere Versionen von NIM verwendet wurden. Diese verschiedenen Versionen werden, wenn nicht speziell vermerkt, unter dem Namen NIM-P zusammengefasst. Wenn von NIM gesprochen wird, so bezeichnet dies im Folgenden die Einsetzungsmethode, welche durch die gleichnamige Software durchgeführt wurde.

6.1 Fragebogen PHE

Jede Person erhielt einen Personenfragebogen. Bei mehreren Wohnsitzen erhielt die Person an jedem ihrer Wohnsitze einen Personenfragebogen. Jeder Haushalt erhielt einen Haushaltsfragebogen, in welchem sämtliche Personen des Haushalts aufgeführt werden mussten. Dabei wurde jedoch nicht spezifiziert, ob die Haushaltsmitglieder, die aufgeführt werden müssen, wirtschaftlich oder zivilrechtlich im Haushalt, respektive in der Wohnung leben.

Die Frage bezüglich der Elternschaft (Frage 9) war in drei Teile aufgeteilt. Im ersten Teil musste auf die Frage geantwortet werden, ob man eines oder mehrerer Kinder hatte. Im zweiten Teil wurde die Anzahl Kinder gefragt, falls im ersten Teil mit ja geantwortet wurde. Im dritten Teil wurden die Geburtsjahre der ersten vier Kinder und des letzten, falls der Befragte mehr als vier Kinder hatte, gefragt. So wie die Frage gestellt wurde, konnten sich Personen ohne Kinder von der Frage nicht betroffen fühlen, so dass sie die gesamte Frage übersprangen.

Die Frage der Stellung im Haushalt vermisst eine klare Definition auf welche Person (Haushaltsvorstand/-stände) sich der Befragte beziehen soll. So ist es zum Beispiel nicht klar auf welchen Haushaltsvorstand sich die Antwort 'Sohn, Tochter, Stiefsohn, Stieftochter, Schwiegersohn, Schwiegertochter' bezieht. Allgemein kann gesagt werden, dass der Verwandtschaftsgrad zweier Personen im gleichen Haushalt nicht immer eindeutig bestimmt werden kann. Ebenso ist die Paarbildung bei Haushalten mit mehr als zwei Personen mit der gleichen Angabe nicht eindeutig durchführbar.

6.2 Einsetzungen mit NIM-P

6.2.1 Daten, Variablen und Schichtung

Daten Die Voraussetzung für die Verarbeitung von Haushaltsdaten mit NIM ist, dass sämtliche Haushalte existieren und jede Person dem entsprechenden Haushalt zugeordnet werden kann. Entweder muss die Person einem Privathaushalt (PHH), einem Kollektivhaushalt (KHH) oder dem Sammelgebäude der Gemeinde zugeordnet sein. Diese Voraussetzung war in der PHE nicht vollständig gegeben, da zum Zeitpunkt, als NIM die Daten verarbeitet hat, 282'109 Personen so genannten **Sammelhaushalten** (SHH) zugeteilt waren. Eine Person wurde einem Sammelhaushalt zugeteilt, wenn sie aufgrund mangelhafter Angaben nicht eindeutig einer Wohnung, respektive einem bestehenden Haushalt in einem Gebäude zugeordnet werden konnte. Diese Personen wurden dann in dem Gebäude, dem sie zugeordnet werden konnten, dem Sammelhaushalt zugeordnet. Konnte eine Person keinem Gebäude zugeordnet werden, so wurde diese Person dem Sammelhaushalt der Gemeinde zugeordnet. Zusätzlich hatte jede Gemeinde ein Sammelgebäude, welchem Personen zugeordnet wurden, die z.B. nur zivilrechtlich, aber ohne Wohnsitz in der Gemeinde, vertreten waren (z.B. Personen die im Pflegeheim einer Gemeinde leben, die nicht dieselbe ist wie ihre ehemalige Wohngemeinde, wo sie ihre Schriften noch deponiert haben aber nicht mehr leben). Ausser den letzt genannten Personen, fehlten die Personen der Sammelhaushalte in bereits bestehenden Haushalten oder der ganze Haushalt befand sich im Sammelgebäude. Im ersten Fall kann die Haushaltstruktur durch das Fehlen einer Person derart beeinträchtigt sein, dass NIM-P eine neue kreiert, die von der ursprünglichen Haushaltstruktur grosse Abweichungen haben kann.

Zur Illustration sei hier ein vereinfachtes Beispiel gegeben. Ein Dreipersonenhaushalt, bestehe aus einem Ehepaar und einem Sohn. Der Vater befinde sich im SHH. Mutter und Sohn bilden für NIM-P somit einen Zweipersonenhaushalt. Die Mutter (42) hat 'Ehefrau' als Stellung im Haushalt angegeben, der Sohn (22) 'Sohn'. Die Inkonsistenzen bezüglich der Stellung im Haushalt können gelöst werden, indem die Mutter zur 'Alleinerziehenden' wird oder indem der Sohn zum Ehemann wird, wobei das Ehepaar in diesem Fall einen Altersunterschied von 20 Jahren aufweisen würde, was aufgrund der vorgegebenen erlaubten Altersunterschiede akzeptiert wird. Die Haushaltstrukturen dieser beiden konsistenten Haushalte hat hingegen nichts mehr mit der 'Ehepaar mit Kind' Haushaltstruktur zu tun, welche unter den beschriebenen Voraussetzungen nicht erreicht werden kann.

Für die Verarbeitung der Personendaten mit NIM-P waren die Haushalte durch sämtliche Personen definiert, die in der VZ2000 zivilrechtlich oder wirtschaftlich dem Haushalt zugeordnet wurden. Somit wurden Personen mit mehreren Wohnsitzen mehrfach und unabhängig voneinander bearbeitet, für jeden Wohnsitz einmal. 7'368'591 Personen in Privathaushalten und Sammelhaushalten wurden verarbeitet, wovon rund 170'000 mehrfach gezählt wurden, da sie mehrere Wohnsitze hatten (normalerweise zwei), und wie schon erwähnt, 282'109 Personen in Sammelhaushalten.

Variablen Die Variablen, welche mit NIM-P eingesetzt wurden, waren: 'Alter', 'Geschlecht', 'Zivilstand', 'Änderungsjahr des Zivilstands', die Variablen bezüglich der Eltern und die 'Stellung im Haushalt' (siehe Anhang A.1 Tabelle 22). Dies entspricht den Fragen 1 bis 3, 9 und 10 des Fragebogens. Die Personen in Sammelhaushalten wurden einzeln und ohne die Variable Stellung im Haushalt verarbeitet. Im Vergleich zu den Variablen, welche STATCAN zur Verfügung stehen, muss vor allem hervorgehoben werden, dass in der VZ2000 vor der Verarbeitung mit NIM weder eine Referenzperson bestimmt werden konnte, noch eine Variable 'Paar' kreiert werden konnte. Diese beiden Variablen wurden erst nach der Verarbeitung mit NIM kreiert. Für beide Variablen war die Stellung im Haushalt die Basis. Die Möglichkeit diese Variablen zu Testzwecken vor der Verarbeitung mit NIM-P zu kreieren, war nicht vorhanden. Deshalb kann nicht abgeschätzt werden, welche Auswirkungen NIM-P auf diese Variablen hatte.

Schichtung Die Daten wurden nach Kanton und Haushaltgrösse getrennt verarbeitet. Aufgrund von Tests wurde festgestellt, dass in gewissen Schichten bereits ab Haushalten mit 7 Personen die Anzahl verfügbarer Spender gering war, deshalb wurde entschieden, die Haushalte mit 7 bis 10 Personen nicht mehr innerhalb der einzelnen Kantone, sondern innerhalb der Grossregionen (NUTS2) zu verarbeiten. Die Haushalte von Liechtenstein wurden der Grossregion Ostschweiz zugeordnet. Bei Haushalten mit mehr als 10 Personen wurde wegen der zu geringen Anzahl Spender keine Schichtung nach Kantonen/NUTS vorgenommen. In einigen Schichten mit Haushalten mit mehr als 10 Personen, verletzten sämtliche Haushalte der Schweiz die Plausibilitätsregeln. Deshalb wurden die Haushalte mit mehr als 10 Personen in zwei Phasen bearbeitet. Zuerst wurde jede Person individuell ohne Einbezug der Haushaltszusammensetzung bearbeitet. In der zweiten Phase wurden dann die Haushalte wie gewohnt als Einheit behandelt. Sämtliche Daten konnten in der zweiten Phase eingesetzt werden, auch jene, welche in der ersten Phase eingesetzt wurden. Mit diesem Vorgehen wurden bei sämtlichen fehlerhaften Haushalten mit bis zu 15 Personen Daten eingesetzt, so dass diese Haushalte sämtlichen verwendeten Regeln genügten. Es wurde ein nicht abschätzbarer Qualitätsverlust in Kauf genommen, da nicht alle Daten auf die gleiche Weise behandelt wurden. Es ist möglich, dass während der ersten Phase Daten eingesetzt wurden, welche unter Einbezug der Haushaltszusammensetzung nicht eingesetzt worden wären. Die Haushalte mit mehr als 15 Personen wurden nicht geliefert, da sie gleichzeitig mit der Verarbeitung mit NIM-P manuell von der ehemaligen Sektion BHS bearbeitet wurden.

6.2.2 Anpassungen an NIM

Gegenüber der Version von NIM, die von STATCAN geliefert wurde, mussten einige Anpassungen an die Charakteristiken der VZ2000 gemacht werden.

- Unter Mithilfe der SMR SA wurde das gesamte Programm auf UNIX portiert.
- Die Definitionen der bestehenden Variablen wurden an die VZ2000 angepasst.
- Weitere Variablen wurden hinzugefügt.
- Distanzen wurden neu definiert (und programmiert)
- Eine Zufallsauswahl aus den besten Einsetzungsaktionen wurde eingeführt.
- Der input/output musste angepasst werden.
- Eine Funktion wurde hinzugefügt, die es erlaubt, aufgrund von Hilfsvariablen (Flags) zu verhindern, dass ein Variablenwert einer bestimmten Person verändert wird - es sei denn, zwei Variablenwerte, die nicht verändert werden dürfen, sind untereinander inkonsistent. Das Ziel war, Variablenwerten, die vorbedruckt waren oder rückgefragt wurden, einen besonderen Qualitätsstatus zu verleihen.
- NIM wurde an Mehrfachangaben angepasst, wobei ein bereits vorhandener Wert als Spenderwert bevorzugt wurde.

Aufgrund unterschiedlicher Definitionen der Population mussten mehrere Versionen von NIM für die Produktion bereitgestellt werden (c.f. Tabelle 16). Unter anderem mussten für das Tessin eigens eine Version erarbeitet werden, da die Tessiner einen fehlenden Wert der Variable 'Anzahl Kinder' mit '0' codierten anstatt mit 'NULL'. Diese fehlerhafte Codierung wurde beim Einlesen der Daten in NIM behoben, wenn 'PAR=nein' und die Variablen 'P1J' bis 'PLJ' leer waren.

Tabelle 16: Verschiedene Versionen von NIM, die sich aufgrund der Population und der Verarbeitung unterscheiden. 'Pers.' steht für Personen, HH für Haushalte.

Population	Unterschiede zum Basisprogramm	Name von NIM
PHH 1-10 Pers., HH	Basis	PunixM
PHH-TI 1-10 Pers., HH	wie PunixM, input verschieden	PunixMTI
SHH, Einzelpers.	nicht alle Regeln	PunixS
SHH-TI, Einzelpers.	wie PunixS, input verschieden	PunixSTI
PHH 11-15 Pers., Einzelpersonen	wie PunixS mit Bevorzugung der Pers. aus gleichem HH als Spender, output verschieden	PunixR1
PHH 11-15 Pers., HH	wie Basis, input verschieden	PunixR2

Distanzen Die verwendeten Distanzen sind im Anhang [A.4](#) aufgeführt.

Kontrollregeln Die verwendeten Regeln stützten sich auf den Regeln der VZ90 (Boruvka, 1994). Diese wurden in einem ersten Schritt von der ehemaligen Sektion BHS an die VZ2000 angepasst und mussten danach zum Teil an die Verarbeitung mit NIM angepasst werden. In Anhang [A.2](#) sind die Regeln, die angewandt wurden, in den Tabellen [23](#) bis [26](#) aufgeführt. In den Testphasen wurden noch zusätzliche Regeln angewandt, die jedoch unerwünschte Nebeneffekte hatten. Diese Regeln sollten die Plausibilität zwischen den Geburtsdaten der Kinder, wie sie von den Eltern angegeben wurden, und den Geburtsdaten der Kinder im Haushalt gewährleisten. Ebenso die Geburtsdaten eines Haushaltsvorstands, wenn mindestens ein Elternteil im Haushalt lebt. Da es aufgrund der Fragestellung der Stellung im Haushalt nicht ohne weiteres möglich ist, die verwandtschaftlichen Beziehungen zwischen den Haushaltsmitgliedern eindeutig zu bestimmen, musste auf diese Regeln verzichtet werden. Somit ist mit Inkonsistenzen auch bei 'eindeutigen' Fällen dieser Art zu rechnen.

Eine fundamentale Regel besagt, dass in jedem Haushalt mindestens ein Haushaltsvorstand vorhanden sein muss. Diesem Anspruch wird in der Fragestellung jedoch nicht Rechnung getragen. Zum Beispiel haben Mitglieder einer Wohngemeinschaft richtigerweise alle 'Mitglied einer Wohngemeinschaft' angekreuzt, womit aber kein Haushaltsvorstand existiert. Durch Anwendung dieser Regel wurden vor allem Wohngemeinschaften 'zerstört'.

Parameter Die verwendeten Parameter sind im Anhang [A.4](#) aufgeführt. Es wurden ähnliche Größen gewählt, wie sie STATCAN 1996 in ihrer Volkszählung verwendet hat, ausser, dass in der VZ2000 mehr Gewicht auf die Einsetzung einer möglichst kleinen Anzahl Variablen gelegt wurde.

Analysekriterien Da auf den Vergleich mit der 'Wahrheit' verzichtet werden musste, konnten die Kriterien des Abschnitts [3.4](#) nicht angewandt werden. Die Analysen wurden anhand von beschreibenden Analysen und Vergleichen mit externen Daten durchgeführt.

Insbesondere wurde der statistische Einfluss von NIM-P analysiert, indem die Änderungen der Verteilungen gegenüber den Daten vor NIM-P überprüft wurden und spezielle Konstellationen analysiert wurden.

6.2.3 Produktion und Analysen von NIM-P

Produktion Die Produktion der PHH und SHH wurde auf mehrere Server (800MHz SUN) verteilt. Auf dem Produktionsserver der VZ2000 standen 12 Prozessoren zur Verfügung, auf zwei weiteren Servern standen zusammen nochmals 6 Prozessoren zur Verfügung. Auf diese Weise konnte die Produktion auf 3

Tage beschränkt werden, wobei dabei nicht immer alle 18 Prozessoren ausgelastet waren. Aus der Tabelle 30 im Anhang A.5 ist ersichtlich, dass die Verarbeitungszeit in den grössten Schichten (ZH 4-Personen Haushalte) mehr als 4.5 Stunden betrug.

Im Anhang des zweiten Bandes befindet sich eine vereinfachte Darstellung des Verarbeitungsschemas. Für die Verarbeitung der Daten mit NIM mussten die Daten aus der ORACLE Datenbank in ASCII Dateien exportiert werden. Dabei musste die Kodierung der Variablen angepasst werden, z.B. der Wert 'NULL' wurde, sofern nicht anders spezifiziert, in '-1' umcodiert. Anhand des Exports wurden für jeden Kanton und für jede Haushaltsmitglieder Anzahl eine separate Text Datei kreiert, wobei die Sammelhaushalte jedes Kantons in eine spezielle Datei exportiert wurden. Die weiteren Schichtungen wie sie im Abschnitt 6.2.1 beschrieben wurden, wurden erst nachträglich anhand von Shell-Scripts erzeugt ('separprod.sh' und 'nutskt.sh'). Weitere Shell-Scripts dienten der automatisierten Verarbeitung der Daten mit NIM. Dies war vor allem nötig, da die Anzahl Haushaltsmitglieder und die Anzahl verwendeter Regeln (resp. 'Decision Logic Tables') aufgrund der zu verarbeitenden Daten in zwei Dateien, die von NIM verwendet werden, geschrieben werden mussten. Zudem musste vor jedem Durchlauf NIM neu kompiliert werden. Dieser Ablauf wurde anhand der Scripts 'run.sh' und 'build.sh' (respektive 'runti.sh' und 'buildti.sh') durchgeführt. Nach der Verarbeitung der Daten wurden die unverarbeiteten Daten und die veränderten Daten auf einem Server zusammen analysiert. Hierfür war ein manueller Transfer der Dateien nötig. Vor der Analyse wurden die Dateien gruppiert ('saskt.sh' - für unbearbeitete Daten, 'moloskt.sh' und 'extfilekt.sh' für die veränderten Daten). Die Analyse wurde mit SAS durchgeführt ('panalyse.sas', 'pmacros.sas'). Sobald die durchgeführten Analysen keine groben Fehlverhalten aufzeigten, wurden die veränderten Daten in ein speziell für die Schlussbereinigung kreiertes Schema auf der Datenbank eingespielen, wobei die Umkodierungen, die während des Exports gemacht wurden, wieder rückgängig gemacht werden mussten. Die Änderungen durch NIM wurden geflagt. Anhand dieses Schemas wurden von der ehemaligen Sektion BHS weitere Analysen direkt auf der Datenbank gemacht. Dabei ging es um die Analyse der Datentransfers und der Resultate der Einsetzungen. Nach Akzeptieren der Resultate von NIM, wurde die Datenbank der Produktion der VZ2000 anhand des Schemas aufgefrischt.

Nebst den Umkodierungen der Variablen mussten die Identifikatoren der Personen und Haushalte aus Gründen des Datenschutzes umcodiert werden. Im Nachhinein stellt sich die Frage, ob dies sinnvoll war, da in den veröffentlichten Mikrodaten, die Identifikatoren die umcodiert worden sind, geliefert werden.

