Das Korrekturverfahren beim Zensus 2011

Dipl.-Geogr. Katrin Hofmeister, Dr. Michael Fürnrohr

Beim Zensus 2011 wurde in Gemeinden mit 10 000 oder mehr Einwohnern eine Haushaltsstichprobe durchgeführt. Zweck dieser Stichprobe war neben der Erhebung von nicht in Registern verfügbaren Daten primär die gemeindeweise Gewinnung von demographischen und haushaltsstatistischen Informationen zu Über- und Untererfassungen (Karteileichen und Fehlbestände) in den Melderegistern. Mit diesen Informationen sollen die potenziellen Fehler einer unkontrollierten Registerauszählung vermieden werden. Um einen qualitativ hochwertigen, fachlich und regional flexibel auswertbaren Zensuseinzeldatensatz zu erhalten, muss eine Bereinigung der Karteileichen und Fehlbestände auf der Basis der Einzeldaten vorgenommen werden.
Zu diesem Zwecke war es erforderlich, ein Verfahren zu entwickeln, welches die gemeindeweise aggregierten Vorgaben aus der Haushaltsstichprobe möglichst genau umsetzt. Es ist zu berücksichtigen, dass eine solche Korrektur der Einzeldaten nur statistisch erfolgen konnte, d. h. nicht die buchhalterisch betrachtete „Richtigkeit“ des Einzelfalls war relevant und auch realisierbar, sondern die strukturelle Qualität der Zensusergebnisse.

1. Einführung

Ein Zensus oder eine Volkszählung ist eine Inventur, die Bestands- und Strukturdaten zu Bevölkerung, Wohnen und Erwerbstätigkeit auch kleinräumig erhebt und damit im Rahmen des statistischen Gesamtsystems neue Basiszahlen für Fortschreibungen und Stichprobenerhebungen ermittelt.

Amtliche Einwohnerzahlen

Insbesondere dient ein Zensus der Feststellung der amtlichen Einwohnerzahlen (Zahl der Personen mit Hauptwohnsitz) für Bund, Länder und Kommunen. Diese haben eine Vielzahl von unmittelbaren Auswirkungen auf die einzelnen Gebietskörperschaften, auch mit direkten finanziellen Folgen, z. B. beim kommunalen Finanzausgleich. Außerdem besitzen die amtlichen Einwohnerzahlen eine hohe Bedeutung über den Zensusstichtag hinaus. Sie bilden die Grundlage für die Bevölkerungsfortschreibung, mit der zwischen den Zensus in regelmäßigen Abständen die amtlichen Einwohnerzahlen für Bund, Länder und Gemeinden nachgewiesen werden. Die amtliche Einwohnerzahl wird in rund 50 Rechtsvorschriften als eine wichtige Bemessungsgrundlage verwendet. Sie ist unter anderem die Richtgröße für den horizontalen und vertikalen Finanzausgleich und dient der Berechnung der Stimmen der Länder im Bundesrat sowie der Sitze in den kommunalen Vertretungskörperschaften. Angesichts ihrer Bedeutung für das demokratische Staatswesen und der Finanzbeziehungen zwischen Bund, Ländern und Gemeinden kommt der Genauigkeit der ermittelten Zahlen eine herausragende Bedeutung zu. Beispielsweise fällt jeder Einwohner beim Länderfinanzausgleich mit ca. 2 000 Euro ins Gewicht. An die Feststellung der amtlichen Einwohnerzahlen sind daher besondere Anforderungen zu stellen, die über die üblichen Anforderungen an die statistischen Verfahren und Ergebnisse hinausgehen.

Erkenntnisse des Zensustests 2001

Im Rahmen des registergestützten Zensus 2011 bilden die Melderegister die Grundlage für die Ermittlung der Einwohnerzahlen und der demographischen Grunddaten zu Alter, Geschlecht, Familienstand und Staatsangehörigkeit. Zur Vorbereitung des Zensus 2011 wurde im Jahr 2001 ein umfangreicher Zensustest durchgeführt. Hierbei hat sich herausgestellt, dass die Melderegister Fehler hinsichtlich ihrer Vollzähligkeit aufweisen. Sie beinhalten sowohl Karteileichen (Personen, die an einer Anschrift gemeldet,

dort aber tatsächlich nicht wohnhaft sind) als auch Fehlbestände (Personen, die an einer Anschrift nicht gemeldet, dort aber tatsächlich wohnhaft sind). Die Ursachen sind im Wesentlichen auf das Meldeverhalten der Bürgerinnen und Bürger zurückzuführen. Beispiele sind Studentinnen und Studenten, die am Studienort leben, aber noch bei den Eltern gemeldet sind, ältere Menschen in Heimen, die noch bei ihren Nachkommen gemeldet sind oder Ausländer, die ohne Abmeldung in ihre Heimatländer zurückgekehrt sind. Defizite im Verwaltungsvollzug, z. B. Personen, die mehrfach mit Hauptwohnung gemeldet sind, machen nach den Erkenntnissen des Zensustests hingegen nur rund ein Fünftel des Registerfehlers aus.

