DAS HOCHRECHNUNGSVERFAHREN FÜR ZUSATZMERKMALE BEIM ZENSUS 2011

Dr. Andreas Berg, Wolf Bihler

↘ **Schlüsselwörter:** Zensus – Zusatzmerkmale – Raking – Baukastenprinzip

ZUSAMMENFASSUNG

Als ein zentrales Ergebnis des Zensus 2011 wurden die Einwohnerzahlen in demografischen Untergliederungen zeitnah zum Zensusstichtag veröffentlicht und den Nutzerinnen und Nutzern zugänglich gemacht.

Darüber hinaus wurden auch Zusatzmerkmale aus den Bereichen Bildung, Erwerbstätigkeit, Religion und Migration hochgerechnet. Diese Hochrechnung erfolgte erst nach Abschluss umfangreicher Plausibilitätsprüfungen und statistischer Korrekturmaßnahmen auf Basis eines dann konsolidierten Datenbestandes. Die Veröffentlichung dieser mithilfe von Hochrechnungsfaktoren erstellten Ergebnisse fand dadurch deutlich später statt: 36 Monate nach dem Zensusstichtag.

Auch das Hochrechnungsverfahren wurde für die Berechnung der Zusatzmerkmale modifiziert: Zum Einsatz kam ein iteratives Kalibrierungsverfahren, das auf einer proportionalen Anpassung der Stichprobenwerte an vorgegebene Bezugsmerkmale beruht.

↘ ***Keywords:*** *census – additional variables – raking – modular principle*

ABSTRACT

As a major result of the 2011 Census, the numbers of inhabitants, broken down by demographic variables, were released and made available to users shortly after the census reference date.

Also, additional variables regarding education, employment, religion and migration were estimated. Estimation was done after comprehensive plausibility checks and statistical corrections had been finished, so that it was based on a consolidated data set. Consequently those results, obtained by weighting, were released much later, that is, 36 months after the census reference date.

Also, the estimation method was modified for the calculation of the additional variables. An iterative calibration method was used that is based on a proportional adjustment of sample values to given reference variables.

Dr. Andreas Berg

ist Diplom-Ökonometriker und promovierter Statistiker und seit 2005 in der Gruppe „Mathematisch-statistische Methoden, Forschungsdatenzentrum“ des Statistischen Bundesamtes tätig. Er beschäftigt sich schwerpunktmäßig mit der methodischen Ausarbeitung und Programmierung von Stichprobenziehungen und Hochrechnungen und ist mit vorbereitenden Arbeiten zum Zensus 2021 betraut.

Wolf Bihler

ist Diplom-Mathematiker und arbeitet seit 1985 im Statistischen Bundesamt. Er leitet das Referat „Mathematisch-statistische Verfahren für Bevölkerung, Finanzen, Steuern; Wahlen“ und befasst sich insbesondere mit stichprobenmethodischen Fragestellungen bei Haushaltserhebungen.

1

Einleitung

Es war ein wichtiges Ziel des Zensus 2011, Merkmale der Bevölkerung, die nicht in Registern vorhanden sind (sogenannte Zusatzmerkmale), zu erheben und nachzuweisen („Ziel 2“)[1]. Dabei handelt es sich um Fragen zur Erwerbstätigkeit, Bildungsstand, Migration und Religion. Diese Fragen wurden nicht durch eine Vollerhebung erhoben, sondern auf Basis der sogenannten Haushaltsstichprobe[2]. Wie bei jeder Stichprobe stellt sich die Frage, wie man von der Stichprobe auf die Grundgesamtheit schließt. Darum soll es im Folgenden gehen. Dargestellt wird das Hochrechnungsverfahren für endgültige Ergebnisse zum Veröffentlichungstermin 2 (Mai 2014).

Zum Veröffentlichungstermin 2 konnten die Zensusergebnisse einerseits im Vergleich zum Termin der Veröffentlichung der Einwohnerzahlen und erster untergliederter Ergebnisse (Veröffentlichungstermin 1, Mai 2013) qualitativ verbessert werden: Der aus den Einwohnermelderegistern stammende Datenbestand war durch diverse statistikinterne Korrekturschritte verändert worden (Hofmeister/Fürnrohr, 2014). Die veröffentlichten Einwohnerzahlen blieben davon jedoch unberührt. Den resultierenden Bestand der Grundgesamtheit aller für den Zensus zu zählenden Personen bezeichnen wir im Folgenden mit dem Begriff „zensustypischer Datensatz“. Dieser stand zum ersten Veröffentlichungstermin noch nicht zur Verfügung. Andererseits konnten erst zu diesem zweiten Termin die durch akribische Plausibilisierungsarbeiten aufbereiteten Ergebnisse bezüglich der Zusatzmerkmale präsentiert werden.

Von den Ergebnissen, bei denen Zusatzmerkmale einbezogen sind, sind Ergebnisse und Analysen zu unterscheiden, zu denen ausschließlich demografische Merkmale[3] beitragen. Diese werden nicht aus der Stichprobe hochgerechnet, sondern aus dem zensustypischen Datensatz ausgezählt.