Beschreibende Analysen Aus der Tabelle 17 ist ersichtlich, dass NIM im Tessin mit rund 8% am wenigsten Änderungen machte. Eine Erklärung dafür ist, dass das Tessin ein eigenes Erhebungssystem hatte und vereinbart war, dass die Plausibilitäten bereits vor der Lieferung ans BFS durchgeführt werden. In der Tabelle 17 sind rund 120'000 Änderungen im Tessin nicht berücksichtigt, welche die Variable Anzahl Kinder betrifft. Diese wurde von den Tessinern nicht gleich codiert, wie im BFS. Dieser Unterschied wurde beim Einlesen in NIM behoben, weshalb diese nicht als Änderungen betrachtet werden. Es besteht keine Erklärung für den relativ hohen Anteil an Änderungen im Fürstentum Liechtenstein. Ansonsten wurde bei 20% bis 25% der Personen mindestens ein Variablenwert geändert. Aus der Tabelle 31 im Anhang A.5 ist ersichtlich, dass NIM 2'017'952 Variablenwerte geändert hat. Dies entspricht 2.28% aller Variablenwerte. Mehr als die Hälfte davon wurde in der Variablen Stellung im Haushalt geändert. Mit beinahe 23% aller veränderten Variablenwerte folgt die Variable Eltern. Die Fragen 9 (Eltern) und 10 (Stellung im Haushalt) mussten derart oft geändert, dass deren Qualität nicht gewährleistet werden kann. Hinzu kommt dass die Frage 9 bereits vor NIM durch deterministische Korrekturen (CORA_Clear) teilweise bereinigt wurde. So wurde zum Beispiel bei Personen, welche jünger als 12 Jahre sind, die Angaben bezüglich der Anzahl Kinder sowie der Geburtsdaten der Kinder gelöscht und Eltern auf 'keine Kinder' gesetzt. Ähnliche Änderungen, die hier nicht genauer beschrieben werden, wurden ebenfalls vor NIM vorgenommen. Die hohe Frequenz der Änderungen im Vergleich zu den anderen Variablen rührt sicherlich daher, dass diese Fragen nicht vorbedruckt waren und dass durch diese Fragen keine Rückfragen

ausgelöst wurden. Hinzu kommt, dass die Fragestellung in der Frage bezüglich der Stellung im Haushalt nicht eindeutig ist und somit der Handlungsspielraum für die Beantwortung zu gross ist. In der Tabelle 30 ist die Verarbeitung in den Schichten zusammengefasst. Von den 282'109 Personen in Sammelhaushalten erfüllten 84'661 Personen die Plausibilitätsregeln nicht. Dies sind rund 30% aller Personen in SHH, was klar über dem Durchschnitt der anderen Schichten liegt. Dies ist umso erstaunlicher, da die Stellung im Haushalt in den SHH keinen Plausibilitätstests unterzogen wurde. Hingegen zeigt dies auch auf, dass sich ein grosser Teil dieser Personen nicht ohne Grund in den SHH befunden haben.

Tabelle 17: Anzahl Personen in Privathaushalten nach Kanton und Bearbeitung (eingesetzt, nicht eingesetzt). Sowie der prozentuale Anteil am Total der Population des Kantons (eins. [%], n.eins. [%]) und der Gesamtbevölkerung der Schweiz und Liechtenstein (eins./CH), wobei Personen mit mehreren Wohnsitzen mehrfach gezählt wurden.

Kanton	Total Pers.	eins.	eins. [%]	n.eins.	n.eins. [%]	eins./CH [%]
ZH	1244660	281307	22.60	963353	77.40	3.82
BE	962778	200830	20.86	761948	79.14	2.73
LU	352018	82949	23.56	269069	76.44	1.13
UR	35823	7619	21.27	28204	78.73	0.1
SZ	129832	28070	21.62	101762	78.38	0.38
OW	33237	9066	27.28	24171	72.72	0.12
NW	38182	8789	23.02	29393	76.98	0.12
GL	38546	7161	18.58	31385	81.42	0.1
ZG	99612	21870	21.96	77742	78.04	0.3
FR	242845	52308	21.54	190537	78.46	0.71
SO	245991	48997	19.92	196994	80.08	0.66
BS	186843	47147	25.23	139696	74.77	0.64
BL	260693	48591	18.64	212102	81.36	0.66
SH	73398	14845	20.23	58553	79.77	0.2
AR	53767	11132	20.70	42635	79.30	0.15
AI	14992	1825	12.17	13167	87.83	0.02
SG	457623	96201	21.02	361422	78.98	1.31
GR	195247	45096	23.10	150151	76.90	0.61
AG	549083	105288	19.18	443795	80.82	1.43
TG	230399	47538	20.63	182861	79.37	0.65
TI	310317	23645	7.62	286672	92.38	0.32
VD	643624	149206	23.18	494418	76.82	2.02
VS	283993	67705	23.84	216288	76.16	0.92
NE	169871	33923	19.97	135948	80.03	0.46
GE	411476	106135	25.79	305341	74.21	1.44
JU	70341	17102	24.31	53239	75.69	0.23
FL	33400	12165	36.42	21235	63.58	0.17
Total	7368591	1576510	21.39	5792081	78.61	21.40

Einfluss auf die Verteilung Aus der Tabelle 32 im Anhang A.5.2 ist ersichtlich, dass die Einsetzungen des Geburtsjahrs nur einen kleinen Einfluss auf die 5-jahresklassen des Alters hatten. Einzig die beiden Klassen 15-19 und 65+ sahen ihren Anteil gegenüber den anderen Veränderungen überdurchschnittlich verändern. Im Durchschnitt wurde die relative Häufigkeit um 0.07 erhöht, mit einer Standardabweichung von 0.13. Da die Abweichungen dennoch gering sind, werden die übrigen Variablen nicht als Hilfsvariablen verwendet, um mehr Klarheit darüber zu erhalten, von welchen Schichten (Kanton, Anzahl Personen im Haushalt, Geschlecht, etc.) diese überproportionalen Veränderungen herrühren.

Vergleich mit anderen Datenquellen Die Vergleiche mit anderen Datenquellen sollten allfällige gravierende Probleme der Verarbeitung mit NIM-P zu Tage bringen. Hierfür wurden die Daten nach der Verarbeitung mit NIM-P mit Daten der VZ90 und der ESPOP vom Dezember 2000 verglichen. Personen mit mehreren Wohnsitzen sind in den Daten, die NIM-P verarbeitet hat, mehrfach vorhanden. Die Definition der Grundgesamtheiten der VZ90 und der ESPOP sind von jener der Daten nach NIM-P verschieden. Deshalb müssen die Vergleiche mit Vorsicht genossen werden.

Vergleich der VZ2000 mit der VZ90 Nach der SB-NIM wurden die Häufigkeiten der Variablen 'Alter' (5-jahresklassen), 'Geschlecht', 'Zivilstand', 'Nationalität' (Schweizer, Ausländer), 'Ausländerausweis' und 'Stellung im Haushalt' mit den Daten der VZ90 verglichen. Die Häufigkeiten jeder Variablen wurden auf Landes- sowie auf der Ebene der Kantone verglichen. Ebenso wurde das Geschlecht mit folgenden Variablen und Kombinationen von Variablen gekreuzt: Zivilstand, Nationalität, Alter, Ausländerausweis, Stellung im Haushalt, Nationalität und Zivilstand, Nationalität und Alter, Zivilstand und Nationalität und Alter. Die Variablen Nationalität und Ausländerausweis wurden erst nach diesen Vergleichen bereinigt.

In Absprache mit den Fachspezialisten wurde ein Unterschied in der Randverteilung von bis zu $\pm 5\%$ als zulässig betrachtet. Auf Landesebene wurden zudem die Fälle untersucht, welche mehr als 1% von den Werten der VZ90 abweichen. Die gesamte Population der jeweiligen VZ des betreffenden Kantons wurde als 100% definiert (bei den Tabellen der ganzen Schweiz, war dies die gesamte Bevölkerung).

Eine 5% Abweichung zu den Werten der VZ90 wurde nirgends festgestellt. 1% Abweichungen und grösser, waren einzig für die Variablen Nationalität (Anstieg der ausländischen Bevölkerung), Ausländerausweis (Anstieg der Ausweise C) und dem Alter feststellbar. Beim Alter handelt es sich sehr wahrscheinlich um Verschiebungen innerhalb der Alterspyramide. Aus Platzgründen wurde darauf verzichtet die Vergleichstabellen aufzuführen.

Aus der Tabelle 18 geht hervor, dass nach NIM-P 5 Charakteristiken in der VZ2000 sind, welche in der VZ90 nicht vorhanden waren. Es handelt sich hierbei um Personen, welche jung verheiratet, geschieden oder verwitwet sind. Diese Fälle entsprechen den verwendeten Plausibilitätsregeln. Verluste von Charakteristiken wurden nicht festgestellt.

Aufgrund der obigen Resultate wurde die Verarbeitung mit NIM-P als genügend eingeschätzt.

Vergleich der VZ2000 mit ESPOP Von der Sektion 'Bevölkerung' (BEV) wurden Kreuztabellen der ESPOP von Ende Dezember 2000 zur Verfügung gestellt. Für jeden Kanton wurden folgende Variablen verglichen:

- Alter (5-jahresklassen)
- Geschlecht
- Nationalität.

Auf Niveau Schweiz wurden dieselben Variablen und zusätzlich noch der Zivilstand verglichen.

Aus der Tabelle 19 geht hervor, dass einzig in fünf Kantonen Unterschiede vorhanden sind, welche 1% übersteigen. Es handelt sich dabei jeweils um die Altersklassen 20-24 Jahre oder 25-29 Jahre. Für Schweizer wurde festgestellt, dass in der ESPOP eine Schweizerin zwischen 10 und 14 verheiratet ist und dies in der VZ2000 nicht der Fall ist. Verwitwete Schweizer existieren in der ESPOP erst ab 20 Jahren, hingegen sind in der VZ2000 vier Frauen und drei Männer vorhanden, welche vor 20 bereits verwitwet sind, siehe Tabelle 18. Die ESPOP führt eine geschiedene Frau zwischen 15 und 19. Diese ist in der VZ2000 nicht zu finden. Hingegen bestehen in der VZ2000 19 geschiedene Männer zwischen 15 und 19. Bezüglich der Ausländer konnten die im letzten Abschnitt erwähnten Unterschiede nicht verglichen

Tabelle 18: Charakteristiken die in der VZ2000 vorkommen aber in der VZ90 nicht vorkommen. Nach Kanton, Geschlecht, Zivilstand, Nationalität und 5-Jahres Altersklasse.

Kanton	Geschlecht	Zivilstand	Nationalität	Alter	Kanton	Geschlecht	Zivilstand	Nationalität	Alter
	Frauen	Verwitwet	CH	15-19		Männer	Verwitwet	CH	15-19
CH		4			CH		3		
FR		1			GL		1		
BL		1			AG		1		
VD		1			GE		1		
GE		1							
	Männer	Geschieden	CH	15-19		Männer	Verheiratet	Ausländer	10-14
CH		19			CH		13		
ZH		4			ZH		1		
BE		4			BE		2		
LU		1			LU		2		
ZG		1			SO		1		
SO		2			BL		2		
AG		3			SH		1		
TI		1			SG		1		
GE		2			AG		2		
JU		1			GE		1		
	Frauen	Verheiratet	Ausländer	10-14					
CH		11							
ZH		3							
BE		2							
LU		1							
FR		1							
SH		2							
VS		1							
GE		1							

werden, da keine diesbezüglichen Daten zur Verfügung standen.

Da sich die Unterschiede in engen Grenzen halten, wurde auf Änderungen in der Verarbeitung mit NIM-P verzichtet. Der Aufwand wäre zu gross gewesen, wenn die neu dazugekommenen Charakteristiken mit NIM-P hätten verhindert werden sollen. Denn alle Privathaushalte hätten nochmals mit NIM-P verarbeitet werden müssen.

Tabelle 19: Unterschiede in den Resultaten der ESPOP und der VZ2000 nach NIM, welche grösser als 1% sind.

Kanton	Altersklasse	ESPOP [%]	VZ2000 [%]	Differenz
OW	25-29	5.71	6.83	1.12
NW	20-24	4.85	5.90	1.15
AR	20-24	4.48	5.89	1.41
SG	20-24	5.91	7.11	1.20
GR	25-29	6.00	7.13	1.13

Änderungsjahr des Zivilstands und Bildung der Paare Da sämtliche Jahrzahlen zu viele Modalitäten haben, um Jahres-Distanzen in NIM-P zu berechnen, wurden hierfür Altersklassen verwendet. Für die Variable Zivstand_jj wurden die 5-Jahresklassen verwendet. Die Bildung von Paaren konnte nicht vor der Verarbeitung mit NIM-P vorgenommen werden, da dieses Merkmal nicht gefragt worden war und erst nach der Verarbeitung mit NIM-P aufgrund der Stellung im Haushalt konstruiert werden konnte. Somit war kein Hilfsmittel zu Verfügung, um die Änderungsjahre des Zivilstands von potentiellen Paaren abzugleichen.

Die Variable Zivstand_jj wird unter anderem bei der Bildung der Paare herangezogen.

Frage 9: Eltern und Geburtsjahre der Kinder NIM-P hat bei 462'554 Personen die Frage 9 eingesetzt. 1996 Personen hatten ein fehlendes Geburtsdatum und 135'404 waren jünger als 16 Jahre. 426'124 Personen in Privathaushalten hatten keinerlei Angaben in der Frage 9. Es wird die Annahme getroffen, dass bei einer Mehrheit der Personen in Privathaushalten davon ausgegangen werden kann, dass sie diese Frage nicht beantwortet haben, weil sie sich nicht betroffen fühlten. Mit anderen Worten wird davon ausgegangen, dass die Mehrheit der Personen, welche die Frage 9 nicht ausgefüllt haben, keine Kinder haben.

Grundsätzlich nimmt NIM die kleinste Anzahl Änderungen vor, damit ein Haushalt keine Regeln mehr verletzt. Die kleinste Änderung, wenn keine Angaben vorhanden sind, besteht darin, die Variable 'Eltern_Cd' auf 1 zu setzen (keine Kinder) und die restlichen Variablen der Frage 9 leer zu lassen. Die Wahrscheinlichkeit, dass Personen dabei sind, welche Kinder haben ist sicherlich nicht 0. Zudem hat NIM bei 32'754 Personen, welche Angaben in der Frage 9 hatten, 'Eltern_Cd' auf 1 gesetzt und die vorhandenen Angaben gelöscht. 18'681 von diesen Fällen hatten nur 'Eltern_Cd'=2 und sonst keine weiteren Angaben. Die Regeln bezüglich des nötigen Altersunterschieds zwischen Kindern und Eltern haben im weiteren dazu geführt, dass 'Eltern_Cd' auf 1 gesetzt wurde und die restlichen Angaben gelöscht wurden.

Im Folgenden werden einige Überlegungen angestellt, ob die Frage 9 mit deterministische Korrekturen besser bereinigt hätte werden können. Eine Übernahme der Geburtsjahre allfälliger Kinder im Haushalt muss ausgeschlossen werden, da es einerseits nicht möglich ist, den Verwandtschaftsgrad zweier Personen im Haushalt zu bestimmen. Andererseits wohnen in vielen Haushalten nicht mehr alle Kinder; diese könnten somit nicht eingesetzt werden. Von einer Übernahme der Angaben des Partners ist ebenfalls abzuraten, da wiederum der Verwandtschaftsgrad nicht bestimmt werden kann, d.h. vor den Einsetzungen der Stellung im Haushalt ist es nicht möglich den Partner im Haushalt zu finden. Die Möglichkeiten für Korrekturen gehen nicht weiter und zusätzliche Korrekturen sind kaum zu bewerkstelligen. Deshalb muss von deterministischen Korrekturen für diese Frage abgeraten werden.

Frage 10: Stellung im Haushalt Nach dem ersten Durchlauf von NIM-P wurde festgestellt, dass die Anzahl Haushalte mit Kindern, welche älter sind als die Eltern massiv zugenommen hat. Diese Haushalte weisen mindestens eine Person auf, welche 'Stellung im Haushalt = Sohn, Tochter, Stiefsohn, Stieftochter, Schwiegersohn, Schwiegertochter' hat, deren Alter grösser ist als das Alter von mindestens einem Haushaltsvorstand. Aufgrund der Fragestellung der Stellung im Haushalt kann diese Kombination nicht ausgeschlossen werden. Deshalb wurde keine Regel verwendet, welche solche Kombinationen verhindert. Rund 500 Haushalte, mit dieser Kombination wurden von NIM-P nicht verändert, d.h. diese Haushalte hatten keinerlei Fehler. Rund 6'000 Haushalte kamen beim Durchlauf mit NIM-P dazu. Gleichfalls wurde die Kombination untersucht, bei der mindestens eine Person im Haushalt 'Stellung im Haushalt = Vater, Mutter, Stiefvater, Stiefmutter, Schwiegervater, Schwiegermutter' hat und die gleichzeitig jünger ist als mindestens ein Haushaltsvorstand. Auch diese Kombination kann vorkommen, weshalb in NIM-P ebenfalls keine Regel verwendet wurde, die solche Kombinationen verhindert. Rund 200 Haushalte mit dieser Kombination wurden von NIM-P nicht verändert. Rund 2'000 Haushalte kamen beim Durchlauf mit NIM-P dazu. Es wurde analysiert, wie NIM-P diese Fälle konstruiert hat. Dabei drängte sich die Schlussfolgerung auf, dass keine Möglichkeit vorhanden war, eindeutige Regeln zu verwenden, um solche Kombinationen in 'plausiblen' Fällen zu akzeptieren, sie sonst aber zu verhindern und abzuändern. Die beiden oben genannten Kombinationen sind beide möglich aber eher unwahrscheinlich. Dennoch gibt es Zweifel daran, dass eine so grosse Anzahl solcher Kombinationen existiert. Schlussendlich wurden diese nicht vollständig verhindert, indem die Haushalte mit diesen Konstellationen, die keinerlei Regeln verletzten, beibehalten wurden. Alle anderen Haushalten wurden diese Konstellationen verwehrt. In der Tabelle 20 ist die Anzahl Haushalte mit solchen Konstellationen nach der endgültigen Verarbeitung mit NIM-P aufgeführt; diese existieren somit in den Daten nach NIM-P.

Tabelle 20: In der ersten Spalte ist die Konstellation des Haushalts. Kind steht für sämtliche möglichen Antworten in der Rubrik Sohn/Tochter/..., HH-Vorstand steht für Haushalts Vorstand und Eltern steht für sämtliche möglichen Antworten in der Rubrik Vater/Mutter/... . In der zweiten Spalte ist die Anzahl Haushalte mit der jeweiligen Konstellation.

Konstellation	Anzahl Haushalte
Alter Kinder > Alter HH-Vorstand	525
Alter Eltern < Alter HH-Vorstand	227

Variablen spezifische Analysen Es wird folgend kurz aufgezeigt, weshalb vor allem NIM-P, aber auch NIM-S Änderungen vorgenommen haben, die nicht zu erwarten und nach der Verarbeitung mit NIM-P als nicht wünschenswert erachtet wurden. Es werden Privathaushalte mit 1 bis 10 Personen und die Sammelhaushalte untersucht.

Änderungen des Geburtsjahres In insgesamt 10'521 Haushalten wurde das Geburtsjahr verändert, wovon in 5'106 Fällen keine Regel bezüglich des Alters verletzt worden war. In mehr als der Hälfte dieser Fälle (2'267 HH) wurden nur Regeln der Stellung im Haushalt verletzt. Vor allem fehlte ein Haushaltsvorstand oder der HH wies eine ungerade Anzahl verheirateter Paare auf. Weiter wurde das Geburtsjahr aufgrund von Inkonsistenzen bezüglich des Jahres der letzten Änderung des Zivilstands und der Geburtsjahre der Kinder (auf dem gleichen Fragebogen) geändert.