Ferner hat sich im Zensustest gezeigt, dass die Höhe der Registerfehler abhängig von der Größe der Gemeinde ist. Wie aus der Tabelle 1 ersichtlich wird, weisen Gemeinden mit weniger als 10 000 Einwohnern deutlich geringere Karteileichen- und Fehlbestandsraten auf als dies bei größeren Gemeinden – vor allem aber bei Großstädten – der Fall ist.

Haushaltsstichprobe zur Qualitätssicherung

Der Zensustest hat letztlich gezeigt, dass die Melderegister grundsätzlich zur Ermittlung der amtlichen Einwohnerzahlen und demographischer Basisdaten in einem Zensus geeignet sind. Er hat aber auch verdeutlicht, dass Maßnahmen zur Qualitätssicherung unumgänglich sind, um die erforderliche Qualität der Ergebnisse zu erreichen. Als wichtigste Maßnahmen für den Zensus 2011 hat der Gesetzgeber in § 7 Gesetz über den registergestützten Zensus im Jahre 2011 (Zensusgesetz 2011 – ZensG 2011) die Durchführung einer primärstatistischen Haushaltebefragung auf Stichprobenbasis in Gemeinden mit mindestens 10 000 Einwohnern im Umfang von knapp 10 % der Gesamtbevölkerung angeordnet. Hauptziel dieser Haushaltebefragung auf Stichprobenbasis ist es, den Umfang der Karteileichen und Fehlbestände im Melderegister je Gemeinde zu schätzen und darauf basierend die Zahl der im Melderegister verzeichneten Personen zu korrigieren. Damit folgt der deutsche Zensus methodisch Vorbildern aus verschiedenen Ländern, u. a. auch Israel.[1]

Problemstellung: Umsetzung der ermittelten Registerfehler in den Einzeldatensatz

Zur Gewinnung eines qualitativ hochwertigen Einzeldatenbestandes sind je Gemeinde die aus den Melderegistern gewonnenen Datensätze um die Ergebnisse zu den Registerfehlern aus der Stichprobe zu korrigieren. Diese Umsetzung der Stichprobenergebnisse kann letztlich nur dadurch erfolgen, dass die Zahl der Personendatensätze im Melderegisterbestand einer Gemeinde durch Löschungen bzw. Imputationen um den aus der Stichprobe geschätzten Wert an Karteileichen und Fehlbeständen reduziert bzw. erhöht wird. Im Ergebnis dieser Maßnahmen entspricht dann die Zahl der Personendatensätze einer Gemeinde der korrigierten Einwohnerzahl. Hierbei ergeben sich jedoch vier Probleme:

- Nach den Ergebnissen des Zensustests weisen Karteileichen und Fehlbestände eine signifikant andere demographische und haushaltsstatistische Struktur auf als die Grundgesamtheit der Bevölkerung einer Gemeinde. Ein rein zufälliges Löschen oder Hinzufügen von Personendatensätzen in den Melderegisterdaten würde implizit unterstellen, dass die Verteilung der Merkmale bei den Karteileichen bzw. Fehlbeständen der Verteilung der Grundgesamtheit entspräche. Ein solches Vorgehen würde zwar zu einer korrekten amtlichen Einwohnerzahl führen, hätte aber je nach Umfang der Registerfehler eine mehr oder weniger starke Ver-

1 Weitere Informationen können in dem Artikel „The 2008 Israel Integrated Census of Population and Housing – Basic conception and procedure“ von Charles S. Kamen nachgelesen werden (veröffentlicht unter www.cbs.gov.il/mifkad/census2008_e.pdf).

Tab. 1 **Karteileichen- und Fehlbestandsraten aus dem Zensustest 2001**

Bundesland bzw. Gemeindegrößenklasse	Bevölkerung am Ort der Hauptwohnung im Zensustest 2001				
	Personen im Melderegister	Karteileichen		Fehlbestände	
	1 000	1 000	%	1 000	%
Bayern	11 957,5	307,9	*2,6*	211,6	*1,8*
Gemeinden mit Einwohnern von … bis unter …					
unter 10 000	22 947,5	459,5	*2,0*	303,6	*1,3*
10 000 bis 50 000	26 112,7	643,4	*2,5*	384,4	*1,3*
50 000 bis 800 000	23 944,5	801,6	*3,4*	509,3	*2,1*
800 000 oder mehr	6 980,2	416,3	*6,0*	207,1	*3,0*
Deutschland	**79 984,9**	**2 320,8**	***2,9***	**1 368,4**	***1,7***

zerrung der demographischen und haushaltsstatistischen Ergebnisse zur Folge.

- Die Haushaltsstichprobe liefert nur eingeschränkte Informationen zu den Registerfehlern. So sind neben den bivariaten Verteilungen Geschlecht/Staatsangehörigkeit und Geschlecht/Alter von den weiteren Merkmalen nur die Randverteilungen der Merkmale mit eingeschränkten Ausprägungen (z. B. nur Altersklassen und keine Einzelaltersjahre) mit einem vertretbaren Stichprobenfehler ermittelbar. Für eine fachlich vollständig verzerrungsfreie Korrektur wäre aber die unbekannte vollständige multivariate Verteilung der Karteileichen und Fehlbestände erforderlich.
- Darüber hinaus treten Karteileichen und Fehlbestände nur in sehr seltenen Fällen in einem Haushaltszusammenhang auf. Es bedarf daher einer getrennten Korrektur der Karteileichen und der Fehlbestände durch Löschungen bzw. Imputationen.
- Ferner hätte eine rein durch Zufallsverfahren gesteuerte Korrektur die Entstehung unplausibler Haushaltsergebnisse zur Folge.