Ziel ist es, Hochrechnungsfaktoren für die Zusatzmerkmale zu erstellen. Die eigentliche Erstellung hochgerechneter Ergebnisse für die Zusatzmerkmale erfolgte anhand dieser Hochrechnungsfaktoren in der Auswertungsdatenbank. Für jede Hochrechnung – mit Ausnahme der freien Hochrechnung, bei der nur die Ziehungswahrscheinlichkeiten der Stichprobenpersonen benötigt werden – benötigt man Bezugsmerkmale[4], für die Totalwerte der Grundgesamtheit (Eckwerte) bekannt sein müssen. Während für die Hochrechnung vorläufiger, erster Ergebnisse (zum Veröffentlichungstermin 1) als Bezugsmerkmale die gemeldeten Personen mit ausgewählten Ausprägungen demografischer Merkmale herangezogen wurden, fungieren für das hier vorliegende Konzept demografische Merkmale der Personen des zensustypischen Datensatzes als Bezugsmerkmale.

Die Hochrechnungsfaktoren sollen so beschaffen sein, dass hieraus berechnete Ergebnisse mit ausgezählten Totalwerten bestimmter demografischer Untergliederungen aus dem zensustypischen Datensatz so gut wie möglich übereinstimmen. Dabei handelt es sich um ein Kalibrierungsproblem: Ausgehend von „ursprünglichen“ Hochrechnungsfaktoren (im Folgenden Eingangsfaktoren) – in unserem Fall die schon für die Ermittlung der Einwohnerzahl berechneten Faktoren (Berg/Bihler, 2011) – werden neue Faktoren berechnet. Diese unterscheiden sich möglichst wenig von den Eingangsfaktoren und genügen gleichzeitig der Bedingung, dass sich die vorgegebenen „Eckwerte“ ergeben, wenn die entsprechenden demografisch untergliederten Positionen mit diesen Faktoren hochgerechnet werden. Durch diesen Ansatz sollen bei den endgültigen Ergebnissen Inkohärenzen zwischen hochgerechneten Randsummen demografischer Merkmale und den Ergebnissen aus rein demografischen Tabellen möglichst vermieden beziehungsweise verringert werden.

Bei den erwerbsstatistischen Zusatzmerkmalen gibt es eine Besonderheit: Einige erwerbsstatistische Merkmale werden nur für einen Teil aus der Stichprobe hochgerechnet und der andere Teil wird aus erwerbsstatistischen Registern ausgezählt (sogenanntes Baukastenprinzip; siehe auch Statistische Ämter des Bundes und der Länder, 2015, hier: Seite 62 ff.).

1 Unter „Ziel 1“ wird beim Zensus 2011 die Ermittlung der Einwohnerzahl und demografisch untergliederter Ergebnisse verstanden.

2 Zum Design dieser Stichprobe siehe Berg/Bihler (2011).

3 Merkmale, die in den Melderegistern geführt werden, wie Alter, Geschlecht, Familienstand und Staatsangehörigkeit.

4 Wir verwenden hier nicht den in der Stichprobentheorie üblichen Begriff Hilfsmerkmal oder Hilfsvariable („auxiliary variable“), da in den zensusgesetzlichen Grundlagen der Begriff „Hilfsmerkmal“ anderweitig verwendet wird.

Die Hochrechnung erfolgte eingeschränkt auf die Bevölkerung am Hauptwohnsitz. Einheit der Hochrechnung ist die Person. Small-Area-Schätzungen wurden nicht eingesetzt, da es noch kein ausgereiftes und einsatzfähiges Verfahren gab, mit dem man Small-Area-Schätzer mit Hochrechnungsfaktoren darstellen kann.[5]

2

Die Zusatzmerkmale im Überblick

Im Zuge der Haushaltebefragung im Zensus 2011 wurden die in die Stichprobe gelangten Haushalte gerade auch zu den Zusatzmerkmalen interviewt. Es handelt sich hierbei um Fragen bezüglich Erwerbstätigkeit, Bildungsstand, Migration und Religion, die nicht oder nicht mit hinreichender Qualität aus bereits vorhandenen Registern für feine regionale Strukturen auswertbar sind.

Durch die – im Vergleich mit anderen Erhebungen – mit sehr hohem Stichprobenumfang ausgestattete Haushaltebefragung des Zensus 2011 kann eine qualitativ hochwertige Datenbasis zu diesen hochrelevanten gesellschaftlichen Fragestellungen angeboten werden. Nähere Erläuterungen zu den Zusatzmerkmalen – auch im Zusammenhang mit der Entwicklung des Haushaltefragebogens zum Zensus 2011 – liefert Gauckler (2011).

5 Eine Abkehr von Hochrechnungsfaktoren und dem damit verbundenen schätzmethodischen Paradigmenwechsel wurde für den Zensus 2011 von den zuständigen Gremien verworfen.