Änderungen des Zivilstands Bei 50'421 Personen in 37'940 Haushalten wurde die Angabe des Zivilstands abgeändert, wovon 4'400 Personen in 2'715 HH fehlende oder ungültige Werte aufwiesen. 5'904 Personen befanden sich in Haushalten, welche keine Regel bezüglich des Zivilstands verletzten (4'409 HH). Hauptverantwortlich für die Änderungen des Zivilstands waren Inkonsistenzen und fehlende Werte in der Stellung im Haushalt.

Stellung im Haushalt In 17'109 Haushalten wurden 18'505 Personen, die ursprünglich 'Sohn / Tochter' als Stellung im Haushalt hatten, zum Haushaltsvorstand. In 14 HH wurde keine Regel bezüglich der Stellung im Haushalt verletzt. Von den 18'505 Personen waren 8'259 alleinlebende Kinder. In 5'336 Haushalten fehlte ein Haushaltsvorstand, wovon 608 Haushalte zu junge Eltern hatten. Weitere 247 HH wiesen eine ungerade Anzahl verheirateter Partner auf. In all diesen Fällen, hatte NIM, wenn überhaupt, nur einen kleinen Handlungsspielraum, um die ursprünglichen Angaben zur Stellung im Haushalt beizubehalten. Offensichtlich wiesen diese Haushalte Phantasieangaben auf oder die Haushaltszusammensetzung war nicht vollständig.

68'850 Personen wurden durch NIM-P zu Konkubinen-Paaren. Hiervon lebten 30'798 Personen in Haushalten ohne Haushaltsvorstand, 20'344 in Haushalten mit einer ungeraden Anzahl Personen mit der Angabe 'Partner'. 7'525 weitere lebten in Haushalten, in denen überhaupt keine Angaben zur Stellung im Haushalt vorhanden waren. Weitere 9'205 hatten die Angabe 'alleinlebend', obwohl sie mit anderen Personen im Haushalt lebten. Aufgrund der Definition der Distanz der Variablen 'Stellung im Haushalt' kostete es NIM-P weniger, 'Partner' einzusetzen, als zum Beispiel 'Alleinerziehende Person' und 'Ehemann, -frau'.

Fazit Abgesehen von nicht erwünschten Randeffekten, konnte im Vergleich mit anderen Datenquellen, auch auf Gemeindeebene kein negativer Einfluss von NIM-P auf die demographischen Variablen nachgewiesen werden (Rais, 2002).

NIM-P musste aber vor allem in den Fragen 'Eltern' und deren Unterfragen sowie in der Frage zur Stellung im Haushalt überaus viele Daten einsetzen und verändern. Von den mit NIM-P verarbeiteten Variablen wurden diese am wenigsten rückgefragt und waren nicht vorbedruckt. Zudem ist in beiden Fragen die Fragestellung nicht genügend klar, was dem Befragten einen Interpretationsspielraum offen lässt. Hinzu kommt eine unzureichend durchgeführte Haushaltsbildung. Aufgrund der Analysen stellte sich heraus, dass die Distanzdefinition für diese beiden Variablen nicht optimal war. Wäre dies der Fall gewesen, hätte jedoch mit einer ungleich höheren Anzahl Restfälle gerechnet werden müssen, da die Spendersuche erschwert worden wäre.

Auf jeden Fall konnten und können Mankos im Fragebogen und der Produktion nicht mit Einsetzungen wett gemacht werden. Zudem müssen Einsetzungsverfahren früher getestet werden können. Hierfür müssen sämtliche Ansprüche an die Qualität der Daten und Prozesse definiert worden sein, sowie die dementsprechenden Ressourcen investiert werden.

6.3 KHH mit NIM-K

Als die SB der KHH begann, standen nicht alle Kollektivhaushalte zur Verfügung. Einerseits wurde nach der Verarbeitung mit NIM-K festgestellt, dass einige Privathaushalte mit mehr als 15 Personen KHH sind, und andererseits konnten aus technischen Gründen nicht alle KHH für die Verarbeitung mit NIM-K geliefert werden. Die Personen der KHH wurden einzeln bearbeitet, da die Stellung im Haushalt in den KHH keinen Einfluss hat. Diese Variable wurde deshalb auch nicht bearbeitet. Ansonsten wurden die gleichen Fragen bearbeitet wie bei den Privathaushalten. Die vorhandenen KHH wurden nach der Oberkategorie der KHH getrennt behandelt (siehe Tabelle 27 im Anhang A.3). Spender aus demselben KHH wie der Empfänger wurden in der Distanzberechnung von NIM-K gegenüber Spendern aus anderen KHH bevorteilt. Gleichermassen wurden die Personen mit dem gleichen Status im KHH (Insasse/Patient/etc., Personal, Angehörige/übrige Personen) als Spender bevorzugt.

Aus der Tabelle 21 ist zu erkennen, dass der Anteil an Empfänger stark variiert, je nach Haushalt Typ, welcher somit einen starken Einfluss auf die Qualität der Resultate hat. Bei 28.06% der Personen in KHH hat NIM-K Änderungen vorgenommen. Aus der Tabelle 43 ist ersichtlich, dass die Variable 'Eltern' (PAR) am meisten eingesetzt wurde. Dies rührt daher, dass die Frage 9 in vielen Fällen nicht beantwortet wurde.

Frage 9: Eltern NIM hat bei rund 37'000 Personen in KHH die Frage 9 geändert. In den meisten Fällen wurde 'Eltern_Cd' auf 1 gesetzt (keine Kinder). Wie schon bei den Privathaushalten, war auch hier die Frage 9 in den meisten Fällen ganz leer. Rund die Hälfte der Änderungen in dieser Frage wurden bei Personen in Altersheimen (Haushalt Typ = 41) gemacht, siehe Tabelle 43. Es wird davon ausgegangen, dass in den meisten Fällen der KHH, bei denen die Frage 9 gänzlich fehlt, die Leitung des KHH die Fragebogen ausgefüllt hat und nicht die betreffenden Personen selbst. Die Leitung konnte die Frage 9 jedoch nicht beantworten, da sie keinerlei diesbezügliche Informationen zur Verfügung hatte. Da keine geeignete andere Datenquelle bezüglich dieser Frage vorhanden ist, war es nicht möglich die Qualität dieser Einsetzungen objektiv zu beurteilen.

6.4 Bemerkungen zur Anwendung von NIM in der PHE

Wohnsitzabgleich Ursprünglich war vorgesehen einen Prozess SB-WSA durchzuführen. Dieser sollte die Übereinstimmung der Daten von Personen mit mehreren Wohnsitzen garantieren. Da keine Kriterien bestehen, welcher Wohnsitz bevorzugt werden sollte wurde dieser Prozess nicht entwickelt. Eine Übernahme der Variablenwerte nach NIM von einem Wohnsitz in den anderen hätte zusätzlich das Risiko mit sich gebracht, dass ganze Haushalte nicht mehr konsistent gewesen wären. Dies war ein weiterer Grund für den Entscheid SB-WSA nicht durchzuführen.

Somit können potenziell sämtliche Variablenwerte von Personen mit mehreren Wohnsitzen, von einem Wohnsitz zum anderen unterschiedlich sein. Es kann keine Garantie abgegeben werden, dass sich die 'beiden' Personen gleichen, da sie unabhängig in zwei unterschiedlichen Haushaltszusammenhängen behandelt wurden. Dies wird noch dadurch verstärkt, dass die Personen vor der SB durchaus unterschiedliche Angaben in den verschiedenen Wohnsitzen haben konnten.

Vorbedruck, Rückfragen NIM hätte ohne weiteres Variablen die vom Vorbedruck oder Rückfragen herrührten, belassen können. Diese Funktionalität wurde in NIM integriert. Dies hätte zur Folge gehabt, dass mehr Restfälle nach NIM übrig geblieben wären, da Inkonsistenzen, die von diesen Quellen stammten, nicht hätten gelöst werden können. Da sich herausstellte, dass weder systematisch auf den Vorbedruck noch auf Rückfragen Verlass ist, und dass für einen beachtlichen Teil der Werte nicht mehr gesagt werden konnte, woher sie stammten (Vorbedruck, Rückfragen, Fragebogen, etc.), wurde diese Möglich-

Tabelle 21: Nach KHH Oberkategorie (Kat), Haushalts Typen (Typ) aufgeteilt: die Anzahl Personen, bei welchen NIM-K mindestens eine Variable eingesetzt hat (eins.), der prozentuale Anteil der Empfänger an der Population (eins. [%]), die Anzahl Personen ohne Fehler (n.eins.), sowie deren Anteil an der Population (n.eins. [%]). In der letzten Spalte ist der prozentuale Anteil der Empfänger an der gesamten KHH Bevölkerung.

Kat	Typ	Total Pers.	eins.	eins. [%]	n.eins.	n.eins. [%]	% eins./Total
1	11	2127	1021	48.00	1106	52.00	0.55
1	12	943	519	55.04	424	44.96	0.28
1	13	86	20	23.26	66	76.74	0.01
2	20	15	0	0.00	15	100.00	0.00
2	21	10597	1821	17.18	8776	82.82	0.98
2	22	10339	1555	15.04	8784	84.96	0.84
2	23	1125	103	9.16	1022	90.84	0.06
3	31	2889	710	24.58	2179	75.42	0.38
3	32	895	329	36.76	566	63.24	0.18
3	33	2445	1254	51.29	1191	48.71	0.68
3	34	1243	609	48.99	634	51.01	0.33
4	40	262	27	10.31	235	89.69	0.01
4	41	75329	22831	30.31	52498	69.69	12.30
4	42	12932	1541	11.92	11391	88.08	0.83
4	43	1567	297	18.95	1270	81.05	0.16
4	44	1668	324	19.42	1344	80.58	0.17
4	45	241	40	16.60	201	83.40	0.02
4	46	2211	184	8.32	2027	91.68	0.10
4	47	2342	281	12.00	2061	88.00	0.15
5	50	7	0	0.00	7	100.00	0.00
5	51	6271	1255	20.01	5016	79.99	0.68
5	52	1100	48	4.36	1052	95.64	0.03
5	53	102	13	12.75	89	87.25	0.01
6	61	5342	975	18.25	4367	81.75	0.53
6	62	1985	374	18.84	1611	81.16	0.20
6	63	54	6	11.11	48	88.89	0.00
6	64	2075	398	19.18	1677	80.82	0.21
7	71	1268	389	30.68	879	69.32	0.21
7	72	4531	973	21.47	3558	78.53	0.52
7	73	13978	8622	61.68	5356	38.32	4.65
7	74	1721	428	24.87	1293	75.13	0.23
8	81	17897	5126	28.64	12771	71.36	2.76
8	82	3	2	66.67	1	33.33	0.00
Total	185590	52075	28.06	133515	71.94	28.06	

keit nicht genutzt. Aufgrund dieses Entscheides kann die Konsistenz zwischen den Daten nach NIM und jenen auf dem Fragebogen oder nach den Rückfragen nicht gewährleistet werden. An dieser Stelle sei daran erinnert, dass die Methode, die in NIM angewandt wird, nicht zum Ziel hat, die 'Wahrheit' auf individuellem Niveau herzustellen, sondern zum Ziel hat, die multivariate Verteilung beizubehalten.

Nachbearbeitung von NIM-P/NIM-K In den nachträglichen Analysen, die von der ehemaligen Sektion BHS durchgeführt wurden, kamen verschiedene Konfigurationen zum Vorschein, die als inakzeptabel deklariert wurden. Es handelte sich vor allem um eine zu grosse Anzahl gleichgeschlechtlicher Partnerschaften, zu junge Eltern, Änderungen der vorbedruckten Werte, etc. In den meisten Fällen stellte sich heraus, dass es sich um Haushalte handelte, deren Zusammensetzung unvollständig war, oder es handelte sich um einen falschen Haushaltstypen. Aus diesen Gründen, und da sich in einigen Gemeinden ein beachtlicher Teil der Bevölkerung in den SHH befand, was z.T. einen Vergleich mit 1990 verunmöglichte, wurden rund 250'000 Personenfragebogen und die zugehörigen Haushalte nochmals manuell bearbeitet.

6.5 Modellbasierte Einsetzungen

Die deterministischen Korrekturen konnten in den Variablen Nationalität, Ausländerausweis / Aufenthaltsstatus und Muttersprache nicht alle fehlenden Werte einsetzen. Zudem wies die Variable Geburtsdatum noch fehlende Werte auf. Für diese Fragen mussten 100% gültige Antworten vorhanden sein. Deshalb wurden diese Restfälle durch MBE vervollständigt.

Geburtsdatum Für die Zufallsauswahl des Tags und des Monats wurde das Geburtsjahr als Hilfsvariable verwendet. Die Verteilung des Geburtsdatums wurde anhand der Verteilung der bestehenden Werte nach Geburtsjahr geschätzt. Insgesamt musste bei 12'984 Personen das Datum eingesetzt werden. Zur Auswahl standen 36'975 Kombinationen von Jahrzahlen mit Monaten und Tagen als Spender zur Verfügung.

Dieses Vorgehen hat bei älteren Personen zu Verzerrungen geführt, da nur wenige Geburtsdaten für Personen, welche vor 1910 geboren sind, vorhanden waren. Der geringen Anzahl wegen, wird an dieser Stelle aber nicht näher auf diese Einsetzungen eingegangen.

Nationalität In der Nationalität mussten 5'515 Fälle zufällig eingesetzt werden. Die Verteilung wurde pro Kanton aufgrund der Variablen Alter, Geschlecht, Zivilstand und Ausländerausweis geschätzt. Sämtlichen Mitgliedern eines Haushalts wurde dieselbe Nationalität eingesetzt, sofern sie fehlte. Die älteste Person im Haushalt wurde als Referenzperson betrachtet. Auf diese Weise wurde die Kreierung von 'Mischnationalitäten-Haushalten' verhindert. Da die Verteilung der Nationalität der Personen verwendet wurde, um sie auf Haushaltsniveau anzuwenden, ist dieser Ansatz nicht optimal. Unter dem vorhandenen Zeitdruck war es nicht möglich, eine komplexere Methode anzuwenden.

Aus der Tabelle 44 geht hervor, dass den beiden Ländern Italien und Sierra Leone, die vor der Einsetzung schon 'zu viele' Personen gegenüber der Statistik 'PETRA' aufwiesen, ebenfalls noch Personen hinzugefügt wurden. Nach Abklärungen mit den thematischen Spezialisten wurden diese akzeptiert. Für drei Personen war es aufgrund von fehlenden Werten in den Hilfsvariablen nicht möglich, die Nationalität einzusetzen. Diese wurden manuell bearbeitet.

Aufenthaltsstatus Der Aufenthaltsstatus wurde aufgrund der Nationalität bei 14'607 Personen durch eine MBE eingesetzt. Der gleiche Aufenthaltsstatus wurde sämtlichen Personen des gleichen Haushalts eingesetzt, sofern er fehlte.

Aus der Tabelle 45 geht hervor, dass die meisten Ausweise in den ersten beiden Kategorien hinzugefügt wurden. Zwei Personen wurde der Aufenthaltsstatus aufgrund der fehlenden Nationalität manuell eingesetzt.

Sprache Es wurden nur Personen die Sprache eingesetzt, die weder Schweizer noch Fürstentum Liechtensteiner waren. Für diese wurden von der ehemaligen Sektion BHS deterministische Korrekturen angewandt, um die fehlenden Werte einzusetzen. Bei 92'182 Personen wurde die Sprache mit einer MBE eingesetzt, wobei die Nationalität als Hilfsvariable diente. Sämtlichen Mitgliedern eines Haushalts, denen die Sprache noch fehlte, wurde die gleiche Sprache eingesetzt, um die Kreierung von 'Mehrsprachen-Haushalten' zu verhindern. In der Tabelle 46 sind die Resultate der Einsetzung ersichtlich. Aufgrund von Abklärungen mit den thematischen Spezialisten, wurden diese akzeptiert.

Anhang

A PHH/SHH/KHH der VZ2000

A.1 Variablenabkürzungen

Tabelle 22: Abkürzungen der Variablen, welche in der PHH/SHH/KHH verwendet wurden. In der letzten Kolonne steht 'E' für Einsetzung, Änderung möglich und 'H' für Hilfsvariable, die nicht geändert oder eingesetzt werden kann. Die Variable 'Stellung im Haushalt' wurde für Personen in Sammelhaushalten und Kollektivhaushalten weder eingesetzt noch verändert, deshalb ist '*E*' kursiv geschrieben.

Variablenname	Abkürzung	Einsetzung/ Hilfsvariable
Alter	GEJA	E
Geschlecht	GESL	E
Zivilstand	ZIVL	E
Alter bei der letzten Zivilstandsänderung	ZIVJ	E
Nationalität	NATI1	H
Alter bei Erhalt der Schweizer Nationalität	CHJJ	H
Eltern	PAR	E
Anzahl Kinder	ZKIND	E
Alter bei Geburt 1. Kind	P1J	E
Alter bei Geburt 2. Kind	P2J	E
Alter bei Geburt 3. Kind	P3J	E
Alter bei Geburt 4. Kind	P4J	E
Alter bei Geburt letztes Kind	PLJ	E
Stellung im Haushalt	STHH	<i>E</i>
Anzahl Personen im Haushalt	ANZG	H

A.2 Kontrollregeln

Tabelle 23: Kontrollregeln die in der PHH/SHH/KHH verwendet wurden. Die erste Kolonne 'Nr.' bezieht sich auf die BFS interne Nummerierung der Regeln, die auch in Dokumenten von der ehemaligen Sektion BHS wiedergefunden werden kann. Unter 'Anwendung' ist aufgeführt, ob die Regel bei den Personenhaushalten ('P'), Sammelhaushalten ('S') oder/und den Kollektivhaushalte ('K') angewandt wurde. Das Fürstentum Liechtenstein wurde wie die Schweiz ('CH') behandelt.