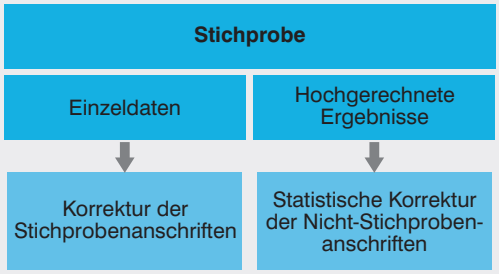
Zur Gewinnung qualitativ hochwertiger Zensusergebnisse benötigte man also ein sehr viel komplexeres Verfahrens als das bloße Löschen und Hinzufügen von Datensätzen. Im Zuge der Vorbereitung des Zensus 2011 hat das Bayerische Landesamt für Statistik und Datenverarbeitung ein Verfahren entwickelt, das eine weitgehend verzerrungsfreie Korrektur ermöglicht. Die Grundzüge dieses Verfahrens werden im Folgenden vorgestellt.

2. Modell und Ablauf des Korrekturverfahrens

Eine optimale Lösung im Sinne völlig verzerrungsfreier demographischer und haushaltsstatistischer Ergebnisse ist nur bei vollständiger Information über alle Karteileichen und Fehlbestände gegeben. Im Modell des Zensus ist diese Information aber nur für die im Rahmen der Haushaltsstichprobe primärstatistisch erhobenen knapp 10 % der Anschriften in Gemeinden mit mindestens 10 000 Einwohnern verfügbar und wird auch unmittelbar genutzt. Dies bedeutet, dass hier eine anschriftenscharfe Korrektur der festgestellten Karteileichen und Fehlbestände stattfindet (vgl. Abbildung 1).

Grundstruktur der Korrektur der Registerfehler Abb. 1

Korrektur der Melderegisterdaten anhand der Ergebnisse der Stichprobe



Die Korrektur bewirkt die Herstellung vollzähliger Zensuseinzeldatensätze, die flexible Auswertungen ermöglichen.

Bei den verbleibenden rund 90 % der Anschriften in Gemeinden mit mindestens 10 000 Einwohnern sind die tatsächlichen Registerfehler im Sinne von Einzelfällen unbekannt. Bekannt sind lediglich die aus der Haushaltsstichprobe (geschätzte) Summe der Fälle sowie deren Randverteilungen zu demographischen und haushaltsstatistischen Merkmalen. Aufgrund dieser unvollständigen Information ist es weder möglich eine buchhalterisch betrachtete „richtige“ Korrektur der Einzeldaten vorzunehmen, noch eine statistisch „optimale“ Korrektur durchzuführen, da hierfür die Kenntnis der vollständigen multivariaten Verteilung aller Merkmale erforderlich wäre. Unter diesen Prämissen ist eine statistisch hinreichende Korrektur der Registerfehler dann gegeben, wenn die Randverteilungen der korrigierten, also der gelöschten bzw. imputierten Einzeldaten den aus der Stichprobe geschätzten Randverteilungen dieser Merkmale entsprechen.

Exkurs Ranking

Eine wichtige Hilfsgröße bei der näherungsweisen Bestimmung der Verteilung von Karteileichen bildet das aus der Haushaltegenerierung gewonnene Merkmal „Ranking“. In der Haushaltegenerierung werden in der sog. Phase A – vereinfacht dargestellt – Haushalte anhand von Verzeigerungen des Melderegisters (Nachweise von Ehepaaren und Kindern) sowie sog. harten Generierungskriterien gebildet und über die in der Gebäude- und Wohnungszählung (GWZ) erhobenen Namen von Bewohnern

von Wohnungen mit den Wohnungsdaten verknüpft.[2] Nach dieser Phase der Haushaltegenerierung ist das Merkmal Ranking bei jeder Person gefüllt und weist eine der folgenden Ausprägungen auf:

1. Person wurde über die Wohnungsnutzerangaben mit der Wohnung (Modul 2 der Phase A) verknüpft.
2. Person wurde über eine andere Person mit der Wohnung (Modul 4 der Phase A) verknüpft.
3. Unverknüpfte Person mit deutscher Staatsangehörigkeit.
4. Unverknüpfte Person mit ausländischer Staatsangehörigkeit.

Verknüpft bzw. unverknüpft zeigt dabei an, ob eine Person bereits mit einer Wohnung zusammengeführt werden konnte (verknüpft mit einer Wohnung) oder nicht (unverknüpft).

Nach den Erkenntnissen des Zensustests weist das Ranking, also der Status nach Phase A „verknüpft" oder „unverknüpft", in Hinblick auf das Vorkommen von Karteileichen bemerkenswerte Unterschiede auf. So waren im Zensustest nur etwa 1 % der verknüpften Personen Karteileichen, während bei den unverknüpften Personen insgesamt rund 17 % und bei den unverknüpften ausländischen Personen sogar etwa 33 % Karteileichen waren. Aufgrund dieser erheblichen Unterschiede kann man sich bei Kenntnis dieser Werte der vollständigen multivariaten Verteilung von Karteileichen sehr viel besser annähern als bei alleiniger Kenntnis der Randverteilungen der rein demographischen Merkmale. Aus diesem Grund erfolgt im Rahmen der Haushaltsstichprobe neben der Schätzung der demographischen Struktur der Karteileichen in einer Gemeinde auch die Schätzung des Rankings.