Die letztlich in einer Auswertungsdatenbank (https://ergebnisse.zensus2011.de) den Nutzerinnen und Nutzern zur Verfügung stehenden Auswertungsmerkmale werden aus dieser Haushaltebefragung abgeleitet und auf relativ kleinräumiger Basis bis hin zur Gemeindeebene (Gemeinden mit 10 000 oder mehr Einwohnerinnen und Einwohnern) mittels variabler Diagramme und Tabellen präsentiert. ↘ Übersicht 1

3

Ausgangssituation vor der Ziel-2-Hochrechnung

Für die Berechnung der Ziel-2-Hochrechnungsfaktoren wird – anders als beim Hochrechnungsverfahren bei der Ermittlung der Einwohnerzahl (Berg/Bihler, 2011; Ziel-1-Hochrechnung) – auf den zensustypischen Datensatz zurückgegriffen. Daneben wurden schon bekanntes Zahlenmaterial und Hilfsdateien aus der Ziel-1-Hochrechnung verwendet: ein Auszug aus dem Anschriften- und Gebäuderegister und ein Auszug aus der personenbezogenen Datei der Stichprobenanschriften (Hirner/Stiglmayr, 2013) sowie zusätzlich die Einwohnerzahldatei als Ergebnis der Ziel-1-Hochrechnung.

Der für die Ziel-2-Hochrechnung neu hinzugekommene zensustypische Datensatz wurde auf Kompatibilität mit dem restlichen Datenmaterial hin überprüft. Hinsichtlich einer einheitlichen Klassifikation und Kodierung der Merkmalsausprägungen wurden Umbenennungen

Übersicht 1
In der Auswertungsdatenbank enthaltene Zusatzmerkmale beim Zensus 2011

Kategorie	Merkmalsname
Religion	Religion (7 Ausprägungen)
Migration	Migrationshintergrund Migrationserfahrung Migration nach ausgewählten Ländern Migrationserfahrung nach Zuzugsjahrzehnt Migration nach Aufenthaltsdauer
Schul- und Berufsbildung	Klassenstufen und Schulform der Schüler/-innen einer allgemeinbildenden Schule Höchster Schulabschluss Höchster beruflicher Abschluss
Beruf	Erwerbsstatus Wirschaftszweig Stellung im Beruf

vorgenommen und die Anzahl der Datensätze mit der amtlichen Einwohnerzahl abgeglichen.

Da die Bezugsmerkmale in Stichprobe und Grundgesamtheit genau gleich definiert sein sollten, sind im Stichprobenmaterial die Merkmalsausprägungen aus dem zensustypischen Datensatz relevant und nicht die Ausprägungen laut Fragebogen.

Nach der Verknüpfung der Datensätze mit dem zensustypischen Datensatz als Basisdatei standen zu Beginn der Hochrechnungsarbeiten mehr als 79,5 Millionen Datensätze zur Verfügung.

4

Das Hochrechnungsverfahren im Detail

Im Gegensatz zu den Schätzungen bei der Ermittlung der Einwohnerzahlen geht es bei den Zusatzmerkmalen zuerst nicht darum, Parameterwerte wie Totalwerte oder Anteilswerte der interessierenden Merkmale zu schätzen. Vielmehr geht es um die Darstellung durch Hochrechnungsfaktoren, mit denen dann relativ einfach Schätzwerte durch geeignete Verknüpfung mit den gewünschten Merkmalskombinationen in einer Auswertungsdatenbank berechnet werden können.

Ausgangsfaktor für die Berechnungen war das schon bei der Hochrechnung der Einwohnerzahl verwendete Designgewicht, welches bereits um Antwortausfälle und Anschriftenzusammenfassungen bereinigt wurde (Berg/Bihler, 2014, hier: Seite 233 ff.).

Durch Modifikation der Gewichte soll eine möglichst genaue Kohärenz zwischen den Stichprobenergebnissen und Randverteilungen in der Grundgesamtheit bezüglich der Bezugsmerkmale hergestellt werden. Dazu wird ein in der Literatur unter verschiedenen Namen, wie beispielsweise „Raking“ oder „Iterative Proportional Fitting (IPF)“, bekanntes iteratives Verfahren genutzt.

Sukzessiv werden für jede Kreuzkombination von Bezugsmerkmalen mithilfe von sogenannten Maximum-Likelihood-Schätzern neue Gewichte berechnet, bis eine gewünschte vorgegebene Kohärenz erreicht wird. Mit zunehmender Zahl von Variablen und deren Ausprägungen wird es immer schwieriger, eine 100-prozentige Kohärenz zu gewährleisten. Dementsprechend wird eine Kohärenz erst nach sehr vielen Schritten erreicht, die mit großem Rechen- und damit auch Zeitaufwand einhergehen.

Bereits der Einsatz von mehr als acht Bezugsmerkmalen führt bei diesem Verfahren oftmals nicht mehr zu zufriedenstellenden Ergebnissen. Zusätzlich sollte darauf geachtet werden, auf schwach besetzte Merkmalsausprägungen weitestgehend zu verzichten[6]. Auch sollten keine Kombinationen von Ausprägungen in der Stichprobe, die stark von der Verteilung in der Grundgesamtheit abweichen, für das Verfahren zugelassen werden: beispielsweise 90 % Frauen und 10 % Männer in der Stichprobe und im Gegensatz dazu 30 % Frauen und 70 % Männer in der Grundgesamtheit.