Nr.	Regel	Anwendung
1.01.000	$(GEJA(X_1) > 120)$	P/S/K
1.01.002/003	$(GEJA(X_1) < 0)$	P/S/K
1.01.006	$(GEJA(X_1) \geq 0) \cap (ZIVL(X_1) = \text{verheiratet}) \cap (GEJA(X_1) < 15) \cap (NATI(X_1) = CH)$	P/S/K
1.01.006	$(GEJA(X_1) \geq 0) \cap (ZIVL(X_1) = \text{verheiratet}) \cap (GEJA(X_1) < 11) \cap (NATI(X_1) \neq CH)$	P/S/K
1.09.009	$(GEJA(X_1) \geq 0) \cap (ZIVL(X_1) = \text{verwitwet}) \cap (GEJA(X_1) < 15) \cap (NATI(X_1) = CH)$	P/S/K
1.09.009	$(GEJA(X_1) \geq 0) \cap (ZIVL(X_1) = \text{verwitwet}) \cap (GEJA(X_1) < 11) \cap (NATI(X_1) \neq CH)$	P/S/K
1.01.012	$(GEJA(X_1) \geq 0) \cap (ZIVL(X_1) = \text{geschieden}) \cap (GEJA(X_1) < 16) \cap (NATI(X_1) = CH)$	P/S/K
1.01.012	$(GEJA(X_1) \geq 0) \cap (ZIVL(X_1) = \text{geschieden}) \cap (GEJA(X_1) < 12) \cap (NATI(X_1) \neq CH)$	P/S/K
1.01.015	$(GEJA(X_1) \geq 0) \cap (ZIVL(X_1) = \text{verheiratet}) \cap (GEJA(X_1) - ZIVJ(X_1) < 15) \cap (NATI(X_1) = CH) \cap (ZIVL(X_1) < CHJJ(X_1))$	P/S/K
1.01.015	$(GEJA(X_1) \geq 0) \cap (ZIVL(X_1) = \text{verheiratet}) \cap (GEJA(X_1) - ZIVJ(X_1) < 11) \cap (NATI(X_1) = CH) \cap (ZIVL(X_1) \leq CHJJ(X_1))$	P/S/K
1.01.015	$(GEJA(X_1) \geq 0) \cap (ZIVL(X_1) = \text{verheiratet}) \cap (GEJA(X_1) - ZIVJ(X_1) < 11) \cap (NATI(X_1) \neq CH)$	P/S/K
1.01.019	$(GEJA(X_1) \geq 0) \cap (ZIVL(X_1) = \text{verwitwet}) \cap (GEJA(X_1) - ZIVJ(X_1) < 15) \cap (NATI(X_1) = CH) \cap (ZIVL(X_1) < CHJJ(X_1))$	P/S/K
1.01.019	$(GEJA(X_1) \geq 0) \cap (ZIVL(X_1) = \text{verwitwet}) \cap (GEJA(X_1) - ZIVJ(X_1) < 11) \cap (NATI(X_1) = CH) \cap (ZIVL(X_1) \leq CHJJ(X_1))$	P/S/K
1.01.019	$(GEJA(X_1) \geq 0) \cap (ZIVL(X_1) = \text{verwitwet}) \cap (GEJA(X_1) - ZIVJ(X_1) < 11) \cap (NATI(X_1) \neq CH)$	P/S/K
1.01.022	$(GEJA(X_1) \geq 0) \cap (ZIVL(X_1) = \text{geschieden}) \cap (GEJA(X_1) - ZIVJ(X_1) < 16) \cap (NATI(X_1) = CH) \cap (ZIVL(X_1) < CHJJ(X_1))$	P/S/K
1.01.022	$(GEJA(X_1) \geq 0) \cap (ZIVL(X_1) = \text{geschieden}) \cap (GEJA(X_1) - ZIVJ(X_1) < 12) \cap (NATI(X_1) = CH) \cap (ZIVL(X_1) \leq CHJJ(X_1))$	P/S/K
1.01.022	$(GEJA(X_1) \geq 0) \cap (ZIVL(X_1) = \text{geschieden}) \cap (GEJA(X_1) - ZIVJ(X_1) < 12) \cap (NATI(X_1) \neq CH)$	P/S/K
1.01.025	$(P1J(X_1) \geq 0) \cap (GEJA(X_1) - P1J(X_1) < 12)$	P/S/K
1.01.025	$(P2J(X_1) \geq 0) \cap (GEJA(X_1) - P2J(X_1) < 12)$	P/S/K
1.01.025	$(P3J(X_1) \geq 0) \cap (GEJA(X_1) - P3J(X_1) < 12)$	P/S/K
1.01.025	$(P4J(X_1) \geq 0) \cap (GEJA(X_1) - P4J(X_1) < 12)$	P/S/K
1.01.025	$(PLJ(X_1) \geq 0) \cap (GEJA(X_1) - PLJ(X_1) < 12)$	P/S/K
1.01.028	$(GESL(X_1) = \text{weiblich}) \cap (P1J(X_1) \geq 0) \cap (GEJA(X_1) - P1J(X_1) > 60)$	P/S/K
1.01.028	$(GESL(X_1) = \text{weiblich}) \cap (P2J(X_1) \geq 0) \cap (GEJA(X_1) - P2J(X_1) > 60)$	P/S/K
1.01.028	$(GESL(X_1) = \text{weiblich}) \cap (P3J(X_1) \geq 0) \cap (GEJA(X_1) - P3J(X_1) > 60)$	P/S/K
1.01.028	$(GESL(X_1) = \text{weiblich}) \cap (P4J(X_1) \geq 0) \cap (GEJA(X_1) - P4J(X_1) > 60)$	P/S/K
1.01.028	$(GESL(X_1) = \text{weiblich}) \cap (PLJ(X_1) \geq 0) \cap (GEJA(X_1) - PLJ(X_1) > 60)$	P/S/K

Tabelle 24: Kontrollregeln die in der PHH/SHH/KHH verwendet wurden. Die erste Kolonne 'Nr.' bezieht sich auf die BFS interne Nummerierung der Regeln, die auch in Dokumenten von der ehemaligen Sektion BHS wiedergefunden werden können. Unter 'Anwendung' ist aufgeführt, ob die Regel bei den Personenhaushalten ('P'), Sammelhaushalten ('S') oder/und den Kollektivhaushalte ('K') angewandt wurde. Das Fürstentum Liechtenstein wurde wie die Schweiz ('CH') behandelt.

Nr.	Regel	Anwendung
1.01.033	$(GEJA(X_1) < 15) \cap (STHH(X_1) = \text{alleinlebend})$	P
1.01.036	$(GEJA(X_1) < 15) \cap (STHH(X_1) = \text{Partner/alleinerziehend/anderer Haushaltsvorstand})$	P
1.01.045	$(GEJA(X_1) < 15) \cap (STHH(X_1) = \text{Ehemann/-frau}) \cap (NAT1(X_1) = CH)$	P
1.01.045	$(GEJA(X_1) < 11) \cap (STHH(X_1) = \text{Ehemann/-frau}) \cap (NAT1(X_1) \neq CH)$	P
1.01.048	$(GEJA(X_1) > 75) \cap (STHH(X_1) = \text{Sohn/Tochter})$	P
1.01.051	$(GEJA(X_1) < 25) \cap (STHH(X_1) = \text{Vater/Mutter})$	P
1.01.055	$(GEJA(X_1) < 15) \cap (STHH(X_1) = \text{Angestellter/Au-pair})$	P
1.02.001	$(GESL(X_1) = \text{leer/ungültig})$	P/S/K
1.03.001/003	$(ZIVL(X_1) = \text{leer/ungültig})$	P/S/K
1.03.007	$(ZIVJ(X_1) = \text{ungültig})$	P/S/K
1.09.003/009	$(PAR(X_1) = \text{leer/ungültig})$	P/S/K
1.09.004	$(PIJ(X_1) = \text{ungültig})$	P/S/K
1.09.004	$(P2J(X_1) = \text{ungültig})$	P/S/K
1.09.004	$(P3J(X_1) = \text{ungültig})$	P/S/K
1.09.004	$(P4J(X_1) = \text{ungültig})$	P/S/K
1.09.004	$(PLJ(X_1) = \text{ungültig})$	P/S/K
1.09.004	$(PIJ(X_1) > 98)$	P/S/K
1.09.004	$(P2J(X_1) > 98)$	P/S/K
1.09.004	$(P3J(X_1) > 98)$	P/S/K
1.09.004	$(P4J(X_1) > 98)$	P/S/K
1.09.004	$(PLJ(X_1) > 98)$	P/S/K
1.09.006	$(ZKIND(X_1) = \text{ungültig})$	P/S/K
1.09.006	$(ZKIND(X_1) > 25)$	P/S/K
1.10.001	$(STHH(X_1) = \text{leer})$	P
1.10.002	$(STHH(X_1) = \text{ungültig})$	P
1.10.XXX	$(STHH(X_1) = \text{'Mehrfachangabe'})$	P
3.01.052	$(\text{Ungerade Anzahl } i \text{ mit } STHH(X_i) = \text{Partner im Haushalt})$	P
3.01.052(5)	$(STHH(X_1) = \text{Ehemann/-frau}) \cap (ZIVL(X_1) \neq \text{verheiratet})$	P
3.01.012	$(\text{Mindestens eine Person im Haushalt muss älter als 14 Jahre sein})$	P
3.01.012	$(\text{Mindestens eine Person im Haushalt muss Haushaltsvorstand sein})$	P
3.01.005	$(ANZG > 1) \cap (STHH(X_1) = \text{alleinlebend})$	P
3.01.006	$(ANZG > 1) \cap (STHH(X_1) = \text{alleinerziehend}) \cap (\text{keine weitere Person mit } STHH(X_2) = \text{Sohn/Tochter})$	P
3.01.051	$(\text{Ungerade Anzahl } i \text{ mit } STHH(X_i) = \text{Ehemann/-frau im Haushalt})$	P

Tabelle 25: Kontrollregeln die in der PHH/SHH/KHH verwendet wurden. Die erste Kolonne 'Nr.' bezieht sich auf die BFS interne Nummerierung der Regeln, die auch in Dokumenten von der ehemaligen Sektion BHS wiedergefunden werden können. Unter 'Anwendung' ist aufgeführt, ob die Regel bei den Personenhaushalten ('P'), Sammelhaushalten ('S') oder/und den Kollektivhaushalte ('K') angewandt wurde.

Nr.	Regel	Anwendung
1.03.008	$(ZIVL(X_1)=verheiratet/verwitwet/geschieden) \cap (ZIVJ(X_1)=leer)$	P/S/K
1.03.009	$(ZIVL(X_1) \neq verheiratet) \cap (STHH(X_1)=Ehemann/-frau)$	P
1.09.001	$(GEJA(X_1) < 12) \cap (PAR(X_1)=ja)$	P/S/K
1.09.001	$(GEJA(X_1) < 12) \cap (ZKIND(X_1) > 0)$	P/S/K
1.09.001	$(GEJA(X_1) < 12) \cap (PIJ(X_1) \geq 0)$	P/S/K
1.09.001	$(GEJA(X_1) < 12) \cap (P2J(X_1) \geq 0)$	P/S/K
1.09.001	$(GEJA(X_1) < 12) \cap (P3J(X_1) \geq 0)$	P/S/K
1.09.001	$(GEJA(X_1) < 12) \cap (P4J(X_1) \geq 0)$	P/S/K
1.09.001	$(GEJA(X_1) < 12) \cap (PLJ(X_1) \geq 0)$	P/S/K
1.09.005	$(ZKIND(X_1) < 1) \cap (PAR(X_1)=ja)$	P/S/K
1.09.005	$(ZKIND(X_1) < 1) \cap (PIJ(X_1) \geq 0)$	P/S/K
1.09.005	$(ZKIND(X_1) < 1) \cap (P2J(X_1) \geq 0)$	P/S/K
1.09.005	$(ZKIND(X_1) < 1) \cap (P3J(X_1) \geq 0)$	P/S/K
1.09.005	$(ZKIND(X_1) < 1) \cap (P4J(X_1) \geq 0)$	P/S/K
1.09.005	$(ZKIND(X_1) < 1) \cap (PLJ(X_1) \geq 0)$	P/S/K
1.09.007	$(ZKIND(X_1) > 0) \cap (PIJ(X_1) \neq leer)$	P/S/K
1.09.007	$(ZKIND(X_1) > 1) \cap (P2J(X_1) \neq leer)$	P/S/K
1.09.007	$(ZKIND(X_1) > 2) \cap (P3J(X_1) \neq leer)$	P/S/K
1.09.007	$(ZKIND(X_1) > 3) \cap (P4J(X_1) \neq leer)$	P/S/K
1.09.007	$(ZKIND(X_1) > 4) \cap (PLJ(X_1) \neq leer)$	P/S/K
1.09.007	$(ZKIND(X_1) < 1) \cap (PIJ(X_1) \geq 0)$	P/S/K
1.09.007	$(ZKIND(X_1) < 2) \cap (P2J(X_1) \geq 0)$	P/S/K
1.09.007	$(ZKIND(X_1) < 3) \cap (P3J(X_1) \geq 0)$	P/S/K
1.09.007	$(ZKIND(X_1) < 4) \cap (P4J(X_1) \geq 0)$	P/S/K
1.09.007	$(ZKIND(X_1) < 5) \cap (PLJ(X_1) \geq 0)$	P/S/K
1.09.008	$(PAR(X_1) \neq nein) \cap (ZKIND(X_1) > 0)$	P/S/K
1.09.008	$(PAR(X_1) \neq nein) \cap (PIJ(X_1) \geq 0)$	P/S/K
1.09.008	$(PAR(X_1) \neq nein) \cap (P2J(X_1) \geq 0)$	P/S/K
1.09.008	$(PAR(X_1) \neq nein) \cap (P3J(X_1) \geq 0)$	P/S/K
1.09.008	$(PAR(X_1) \neq nein) \cap (P4J(X_1) \geq 0)$	P/S/K
1.09.008	$(PAR(X_1) \neq nein) \cap (PLJ(X_1) \geq 0)$	P/S/K
1.09.010	$(PAR(X_1) \neq ja) \cap (PIJ(X_1) \neq leer)$	P/S/K
1.09.010	$(ZKIND(X_1) > 0) \cap (PIJ(X_1) \neq leer)$	P/S/K

Tabelle 26: Kontrollregeln die in der PHH/SHH/KHH verwendet wurden. Die erste Kolonne 'Nr.' bezieht sich auf die BFS interne Nummerierung der Regeln, die auch in Dokumenten von der ehemaligen Sektion BHS wiedergefunden werden können. Unter 'Anwendung' ist aufgeführt, ob die Regel bei den Personenhaushalten ('P'), Sammelhaushalten ('S') oder/und den Kollektivhaushalten ('K') angewandt wurde. 'Hh.vorstand' steht für Haushaltsvorstand.

Nr.	Regel	Anwendung
Nr.	Regel	Anwendung
1.03.XX1	$(ZIVL(X_1)=\text{ledig}) \cap (ZIVJ(X_1) \neq \text{leer})$	P/S/K
1.09.XX1	$(ZKIND(X_1)=\text{leer/ungültig}) \cap (STHH(X_1)=\text{alleinerziehend})$	P
1.10.XX1	$(ANZG = 1) \cap (STHH(X_1)=\text{Ehemann/-frau/Partner/alleinerziehend/and. Hh.vorstand/Sohn/Tochter/Vater/Mutter/Bruder/Schwester/and. Verwandter von Hh.vorstand})$	P
1.01.XX1	Im Haushalt muss mindestens ein Hh.vorstand sein, der 15 Jahre älter ist als jedes Kind im Haushalt (STHH wird verändert, das Alter nicht)	P
1.01.XX2	Im Haushalt muss mindestens ein Hh.vorstand sein, der 15 Jahre jünger ist als jede/r Mutter/Vater im Haushalt (STHH wird verändert, das Alter nicht)	P

A.3 KHH-Typen

Tabelle 27: Haushaltstypen der Kollektivhaushalte nach Oberkategorie und Typ, wie sie in der SB verwendet wurden.

Oberkategorie		Kollektivhaushaltstyp	
Kat	Oberkat	Typ	KHH-Typ
1	Institution des Straf- und Massnahmenvollzugs	11	Strafanstalt, reine Massnahmenanstalt
		12	Haftanstalt, Bezirks- oder Untersuchungsgefängnis
		13	Institution, die auch Strafgefängnisse aufnehmen kann, aber nicht zum Strafvollzug im engen Sinne gehört
2	Internat, Wohnheim einer Bildungseinrichtung	20	nicht näher bestimmtes Internat, Wohnheim einer Bildungseinrichtung/Zweifelsfälle
		21	Internat
		22	Studenten-, Schüler- oder Lehrlingswohnheim
3	Krankenanstalt, Heilstätte, Spital	23	Erziehungsheim
		30	nicht näher bestimmbarer Krankenanstalt, Heilstätte, Spital/Zweifelsfälle
		31	Allgemeinkrankenhaus
		32	Spezialklinik
		33	Psychiatrische Klinik
		34	Geriatrische Klinik
4	Heim der Wohlfahrtspflege, Sozialmedizinische Institution	40	nicht näher bestimmbarer Heim der Wohlfahrtspflege, Sozialmedizinische Institution/Zweifelsfälle
		41	Altersheim, Alters- und Pflegeheim, Pflegeheim
		42	Institution für behinderte Personen
		43	Institution für Suchtkranke
		44	Heim für psychosoziale Fälle
		45	Kur- und Erholungsheim
		46	Heim für Kinder und Jugendliche, Waisenhaus
		47	Sonstiges Wohnheim, therapeutische Wohngemeinschaft
5	Kloster, Ordenshaus, Unterkunft einer religiösen Vereinigung	50	nicht näher bestimmbarer Kloster, Ordenshaus, Unterkunft einer religiösen Vereinigung/Zweifelsfälle
		51	Katholisch
		52	Evangelisch, freikirchlich
		53	Andere religiös-weltanschauliche Gruppe
6	Hotel, Pension oder anderer Beherbergungsbetrieb	60	nicht näher bestimmbarer/-s Hotel, Pension oder anderer Beherbergungsbetrieb/Zweifelsfälle
		61	Hotel, Hotel- und Kurbetrieb, Aparthotel
		62	Gasthaus, Pension
		63	Zelt- und/oder Wohnwagenplatz
		64	Anderer Beherbergungsbetrieb, Gruppenunterkunft
7	Sonstiger Kollektivhaushalt oder Gemeinschaftsunterkunft	70	nicht näher bestimmbarer sonstiger Kollektivhaushalt oder Gemeinschaftsunterkunft/Zweifelsfälle
		71	Baracke für Bauarbeiterinnen und Bauarbeiter
		72	Wohnheim für Arbeiterinnen und Arbeiter
		73	Unterkunft für Asyl Suchende
		74	Anderer Kollektivhaushalt
8	Sammelhaushalt	81	Karteinwohner ohne physischen Wohnsitz in der Gemeinde
		82	Fahrende und Obdachlose (physische Präsenz; Unterkunft ist kein festes Gebäude)

A.4 Distanzen und Parameter

Distanzen

Geburtsjahr Die Geburtsjahre wurden in NIM-P/K in ein Alter umgewandelt, da die Tests in dieser Form einfacher und einheitlicher waren. Um die Altersdistanz zu berechnen wurden 22 Altersgruppen gebildet (siehe Tabelle 28). Wechsel des Alters von Kindern (<16) zu Erwachsenen und umgekehrt wurden mit einem Strafgewicht geahndet. Ebenso wurde der Wechsel des Alters vom Rentenalter, Frauen 63 und Männer 65, zu einem tieferen Alter mit einem Strafgewicht geahndet.

Tabelle 28: Altersgruppen, die für die Distanzberechnung des Alters verwendet wurden.

Wert	Altersbereich
1	0-6
2	7-15
3	16-19
4	20-24
5	25-29
6	30-34
7	35-39
8	40-44
9	45-49
10	50-54
11	55-59
12	60-64 ¹²
13	65-69
14	70-74
15	75-79
16	80-84
17	85-89
18	90-94
19	95-99
20	100-104
21	105-109
22	>109

Jahrzahlen-Distanz Die **Jahrzahlen-Distanz** wurde für die Variablen ZIVJ, P1J - P4J und PLJ verwendet. Sie basiert auf den 5-Jahresaltersklassen des BFS. Gehören zwei Personen derselben Altersklasse an, so ist die Distanz der entsprechenden Variablen zwischen den beiden 0 ansonsten ist sie 1.