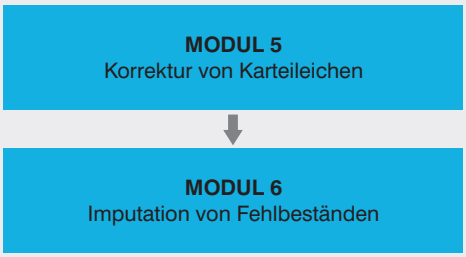
Getrennte Behandlung von Karteileichen und Fehlbeständen

Die Ergebnisse des Zensustests haben auch gezeigt, dass Karteileichen und Fehlbestände sehr unterschiedliche demographische Strukturen aufweisen. So sind Fehlbestände im Durchschnitt deutlich jünger als Karteileichen, ein Indiz für eine mobile Bevölkerungsgruppe. Demgegenüber wurden im Zensustest Karteileichen in der Gruppe der über 60-Jährigen nachgewiesen (z. B. Personen, die in Altenheimen untergebracht sind, aber noch bei ihren Nachkommen gemeldet sind), während es kaum Fälle gab, bei denen über 60-Jährige an Anschriften lebten, an denen sie nicht gemeldet waren.

Aufgrund dieser signifikanten Verteilungsunterschiede würde eine saldierte Korrektur von Karteileichen und Fehlbeständen zu erheblichen demographischen Verzerrungen führen. Es bedarf also eines zweigleisigen Verfahrensansatzes. In Abbildung 2 kann man die beiden Komponenten des Korrekturverfahrens erkennen.

Zweigleisige Korrektur der Karteileichen und Fehlbestände Abb. 2

Korrekturverfahren für Gemeinden mit 10 000 oder mehr Einwohnern



Um eine Löschung/Imputation einzelner Personendatensätze so zu realisieren, dass die Randverteilungen der Gesamtzahl der Löschungen/Imputationen den aus der Stichprobe geschätzten Randverteilungen entsprechen, bedarf es der Kenntnis der vollständigen multivariaten Verteilung. Da diese, wie eingangs erwähnt, nicht vorliegt, bedarf es vor der eigentlichen Korrektur der Schätzung der multivariaten Verteilung mittels eines Näherungsverfahrens.

Approximation der multivariaten Verteilung

Das hierzu verwendete Verfahren lehnt sich an die aus dem Operations Research bekannte Monte-Carlo-Methode an, die auf einer Zufallsauswahl basiert. Ausgangspunkt bildet zunächst die Annahme, dass die demographischen Merkmale statistisch unabhängig und somit die Wahrscheinlichkeiten multiplikativ verknüpfbar sind. Dies erscheint zunächst nicht sinnvoll, da Fälle entstehen, die zwar rechnerisch eine Wahrscheinlichkeit größer Null aufweisen, real aber nicht existieren. So sei beispielsweise die Wahrscheinlichkeit für die Altersklasse unter sechs Jah-

2 Weitere Informationen zu diesem Verfahren können in dem Artikel von Ingrid Kreuzmair und Marco Reisch „Zensus 2011: Ablauf der Haushaltegenerierung" in Bayern in Zahlen 9/2012 nachgelesen werden.

re gleich p1 und die Wahrscheinlichkeit für verwitwet p2. Dann ergäbe sich bei Unabhängigkeit die positive Wahrscheinlichkeit p1 x p2, obgleich verwitwete Kinder unter sechs Jahren real nicht vorkommen. Um diese Unzulänglichkeit der Unabhängigkeitsannahme auszugleichen, wird in dem iterativen Prozess jede Merkmalskombination dahingehend überprüft, ob es in der Grundgesamtheit eine Person gibt, die diese Merkmalskombination aufweist. Letztlich können nur Personendatensätze gelöscht oder gedoppelt werden, die in der Realität auch existieren. Auf diese Weise werden die unbekannten Kovarianzen zwischen den Merkmalen näherungsweise modelliert. Dieser Verfahrensteil wird in Kapitel 3 näher erläutert.

Löschung der Karteileichen/Doppelung der zu imputierenden Fehlbestände

Bei der eigentlichen Löschung der Karteileichen werden die Karteileichen anhand der Ergebnisse der Approximation und des haushaltsstatistischen Anpassungsrahmens durch ein iteratives Verfahren statistisch ausfindig gemacht und gelöscht. Methodisch analog dazu erfolgt die Doppelung der zu imputierenden Datensätze. Hierbei werden real in dem Datensatz existierende Personen gedoppelt und in einem späteren Verfahrensschritt an eine bestehende Anschrift in der Gemeinde imputiert.

Im Gegensatz zur Löschung der Karteileichen ist es bei der Korrektur der Fehlbestände notwendig, wieder eine geeignete Wohnung bzw. einen geeigneten Teilhaushalt im Datenbestand ausfindig zu machen.