Ein einfaches Beispiel für die Anpassung an zwei Merkmale A und B mit je zwei Ausprägungen soll hier die Vorgehensweise anhand einer Kontingenztafel illustrieren:

› Schritt 0: Ausgehend von unveränderlichen Eckwerten (Randwerten) $Eck_{1.}$, $Eck_{2.}$, $Eck_{.1}$, $Eck_{.2}$ (verallgemeinert $Eck_{i.}$ und $Eck_{.j}$), wobei zum Beispiel $Eck_{.2}$ für den Totalwert der Ausprägung 2 bei Merkmal B steht, werden die Stichprobenwerte als Merkmalskombinationsausprägungen beider Variablen mithilfe der Bezeichnungen z_{11}, z_{12}, z_{21}, und z_{22}, (allgemein z_{ij},) in die Tafel übernommen. N ist der Totalwert, der sich durch $Eck_{1.}$ + $Eck_{2.}$, beziehungsweise $Eck_{.1}$ + $Eck_{.2}$ ergibt, also der Totalwert in der Grundgesamtheit über alle Merkmalsausprägungen einer Variablen. ↘ Tabelle 1

Tabelle 1
Ausgangssituation nach Schritt 0

z_{11}	z_{12}	$Eck_{1.}$
z_{21}	z_{22}	$Eck_{2.}$
$Eck_{.1}$	$Eck_{.2}$	N

› Schritt 1: Es wird zeilenweise (man kann natürlich auch zuerst mit den Spalten beginnen) eine Anpassung an die Eckwerte $Eck_{1.}$ beziehungsweise $Eck_{2.}$ vorgenommen und die Zellwerte werden proportional mithilfe der Formel

$$z_{i,j}^{(1)} = \frac{z_{i,j}^{(0)} \cdot Eck_{i.}}{\sum_j z_{i,j}^{(0)}}$$

verändert.

6 Als problematisch erwiesen sich in diesem Verfahren häufig Ausprägungen, die weniger als 5 % des Stichprobenumfangs enthielten. Weitere Details hierzu siehe beispielsweise bei Bishop und andere (2007).

› Schritt 2: Durch analoges Vorgehen wird auf die Eckwerte in den Spalten, also $Eck_{.1}$ und $Eck_{.2}$, hin eine Anpassung vorgenommen und die in Schritt 1 ermittelten Zellenwerte werden bezüglich der anderen Variable proportional verändert.

$$z_{i,j}^{(2)} = \frac{z_{i,j}^{(1)} \cdot Eck_{.j}}{\sum_i z_{i,j}^{(1)}}$$

Die Schritte 1 und 2 werden nun so lange abwechselnd wiederholt, bis sich die gewünschte Kohärenz eingestellt hat, im Optimalfall also mit der vollständigen Kohärenz $\sum_i z_{i,j} = Eck_{.j}$ und $\sum_j z_{i,j} = Eck_{i.}$.

Diese Vorgehensweise ist ebenfalls mithilfe statistischer Verteilungsannahmen modellierbar, indem man von einer multinomialen Verteilung für die Zelleneinträge ausgeht und diese in logarithmierter Version darstellt. Die, wie oben erläutert, sukzessiv erzeugten Schätzungen für die Zelleinträge sind unter diesen Voraussetzungen dann stets Maximum-Likelihood-Schätzungen.

Alle Berechnungen wurden in SAS durchgeführt, wobei das Iterativ-Proportional-Fitting-Verfahren mithilfe der CALL IPF-Funktion im IML-Modul durchgeführt wurde.[7]

Eckwerte und Bezugsmerkmale

Besonders geeignet erschienen für die Hochrechnung folgende Variablen:

› Geschlecht und Nationalität (in jeweils zwei Ausprägungen),

› Familienstand (in vier Ausprägungen)

sowie die Merkmale

› Alter (in neun Klassen),

› Erwerbstätigkeit (einmal in fünf Ausprägungen mit Variablennamen ERW1 und einmal in zwei Ausprägungen mit Variablennamen ERW2 untergliedert[8]),

› Wirtschaftszweige WZ (in elf Ausprägungen),

› Regierungsbezirk RB (in 38 Ausprägungen, dabei wird bei Bundesländern ohne Regierungsbezirke das Bundesland selbst als Regierungsbezirk übernommen), sowie

› eine weitere regionale Einteilung in eine Variable DOMAIN (2 020 Ausprägungen). Diese regionale Einteilung orientiert sich an der gewünschten regionalen Gliederungstiefe der Ergebnisse für Kreise und kreisfreie Städte, Gemeinden mit 10 000 oder mehr Einwohnerinnen und Einwohnern, Stadtteilen in Großstädten mit 400 000 oder mehr Einwohnerinnen und Einwohnern sowie Gemeindeverbände in Rheinland-Pfalz mit 10 000 oder mehr Einwohnerinnen und Einwohnern. Daraus wurde eine erschöpfende Gliederung nach Stadtteilen von Großstädten, großen Gemeinden, großen Verbandsgemeinden in Rheinland-Pfalz und Kreisresten abgeleitet. Das Konstruktionsprinzip ist angelehnt an die Bildung der „Sampling Points“ bei der Stichprobenziehung (Berg/Bihler, 2011).