Stellung im Haushalt Bei der Distanz für die Stellung im Haushalt wurden Spender bevorzugt, die eine der Mehrfachangaben des Empfängers hatten, indem nur die halbe Distanz (0.5) verrechnet wurde. Ansonsten wurde eine Indikatordistanz verwendet.

Indikatordistanz Für die beiden Variablen PAR und ZKIND wurde eine Indikatordistanz verwendet, die 1 ist, wenn der Empfänger und der Spender die gleiche Antwort aufweisen und 0 ansonsten.

¹²Dem Rentenalter wurde in der Distanzdefinition speziell Rechnung getragen.

Parameter Die Erklärung der Bedeutung der Parameter befindet sich im Abschnitt 4.4.

Tabelle 29: Parameter von NIM-P und NIM-K, ihre Abkürzung und ihr Wert in der Produktion. 'Max.' steht für maximal, '#' steht für Anzahl.

Beschreibung	Abkürzung	Wert
Max. # potentieller Spender	NUM_DONORS_DIST_CAL	1'000
Max. # Empfänger	MAX_FAILED_IMPUTE	1'000
Max. # Einsetzungen, die getestet werden	NBRNMCIAS	50
Max. # DLTs	MAX_DLTS	50
Max. # Behauptungen in den DLTs	MAXPROPOSITIONS	600
Max. # Regeln	MAX_NUMBER_EDITS	60
α	ALPHA	0.95
γ	GAMMA	1.05
Max. Distanz	MAX_DIST	50.0
Strafgewicht für den Wechsel der Alterskategorien Kind vs. Erwachsen, Rentner vs. Nicht-Rentner	GEBJw	1.2
Gewicht der Bit-Distanz	BIT_WEIGHT	0.1

A.5 Einsetzungen

A.5.1 Einsetzung mit NIM-P: Schichten und Variablen

Tabelle 30: In der ersten Spalte sind die Abkürzungen der Kantone, sowie Liechtenstein (FL) und der NUTS Regionen: Région Lémanique (RL), Espace Mittelland (EM), Nordschweiz (NS), Ostschweiz (OS) und Zentralschweiz (ZS). Die Abkürzung des Kantons mit einem 'S1' angehängt steht für die Sammelhaushalte des betreffenden Kantons. CH_R1 steht für die erste Phase der Haushalte mit 10 bis 15 Mitgliedern. CH_R2 steht für die zweite Phase der Haushalte mit 10 bis 15 Personen. In der zweiten Spalte steht die Anzahl Personen im Haushalt (APER). 'Empfänger HH' steht für die Anzahl fehlerhafter Haushalte. Die Anzahl potentieller Spender (pot. Spender) ist die Anzahl fehlerfreier Haushalte. CPU_{tot} ist die Verarbeitungszeit, CPU/HH ist die durchschnittliche Verarbeitungszeit pro Haushalt.

Kanton	APER	Empfänger HH	pot. Spender	Empfänger [%]	CPU _{tot} [s]	CPU/HH [s]
ZH	1	64936	160689	28.78	7859.73	3.48
ZH	10	64	21	75.29	458.47	539.38
ZH	2	47143	135786	25.77	12763.71	6.98
ZH	3	27583	38895	41.49	12915.1	19.43
ZH	4	24286	41414	36.96	16854.85	25.65
ZH	5	9366	11546	44.79	9212.86	44.06
ZH	6	3099	2360	56.77	4166.84	76.33
ZH	7	869	391	68.97	830.63	65.92
ZH	8	266	92	74.3	279.07	77.95
ZH	9	114	31	78.62	473.12	326.29
ZHS1	1	11900	27030	30.57	336.24	0.86
BE	1	41367	110602	27.22	5045.93	3.32
BE	2	33290	104208	24.21	9345.5	6.8
BE	3	20349	29424	40.88	9512.96	19.11
BE	4	18947	34341	35.56	13077.36	24.54
BE	5	7753	10740	41.92	7565.41	40.91
BE	6	2589	2465	51.23	3470.13	68.66
BES1	1	11620	26389	30.57	327.92	0.86
LU	1	14802	32885	31.04	1814.78	3.81
LU	2	11912	29735	28.6	3391.01	8.14
LU	3	7797	9810	44.28	3717.05	21.11
LU	4	7907	12750	38.28	5589.4	27.06
LU	5	4191	5167	44.79	4132.95	44.16
LU	6	1640	1374	54.41	2188.88	72.62
LUS1	1	3086	7801	28.35	86.37	0.79
UR	1	1299	2699	32.49	161.07	4.03
UR	2	1199	2849	29.62	355.11	8.77
UR	3	809	1111	42.14	390.75	20.35
UR	4	790	1349	36.93	562.55	26.3
UR	5	463	646	41.75	337.12	30.4
UR	6	149	177	45.71	79.63	24.43
URS1	1	168	921	15.43	4.5	0.41
SZ	1	4314	10915	28.33	532.37	3.5
SZ	2	3864	11103	25.82	1126.21	7.52
SZ	3	2947	4065	42.03	1414.25	20.17
SZ	4	2987	5106	36.91	2103.27	25.99
SZ	5	1548	2002	43.61	1575.6	44.38
SZ	6	592	498	54.31	538.82	49.43

Fortsetzung auf der nächsten Seite ...

Fortsetzung der Tabelle 30.						
Kanton	APER	Empfänger HH	pot. Spender	Empfänger [%]	CPU _{tot} [s]	CPU/HH [s]
SZS1	1	840	3035	21.68	22.95	0.59
OW	1	1503	2413	38.38	185.25	4.73
OW	2	1305	2303	36.17	387.33	10.74
OW	3	832	792	51.23	343.31	21.14
OW	4	915	1104	45.32	695.46	34.45
OW	5	493	510	49.15	305.59	30.47
OW	6	196	131	59.94	105.02	32.12
OWS1	1	332	1110	23.02	9.21	0.64
NW	1	1666	3399	32.89	201.58	3.98
NW	2	1249	3328	27.29	357.33	7.81
NW	3	915	1065	46.21	441.27	22.29
NW	4	866	1442	37.52	618.23	26.79
NW	5	436	493	46.93	261.32	28.13
NW	6	164	140	53.95	73.6	24.21
NWS1	1	346	988	25.94	9.48	0.71
GL	1	1152	3772	23.4	140.32	2.85
GL	2	1116	3650	23.42	326.57	6.85
GL	3	788	1296	37.81	372.48	17.87
GL	4	816	1553	34.44	574.89	24.27
GL	5	377	581	39.35	250.29	26.13
GL	6	128	150	46.04	57.24	20.59
GLS1	1	225	966	18.89	6.15	0.52
ZG	1	4944	9665	33.84	611.65	4.19
ZG	2	3120	9425	24.87	900.31	7.18
ZG	3	2192	3081	41.57	1049.18	19.9
ZG	4	2011	3765	34.82	1426.01	24.69
ZG	5	906	1123	44.65	906.6	44.68
ZG	6	277	232	54.42	189.16	37.16
ZGS1	1	1196	5382	18.18	34.44	0.52
FR	1	8867	20013	30.7	1083.49	3.75
FR	2	7908	19772	28.57	2275.33	8.22
FR	3	5983	8110	42.45	2850.92	20.23
FR	4	5551	9998	35.7	3921.94	25.22
FR	5	2414	3641	39.87	2395.82	39.57
FR	6	778	807	49.09	904.21	57.05
FRS1	1	2634	8171	24.38	72.49	0.67
SO	1	8608	24529	25.98	1058.36	3.19
SO	2	7900	25938	23.35	2252.8	6.66
SO	3	5426	8400	39.24	2566.55	18.56
SO	4	5222	9755	34.87	3660.46	24.44
SO	5	2240	3132	41.7	2197.35	40.9
SO	6	642	649	49.73	651.26	50.45
SOS1	1	1597	4049	28.29	44.07	0.78
BS	1	13100	35278	27.08	1584.45	3.28
BS	2	8021	20212	28.41	2167.05	7.68
BS	3	4173	4856	46.22	1961.79	21.73
BS	4	3317	4229	43.96	2339.1	31
BS	5	1281	1149	52.72	1271.74	52.33
BS	6	407	215	65.43	219.61	35.31
BSS1	1	1854	5386	25.61	26.48	0.37
BL	1	9238	26850	25.6	1118.25	3.1
BL	2	8295	30835	21.2	2280.11	5.83
Fortsetzung auf der nächsten Seite ...						

Fortsetzung der Tabelle 30.						
Kanton	APER	Empfänger HH	pot. Spender	Empfänger [%]	CPU _{tot} [s]	CPU/HH [s]
BL	3	5642	9390	37.53	2629.5	17.49
BL	4	5153	10353	33.23	3552.2	22.91
BL	5	1843	2882	39.01	1791.65	37.92
BL	6	564	553	50.49	501.21	44.87
BLS1	1	1957	4464	30.48	29.18	0.45
SH	1	2917	8276	26.06	358.37	3.2
SH	2	2384	7849	23.3	684.83	6.69
SH	3	1492	2296	39.39	715.13	18.88
SH	4	1422	2720	34.33	995.05	24.02
SH	5	658	908	42.02	612.41	39.11
SH	6	206	201	50.61	108.22	26.59
SHS1	1	792	1696	31.83	21.83	0.88
AR	1	2003	5337	27.29	241.66	3.29
AR	2	1644	4948	24.94	471.72	7.16
AR	3	1145	1539	42.66	542.5	20.21
AR	4	1144	1948	37	816.84	26.42
AR	5	594	868	40.63	521.76	35.69
AR	6	247	224	52.44	126.62	26.88
ARS1	1	360	1278	21.98	9.64	0.59
AI	2	349	1066	24.66	101.91	7.2
AI	3	199	472	29.66	59.9	8.93
AI	4	164	663	19.83	95.62	11.56
AI	5	145	452	24.29	93.6	15.68
AI	6	69	170	28.87	37.66	15.76
AIS1	1	19	243	7.25	0.32	0.12
SG	1	16217	45755	26.17	1984.79	3.2
SG	2	13560	40736	24.97	3858.44	7.11
SG	3	9914	13893	41.64	4711.11	19.79
SG	4	9859	17159	36.49	6876.99	25.45
SG	5	5048	7235	41.1	4931.67	40.15
SG	6	1878	1902	49.68	2509.88	66.4
SGS1	1	3163	10165	23.73	86.88	0.65
GR	1	9907	18981	34.29	1218.73	4.22
GR	2	7033	15611	31.06	2001.83	8.84
GR	3	4787	5578	46.18	2299.75	22.19
GR	4	4472	7060	38.78	3258.48	28.26
GR	5	1909	2528	43.02	1967	44.33
GR	6	519	575	47.44	471.1	43.06
GRS1	1	2314	10954	17.44	65.02	0.49
AG	1	17832	51570	25.69	2154.32	3.1
AG	2	15938	56715	21.94	4392.33	6.05
AG	3	11568	19077	37.75	5439.43	17.75
AG	4	11712	23124	33.62	8069.62	23.16
AG	5	5006	7389	40.39	4818.48	38.87
AG	6	1687	1616	51.07	2219.89	67.21
AGS1	1	3742	9188	28.94	56.46	0.44
TG	1	7936	21374	27.08	976.13	3.33
TG	2	6646	21176	23.89	1897.1	6.82
TG	3	4841	6920	41.16	2292.19	19.49
TG	4	4995	8954	35.81	3532.18	25.32
TG	5	2529	3676	40.76	2484.93	40.05
TG	6	905	982	47.96	1215.3	64.4
Fortsetzung auf der nächsten Seite ...						

Fortsetzung der Tabelle 30.						
Kanton	APER	Empfänger HH	pot. Spender	Empfänger [%]	CPU _{tot} [s]	CPU/HH [s]
TGS1	1	2138	5569	27.74	59.01	0.77
TI	1	1254	46941	2.6	124.74	0.26
TI	10	2	5	28.57	0.75	10.71
TI	2	7333	31860	18.71	1508.21	3.85
TI	3	4097	18763	17.92	1499.53	6.56
TI	4	2799	16825	14.26	1522.69	7.76
TI	5	952	4160	18.62	726.8	14.22
TI	6	247	796	23.68	213.16	20.44
TI	7	78	132	37.14	39.78	18.94
TI	8	28	36	43.75	17.19	26.86
TI	9	6	11	35.29	7.88	46.35
TIS1	2	1097	1151	48.80	32.13	1.43
VD	1	32979	73822	30.88	4014.66	3.76
VD	2	25165	58887	29.94	7191.07	8.56
VD	3	15960	21031	43.15	7586.65	20.51
VD	4	14069	22831	38.13	9948.01	26.96
VD	5	5097	6601	43.57	5145.12	43.98
VD	6	1455	1341	52.04	2013.15	72
VDS1	1	7526	21262	26.14	209.19	0.73
VS	1	11847	22094	34.9	1448.47	4.27
VS	2	9762	20576	32.18	2817.56	9.29
VS	3	7702	8639	47.13	3597.8	22.02
VS	4	7493	10695	41.2	5232.87	28.77
VS	5	3021	3824	44.13	3008.76	43.96
VS	6	753	718	51.19	783.51	53.26
VSS1	1	4460	16139	21.65	126.06	0.61
NE	1	7887	20441	27.84	958.17	3.38
NE	2	5720	17294	24.85	1592.98	6.92
NE	3	3851	5716	40.25	1808.55	18.9
NE	4	3436	6538	34.45	2394.23	24
NE	5	1123	1829	38.04	1118.54	37.89
NE	6	316	394	44.51	230.23	32.43
NES1	1	1626	4970	24.65	44.79	0.68
GE	1	23636	53162	30.78	2839.13	3.7
GE	2	14592	34542	29.7	4037.39	8.22
GE	3	10401	13694	43.17	4897.99	20.33
GE	4	8413	13317	38.72	5933.46	27.31
GE	5	2792	3283	45.96	2776.12	45.7
GE	6	783	665	54.07	809.55	55.91
GES1	1	18181	16074	53.08	405.24	1.18
JU	1	2868	5700	33.47	382.06	4.46
JU	2	2948	5359	35.49	849.7	10.23
JU	3	2008	2033	49.69	971.07	24.03
JU	4	1874	2238	45.57	1356.63	32.99
JU	5	924	980	48.53	916	48.11
JU	6	264	222	54.32	142.93	29.41
JUS1	1	753	2281	24.82	21.41	0.71
FL	1	1808	2540	41.58	222.57	5.12
FL	2	1864	1658	52.92	501.59	14.24
FL	3	1319	732	64.31	441.76	21.54
FL	4	1267	842	60.08	712.19	33.77
FL	5	521	313	62.47	245.87	29.48
Fortsetzung auf der nächsten Seite ...						

Fortsetzung der Tabelle 30.						
Kanton	APER	Empfänger HH	pot. Spender	Empfänger [%]	CPU _{tot} [s]	CPU/HH [s]
FL	6	128	67	65.64	44.59	22.87
FLS1	1	735	786	48.32	21.05	1.38
RL	10	29	20	59.18	233.45	476.43
RL	7	834	448	65.05	839.54	65.49
RL	8	257	93	73.43	409.23	116.92
RL	9	97	32	75.19	345.66	267.95
EM	10	42	8	84	85.37	170.74
EM	7	1291	687	65.27	1509.36	76.31
EM	8	445	169	72.48	363.82	59.25
EM	9	151	57	72.6	185.11	89
NS	10	43	11	79.63	179.08	331.63
NS	7	715	359	66.57	638.81	59.48
NS	8	267	84	76.07	259.63	73.97
NS	9	78	29	72.9	278.66	260.43
OS	10	56	17	76.71	216.34	296.36
OS	7	1187	683	63.48	1420.95	75.99
OS	8	369	180	67.21	322.3	58.71
OS	9	110	50	68.75	232.97	145.61
ZS	10	44	3	93.62	22.34	47.53
ZS	7	1055	472	69.09	1165.77	76.34
ZS	8	313	115	73.13	315.58	73.73
ZS	9	126	35	78.26	296.83	184.37
CH_R2	11	102	48	68	203.33	135.55
CH_R2	12	61	36	62.89	176.65	182.11
CH_R2	13	28	15	65.12	220.33	512.4
CH_R2	14	19	11	63.33	130.91	436.37
CH_R2	15	11	3	78.57	16.2	115.71
Total		1060023	2359849	31.00	390869.82	
CH_R1	11	276	1374	16.73	8.1	0.49
CH_R1	12	150	1014	12.89	4.53	0.39
CH_R1	13	103	456	18.43	2.09	0.37
CH_R1	14	72	348	17.14	1.23	0.29
CH_R1	15	52	158	24.76	0.7	0.33

Tabelle 31: Die Kantone sind mit ihren Autokennzeichen abgekürzt. Die Anzahl Personen in Privathaushalten, bei welchen NIM-P das Geburtsjahr eingesetzt oder geändert hat, sind in der Spalte GEJA aufgeführt. Desgleichen für die Variable Geschlecht (GESL), Zivilstand (ZIVL), Änderungsjahr des Zivilstands (ZIVJ), Eltern, Anzahl Kinder (ZKIND), die Geburtsjahre der ersten vier Kinder sowie des letzten (GjKind1-GjKind4, GjKindL) und die Stellung im Haushalt (STHH). In der zweit letzten Zeile ist das Total der Einsetzungen pro Variable aufgeführt. In der letzten Zeile steht der prozentuale Anteil der eingesetzten Variablen an der gesamten Bevölkerung der Privathaushalte.

Kanton	GEJA	GESL	ZIVL	ZIVJ	Eltern	ZKIND	GjKind1	GjKind2	GjKind3	GjKind4	GjKindL	STHH
ZH	2572	58	7147	30583	82816	13078	13384	4827	1448	824	1131	198852
BE	2796	84	6144	13387	62674	11834	9793	4009	1197	670	863	139989
LU	772	12	1852	10609	22390	6262	4002	1880	708	397	411	57228
UR	58	5	260	1479	1564	562	264	132	66	29	40	4889
SZ	278	40	687	1803	7548	2236	1387	643	238	123	132	20136
OW	67	4	314	2351	2202	626	293	148	41	21	46	5509
NW	83	2	271	1468	2247	615	353	183	46	30	46	5740
GL	173	16	259	884	1680	469	372	191	64	34	30	4811
ZG	191	9	715	2009	6345	1696	1655	892	391	202	160	15235
FR	842	52	2034	7458	14764	3079	1841	858	298	177	213	33455
SO	495	43	1466	4019	13814	3481	2717	1079	325	172	218	33717
BS	421	9	925	1920	14821	2329	2773	743	220	113	160	35038
BL	504	18	1404	4692	13790	2730	2592	928	296	169	239	33343
SH	228	19	509	1717	4237	883	778	306	96	43	59	9877
AR	84	1	197	521	3316	797	581	269	103	48	45	7978
AI	27	1	60	67	298	261	81	56	29	21	22	1305
SG	791	11	1544	3518	26700	7501	5506	2534	893	414	448	70720
GR	495	31	1868	4774	11830	3096	2472	1114	409	213	193	31302
AG	1526	34	2576	7008	29541	7033	5605	2424	766	431	529	74493
TG	511	13	954	2442	14320	3118	2412	1134	364	174	228	33839
TI	370	0	1229	2195	149	189	4025	2892	1026	316	99	17131
VD	1459	62	6498	32348	42038	5206	4655	2360	717	371	494	90622
VS	1194	50	2504	10040	19423	4236	3153	1635	539	266	282	43670
NE	241	4	804	2524	10571	1399	1270	614	161	83	124	23602
GE	786	16	6817	7857	46283	3822	3863	1843	556	250	234	68055
JU	299	18	1092	5512	4227	894	517	282	124	68	64	8682
FL	81	6	291	5698	2966	558	428	188	65	28	40	5538
Total	17344	618	50421	168883	462554	87990	76772	34164	11186	5687	6550	1074756
Total [%]	0.24	0.01	0.68	2.29	6.28	1.19	1.04	0.46	0.15	0.08	0.09	14.59

A.5.2 Einfluss von NIM-P auf die Verteilungen

Tabelle 32: Einfluss der Einsetzungen durch NIM-P auf die Verteilung der 5-Jahres Altersklassen ('Alterskl.'). 'GEJA' und 'f_GEJA' bezeichnen die absolute und die relative Häufigkeit vor den Einsetzungen. 'iGEJA' und 'f_iGEJA' bezeichnen die absolute und die relative Häufigkeit nach den Einsetzungen. 'diff' steht für die Differenz in Promillen der relativen Häufigkeit nach den Einsetzungen und der relativen Häufigkeit vor den Einsetzungen.