Bei der Allokation der reinen Fehlbestandshaushalte (ganze Haushalte werden in noch nicht belegte Wohnungen imputiert) werden neben der Wohnungsgröße Anschrifteninformationen genutzt, um eine möglichst genaue Zuordnung zu Wohnungen möglich zu machen. Hierbei wird vor allem die Relation von momentan vorhandenen Wohnungen zu momentan vorhandenen Haushalten pro Anschrift betrachtet, um potentielle Anschriften für die Imputation zu identifizieren.

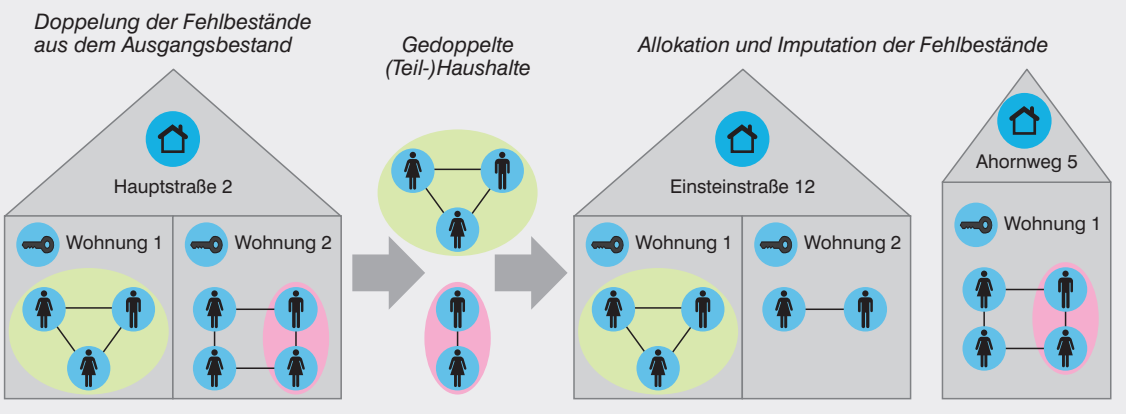
Werden gemischte Fehlbestandshaushalte imputiert (einzelne Personen bzw. Teilhaushalte), so werden diese mit bestehenden Haushalten zusammengeführt. Dabei wird versucht, möglichst die Informationen des „Herkunftshaushaltes“ zu nutzen, um eine möglichst gute Nachbildung zu erlangen.

In Abbildung 3 wurden an der Anschrift „Hauptstraße 2“ zwei (Teil)Haushalte gedoppelt:

- Ein kompletter Dreipersonenhaushalt
- Zwei Personen aus einem Vierpersonenhaushalt

Die Haushalte an der Anschrift „Hauptstraße 2“ bleiben erhalten. Für die beiden duplizierten Haushalte werden nun geeignete Anschriften gesucht und anschließend erfolgt die Imputation. In diesem Beispiel wird der komplette Dreipersonenhaushalt in eine noch nicht durch einen Haushalt belegte, jedoch in der GWZ als bewohnt gemeldete Wohnung imputiert. Die zwei Personen aus dem Vierpersonenhaushalt werden an einem bestehenden Zweipersonenhaus-

Doppelung und Allokation der Fehlbestandshaushalte Abb. 3



halt im Ahornweg 5 angefügt. Damit wird die Struktur des Herkunftshaushaltes nachgebildet.

Im Folgenden wird der Schwerpunkt des Artikels auf die Methodik der Approximation gelegt.

3. Darstellung des Verfahrensablaufs der Approximation der multivariaten Verteilungen an einem Beispiel

3.1 Ausgangsdaten aus der Haushaltsstichprobe

Zur verständlicheren Darstellung des Verfahrens wird eine fiktive Gemeinde („Beispielgemeinde“) gewählt. Die Daten der Beispielgemeinde sind in Tabelle 2 dargestellt. Die Darstellung des Verfahrens der Approximation erfolgt anhand der Hauptwohnungs-Karteileichen.

3.2 Approximation

Definition der Merkmalsklassen

Die Approximation hat die Aufgabe, aus den in Tabelle 2 dargestellten uni- und bivariaten Merkmalen die multivariaten Merkmale zu ermitteln. Betrachtet man in unserem Beispiel die Ausprägungen der fünf demographischen Merkmale und des Merkmals Ranking, so besteht die vollständige multivariate Verteilung unter der Annahme der Unabhängigkeit theoretisch aus