Ausgangspunkt war hier der Wunsch einer ausgewogenen Auswahl von demografischen und erwerbsstatistischen Anpassungsmerkmalen, wobei die Anzahl und Kreuzkombinationen für das Modell den Verfahrensrestriktionen angepasst werden sollten.

Ein erster Vorschlag wurde mithilfe des folgenden Modells unterbreitet:

DOMAIN*ERW2*ALTER*GESCHL +

DOMAIN*ERW2*NAT*GESCHL+

DOMAIN*ERW2*NAT*FAMST+

DOMAIN*ERW1+

RB*ERW2*ALTER*WZ.

Dabei bedeutet „*“ eine Kreuzkombination von Merkmalen und „+“ eine getrennte Anpassung an Randverteilungen, wobei die Randverteilungen auch mehrdimensional sein können. Die Bezeichnung „NAT+GESCHL“ beispielsweise erfordert demnach eine getrennte Betrachtung von Deutschen und Nichtdeutschen sowie Männern und Frauen.

In der Testphase dieses Modells wurde schnell deutlich, dass die obersten und untersten Altersklassen zusammengeführt werden müssen, da in diesen Altersklassen ein sehr einseitiges Erwerbsprofil festzustellen ist. Das Merkmal ALTER wurde somit auf lediglich sieben Klassen reduziert.

7 Für eine ausführliche Beschreibung und Erweiterungen der CALL IPF-Funktion in SAS siehe Izrael und andere (2000).

8 Der Hintergrund für diese zweifache Einteilung wird aus der Beschreibung des Baukastenprinzips ersichtlich (siehe unten).

Des Weiteren stellten sich Kreuzkombinationen mit dem regionalen Merkmal DOMAIN als zu ambitioniert heraus. Insbesondere in Kombination mit mehrdimensionalen Merkmalen erwiesen sich die sich ergebenden Besetzungszahlen oftmals als sehr gering.

Aus diesen Erkenntnissen heraus wurde eine ausgewogene Mischung aus Bezugsmerkmalen, die mehrdimensional mit der regionalen Variablen RB und eindimensional mit der Variable DOMAIN kombiniert werden, entwickelt.

Das final zum Einsatz gekommene Bezugsmerkmalmodell lautet schließlich:

RB*ERW2*ALTER*GESCHL +
RB*ERW2*NAT*GESCHL+
RB*ERW2*NAT*FAMST+
RB*WZ+
DOMAIN*ALTER+
DOMAIN*GESCHL+
DOMAIN*NAT+
DOMAIN*FAMST+
DOMAIN*ERW1.

↘ **Übersicht 2** gibt einen Überblick über die letztlich verwendeten Bezugsmerkmale und deren Ausprägungen.

Datensituation

Aus dem durch die oben beschriebenen Verknüpfungen erweiterten zensustypischen Datensatz heraus wird nunmehr die Abgrenzung der für die Hochrechnung zur Verfügung stehenden Grundgesamtheit vollzogen. Ziel ist es, auf einen Personenbestand zurückzugreifen, der nach Durchführung des Korrekturverfahrens alle Perso-

Übersicht 2
Für die Hochrechnung der Zusatzmerkmale beim Zensus 2011 verwendete Bezugsmerkmale

Variablenname	Merkmal	Ausprägungen
GESCHL	GESCHLECHT	männlich, weiblich
NAT	NATIONALITÄT	deutsch, nichtdeutsch
FAMST	FAMILIENSTAND	(1) ledig oder nicht bekannt (2) verheiratet, in eingetragener Lebenspartnerschaft (3) verwitwet, durch Tod aufgelöste Lebenspartnerschaft, durch Todeserklärung aufgelöste Lebenspartnerschaft (4) geschieden, Ehe aufgehoben, aufgehobene Lebenspartnerschaft
ALTER	ALTER	(1) ALTER ≤ 19 (2) 20 ≤ ALTER ≤ 29 (3) 30 ≤ ALTER ≤ 39 (4) 40 ≤ ALTER ≤ 49 (5) 50 ≤ ALTER ≤ 59 (6) 60 ≤ ALTER ≤ 69 (7) ALTER ≥ 70
ERW1	ERWERBSTÄTIGKEIT mit Kategorisierung 1	(1) sozialversicherungspflichtig beschäftigt (2) Beamte, Richter und Soldaten (3) Arbeitslose und Personen in Umschulungsmaßnahmen (4) sonstige Personen (5) nicht in Erwerbstätigenregistern enthalten
ERW2	ERWERBSTÄTIGKEIT mit Kategorisierung 2	(1) aus ERW1 die Ausprägungen (1) und (2) (2) aus ERW1 die Ausprägungen (3) bis (5)
WZ	WIRTSCHAFTSZWEIG	(1) Land- und Forstwirtschaft (2) Verarbeitendes Gewerbe (3) Baugewerbe (4) Handel, Verkehr, Gastgewerbe (5) IT, Kommunikation (6) Finanzen, Versicherungen (7) Immobilien (8) Unternehmensbezogene Dienstleistungen (9) Öffentliche Verwaltung, Erziehung, Gesundheit (10) Sonstige Dienstleistungen (11) keine
DOMAIN	regionales Merkmal DOMAIN	2 020 Ausprägungen
RB	Regierungsbezirk RB	38 Ausprägungen

nen am Hauptwohnsitz umfasst, mit Ausnahme derjenigen, die an sensiblen Sonderanschriften anzutreffen sind. Insgesamt umfasst diese Ziel-2-Grundgesamtheit ein Datenvolumen von 79 652 357 Datensätzen.

