I B - Forschungsdatenzentrum
Anja Crößmann

Wiesbaden, September 2009

Konzept zur Anonymisierung der Volkszählung der Bundesrepublik Deutschland im Jahre 1987 zur Verwendung als Public-Use-File (PUF)

I. Vorbemerkung

Die Volkszählung der Bundesrepublik Deutschland (BRD) im Jahre 1987 diente als Mehrzweckerhebung zur Bevölkerungs- und Berufszählung sowie zur Gebäude- und Wohnungszählung. Als stichtagsbezogene Erhebung ermittelte sie die wichtigsten demographischen, sozialen und ökonomischen Merkmale der Einwohner und Haushalte in der Bundesrepublik. Rechtsgrundlage war das Gesetz über eine Volks-, Berufs-, Gebäude-, Wohnungs- und Arbeitsstättenzählung (Volkszählungsgesetz 1987) vom 8. November 1985 (BGBl. I S. 2078).
Dem Forschungsdatenzentrum des Statistischen Bundesamtes wurden die Volkszählungsdaten der BRD von 1987 von den Statistischen Landesämtern zum Zwecke der Aufbereitung und Anonymisierung übermittelt. Gemäß § 16 Abs. 1 BStatG sind Einzelangaben über persönliche und sachliche Verhältnisse nicht geheim zu halten, soweit diese dem Befragten oder Betroffenen nicht zuzuordnen, d.h. absolut anonym sind.
Vorliegendes Konzept beschreibt die Vorgehensweise des Forschungsdatenzentrums des Statistischen Bundesamtes bei der Aufbereitung und Anonymisierung der Daten der Volkszählung der BRD von 1987 zur Erstellung eines absolut anonymisierten Mikrodaten-Files (Public-Use-File).

II. Basismaterial

Das Basismaterial der Volkszählung der BRD von 1987 umfasst ca. 63,9 Mio. Personen, 26,7 Mio. Haushalte, 25,9 Mio. Wohnungen und 177 Merkmale. Die Personenangaben enthalten demografische Informationen, Angaben zu den Quellen des Lebensunterhalts, der Bildung, der Erwerbstätigkeit sowie der Haushaltszusammensetzung. Die Wohnungsangaben geben Auskunft über die Wohnungsbelegung sowie deren Ausstattung. Angaben über Gebäude informieren über deren Nutzung und Ausstattung.

III. Plausibilisierung der Daten

Die Plausibilisierung der Daten erfolgt auf der Grundlage von Häufigkeitsauszählungen der Ausprägungen aller Variablen. Hier werden Ausreißerwerte identifiziert und fehlerhafte Angaben gelöscht bzw. durch richtige, falls diese aus dem Material ableitbar sind, ersetzt.
Datensätze von leerstehenden, als Nebenwohnung oder zweckentfremdet genutzten Wohnungen werden aus dem Datenmaterial gelöscht, da für diese Wohnungen keine Personeninformationen vorhanden sind.

IV. Anonymisierungsmaßnahmen

Folgendes Bündel an Anonymisierungsmaßnahmen führt zur absoluten Anonymität der Volkszählungsdaten der BRD von 1987:

1. Alter der Daten

Da die Volkszählung der BRD von 1987 mittlerweile mehr als zwei Jahrzehnte zurückliegt, kann angenommen werden, dass Zusatzinformationen nur in eingeschränktem Umfang verfügbar sind und wenn sie vorliegen, nur von geringer Verlässlichkeit sind. Insbesondere kann davon ausgegangen werden, dass viele der befragten Haushalte in ihrer damaligen Zusammensetzung und Struktur nicht mehr existieren sowie Informationen zu Haushaltsmitgliedern nicht mehr aktuell sind. Das Alter der Daten stellt somit ein erhebliches Anonymitätskriterium dar.

2. Stichprobenziehung

Als erster Schritt der Anonymisierung wird aus dem Originalmaterial eine systematische 5% Zufallsstichprobe auf der Haushaltsebene mit Hilfe des Schlussziffernverfahrens gezogen. Als Vorbedingung der Stichprobenziehung wird das Originalmaterial nach Bundesland, Regierungsbezirk, Gemeindegrößenklasse der Wohnsitzgemeinde, Zahl der Personen in Privathaushalten, lfd. Nummer der Wohnung im Gebäude und lfd. Nummer des Haushalts in der Wohnung sortiert. Durch diese Anordnung ist gewährleistet, dass die Stichprobe hinsichtlich dieser Merkmale nur geringe zufallsbedingte Abweichungen aufweist. Anschließend wird allen Haushalten eine laufende Haushaltsnummer erteilt. Jede Person in einer Gemeinschaftsunterkunft erhält hierbei eine eigene fortlaufende Haushaltsnummer.
Zur Ziehung der 5% Haushaltsstichprobe werden die letzten zwei Endziffern der Haushaltsnummer verwendet. Die Auswahlwahrscheinlichkeit beträgt 5 aus 100. Daher wird in einem Intervall zwischen 0 und 100/5 eine Zahl Z zufällig gewählt. Ausgehend von diesem zufällig ausgewählten Startwert Z werden fünf Werte X_i im Intervall von 00 bis 99 nach der Formel:

$$X_i = Z + ganzzahl\left(i * \frac{100}{5} \right)$$, mit i= 0 bis 99

ermittelt. Alle Haushalte mit den Endziffernkombinationen X_i (d.h. 5 aus 100) werden in die Stichprobe aufgenommen. Durch die Stichprobenziehung kann ein potenzieller Datenangreifer nicht sicher sein, ob die gesuchte Person oder der gesuchte Haushalt sich in der Stichprobe befindet.

3. Löschung von regionalen Informationen

Als weitere Anonymisierungsmaßnahme werden alle regionalen Informationen bis auf das Bundesland und das Pendlerziel-Land aus dem Datenmaterial gelöscht. Im Einzelnen handelt es sich

dabei um die Variablen: Gemeindeschlüssel, Regierungsbezirk, Kreis, Gemeinde, Gemeindeteil, Block/Blockseite, Regionallisten-Nummer sowie die Pendlerzielangaben Ziel-Gemeinde, Straße bzw. Gemeindeteil, Haus-Nummer und Gemeindegrößenklasse der Pendler-Zielgemeinde.
Um dennoch eine eindeutige Identifizierung der Gebäude, Wohnungen und Haushalte zu ermöglichen werden die Gebäude im Bundesland mit einer eindeutigen laufenden Nummer versehen (s. hierzu den Abschnitt „systemfreie Sortierung“).

Die Variable Gemeindegrößenklasse der Wohnsitzgemeinde (ef10) wird – angelehnt an das Anonymisierungskonzept zur Erstellung des Mikrozensus als Scientific-Use-File – neu generiert. Keine Gemeinde ist hiernach in der Grundgesamtheit mit weniger als 500 000 Einwohnern identifizierbar und in jeder Gemeindegrößenklasse eines Bundeslands sind mindestens 400 000 Einwohner in der Grundgesamtheit vertreten. Die Variable der Gemeindegrößenklasse der Wohnsitzgemeinde teilt nun die Gemeinden in den Bundesländern in folgende Größenklassen ein:

1:		unter	5 000 Einwohner
2:	5 000	bis unter	20 000 Einwohner
3:	20 000	bis unter	100 000 Einwohner
4:	100 000	bis unter	500 000 Einwohner
5:	500 000 Einwohner und mehr.		

Davon abweichend werden zusätzlich folgende Größenklassen ausgewiesen:

für Nordrhein-Westfalen und Saarland:
6: unter 20 000 Einwohner
für Saarland:
7: 20 000 und mehr Einwohner
für Berlin (West):
8: unter 500 000 Einwohner
für Bremen:
9: 100 000 und mehr Einwohner.

4. Löschung von weiteren Variablen

Folgende Variablen gingen aus Anonymisierungsgründen nicht in das Datenmaterial des PUF ein:

ef1	Satzart
ef2	Gemeindeschlüssel
ef2u2	- Regierungsbezirk
ef2u3	- Kreis
ef2u4	- Gemeinde
ef3	Gemeindeteil
ef4	Block/Blockseite
ef4u1	- Block-Nummer
ef4u2	- Blockseite

ef5 Regionallisten-Nummer
ef9 Lfd. Nummer des Pendlers in der Gemeinde
ef11 Kennzeichnung der maßgeblichen Sätze für die Zählung von Regionallisten
ef17 Schicht-Nummer (leer)
ef18 Hochrechnungs-Faktor (leer)
ef23 Wohnungsausstattung (geht aus ef23u1-ef23u4 hervor)
ef25 Brennstoff/Wärmequelle (geht aus ef25u1-ef25u6 hervor)
ef34 Seit wie vielen Monaten steht die Wohnung leer?
ef35 Wohnung steht leer
ef36 Wohnung ist von ausländischen Streitkräften, diplomatischen Vertretungen usw. privatrechtlich gemietet
ef62 Anzahl der selbstbewohnten Räume
ef71 Geburtsmonat
ef80 Erwerbstätigkeit (geht aus ef80u1-ef80u6 hervor)
ef87 Pendlersignierungen
ef87u2 Pendlersignierungen (Ziel) – Gemeinde
ef87u3 Pendlersignierungen (Ziel) – Straße
ef87u4 Pendlersignierungen (Ziel) – Haus-Nr.
ef95 Kennzeichen für Zuordnung bzw. Ergänzung bei der Personentypisierung
ef96 Leer
ef120 Lfd. Nummer des zugehörigen verheirateten Paares im Haushalt
ef143 Weiblicher Partner ohne Schulabschluss, d.h. noch Schülerin/Studentin (leer)
ef150 Pendler-Zielangaben
ef150u1 - Ziel-Land
ef150u2 - Ziel-Regierungsbezirk
ef150u3 - Ziel-Kreis
ef150u4 - Ziel-Gemeinde (amtl. Gemeindeschlüssel)
ef150u5 Straße bzw. Gemeindeteil
ef150u6 Haus-Nummer
ef151 Gemeindegrößenklasse der Pendler-Zielgemeinde
ef200 Satzstellen 183-200
ef201 Leer