Tab. 2 **Demographischer Anpassungsrahmen und Ranking in der Beispielgemeinde**

Merkmale	Bevölkerung am Ort der Hauptwohnung aus dem Melderegister	In der Stichprobe ermittelte Karteileichen am Ort der Hauptwohnung
	Anzahl	
Personen insgesamt	**29 461**	**1 813**
Geschlecht/Staatsangehörigkeit		
Männer		
deutsch	11 632	626
nicht-deutsch	2 772	431
Frauen		
deutsch	12 706	509
nicht-deutsch	2 351	247
Familienstand		
ledig bzw. unbekannt	12 809	1 030
verheiratet bzw. Lebenspartnerschaft	12 435	542
verwitwet bzw. Lebenspartner verstorben	1 772	85
geschieden bzw. Lebenspartnerschaft aufgehoben	2445	156
Geschlecht/Alter		
Männer		
unter 6 Jahre	1 550	50
6 bis unter 18 Jahre	941	71
18 bis unter 25 Jahre	1 372	140
25 bis unter 30 Jahre	1 195	170
30 bis unter 40 Jahre	2 748	281
40 bis unter 50 Jahre	2 058	131
50 bis unter 60 Jahre	1 322	76
60 bis unter 65 Jahre	1 562	80
65 Jahre oder älter	1 656	58
Frauen		
unter 6 Jahre	1 111	41
6 bis unter 18 Jahre	1 285	60
18 bis unter 25 Jahre	1 464	150
25 bis unter 30 Jahre	1 126	105
30 bis unter 40 Jahre	2 369	148
40 bis unter 50 Jahre	2 072	70
50 bis unter 60 Jahre	1 453	44
60 bis unter 65 Jahre	1 466	45
65 Jahre oder älter	2 711	93
Erwerbstätigkeit		
sozialversicherungspflichtig Beschäftigte	18 231	650
Beamte, Richter und Soldaten	2 850	30
Arbeitslose und Personen in Umschulung	2 625	750
sonstige Personen	5 755	383
Ranking		
vor Modul 4 verknüpft	24 440	536
in Modul 4 verknüpft	1 286	173
unverknüpfte Deutsche	2 598	639
unverknüpfte Nicht-Deutsche	1 137	465

2 (Geschlecht) x 4 (Familienstand) x 9 (Alter)
x 2 (Staatsangehörigkeit) x 4 (Erwerbstätigkeit)
x 4 (Ranking) = 2 304 Werten.

Diese Werte werden nachfolgend als Klassen bezeichnet. Jede Klasse lässt sich numerisch als sechsstellige Zahlenkombination darstellen. Die Klasse 113111 bei Hauptwohnsitz bedeutet zum Beispiel männlich, ledig, 18 bis unter 25 Jahre, deutsch, sozialversicherungspflichtig Beschäftigter, vor Modul 4 verknüpft. Auf diese Weise lässt sich jeder Personensatz in den Registerdaten durch eine Merkmalskombination eindeutig charakterisieren.

Tatsächlich ist die Zahl der Klassen deutlich geringer. Zum einen, weil sich bestimmte Kombinationen ausschließen (z. B. Staatsangehörigkeit deutsch und unverknüpfter Nicht-Deutscher), zum anderen, weil bestimmte Kombinationen extrem selten sind und ggf. in dem jeweiligen Datenbestand gar nicht vorkommen (z. B. verwitwete Person zwischen 6 und unter 18 Jahren).

Verfahrensablauf

Schritt 1:

Für die Registerdaten in der Beispielgemeinde werden für die fünf demographischen Merkmale und das Merkmal Ranking die Häufigkeiten aller Klassen ermittelt. Die Anzahl der für jede Klasse im Register festgestellten Personen (Grundgesamtheit) bilden im weiteren Verfahrensablauf Grenzwerte, da nur maximal so viele Personen gelöscht werden können, wie in der jeweiligen Grundgesamtheit vorkommen.

Schritt 2:

Es soll nun zufällig eine der Klassen gezogen werden. Ausschlaggebend hierbei ist, dass die Wahrscheinlichkeit für die Ziehung einer Klasse nicht der Häufigkeit in der Grundgesamtheit, sondern der Häufigkeit in den Karteileichen entsprechen soll. Da diese allerdings nicht bekannt ist, wird davon ausgegangen, dass die aus der Stichprobe bekannten uni- oder bivariaten Verteilungen der Merkmalsausprägungen voneinander statistisch unabhängig und folglich multiplikativ verknüpfbar sind.

Ausgehend von dieser Annahme ist es nun möglich, für jedes Merkmal bzw. jede Merkmalskombination, für das bzw. die aus der Stichprobe Informationen vorliegen, unabhängig voneinander eine entsprechende Zufallsauswahl zu treffen.

Schritt 2.1: Berechnung der Ziehungswahrscheinlichkeit

Es seien nun GG die Grundgesamtheit und M_1, ..., M_5 die in Tabelle 2 aufgeführten Merkmale bzw. Merkmalskombinationen für Hauptwohnungspersonen. Mit m_{ij} sei die Anzahl der Einheiten in der Grundgesamtheit und als k_{ij} die zu löschenden Einheiten (Anzahl der Karteileichen) eines Merkmals i mit der Ausprägung j bezeichnet. Die Gesamtheit aller zu löschenden Sätze (Karteileichen) wird mit KL bezeichnet.

Bei zufälligem (gleichverteilten) Ziehen in der Grundgesamtheit beläuft sich die relative Häufigkeit $h_{ij}(GG)$ einer zu ziehenden Merkmalsausprägung auf:

$$h_{ij}(GG) = m_{ij}/GG$$

Für die relative Häufigkeit einer Merkmalsausprägung in den Karteileichen $h_{ij}(KL)$ gilt:

$$h_{ij}(KL) = k_{ij}/KL$$

Somit gilt für den Anpassungsfaktor a_{ij}, der angibt, um wie viel häufiger (oder auch seltener) als in der Grundgesamtheit vorhanden eine bestimmte Merkmalsausprägung ausgewählt werden soll:

$$a_{ij} = h_{ij}(KL)/h_{ij}(GG) = (k_{ij}/KL)/(m_{ij}/GG) = (k_{ij}/m_{ij}) * (GG/KL)$$

D. h. der Anpassungsfaktor ergibt sich aus der merkmalsspezifischen Karteileichenrate einer Merkmalsausprägung multipliziert mit dem Quotienten aus Grundgesamtheit und Karteileichenzahl.