Aus der Stichprobendatei wurden für die weiteren Berechnungen lediglich alle existenten Personen mit Hauptwohnsitz an Nichtsonderanschriften verwendet, zuzüglich aller – gemäß Sonderbereichserhebung – existenten Personen an nichtsensiblen Sonderanschriften. Stichprobenausfälle wurden nicht berücksichtigt. Dieser so zusammengestellte Stichprobendatensatz enthält 7 452 833 Personendatensätze.

Das Baukastenprinzip

Erwerbsstatistische Merkmale im Zensus 2011 besitzen die besondere Eigenschaft, dass einige gemäß der obigen Beschreibung einzig auf Basis der Stichprobe hochgerechnet werden. Andere dagegen – die sogenannten Baukastenmerkmale – werden jedoch nur für eine Teilmenge der Stichprobe hochgerechnet und die Restmenge wird ausgezählt.[9] Die Abgrenzung dieser beiden Mengen erfolgt über die Variable ERWERBSTÄTIGKEIT AUS REGISTERN. Zum sogenannten Auszählungsteil (Baukasten 1) gehören die gemeldeten Personen, die die Ausprägungen „sozialversicherungspflichtig beschäftigt“ und „Soldat, Richter, Beamter“ der Variable ERWERBSTÄTIGKEIT AUS REGISTERN besitzen. Die Daten zur ersten Ausprägung, bei der im Übrigen geringfügig Beschäftigte nicht enthalten sind, stammen aus den Registerangaben der Bundesagentur für Arbeit. Für die zweite Ausprägung wurde der Personalstand aus Registerangaben der öffentlichen Arbeitgeber ausgewertet.

Die restlichen Ausprägungen des Merkmals ERWERBSTÄTIGKEIT AUS REGISTERN bilden den Baukasten 2 oder Hochrechnungsteil des Baukastens. Für die Hochrechnung der Baukastenmerkmale müssen Grundgesamtheit und Stichprobe auf den Hochrechnungsteil eingeschränkt werden.[10]

Mit den zu Baukasten 1 gehörenden Einschränkungen fungieren folgende Merkmale als Baukastenmerkmale:

- Erwerbsstatus (nach ILO/Labour-Force-Konzept)
- Stellung im Beruf
- Wirtschaftszweig (nach WZ 2008[11])
- Arbeitsort (nach dem amtlichen Gemeindeschlüssel)

Das Anpassungsmodell für die Baukastenmerkmale wurde abweichend von den übrigen Zusatzmerkmalen auf folgende Merkmalskombinationen aufgesetzt:

RB*ALTER*GESCHL +
RB*NAT*GESCHL+
RB*NAT*FAMST+
RB*NAT*ERW1+
RB*WZ+
DOMAIN*ALTER+
DOMAIN*GESCHL+
DOMAIN*NAT+
DOMAIN*FAMST+
DOMAIN*ERW1

Die dadurch entwickelten Hochrechnungsfaktoren sind aufgrund der unterschiedlichen Basis nicht mit den Hochrechnungsfaktoren im bisherigen Verlauf der Berechnungen vergleichbar. Um trotzdem eine gewisse Kohärenz zu den sonstigen Zusatzmerkmalen herstellen zu können, haben wir zur Bedingung gemacht, dass die zur Hochrechnung verwendeten kombinierten Bezugsmerkmale für die sonstigen Zusatzmerkmale jeweils eine erwerbsstatistische Merkmalskomponente aus dem Auszählungsteil enthalten.

Unter Berücksichtigung des Baukastensystems verteilt sich der Datenumfang wie in ↘ Tabelle 2 dargestellt.

Tabelle 2
Verteilung des Datenumfangs

	Datensätze
Grundgesamtheit	79 652 357
Auszählungsteil Baukasten	29 226 746
Stichprobenumfang	7 452 833
Stichprobenteil des Baukastens	4 844 074

9 Ziel ist es, die Ergebnisse, die aus einer reinen Hochrechnung erhältlich wären, durch die Nutzung von Informationen aus erwerbsstatistischen Registern zu verbessern, unter anderem durch den dadurch reduzierten Gesamtzufallsfehler. Voraussetzung ist, dass die Baukastenmerkmale im Auszählungsteil hinreichend zuverlässig sind und der Messfehler nicht höher als bei der Befragung ist.