Alterskl.	GEJA	f_GEJA [%]	iGEJA	f_iGEJA [%]	diff [‰]
fehlend	7630	0.10	2	0.00	-1.04
jünger als 1	69569	0.94	69450	0.94	-0.02
01-04	319984	4.34	320321	4.35	0.05
05-09	429942	5.83	430227	5.84	0.04
10-14	433303	5.88	432782	5.87	-0.07
15-19	431782	5.86	433388	5.88	0.22
20-24	466854	6.34	467674	6.35	0.11
25-29	515515	7.00	516441	7.01	0.13
30-34	603524	8.19	604362	8.20	0.11
35-39	636146	8.63	636761	8.64	0.08
40-44	570804	7.75	571336	7.75	0.07
45-49	509464	6.91	509972	6.92	0.07
50-54	501953	6.81	502430	6.82	0.06
55-59	449975	6.11	450384	6.11	0.06
60-64	355243	4.82	355490	4.82	0.03
65-69	316331	4.29	316407	4.29	0.01
70-74	274426	3.72	274576	3.73	0.02
75-79	225573	3.06	225771	3.06	0.03
80-84	140259	1.90	140348	1.90	0.01
85-89	78907	1.07	79039	1.07	0.02
90-94	26629	0.36	26665	0.36	0.00
95-	4778	0.06	4765	0.06	0.00
Total	7368591	100.00	7368591	100.00	

Tabelle 33: Einfluss der Einsetzungen durch NIM-P auf die Verteilung des Geschlechts. 'GESL' und 'f_GESL' bezeichnen die absolute und die relative Häufigkeit vor den Einsetzungen. 'iGESL' und 'f_iGESL' bezeichnen die absolute und die relative Häufigkeit nach den Einsetzungen. 'diff' steht für die Differenz in Promillen der relativen Häufigkeit nach den Einsetzungen und der relativen Häufigkeit vor den Einsetzungen.

Geschlecht	GESL	f_GESL [%]	iGESL	f_iGESL [%]	diff [‰]
männlich	3623486	49.17	3623824	49.18	0.05
weiblich	3744566	50.82	3744766	50.82	0.03
ungültig	103	0.00	1	0.00	-0.01
fehlend	436	0.01	0	0.00	-0.06
Total	7368591	100.00	7368591	100.00	

Tabelle 34: Einfluss der Einsetzungen durch NIM-P auf die Verteilung des Zivilstands. 'ZIVL' und 'f_ZIVL' bezeichnen die absolute und die relative Häufigkeit vor den Einsetzungen. 'iZIVL' und 'f_iZIVL' bezeichnen die absolute und die relative Häufigkeit nach den Einsetzungen. 'diff' steht für die Differenz in Promillen der relativen Häufigkeit nach den Einsetzungen und der relativen Häufigkeit vor den Einsetzungen.

Zivilstand	ZIVL	f_ZIVL [%]	iZIVL	f_iZIVL [%]	diff [‰]
ledig	3134682	42.54	3167512	42.99	4.46
verheiratet	3432918	46.59	3426909	46.51	-0.82
verwitwet	380934	5.17	375245	5.09	-0.77
geschieden	408446	5.54	398921	5.41	-1.29
ungültig	5420	0.07	1	0.00	-0.74
fehlend	6191	0.08	3	0.00	-0.84
Total	7368591	100.00	7368591	100.00	

Tabelle 35: Einfluss der Einsetzungen durch NIM-P auf die Verteilung des Jahres der letzten Änderung des Zivilstands. 'ZIVJ' und 'f_ZIVJ' bezeichnen die absolute und die relative Häufigkeit vor den Einsetzungen. 'iZIVJ' und 'f_iZIVJ' bezeichnen die absolute und die relative Häufigkeit nach den Einsetzungen. 'diff' steht für die Differenz in Promillen der relativen Häufigkeit nach den Einsetzungen und der relativen Häufigkeit vor den Einsetzungen.

Jahr der letzten Änderung des Zivilstands	ZIVJ	f_ZIVJ [%]	iZIVJ	f_iZIVJ [%]	diff [‰]
leer	3322719	45.09	3167677	42.99	-21.04
2000	104897	1.42	110315	1.50	0.74
1996-1999	563146	7.64	589220	8.00	3.54
1991-1995	612435	8.31	637244	8.65	3.37
1986-1990	532927	7.23	554165	7.52	2.88
1981-1985	423225	5.74	440129	5.97	2.29
1976-1980	343503	4.66	358261	4.86	2.00
1971-1975	342153	4.64	355975	4.83	1.88
1966-1970	324545	4.40	335868	4.56	1.54
1961-1965	270241	3.67	278324	3.78	1.10
1956-1960	217256	2.95	223237	3.03	0.81
1951-1955	157276	2.13	161250	2.19	0.54
1946-1950	97895	1.33	100319	1.36	0.33
1941-1945	41997	0.57	42990	0.58	0.13
1936-1940	11437	0.16	11503	0.16	0.01
1931-1935	2111	0.03	1945	0.03	-0.02
1926-1930	309	0.00	155	0.00	-0.02
1921-1925	164	0.00	13	0.00	-0.02
1916-1921	93	0.00	0	0.00	-0.01
1911-1915	77	0.00	1	0.00	-0.01
1906-1910	79	0.00	0	0.00	-0.01
vor 1906	106	0.00	0	0.00	-0.01
Total	7368591	100.00	7368591	100.00	

Tabelle 36: Einfluss der Einsetzungen durch NIM-P auf die Verteilung der Variablen Eltern. 'PAR' und 'f_PAR' bezeichnen die absolute und die relative Häufigkeit vor den Einsetzungen. 'iPAR' und 'f_iPAR' bezeichnen die absolute und die relative Häufigkeit nach den Einsetzungen. 'diff' steht für die Differenz in Promillen der relativen Häufigkeit nach den Einsetzungen und der relativen Häufigkeit vor den Einsetzungen.

Elternschaft	PAR	f_PAR [%]	iPAR	f_iPAR [%]	diff [‰]
Nein	3391595	46.03	3841934	52.14	61.12
Ja	3541598	48.06	3526486	47.86	-2.05
leer	427738	5.80	171	0.00	-58.03
ungültig	7660	0.10	0	0.00	-1.04
Total	7368591	100.00	7368591	100.00	

Tabelle 37: Einfluss der Einsetzungen durch NIM-P auf die Verteilung der Variablen Anzahl Kinder. 'ZKIND' und 'f_ZKIND' bezeichnen die absolute und die relative Häufigkeit vor den Einsetzungen. 'iZKIND' und 'f_iZKIND' bezeichnen die absolute und die relative Häufigkeit nach den Einsetzungen. 'diff' steht für die Differenz in Promillen der relativen Häufigkeit nach den Einsetzungen und der relativen Häufigkeit vor den Einsetzungen. In der untersten Zeile ist das Total der Anzahl Kinder angegeben.

Anz. Kinder	ZKIND	f_ZKIND [%]	iZKIND	f_iZKIND [%]	diff [‰]
leer	3755636	50.97	3841385	52.13	11.64
ungültig	126779	1.72	726	0.01	-17.11
1	779391	10.58	803674	10.91	3.30
2	1586869	21.54	1592495	21.61	0.76
3	741417	10.06	740252	10.05	-0.16
4	260890	3.54	261845	3.55	0.13
5	66712	0.91	75411	1.02	1.18
6	26319	0.36	28612	0.39	0.31
7	11767	0.16	12195	0.17	0.06
8	6033	0.08	6008	0.08	0.00
9	2988	0.04	2835	0.04	-0.02
10	1518	0.02	1405	0.02	-0.02
11	991	0.01	697	0.01	-0.04
12	530	0.01	407	0.01	-0.02
13	289	0.00	244	0.00	-0.01
14	154	0.00	136	0.00	0.00
15	156	0.00	154	0.00	0.00
16	64	0.00	63	0.00	0.00
17	34	0.00	33	0.00	0.00
18	16	0.00	10	0.00	0.00
19	3	0.00	3	0.00	0.00
20	6	0.00	1	0.00	0.00
21+	29	0.00	0	0.00	0.00
Total	7368591	100.00	7368591	100.00	
Total Kinder	7913884		8000284		10.92

Tabelle 38: Einfluss der Einsetzungen durch NIM-P auf die Verteilungen der Variablen Geburtsjahr der Kinder, 1. und 2. Kind. 'P_nJ' und 'f_P_nJ' bezeichnen die absoluten und die relativen Häufigkeiten vor den Einsetzungen, $n = \{1, 2\}$. 'iP_nJ' und 'f_iP_nJ' bezeichnen die absoluten und die relativen Häufigkeiten nach den Einsetzungen, $n = \{1, 2\}$. 'diff_n' steht für die Differenz in Promillen der relativen Häufigkeiten nach den Einsetzungen und der relativen Häufigkeiten vor den Einsetzungen, $n = \{1, 2\}$. In der untersten Zeile ist das Total der Anzahl Personen, die eine Angabe in 'P_nJ', resp. 'iP_nJ' haben, $n = \{1, 2\}$.

Jahr	P1J	f_P1J [%]	iP1J	f_iP1J [%]	diff1 [%]	P2J	f_P2J [%]	iP2J	f_iP2J [%]	diff2 [%]
leer	3905125		3842118			4674267		4645786		
2000	62912	1.82	65218	1.85	0.33	52130	1.93	53091	1.95	0.15
1996-1999	273971	7.91	280858	7.96	0.54	226955	8.42	230747	8.47	0.51
1991-1995	362974	10.48	372068	10.55	0.71	302810	11.24	308318	11.32	0.85
1986-1990	370071	10.68	377908	10.72	0.31	300221	11.14	305054	11.20	0.61
1981-1985	358028	10.34	366002	10.38	0.41	286373	10.63	290783	10.68	0.51
1976-1980	332563	9.60	338965	9.61	0.10	275514	10.23	278503	10.23	0.03
1971-1975	335546	9.69	340818	9.66	-0.24	267018	9.91	268956	9.88	-0.32
1966-1970	341879	9.87	346663	9.83	-0.41	261533	9.71	263054	9.66	-0.46
1961-1965	311465	8.99	315286	8.94	-0.52	240577	8.93	241710	8.88	-0.52
1956-1960	253500	7.32	256706	7.28	-0.40	189174	7.02	189955	6.98	-0.45
1951-1955	190984	5.51	193637	5.49	-0.23	134986	5.01	135545	4.98	-0.32
1946-1950	141763	4.09	143856	4.08	-0.14	92403	3.43	92769	3.41	-0.22
1941-1945	81830	2.36	82964	2.35	-0.10	45519	1.69	45660	1.68	-0.12
1936-1940	31553	0.91	31709	0.90	-0.12	13801	0.51	13759	0.51	-0.07
1931-1935	11083	0.32	10984	0.31	-0.09	4199	0.16	4097	0.15	-0.05
1926-1930	2638	0.08	2498	0.07	-0.05	817	0.03	732	0.03	-0.03
1921-1925	478	0.01	320	0.01	-0.05	174	0.01	71	0.00	-0.04
1916-1921	103	0.00	12	0.00	-0.03	70	0.00	1	0.00	-0.03
1911-1915	91	0.00	1	0.00	-0.03	37	0.00	0	0.00	-0.01
1906-1910	34	0.00	0	0.00	-0.01	13	0.00	0	0.00	0.00
Total	3463466	100.00	3526473	100.00		2694324	100.00	2722805	100.00	

Tabelle 39: Einfluss der Einsetzungen durch NIM-P auf die Verteilungen der Variablen Geburtsjahr der Kinder, 3. und 4. Kind. 'P_nJ' und 'f_P_nJ' bezeichnen die absoluten und die relativen Häufigkeiten vor den Einsetzungen, $n = \{3, 4\}$. 'iP_nJ' und 'f_iP_nJ' bezeichnen die absoluten und die relativen Häufigkeiten nach den Einsetzungen, $n = \{3, 4\}$. 'diff_n' steht für die Differenz in Promillen der relativen Häufigkeiten nach den Einsetzungen und der relativen Häufigkeiten vor den Einsetzungen, $n = \{3, 4\}$. In der untersten Zeile ist das Total der Anzahl Personen, die eine Angabe in 'P_nJ', resp. 'iP_nJ' haben, $n = \{3, 4\}$.

Jahr	P3J	f_P3J [%]	iP3J	f_iP3J [%]	diff3 [%]	P4J	f_P4J [%]	iP4J	f_iP4J [%]	diff4 [%]
leer	6243870		6238281			6978204		6978530		
2000	19008	1.69	19301	1.71	0.18	5278	1.35	5289	1.36	0.04
1996-1999	85564	7.61	86485	7.65	0.44	23776	6.09	23877	6.12	0.31
1991-1995	118078	10.50	119346	10.56	0.60	33931	8.69	34140	8.75	0.61
1986-1990	119499	10.62	120627	10.67	0.47	32805	8.40	32944	8.45	0.43
1981-1985	106261	9.45	107067	9.47	0.25	29777	7.63	29811	7.64	0.15
1976-1980	95726	8.51	96139	8.51	-0.06	26883	6.89	26760	6.86	-0.26
1971-1975	93978	8.36	94136	8.33	-0.27	33156	8.49	33073	8.48	-0.14
1966-1970	124024	11.03	124265	10.99	-0.33	50658	12.98	50575	12.97	-0.10
1961-1965	129641	11.53	129924	11.49	-0.32	56633	14.51	56539	14.49	-0.12
1956-1960	98121	8.72	98266	8.69	-0.30	42724	10.94	42599	10.92	-0.23
1951-1955	67629	6.01	67744	5.99	-0.20	29245	7.49	29150	7.47	-0.18
1946-1950	42885	3.81	42927	3.80	-0.15	17094	4.38	17008	4.36	-0.18
1941-1945	17996	1.60	17986	1.59	-0.09	6350	1.63	6296	1.61	-0.12
1936-1940	4711	0.42	4674	0.41	-0.05	1586	0.41	1565	0.40	-0.05
1931-1935	1242	0.11	1191	0.11	-0.05	393	0.10	376	0.10	-0.04
1926-1930	247	0.02	217	0.02	-0.03	66	0.02	56	0.01	-0.03
1921-1925	51	0.00	12	0.00	-0.03	15	0.00	3	0.00	-0.03
1916-1921	39	0.00	3	0.00	-0.03	12	0.00	0	0.00	-0.03
1911-1915	17	0.00	0	0.00	-0.02	4	0.00	0	0.00	-0.01
1906-1910	4	0.00	0	0.00	0.00	1	0.00	0	0.00	0.00
Total	1124721	100.00	1130310	100.00		390387	100.00	390061	100.00	

Tabelle 40: Einfluss der Einsetzungen durch NIM-P auf die Verteilung der Variablen Geburtsjahr des letzten Kindes. 'PLJ' und 'f_PLJ' bezeichnen die absolute und die relativen Häufigkeit vor den Einsetzungen. 'iPLJ' und 'f_iPLJ' bezeichnen die absolute und die relative Häufigkeit nach den Einsetzungen. 'diffL' steht für die Differenz in Promillen der relativen Häufigkeit nach den Einsetzungen und der relativen Häufigkeit vor den Einsetzungen. In der untersten Zeile ist das Total der Anzahl Personen, die eine Angabe in 'PLJ', resp. 'iPLJ' haben.

Jahr	PLJ	f_PLJ [%]	iPLJ	f_iPLJ [%]	diffL [%]
leer	7236814		7240374		
2000	1800	1.37	1762	1.37	0.08
1996-1999	6811	5.17	6714	5.24	0.68
1991-1995	8432	6.40	8290	6.47	0.67
1986-1990	8276	6.28	8089	6.31	0.29
1981-1985	7921	6.01	7554	5.89	-1.19
1976-1980	9065	6.88	8605	6.71	-1.68
1971-1975	14960	11.35	14476	11.29	-0.62
1966-1970	23548	17.87	23027	17.96	0.90
1961-1965	22209	16.85	21744	16.96	1.05
1956-1960	14742	11.19	14424	11.25	0.63
1951-1955	8688	6.59	8454	6.59	0.01
1946-1950	4001	3.04	3849	3.00	-0.34
1941-1945	1051	0.80	991	0.77	-0.25
1936-1940	212	0.16	198	0.15	-0.06
1931-1935	48	0.04	39	0.03	-0.06
1926-1930	6	0.00	1	0.00	-0.04
1921-1925	3	0.00	0	0.00	-0.02
1916-1921	2	0.00	0	0.00	-0.02
1911-1915	1	0.00	0	0.00	-0.01
1906-1910	1	0.00	0	0.00	-0.01
Total	131777	100.00	128217	100.00	

Tabelle 41: Einfluss der Einsetzungen durch NIM-P auf die Verteilung der Variablen Stellung im Haushalt **inklusive** der Sammelhaushalte. 'STHH' und 'f_STHH' bezeichnen die absolute und die relative Häufigkeit vor den Einsetzungen. 'iSTHH' und 'f_iSTHH' bezeichnen die absolute und die relative Häufigkeit nach den Einsetzungen. Die relativen Häufigkeiten beziehen sich auf die **vorhandenen** Angaben (gültig und nicht leer). 'diff' steht für die Differenz der relativen Häufigkeit nach den Einsetzungen und der relativen Häufigkeit vor den Einsetzungen. In der untersten Zeile ist das Total der gültigen Antworten angegeben. Leere und ungültige Werte kommen nach der Einsetzung ausschliesslich bei Personen in Sammelhaushalten vor.