Für die Merkmalsausprägung m_{11} „Männer, deutsch“ der Merkmalskombination „Geschlecht/Staatsangehörigkeit“ ergibt sich in unserem Beispiel:

$$h_{11}(GG) = 11\,632/29\,461 = 0{,}39$$
$$h_{11}(KL) = 626/1\,813 = 0{,}35$$
$$a_{11} = h_{11}(KL)/h_{11}(GG) = 0{,}87$$

Deutsche Männer sind also 0,87 mal so oft (und damit um den Faktor 0,13 seltener) auszuwählen, als es ihrem Anteil in der Grundgesamtheit entspricht.

Besonders prägnant ist der Anpassungsfaktor bei der Merkmalsausprägung m_{44} „unverknüpfte Nicht-Deutsche“:

$$h_{44}\ (GG) = 1\ 137/29\ 461 = 0{,}04$$
$$h_{44}(KL) = 465/1\ 813 = 0{,}26$$
$$a_{44} = h_{44}(KL)/h_{44}(GG) = 6{,}65$$

Unverknüpfte Nicht-Deutsche sind demnach um den Faktor 6,65 und damit häufiger auszuwählen, als es ihrem Anteil in der Grundgesamtheit entspricht.

Schritt 2.2: Zufallsziehung

Sind aus der Stichprobe nur die Randverteilungen bekannt, so kann nun für jedes Merkmal einzeln eine Zufallsziehung der Merkmalsausprägung vorgenommen werden. Hierzu werden die Anpassungsfaktoren a_{ij} für alle Ausprägungen j eines Merkmals i errechnet. Zur Erläuterung soll das Merkmal Familienstand herangezogen werden. In Tabelle 3 sind Beispieldaten für den Familienstand aufgeführt.

Tab. 3 **Anpassungsfaktoren für die Ausprägungen des Familienstands der Beispieldaten**

Familienstand	Anzahl Karteileichen	Grundgesamtheit	Anpassungsfaktor
	1	2	3
Ledig	1 030	12 809	1,31
Verheiratet	542	12 435	0,71
Verwitwet	85	1 772	0,78
Geschieden	156	2 445	1,04
Insgesamt	**1 813**	**29 461**	

Ebenfalls in Tabelle 3 sind die Daten für die Grundgesamtheit und die Karteileichen nach den Familienständen aufgelistet. In Spalte 3 sind die Faktoren enthalten, die angeben, um wieviel mal häufiger oder geringer ein Familienstand als Karteileiche auftritt. Die Zahlen wurden nach der oben angeführten Formel berechnet.

Sind, wie im oben angeführten Beispiel, auch bivariate Verteilungen der Karteileichen bekannt, erfolgt die Ziehung sukzessive, d. h. es wird zunächst aus der Merkmalskombination Geschlecht/Staatsangehörigkeit eine Ausprägung entsprechend der errechneten Verteilung zufällig gezogen und damit zwei der Klassenziffern bestimmt. Das Ziehungsergebnis determiniert, ob aus der Kombination Männer/Alter oder Frauen/Alter die nächste Zufallsziehung vorgenommen wird.

Sind für alle Merkmale/Merkmalskombinationen anhand der Anpassungsfaktoren die Ziehungen durchgeführt, ist die potenziell in Frage kommende Klasse bestimmt.

Schritt 2.3: Prüfung auf Zulässigkeit

Nach der Zufallsziehung einer Klasse ist die ausgewählte Klasse hinsichtlich ihrer Zulässigkeit zu prüfen. Diese Prüfung enthält u. a. die Kontrolle, ob die gezogene Klasse unter Berücksichtigung der bereits gezogenen Fallzahlen in der Grundgesamtheit überhaupt existiert, sowie die Kontrolle, ob die Zahl der ausgewählten Einheiten mit der Ausprägung „ledig“ des Merkmals Familienstand die Zahl der ausgewählten Personen unter 18 Jahren nicht unterschreitet. Diese Einschränkung hat sich als notwendig erwiesen, weil die unter 18-Jährigen nahezu alle ledig sind und aufgrund der Auswahl zu vieler Lediger über 17 Jahre die Anzahl der zu löschenden unter 18-Jährigen nicht mehr erreicht werden kann.

Ist eine Klasse nicht gültig, erfolgt eine neue Zufallsauswahl. Durch diese einschränkenden Bedingungen werden – wie bereits erwähnt – die Kovarianzen näherungsweise in dem Modell berücksichtigt.