10 Zum Hochrechnungsteil gehören beispielsweise Selbstständige und Kinder.

11 Klassifikation der Wirtschaftszweige, Ausgabe 2008 (WZ 2008).

5

Beispiele und weitere Besonderheiten

Nach der Definition von Stichprobe und Grundgesamtheit hat sich die Besonderheit ergeben, dass die Stichprobe keine echte Teilmenge der Grundgesamtheit darstellt. Es fehlen in der Grundgesamtheit nämlich weitestgehend die Fehlbestände der Stichprobe aus kleinen Gemeinden. Durch das vorgestellte IPF-Verfahren stören diese zusätzlichen Stichprobeneinheiten nicht, sondern liefern sogar noch einen kleinen Informationsgewinn.

Eine kleine Gruppe von 39 Personen aus der Stichprobe besitzt Ausprägungen der Bezugsmerkmale, die in der Grundgesamtheit überhaupt nicht vorkommen.

Diese Personen fallen aus dem Kalibrierungsverfahren heraus und erhalten einen Hochrechnungsfaktor, der sich aus deren Eingangsgewicht ergibt und mit der Summe aller anderen Hochrechnungsfaktoren aus der zugehörigen Domain wieder kohärente Ergebnisse auf Domainebene liefert.

Trotz aller Bemühungen gab es immer wieder einige „Problemausprägungen“, die in Einzelfällen zu sehr geringen Stichprobengrößen führen. Insbesondere gelangten in einem Kreisrest in Thüringen zufällig keine Nichtdeutschen in die Stichprobe, was zur Folge hatte, dass aufgrund dieser Nichtbesetzung der Modellterm DOMAIN*NAT in Thüringen entfällt.

Kleine Besetzungszahlen lieferte beispielweise auch die auf Regierungsbezirksebene kombinierte Ausprägung „nichtdeutsch/verwitwet/sozialversicherungspflichtig beschäftigt“. In insgesamt acht Fällen lag die Anzahl der Stichprobeneinheiten unter zehn Personen (im Minimalfall drei Personen). Ebenfalls wiesen Kombinationen mit dem Wirtschaftszweig „Land- und Forstwirtschaft“ in wenigen extrem urbanen Regierungsbezirken nur sehr geringe Stichprobenbesetzungszahlen auf (vier Stichproben mit Besetzungszahlen zwischen sechs und zehn).

Trotz der sporadisch auftretenden kleinen Besetzungszahlen in speziellen Ausprägungskombinationen wurde deren Einfluss als sehr gering eingestuft und am oben beschriebenen Modell festgehalten.

6

Auswertung mithilfe der Hochrechnungsfaktoren

Nach erfolgter Berechnung der beiden Hochrechnungsfaktoren werden diese dem zensustypischen Datensatz zugefügt und in einer Auswertungsdatenbank den Nutzerinnen und Nutzern zur Verfügung gestellt.

Um diese Hochrechnungsfaktoren dann ordnungsgemäß zu nutzen, muss auf drei verschiedene Merkmalsgruppen hingewiesen werden:

- Sämtliche beteiligten Merkmale sind demografischen Ursprungs:

  Bei einer Auswertung erfolgt eine klassische Auszählung (zum Beispiel alle Frauen in einer bestimmten Gemeinde mit einem bestimmten Alter und bestimmtem Familienstand).

- Sonstige Zusatzmerkmale, das sind Merkmale, die lediglich aus der Haushaltsstichprobe gewonnen wurden, wie beispielsweise „höchster Bildungsabschluss“ oder „Beruf“:

  Personenzahlberechnungen werden mithilfe des Hochrechnungsfaktors für Zusatzmerkmale durch Multiplikation mit diesem Faktor und anschließender Aggregation durchgeführt. Diese Vorgehensweise erfolgt auch bei Auswertungsmerkmalen, die aus Kombinationen von demografischen und sonstigen Zusatzmerkmalen bestehen.

- Baukastenmerkmale, also die Zusatzmerkmale, die sowohl aus einem Auszählungsteil, der aus Registern bestimmt wird, sowie einem Hochrechnungsteil, der nur aus der Stichprobe stammt, bestehen:

  Hier erfolgt eine kombinierte Auswertung bezüglich der beiden Teilmengen. Die Personen, die zum Hochrechnungsteil gehören, werden dann mit dem Hochrechnungsfaktor für Baukastenmerkmale hochgerechnet und zu den ausgezählten Personen, die mit Ausprägungen des Auszählungsteils verknüpft sind, addiert. Die gleiche Vorgehensweise kommt zum Tragen, wenn es um Auswertungen bezüglich Merkmalskombinationen von demografischen und Baukastenmerkmalen geht.

› Schließlich werden Auswertungen, die Kombinationen von Baukasten- und Zusatzmerkmalen enthalten, unter Nutzung des Hochrechnungsfaktors für Zusatzmerkmale erstellt.

Für die Datensätze, bei denen der Hochrechnungsfaktor für Zusatzmerkmale existiert und für die der Hochrechnungsfaktor für die Baukastenmerkmale nicht berechnet wurde, wurde dieser auf „1“ gesetzt.

7

Zusammenfassung und Ausblick auf den Zensus 2021

Nach der Ermittlung der Einwohnerzahl und erster demografisch untergliederter Ergebnisse stellte die Hochrechnung der Zusatzmerkmale die zweite Komponente zum Erreichen einer tiefgegliederten Auswertungsbasis für den Zensus 2011 dar.