Stellung im Haushalt	STHH	f_STHH [%]	iSTHH	f_iSTHH [%]	diff [%]
Alleinlebend	1038429	16.04	1168505	16.19	0.15
Ehemann/-frau	2997328	46.29	3186412	44.14	-2.15
Partner	382440	5.91	464700	6.44	0.53
Alleinerz.	103923	1.60	142895	1.98	0.37
Anderer HH. Vorstand	57759	0.89	96366	1.33	0.44
Sohn/Tochter	1697082	26.21	2027948	28.09	1.88
Vater/Mutter	51149	0.79	22827	0.32	-0.47
Bruder/Schwester	35143	0.54	32456	0.45	-0.09
And. Verwandter	9314	0.14	10719	0.15	0.00
WG	71074	1.10	35410	0.49	-0.61
Angestellter	4616	0.07	5284	0.07	0.00
Zimmer-/Untermieter	14138	0.22	11203	0.16	-0.06
Andere	12969	0.20	14783	0.20	0.00
leer	626476	9.67	139237	1.93	-7.75
ungültig	266751	4.12	9846	0.14	-3.98
Total (gültig)	6475364	100.00	7219508	100.00	

Tabelle 42: Einfluss der Einsetzungen durch NIM-P auf die Verteilung der Variablen Stellung im Haushalt **ohne** Sammelhaushalte. 'STHH' und 'f_STHH' bezeichnen die absolute und die relative Häufigkeit vor den Einsetzungen. 'iSTHH' und 'f_iSTHH' bezeichnen die absolute und die relative Häufigkeit nach den Einsetzungen. Die relativen Häufigkeiten beziehen sich auf die **vorhandenen** Angaben. 'diff' steht für die Differenz der relativen Häufigkeit nach den Einsetzungen und der relativen Häufigkeit vor den Einsetzungen. In der untersten Zeile ist das Total der gültigen Antworten angegeben.

Stellung im Haushalt	STHH	f_STHH [%]	iSTHH	f_iSTHH [%]	diff [%]
Alleinlebend	1006068	15.86	1136144	16.03	0.17
Ehemann/-frau	2961346	46.69	3150430	44.46	-2.24
Partner	371489	5.86	453749	6.40	0.55
Alleinerz.	101215	1.60	140187	1.98	0.38
Anderer HH. Vorstand	56082	0.88	94689	1.34	0.45
Sohn/Tochter	1664893	26.25	1995759	28.16	1.91
Vater/Mutter	49770	0.78	21448	0.30	-0.48
Bruder/Schwester	34003	0.54	31316	0.44	-0.09
And. Verwandter	8841	0.14	10246	0.14	0.01
WG	63293	1.00	27629	0.39	-0.61
Angestellter	4141	0.07	4809	0.07	0.00
Zimmer-/Untermieter	12337	0.19	9402	0.13	-0.06
Andere	8685	0.14	10499	0.15	0.01
leer	489590	7.72	144	0.00	-7.72
ungültig	254729	4.02	31	0.00	-4.02
Total (gültig)	6475364	100.00	7219508	100.00	

A.5.3 KHH

Tabelle 43: Anzahl Einsetzungen und Änderungen nach Haushalts Typ, 'Typ'. Die Abkürzungen der Variablen entsprechen den in der Tabelle 22 definierten. In den untersten zwei Zeilen stehen das Total der Einsetzungen und deren prozentualer Anteil an der Bevölkerung der KHH.

Typ	GEJA	GESL	ZIVL	ZIVJ	PAR	ZKIND	P1J	P2J	P3J	P4J	PLJ
11	142	20	291	92	769	164	64	9	4	2	2
12	5	134	230	43	459	21	18	1	0	0	1
13	0	0	4	6	10	2	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0
21	83	10	86	36	1584	193	36	14	6	0	1
22	64	11	94	68	900	568	34	18	10	4	3
23	5	1	4	0	97	4	3	3	2	1	0
31	28	6	134	116	455	132	69	19	1	2	3
32	28	1	27	24	228	85	6	4	0	0	0
33	60	22	306	94	1030	153	48	11	4	0	2
34	11	0	128	118	434	95	118	25	3	4	2
40	0	0	2	6	10	12	0	0	0	0	0
41	530	13	2808	3956	14089	5578	3535	876	230	137	101
42	49	9	96	75	1018	417	46	19	10	3	0
43	2	1	30	47	183	67	22	1	0	0	0
44	9	1	40	35	177	92	32	11	5	5	3
45	0	0	8	7	9	28	2	0	1	0	0
46	5	2	21	4	120	44	1	1	0	0	0
47	23	2	38	33	191	43	24	10	7	6	2
50	0	0	0	0	0	0	0	0	0	0	0
51	35	2	24	12	886	335	4	2	2	3	2
52	0	0	2	3	42	0	2	2	0	0	0
53	0	0	0	0	8	4	1	0	0	0	0
61	39	6	184	259	523	121	31	11	8	7	1
62	18	4	59	74	208	67	34	7	1	1	3
63	0	0	0	2	3	1	0	0	0	1	1
64	9	2	87	86	213	70	17	9	4	1	3
71	2	1	143	148	197	39	29	7	3	3	2
72	21	0	147	304	370	269	59	13	2	0	1
73	626	21	2779	602	7838	169	34	6	3	3	6
74	1	0	31	46	356	48	15	2	0	0	1
81	36	2	290	571	4306	544	390	97	20	10	12
82	0	0	0	0	2	0	0	0	0	0	0
Total	1831	271	8093	6867	36715	9365	4674	1178	326	193	152
[%]	0.99	0.15	4.36	3.7	19.78	5.05	2.52	0.63	0.18	0.1	0.08

A.6 MBE in der PHE

Nationalität

Tabelle 44: Nationencode, Zahlen der VZ2000 vor der Restbereinigung, Name des Landes, Vergleichsstatistik Petra, die Differenz zwischen der zweiten und der vierten Kolonne und die Anzahl Nationalitäten, welche mit einer MBE durch Zufallsauswahl eingesetzt wurde.

NATION	VZ2000	Land	Petra	Differenz	MBE
0	3795				3
8100	6008595	Schweiz			1736
8201	2057	Albanien	1487	570	3
8204	7971	Belgien	8164	-193	15
8205	1910	Bulgarien	2006	-96	7
8206	3302	Dänemark	3388	-86	9
8207	115531	Deutschland	118360	-2829	312
8211	2551	Finnland	2561	-10	8
8212	63589	Frankreich	65281	-1692	224
8214	6355	Griechenland	6236	119	6
8215	22706	Vereinigtes Königreich	23718	-1012	96
8216	1614	Irland	1620	-6	9
8218	326806	Italien	326679	127	751
8222	23287	Liechtenstein	1619		2
8223	884	Luxemburg	900	-16	1
8227	14907	Niederlande	15129	-222	38
8228	1567	Norwegen	1596	-29	6
8229	32402	Österreich	30957	1445	91
8230	5021	Polen	4545	476	19
8231	143743	Portugal	147798	-4055	250
8232	3079	Rumänien	3040	39	10
8234	5974	Schweden	5999	-25	21
8236	86021	Spanien	85739	282	133
8239	85061	Türkei	83124	1937	111
8240	3804	Ungarn	3675	129	13
8243	2123	Slowakei	2152	-29	5
8244	3739	Tschechische Republik	3606	133	1
8249	206151	Jugoslawien	211110	-4959	302
8250	43552	Kroatien	43998	-446	66
8251	2875	Slowenien	2800	75	13
8252	47927	Bosnien und Herzegowina	50496	-2569	90
8254	50	Montenegro		50	1
8255	56250	Mazedonien	56154	96	42
8256	8906	Kosovo		8906	22
8261	370	Lettland	387	-17	10
8262	267	Litauen	262	5	1
8264	6994	Russland	6457	537	57
8265	1853	Ukraine	2038	-185	3
8266	307	Weissrussland	330	-23	2
8302	2146	Äthiopien	2234	-88	26
8304	3846	Algerien	3826	20	23
8305	4161	Angola	4305	-144	12
8308	348	Burundi	359	-11	4
Fortsetzung auf nächster Seite . . .					

... Fortsetzung der Tabelle 44					
NATION	VZ2000	Land	Petra	Differenz	MBE
8309	161	Benin	149	12	2
8310	913	Côte D'Ivoire	907	6	5
8311	75	Gabun	63	12	5
8313	1211	Ghana	1198	13	1
8315	605	Guinea	680	-75	1
8317	2058	Kamerun	1960	98	5
8319	994	Kapverden	988	6	9
8320	1053	Kenia	1038	15	8
8322	733	Kongo (Brazzaville)	460	273	8322
8323	4139	Kongo (Kinshasa)	4583	-444	33
8325	191	Liberia	179	12	4
8326	730	Libyen	711	19	4
8327	448	Madagaskar	420	28	1
8331	5759	Marokko	5811	-52	25
8332	122	Mauretanien	114	8	1
8336	1340	Nigeria	1299	41	8
8337	214	Burkina Faso	233	-19	1
8340	197	Zimbabwe	207	-10	6
8341	597	Rwanda	604	-7	6
8345	918	Senegal	901	17	14
8347	754	Sierra Leone	636	118	29
8348	4764	Somalia	5666	-902	77
8349	1165	Südafrika	1102	63	1
8350	614	Sudan	621	-7	6
8353	218	Tansania	213	5	7
8354	430	Togo	439	-9	1
8356	72	Tschad	66	6	1
8357	4488	Tunesien	4464	24	7
8358	294	Uganda	288	6	1
8359	1660	Ägypten	1623	37	12
8362	1333	Eritrea	1296	37	10
8401	1337	Argentinien	1289	48	1
8405	722	Bolivien	708	14	3
8406	8830	Brasilien	8473	357	32
8407	3943	Chile	3878	65	6
8408	230	Costa Rica	180	50	6
8409	4310	Dominikanische Republik	4626	-316	7
8410	1126	Ecuador	1066	60	6
8415	222	Guatemala	226	-4	3
8417	40	Guyana	39	1	2
8421	321	Jamaika	282	39	2
8423	5118	Kanada	5002	116	41
8424	3102	Kolumbien	3089	13	9
8425	1047	Kuba	1042	5	7
8427	1371	Mexiko	1315	56	11
8429	94	Nicaragua	100	-6	2
8430	103	Panama	93	10	2
8432	2786	Peru	2721	65	22
8437	574	Uruguay	556	18	1
8438	813	Venezuela	768	45	
8439	14297	Vereinigte Staaten	14422	-125	90
8440	471	Dominica	49	422	3
Fortsetzung auf nächster Seite ...					

... Fortsetzung der Tabelle [44](#)

NATION	VZ2000	Land	Petra	Differenz	MBE
8445	13	St. Kitts und Nevis	14	-1	1
8501	2309	Afghanistan	2317	-8	4
8505	109	Myanmar	98	11	4
8506	32870	Sri Lanka	33246	-376	85
8507	402	Taiwan	438	-36	7
8508	6593	China	6756	-163	24
8510	6211	Indien	6250	-39	30
8511	1181	Indonesien	1186	-5	2
8512	4591	Irak	4773	-182	35
8513	4129	Iran	4105	24	28
8514	1070	Israel	1082	-12	4
8515	3476	Japan	3434	42	19
8517	362	Jordanien	363	-1	3
8518	1234	Kambodscha	1198	36	2
8523	2248	Libanon	2209	39	5
8525	934	Malaysia	932	2	5
8526	28	Malediven	28	0	
8529	253	Nepal	271	-18	25
8530	236	Korea (Nord-)	103	133	
8533	2334	Pakistan	2401	-67	7
8534	5032	Philippinen	4967	65	29
8535	300	Saudi-Arabien	303	-3	3
8539	1078	Korea (Süd-)	1226	-148	5
8541	1043	Syrien	1078	-35	12
8542	5877	Thailand	5669	208	8
8545	4951	Vietnam	4764	187	10
8546	997	Bangladesh	1001	-4	4
8550	252	Palästina	72	180	1
8560	673	Armenien	767	-94	13
8561	176	Aserbaidshan	169	7	1
8562	408	Georgien	438	-30	4
8567	91	Usbekistan	87	4	2
8601	2298	Australien	2191	107	16
8607	680	Neuseeland	663	17	6
8998	332	Staatenlos	235	97	1

Aufenthaltsstatus

Tabelle 45: Typ des Aufenthaltsstatus (Ausweis), Anzahl mit einer MBE eingesetzter Werte, Prozent der eingesetzten Werte.

Ausweis	MBE	% der eingesetzten Werte
C	6573	45.01
B	4714	32.28
A	68	0.47
N	1354	9.27
S	1239	8.48
F	158	1.08
L	438	3.00
EDA	61	0.42
Total	14605	100.00

Muttersprache

Tabelle 46: In der ersten Spalte steht der Sprachcode, in der zweiten die Anzahl mit einer MBE eingesetzter Werte und in der dritten Spalte steht der Anteil an sämtlichen Einsetzungen. Dies wiederholt sich drei mal.

Sprache	MBE	[%]	Sprache	MBE	[%]	Sprache	MBE	[%]
110100	22198	24.08	123400	7	0.01	134100	23	0.02
110104	2	0	123401	5	0.01	134101	96	0.1
110105	50	0.05	124100	5457	5.92	134102	1	0
110106	6	0.01	124101	48	0.05	134200	213	0.23
110200	32	0.03	125100	225	0.24	134201	2	0
111100	5561	6.03	125101	8	0.01	135100	4	0
111101	22	0.02	125102	1	0	139903	3	0
111102	1	0	125103	1	0	140100	21	0.02
111103	3	0	125104	1	0	140200	24	0.03
112100	240	0.26	125107	1	0	149900	1	0
112101	174	0.19	129900	3	0	150100	6082	6.6
112102	16	0.02	130100	1657	1.8	150101	1	0
112200	30	0.03	130101	1	0	150105	128	0.14
113100	1	0	130200	3	0	150106	7	0.01
114100	10	0.01	130202	6	0.01	150107	2	0
115100	161	0.17	130300	68	0.07	150108	8	0.01
115101	2	0	131100	159	0.17	160100	207	0.22
115300	25	0.03	131200	104	0.11	170100	2	0
115400	103	0.11	131201	289	0.31	170101	1	0
115500	268	0.29	131202	1	0	210100	152	0.16
120100	17700	19.2	132100	1279	1.39	210200	23	0.02
120103	1	0	132101	837	0.91	210300	2	0
120110	1	0	132102	3	0	220101	1	0
120206	35	0.04	132103	1669	1.81	240100	271	0.29
120302	1	0	132104	9	0.01	250101	1	0
121200	6887	7.47	132105	1	0	312100	2	0
121210	1	0	132106	745	0.81	312200	6	0.01
Fortsetzung auf nächster Seite ...								

... Fortsetzung der Tabelle 46

Sprache	MBE	[%]	Sprache	MBE	[%]	Sprache	MBE	[%]
121400	2	0	132107	2	0	313100	8	0.01
122300	67	0.07	132108	10	0.01	313300	2248	2.44
122301	26	0.03	132109	40	0.04	313301	10	0.01
123100	4314	4.68	132110	839	0.91	313400	9	0.01
123101	4	0	132111	105	0.11	314400	8	0.01
123105	2	0	132200	65	0.07	314500	7	0.01
123108	1	0	132201	2	0	399900	2	0
123300	22	0.02	133100	340	0.37	410100	2	0
123301	1	0	133101	1	0	420100	2	0
510101	8	0.01	712500	1	0	738400	4	0
510111	1	0	712700	1	0	740500	1	0
510151	3	0	714200	44	0.05	799900	4	0
521100	10	0.01	714201	33	0.04	799901	1	0
521300	468	0.51	714700	2	0	799902	4	0
521700	1	0	714701	30	0.03	999900	19	0.02
523200	1	0	715200	2	0	999901	6	0.01
524300	1	0	715300	1	0	999902	1	0
525300	3	0	715500	1	0	1010100	1747	1.9
525800	2	0	715801	5	0.01	1010101	32	0.03
526100	1	0	716300	9	0.01	1010112	17	0.02
530100	162	0.18	716301	12	0.01	1010114	1	0
530102	25	0.03	716701	6	0.01	1011100	51	0.06
530300	80	0.09	717800	1	0	1011102	5	0.01
530400	4	0	717801	3	0	1011105	6	0.01
599900	1	0	717803	2	0	1020100	70	0.08
599901	13	0.01	718000	2	0	1020102	3	0
599902	9	0.01	718503	5	0.01	1020103	1	0
599903	20	0.02	720400	1	0	1099900	1	0
610100	1	0	720800	1	0	1120200	1	0
610400	1	0	720900	5	0.01	1120300	4	0
610900	1	0	721000	2	0	1140101	4	0
611000	1	0	721100	1	0	1150100	73	0.08
611500	1	0	721501	2	0	1199900	1	0
613300	1	0	721800	2	0	1211400	68	0.07
613601	1	0	732500	12	0.01	1212200	3	0
617100	1	0	732600	9	0.01	1212400	2	0
699900	1	0	735500	1	0	1213100	135	0.15
710200	6	0.01	736000	4	0	1213200	1	0
710202	8	0.01	736001	20	0.02	1213400	2	0
710203	1	0	736100	3	0	1214200	5	0.01
710204	39	0.04	736300	7	0.01	1215100	4	0
710205	7	0.01	736400	1	0	1215101	54	0.06
710207	84	0.09	736500	1	0	1215200	1	0
710300	43	0.05	737000	12	0.01	1215300	92	0.1
710400	3	0	737100	1	0	1215301	4	0
710600	6	0.01	737200	38	0.04	1215400	228	0.25
711100	1	0	737400	7	0.01	1215600	7	0.01
711200	9	0.01	737500	2	0	1215601	7	0.01
711400	34	0.04	737800	53	0.06	1216100	33	0.04
711500	66	0.07	737900	13	0.01	1216104	3	0
711501	2	0	738000	2	0	1217100	46	0.05
712400	9	0.01	738100	5	0.01	1217101	4	0

Fortsetzung auf nächster Seite ...