Schritt 2.4: Neuberechnung der Auswahlwahrscheinlichkeiten

Die Auswahl einer Klasse wird als potenzielle Löschung einer Person aus den Registerdaten betrachtet und damit reduziert sich für die ausgewählten Merkmalsausprägungen sowohl die Zahl der zu löschenden Einheiten als auch die Zahl der jeweiligen

Tab. 4 **Auszugsweises Ergebnis einer Approximation der Klassenbesetzungen für Hauptwohnsitzkarteileichen**

Klasse	Grundgesamtheit Registerdaten	Approximierte Karteileichen Besetzung der Klassen = zu löschende Personen in dieser Klasse
	1	2
113221	30	10
115211	34	6
126222	12	1
126242	69	28
144241	12	12
147133	13	5
223132	23	14
237131	99	4

Einheiten in der Grundgesamtheit um jeweils Eins. Aufgrund dieses dynamischen Effekts (Ziehen ohne Zurücklegen) muss nach jeder Auswahl einer Klasse für den erneuten Ziehungsvorgang eine Neuberechnung der Anpassungsfaktoren stattfinden.

Diese Schritte werden solange durchgeführt, bis für alle Merkmalsausprägungen die Zahl der zu löschenden Einheiten erfüllt ist. In Tabelle 4 wird auszugsweise das Ergebnis einer Approximation gezeigt. Die sechsstellige Zahlenkombination beschreibt die jeweilige Klasse.

In der Klasse 113221 existieren demnach 30 Personen in der Grundgesamtheit. Die approximierte Besetzung dieser Klasse beläuft sich auf zehn Personen; damit müssen letztlich zehn Personen dieser Klasse aus dem Datensatz gelöscht werden.

4. Zusammenfassung und Bewertung des Verfahrens

Mit dem vorliegenden Verfahren, das eher als Heuristik bezeichnet werden kann, ist es im Rahmen des Zensus möglich, die aus der Haushaltsstichprobe geschätzten Umfänge der Registerfehler in den Einzeldatenbestand des Zensus zu integrieren, um so einen fachlich und regional in beliebiger Tiefe auswertbaren Zensuseinzeldatenbestand zu erhalten. Das Verfahren gewährleistet hierbei, dass die aus Melderegister und Stichprobe ermittelte Einwohnerzahl unverändert bleibt.

In Hinblick auf die Güte des Verfahrens, im Sinne von Abweichungen der aus dem erzeugten Zensusdatenbestand gewonnenen demographischen und haushaltsstrukturellen Ergebnisse zu den „wahren“ demographischen und haushaltsstrukturellen Ergebnissen einer Gemeinde, ist Folgendes zu bemerken.

Der Gesamtfehler setzt sich aus zwei Komponenten zusammen: Die erste Komponente ist der Stichprobenfehler aus den geschätzten Strukturdaten der Registerfehler in einer Gemeinde, an das der Einzeldatenbestand angepasst wird. Dieser ist letztlich durch das Zensusmodell bedingt und nicht durch das Verfahren per se verursacht. Die zweite Fehlerkomponente, der eigentliche Verfahrensfehler, resultiert aus der nur näherungsweise ermittelbaren unbekannten vollständigen multivariaten Verteilung der Registerfehler.

Während der Stichprobenfehler durch entsprechende Fehlerrechnungen quantifizierbar ist, ist eine analytische Quantifizierung des Verfahrensfehlers nicht möglich. Beide Fehlerkomponenten sind aber nicht unabhängig. Bei Gemeinden mit vergleichsweise großen Registerfehlern steigt – bedingt durch die höheren Fallzahlen in der Stichprobe – die Qualität der Stichprobenergebnisse zu den Registerfehlern. Demgegenüber verursacht das höhere Lösch- bzw. Imputationsvolumen zwangsläufig höhere strukturelle Abweichungen zur realen demographischen Struktur. Kurz gesagt: je höher der Registerfehler, desto kleiner der Stichproben- und desto größer der Verfahrensfehler und vice versa.

Anhand des Zensustestdatenmaterials wurde eine Reihe von empirischen Untersuchungen zur Güte des Verfahrens vorgenommen. Im Ergebnis hat sich gezeigt, dass die Auswirkungen auf die demographischen Ergebnisse einer Gemeinde insgesamt gering ausfallen. Nennenswerte relative Abweichungen waren erwartungsgemäß nur bei schwach besetzten Tabellenfeldern, wie z. B. verwitweten Ausländern unter 65 Jahre festzustellen. Etwas stärker fielen die Abweichungen bei den haushaltsstrukturellen Ergebnissen ins Gewicht.

Die Haushaltegenerierung einschließlich des Korrekturverfahrens konnten Ende des Jahres 2013 erfolgreich abgeschlossen werden. Anschließend erfolgt nun eine intensive Evaluationsphase, in der die Verfahren in Hinblick auf eine Verwendung im Zensus 2021 geprüft und weiterentwickelt werden müssen.

Literaturverzeichnis:

Hillier, Frederick; Lieberman, Gerald (1996), Operations Research. Einführung. 5. Auflage, München.

Kreuzmair, Ingrid; Reisch, Marco (2012), Ablauf der Haushaltegenerierung. In: Bayern in Zahlen, Ausgabe 9/2012. S. 615-624.

Kamen, Charles (2005), The 2008 Israel Integrated Census of Population and Housing – Basic conception and procedure. www.cbs.gov.il/mifkad/census2008_e.pdf (28.02.2014).