Im Vergleich zum zeitlich deutlich früher durchgeführten Ziel-1-Verfahren, welches auf einer Anwendung eines verallgemeinerten Regressionsschätzers beruhte, sollten die Auswirkungen möglicherweise in hoher Anzahl auftretender negativer Regressionsgewichte vermieden werden. Daher sollte ein Iterative-Proportional-Fitting-Ansatz zum Einsatz kommen. Aufgrund der Beschaffenheit dieses Verfahrens war eine Reihe von Anpassungen an den gewählten Bezugsmerkmalen notwendig, bevor die realisierten Ergebnisse für Zusatzmerkmale in Form von Hochrechnungsfaktoren einer Auswertungsdatenbank übergeben werden konnten.

Mit dem Anspruch, die Daten aus erwerbsstatistischen Registern effizient in die Ergebnisermittlung einfließen zu lassen, wurde ein „Baukastensystem“ konstruiert. Dieses führte für einige erwerbsstatistische Merkmale durch eine Trennung in einen Auszählungs- und einen Hochrechnungsteil zu besonders zufallsfehlerarmen Schätzungen.

Das hier vorgestellte Verfahren hat sich in der Praxis als anerkannte und bewährte Vorgehensweise insbesondere hinsichtlich einer kohärenten Ergebnisstruktur erwiesen. Trotzdem versucht das Statistische Bundesamt, die gewonnenen Erkenntnisse zur weiteren Modernisierung und Effizienzgewinnen zu nutzen.

Um den kommenden Zensus 2021 vorzubereiten, wurden diesbezüglich bereits zahlreiche Untersuchungen durchgeführt. Hinsichtlich der künftigen Hochrechnungsmöglichkeiten konnten beispielsweise Kombinationen einer Ziel-2-Variable mit Geschlecht und Altersklassen verbunden und Variationen eines G-SPREE-Ansatzes („Generalised Structure Preserving Estimation“) getestet werden. Erfolgversprechende Ergebnisse bezüglich vernachlässigbarer Verzerrungen und deutlich verringerter Standardfehler deuten auf eine Neukonzeption des Hochrechnungsverfahrens hin, bei der Small-Area-Schätzverfahren eingesetzt werden könnten.

LITERATURVERZEICHNIS

Berg, Andreas/Bihler, Wolf. *Das Stichprobendesign der Haushaltsstichprobe des Zensus 2011*. In: Wirtschaft und Statistik. Ausgabe 4/2011, Seite 317 ff.

Berg, Andreas/Bihler, Wolf. *Das Hochrechnungsverfahren zur Ermittlung der Einwohnerzahl im Zensus 2011*. In: Wirtschaft und Statistik. Ausgabe 4/2014, Seite 229 ff.

Bishop, Yvonne M. M./Fienberg, Stephen E./Holland, Paul W. *Discrete Multivariate Analysis: Theory and Applications*. 2007.

Gauckler, Britta. *Die Entwicklung des Fragebogens zur Haushaltebefragung des Zensus 2011*. In: Wirtschaft und Statistik. Ausgabe 8/2011, Seite 718 ff.

Hirner, Stephanie/Stiglmayr, Susanne. *Der Referenzdatenbestand im Zensus 2011*. In: Wirtschaft und Statistik. Ausgabe 1/2013, Seite 30 ff.

Hofmeister, Katrin/Fürnrohr, Michael. *Das Korrekturverfahren beim Zensus 2011*. In: Bayern in Zahlen. Ausgabe 6/2014, Seite 310 ff.

Izrael, David/Hoaglin, David C./Battaglia, Michael P. *A SAS Macro for Balancing a Weighted Sample*. In: Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference. Paper 275. 2010.

Statistische Ämter des Bundes und der Länder (Herausgeber). *Zensus 2011. Methoden und Verfahren.* Wiesbaden 2015.

Herausgeber
Statistisches Bundesamt (Destatis), Wiesbaden

Schriftleitung
Dr. Sabine Bechtold
Redaktionsleitung: Juliane Gude
Redaktion: Ellen Römer

Ihr Kontakt zu uns
www.destatis.de/kontakt

Erscheinungsfolge
zweimonatlich, erschienen im August 2018
Das Archiv aller Ausgaben ab Januar 2001 finden Sie unter www.destatis.de/publikationen

Print
Einzelpreis: EUR 18,– (zzgl. Versand)
Jahresbezugspreis: EUR 108,– (zzgl. Versand)
Bestellnummer: 1010200-18004-1
ISSN 0043-6143
ISBN 978-3-8246-1071-6

Download (PDF)
Artikelnummer: 1010200-18004-4, ISSN 1619-2907

Vertriebspartner
IBRo Versandservice GmbH
Bereich Statistisches Bundesamt
Kastanienweg 1
D-18184 Roggentin
Telefon: +49 (0) 382 04 / 6 65 43
Telefax: +49 (0) 382 04 / 6 69 19
destatis@ibro.de

Papier: Metapaper Smooth, FSC-zertifiziert, klimaneutral, zu 61% aus regenerativen Energien