... Fortsetzung der Tabelle 46

Sprache	MBE	[%]	Sprache	MBE	[%]	Sprache	MBE	[%]
1217102	2	0	1610101	61	0.07	1899900	3	0
1217200	2	0	1610106	31	0.03	2110100	482	0.52
1219900	56	0.06	1610107	2	0	2120100	234	0.25
1221100	9	0.01	1610108	2	0	2210301	1	0
1222100	23	0.02	1620100	109	0.12	2210500	106	0.11
1222101	18	0.02	1620103	6	0.01	2220202	1	0
1222103	1	0	1620104	1	0	2220500	2	0
1223200	73	0.08	1623100	15	0.02	2310200	2	0
1223201	293	0.32	1623101	16	0.02	2712400	1	0
1223202	3	0	1623200	1	0	2910400	13	0.01
1223203	54	0.06	1623203	1	0	2911200	1	0
1223204	5	0.01	1623700	1	0	2911400	4	0
1223300	595	0.65	1699900	3	0	2911500	11	0.01
1223303	3	0	1710302	1	0	2911600	1	0
1223304	10	0.01	1712700	212	0.23	2912200	5	0.01
1223305	2	0	1713100	1	0	2913200	1	0
1229900	33	0.04	1713101	20	0.02	2913500	1	0
1230100	177	0.19	1713200	1	0	2916100	2	0
1311300	1	0	1810200	1	0	2917100	4	0
1312700	16	0.02	1810300	222	0.24	2918100	1	0
1312701	1	0	1810500	7	0.01	2920100	1	0
1313200	7	0.01	1810800	3	0	2923100	3	0
1313400	99	0.11	1810900	1	0	2923101	1	0
1313500	1592	1.73	1811200	1	0	2999901	49	0.05
1313501	3	0	1811500	1	0	2999902	3	0
1313800	1	0	1811600	1	0	3010900	29	0.03
1410100	12	0.01	1811900	2	0	3011000	2	0
1410101	1	0	1813000	17	0.02	3100999	2	0
1410300	178	0.19	1813001	4	0	3110100	3	0
1410301	191	0.21	1813003	1	0	3110200	2	0
1410302	1	0	1815100	3	0	3110900	1	0
1410700	1	0	1821000	1	0	3199900	3	0
1520100	1	0	1821601	2	0			
1610100	730	0.79	1872000	2	0			

B Programme PHE

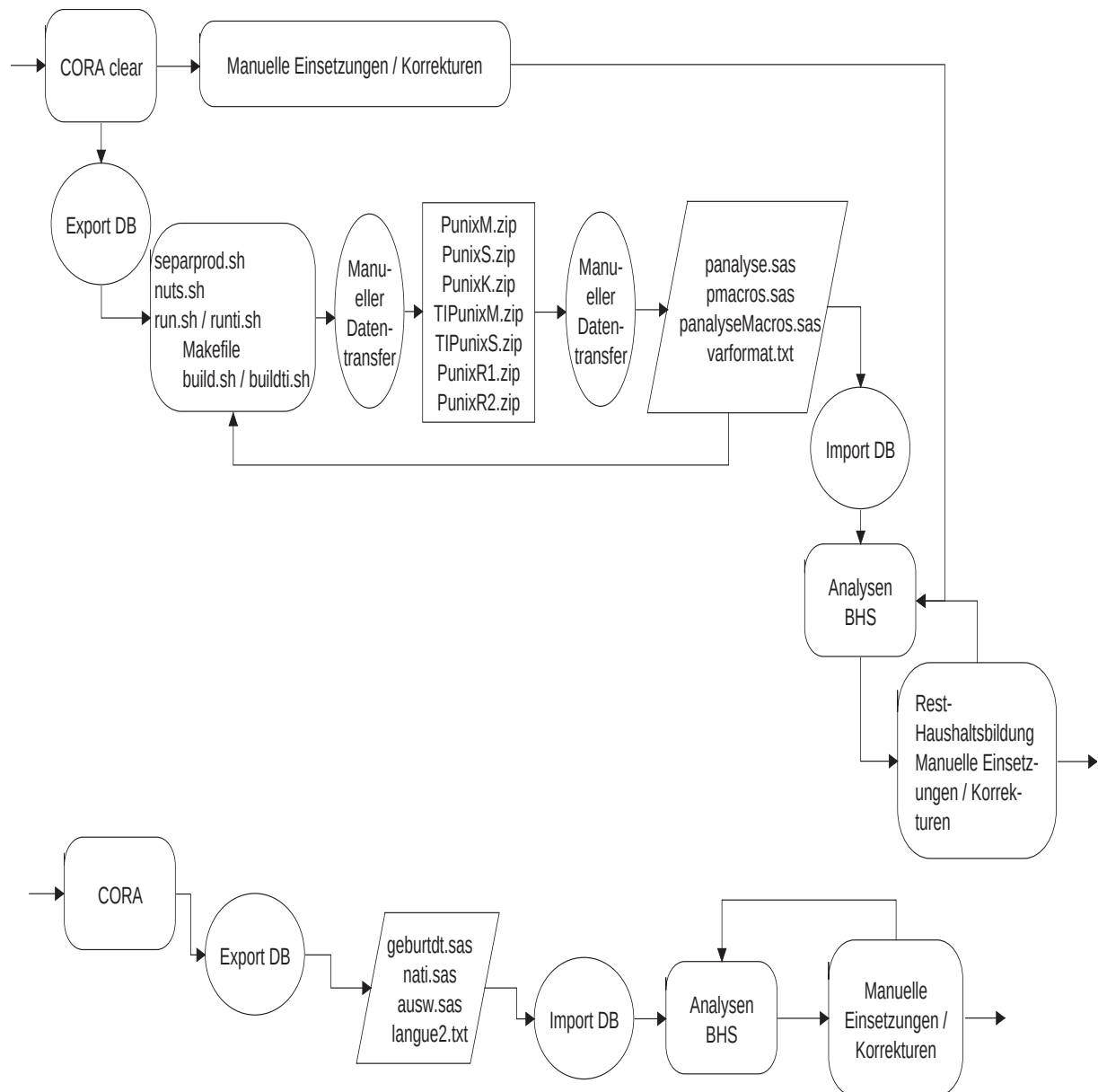


Abbildung 1: Vereinfachte Verarbeitungsschemata der Einsetzungen mit NIM und modellbasierten Einsetzungen in der PHE.

Methodenberichte des Dienstes Statistische Methoden des BFS Rapports de méthodes du Service de méthodes statistiques de l'OFS Methodology reports of the Statistical Methods Unit of SFSO

- Kilchmann D., Eichenberger, P., Potterat, J. (2005). Volkszählung 2000. Statistische Einsetzungsverfahren, Band 2. Bestellnummer: 338-0033
- Kilchmann D., Eichenberger, P., Potterat, J. (2005). Volkszählung 2000. Statistische Einsetzungsverfahren, Band 1. Bestellnummer: 338-0032
- Graf, M., Matei, A. (2005). Enquête suisse sur la structure des salaires 2002. La précision du salaire brut standardisé médian. Numéro de commande: 338-0031
- Graf, E., Renfer, J.-P. (2005). Enquête suisse sur la santé 2002. Plan d'échantillonnage, pondération et estimation de la précision. Numéro de commande: 338-0030
- Potterat, J. (2005). Mietpreis-Strukturerhebung 2003. Gewichtung und Schätzverfahren. Bestellnummer: 338-0029
- Potterat, J. (2005). Landwirtschaftliche Betriebszählung 2003. Schätzverfahren für die Zusatzerhebung. Bestellnummer: 338-0028
- Renaud, A. (2004). Coverage estimation for the Swiss population census 2000. Estimation methodology and results. Order number: 338-0027
- Kilchmann, D. (2004). Revision des Schweizerischen Lohnindex. Schätzmethoden der Lohnindizes und deren Varianzschätzer. Bestellnummer: 338-0026
- Graf, M. (2004). Enquête suisse sur la structure des salaires 2002. Plan d'échantillonnage et extrapolation pour le secteur privé. Numéro de commande: 338-0025
- Renaud, A. (2004). Analyse de données d'enquêtes. Quelques méthodes et illustration avec des données de l'OFS. Numéro de commande 338-0024
- Renaud, A., Potterat, J. (2004). Estimation de la couverture du recensement de la population de l'an 2000. Echantillon pour l'estimation de la sous-couverture (P-sample) et qualité du cadre de sondage des bâtiments. Numéro de commande: 338-0023
- Graf, M. (2004). Fusion de données. Etude de faisabilité. Numéro de commande: 338-0022
- Potterat, J. (2003). Mietpreis-Strukturerhebung 2003. Entwicklung des Stichprobenplans und Ziehung der Stichprobe. Bestellnummer: 338-0021
- Potterat, J. (2003). Landwirtschaftliche Betriebszählung 2003. Stichprobenplan der Zusatzerhebung. Bestellnummer: 338-0020
- Renaud, A. (2003). Estimation de la couverture du recensement de la population de l'an 2000. Echantillon pour l'estimation de la sur-couverture (E-sample). Numéro de commande: 338-0019
- Hulliger, B. (2003). Erhebung über Forschung und Entwicklung in der schweizerischen Privatwirtschaft 2000. Bereinigung der Stichprobe, Ersatz fehlender Werte und Schätzverfahren. Bestellnummer: 338-0018
- Renfer, J.-P. (2003). Enquête 2000 sur la recherche et le développement dans l'économie privée en Suisse. Plan d'échantillonnage. Numéro de commande: 338-0017

- Potterat, J. (2003). Kosten und Nutzen der Berufsbildung aus Sicht der Betriebe. Schätzverfahren. Bestellnummer: 338-0016
- Graf, M., Matei, A. (2003). Stratégie de choix des modèles de désaisonnalisation. Application aux séries de l'emploi total. Numéro de commande : 338-0015
- Potterat, J., Salamin, P.A. (2002). Betriebszählung 2001. Methoden für die Datenbereinigung. Bestellnummer: 338-0014
- Renaud, A. (2002). Programme international pour le suivi des acquis des élèves (PISA). Plans d'échantillonnage pour PISA 2000 en Suisse. Numéro de commande: 338-0013
- Renfer, J.-P. (2002). Enquête 2001 sur les coûts et l'utilité de la formation des apprentis du point de vue des établissements. Plan d'échantillonnage. Numéro de commande: 338-0012
- Potterat, J., Salamin, P.A. (2002). Betriebszählung 2001. Stichprobenplan und Schätzverfahren für die provisorischen Ergebnisse. Bestellnummer: 338-0011
- Graf, M. (2002). Enquête suisse sur la structure des salaires 2000. Plan d'échantillonnage, pondération et méthode d'estimation pour le secteur privé. Numéro de commande: 338-0010
- Renaud, A., Eichenberger P. (2002). Estimation de la couverture du recensement de la population de l'an 2000. Procédure d'enquête et plan d'échantillonnage de l'enquête de couverture. Numéro de commande: 338-0009
- Kilchmann, D., Hulliger, B. (2002). Stichprobenplan für die Obstbaumzählung 2001. Bestellnummer: 338-0008
- Graf, M. (2002). Passage du concept établissement au concept entreprise. Numéro de commande: 338-0007
- Salamin, P.A. (2001). La technique de la double enquête pour la statistique du transport routier de marchandise. Numéro de commande: 338-0006
- Peters, R., Renfer, J.-P. et Hulliger, B. (2001). Statistique de la valeur ajoutée 1997-1998. Procédure d'extrapolation des données. Numéro de commande: 338-0005
- Potterat, J., Hulliger, B. (2001). Schätzung der Sägereiproduktion mit der Sägerei-Erhebung PAUL. Bestellnummer: 338-0004
- Graf, M. (2001). Désaisonnalisation. Aspects méthodologiques et application à la statistique de l'emploi. Numéro de commande: 338-0003
- Hüsler, J., Müller, S. (2001). Schlussbericht Betriebszählung 1995 (BZ 95), Mehrfach imputierte Umsatzzahlen. Bestellnummer: 338-0002
- Renaud, A. (2001). Statistique suisse des bénéficiaires de l'aide sociale. Plan d'échantillonnage des communes. Numéro de commande: 338-0001
- Hulliger, B., Eichenberger, P. (2000). Stichprobenregister für Haushalterhebungen: Umstellung auf Telefonnummern ohne Namen und Adressen, Abläufe für Erstellung und Stichprobenziehung. Bestellnummer: 338-0000
- de Rossi, F.-X. (1998). Méthodes statistiques pour le compte routier suisse.
- Hulliger, B., Kassab, M. (1998). Evaluation of Estimation Methods for the Survey on Environment Protection Expenditures of Swiss Communes.

- Salamin, P.A. (1998). Etablissement d'une clef de passage pondérée entre l'ancienne (NGAE 85) et la nouvelle nomenclature (NOGA 95) générale des activités économiques.
- Peters, R. (1998). Extrapolation des données de l'enquête de structure sur les loyers.
- Bender, A., Hulliger, B. (1997). Enquête suisse sur la population active: rapport de pondération pour 1996.
- Salamin, P.A. (1997). Evaluation de la Statistique de l'emploi.
- Peters, R. (1997). Etablissement du plan d'échantillonnage pour l'enquête 1996 sur la recherche et le développement dans l'économie privée en Suisse.
- Peters, R. (1997). Enquête 1996 sur la structure des salaires en Suisse: établissement du plan d'échantillonnage.
- Peters, R. (1996). Pondération des données de l'enquête sur la famille en Suisse.
- Comment, T., Hulliger, B., Ries, A. (1996). Gewichtungsverfahren für die Schweizerische Arbeitskräfteerhebung (1991-1995).
- Hulliger, B. (1996). Haushalterhebung Familie 1994: Stichprobenplan, Stichprobenziehung und Reservestichproben.
- Peters, R., Hulliger, B. (1996). Schätzverfahren für die Lohnstruktur-Erhebung 1994 / Procédure d'estimation pour l'enquête de 1994 sur la structure des salaires.
- Peters, R. (1996). Schéma de pondération des indices PAUL.
- Hulliger, B., Peters, R. (1996). Enquête sur le comportement de la population suisse en matière de transport en 1994: plan d'échantillonnage et pondération.
- Hulliger, B. (1996). Gütertransportstatistik 1993: Schätzverfahren mit Kompensation der Antwortausfälle.
- Salamin, P.A. (1995). Estimation des flux pour le module II des comptes globaux du marché de travail.
- Peters, R. (1995). Enquête de structure sur les loyers: établissement d'un plan d'échantillonnage stratifié.
- Hulliger, B. (1995). Konjunktuelle Mietpreiserhebung: Stichprobenplan und Schätzverfahren.
- Schwendener, P. (1995). Verbrauchserhebung 1990 - Vertrauensintervalle.
- Peters, R., Hulliger, B. (1994). La technique de pondération des données: application à l'enquête suisse sur la santé.
- Hulliger, B., Peters, R. (1994). Enquête sur la structure des salaires en Suisse: stratégie d'échantillonnage pour le secteur privé.

Publikationsprogramm BFS

Das Bundesamt für Statistik (BFS) hat – als zentrale Statistikstelle des Bundes – die Aufgabe, statistische Informationen breiten Benutzerkreisen zur Verfügung zu stellen.

Die Verbreitung der statistischen Information geschieht gegliedert nach Fachbereichen (vgl. Umschlagseite 2) und mit verschiedenen Mitteln

Programme des publications de l'OFS

En sa qualité de service central de statistique de la Confédération, l'Office fédéral de la statistique (OFS) a pour tâche de rendre les informations statistiques accessibles à un large public.

L'information statistique est diffusée par domaine (cf. verso de la première page de couverture); elle emprunte diverses voies:

<i>Diffusionsmittel</i>	<i>Kontakt N° à composer</i>	<i>Moyen de diffusion</i>
Individuelle Auskünfte	032 713 60 11 info@bfs.admin.ch	Service de renseignements individuels
Das BFS im Internet	www.statistik.admin.ch	L'OFS sur Internet
Medienmitteilungen zur raschen Information der Öffentlichkeit über die neusten Ergebnisse	www.news-stat.admin.ch	Communiqués de presse: information rapide concernant les résultats les plus récents
Publikationen zur vertieften Information (zum Teil auch als Diskette/CD-Rom)	032 713 60 60 order@bfs.admin.ch	Publications: information approfondie (certaines sont disponibles sur disquette/CD-Rom)
Online-Datenbank	032 713 60 86 www.statweb.admin.ch	Banque de données (accessible en ligne)

Nähere Angaben zu den verschiedenen Diffusionsmitteln im Internet unter der Adresse www.statistik.admin.ch → Dienstleistungen → Publikationen Statistik Schweiz.

Informations sur les divers moyens de diffusion sur Internet à l'adresse www.statistique.admin.ch → Services → Les publications de Statistique suisse.

Methodenberichte des Dienstes Statistische Methoden Rapports de méthodes du Service de méthodes statistiques Methodological reports of the Statistical Methods Unit

Die Methodenberichte beschreiben die mathematischen und statistischen Methoden, die den Resultaten und Analysen der öffentlichen Statistik zu Grunde liegen. Sie enthalten ausserdem die Evaluation und Entwicklung von neuen Methoden im Hinblick auf eine zukünftige Anwendung. Diese Publikationen sollen einerseits die verwendeten Methoden dokumentieren, um Transparenz und Wissenschaftlichkeit sicher zu stellen, und sie sollen andererseits die Zusammenarbeit mit den Hochschulen und der Wissenschaft fördern.

Zur Illustration der beschriebenen mathematischen Konzepte, werden im Bericht numerische Resultate aufgeführt. Diese sind allerdings nicht als offizielle Resultate der betreffenden Erhebungen zu verstehen. Ebenfalls können die tatsächlich angewendeten Methoden leicht von den hier beschriebenen abweichen.

Die Methodenberichte sind auf der Internetseite des BFS in elektronischer Form verfügbar.

Les rapports de méthodes décrivent les méthodes mathématiques et statistiques à la base des résultats et des analyses de la statistique publique. Ils présentent également l'évaluation et le développement de nouvelles méthodes en vue d'une application future. Ces publications visent d'une part à documenter les méthodes utilisées ou envisagées dans un souci de transparence et de rigueur scientifique, et d'autre part à favoriser la collaboration avec le monde scientifique et universitaire.

Les résultats numériques présentés dans les rapports de méthodes illustrent les concepts mathématiques décrits, mais ne sont pas des résultats officiels des enquêtes concernées. De même, les méthodes réellement appliquées peuvent différer légèrement de celles décrites dans ces rapports.

Les rapports de méthodes sont disponibles sous forme électronique sur le site internet de l'OFS.

In der Volkszählung 2000 (VZ2000) wurden fehlende Werte und widersprüchliche Daten durch Rückfragen, manuelle Bearbeitungen, deterministische Korrekturen, die Einsetzung eines nahen Nachbarn und durch modellbasierte Einsetzungsmethoden bearbeitet. Die Arbeiten des Dienstes Statistische Methoden (METH) beschränkten sich auf die statistischen Einsetzungsverfahren (Naher Nachbar Einsetzung, Modellbasierte Einsetzung).

Der vorliegende erste Band des Methodenberichts befasst sich mit den Definitionen der Lokalisierung fehlender und fehlerhafter Daten sowie von Einsetzungsverfahren. Die Erarbeitung und Anwendung statistischer Einsetzungsverfahren innerhalb der Personen- und Haushaltserhebung der VZ2000 wird ebenso im ersten Band beschrieben wie das Vorgehen und die Evaluation der Einsetzungen.

Im zweiten Band werden die statistischen Einsetzungsverfahren, die in der Gebäude- und Wohnungserhebung angewandt wurden, beschrieben. Zudem enthält der zweite Band eine Zusammenfassung der von METH durchgeführten Einsetzungsverfahren in der VZ2000 sowie die Schlussfolgerung beider Bände.

Es hat sich gezeigt, dass statistische Einsetzungsverfahren eine beträchtliche Vorbereitung voraussetzen. In der VZ2000 konnten vollständige Daten nur anhand von statistischen Einsetzungsverfahren erreicht werden. Der Einfluss der Einsetzungen hängt stark von der Qualität der vorangegangenen Prozesse der Erhebung ab. Unter Umständen können ungewollte Merkmalskombinationen durch statistische Einsetzungsverfahren vermehrt zu Tage treten.