5. Systemfreie Sortierung

Aus der Anordnung der Datensätze im Originalmaterial lassen sich indirekt Regionalinformationen ableiten. Um diese Möglichkeit auszuschließen, wird das Datenmaterial systemfrei (d.h. nach einem nicht nachvollziehbaren System) sortiert und anschließend die Variablen Gebäude-, Wohnungs- und Haushaltsnummer mit einer eindeutig systemfreien Nummerierung versehen.

6. Vergröberung von Merkmalsausprägungen

Für alle Variablen des Public-Use-Files der Volkszählung der BRD von 1987 gilt, dass jede ausgewiesene Merkmalsausprägung in der univariaten Verteilung der Grundgesamtheit mindestens

10 000 Fälle umfassen muss. Ausgenommen hiervon sind die Variablen der Staatsangehörigkeit (mindestens 100 000 Fälle) sowie die Gemeindegrößenklasse der Wohnsitzgemeinde (s. hierzu den Abschnitt „Löschung von regionalen Informationen“). Um diese Voraussetzung zu erfüllen wird eine sachgerechte Vergröberung der betroffenen Merkmalsausprägungen vorgenommen. Folgende Variablen sind von Vergröberungen betroffen, deren Umsetzungen dem Schlüsselverzeichnis des Public-Use-Files der Volkszählung der BRD von 1987 zu entnehmen sind (s. Anhang):

ef10 Gemeindegrößenklasse der Wohnsitzgemeinde
ef12 Kennzeichnung der maßgeblichen Sätze für die Zählung von Gebäuden
ef29 Fläche der gesamten Wohnung in m²
ef31 Monatsmiete
ef50 Gebäudetyp
ef51 Wohnungstyp
ef52 Anzahl der Wohnungen im Gebäude
ef53 Anzahl der sonstigen Wohneinheiten im Gebäude
ef54 Anzahl der Freizeitwohneinheiten im Gebäude
ef56 Wohnfläche des Gebäudes in m²
ef64 Anzahl der Haushalte in der Wohneinheit
ef65 Anzahl der Personen in der Wohneinheit
ef70 Geburtsjahr
ef84 Hauptfachrichtung des Abschlusses an einer berufsbildenden Schule oder Hochschule
ef85 Erlernter Beruf (Ausprägungen im Originalmaterial dreistellig, im PUF zweistellig)
ef91 Wirtschaftszweig des Betriebes
ef92 Ausgeübte Tätigkeit/Beruf (Ausprägungen im Originalmaterial sechsstellig, im PUF dreistellig)
ef106 Alter
ef109 Zahl der Personen im Haushalt, die zur wohnberechtigten Bevölkerung gehören
ef110 Zahl der Personen im Haushalt, die zur Bevölkerung am Ort der Hauptwohnung gehören
ef111 Zahl der Personen im Haushalt, die zur Bevölkerung in Privathaushalten gehören
ef112 Zahl der Personen im Haushalt, die zur Wohnbevölkerung gehören
ef113 Anzahl der ledigen Personen unter 3 Jahren im Haushalt
ef114 Anzahl der ledigen Personen unter 6 Jahren im Haushalt
ef115 Anzahl der ledigen Personen unter 10 Jahren im Haushalt
ef116 Anzahl der ledigen Personen unter 15 Jahren im Haushalt
ef117 Anzahl der ledigen Personen unter 18 Jahren im Haushalt
ef118 Anzahl der verheirateten Personen im Haushalt insgesamt
ef119 Anzahl der verheirateten Paare im Haushalt
ef121 Anzahl der Einkommensbezieher (15 Jahre und älter) im Haushalt insgesamt
ef122 Anzahl der Einkommensbezieher (15 Jahre und älter) aus Erwerbstätigkeit im Haushalt
ef129 Anzahl der Erwerbspersonen im Haushalt
ef130 Anzahl der Erwerbstätigen im Haushalt
ef131 Anzahl der Personen von 15 bis unter 65 Jahre im Haushalt
ef132 Anzahl der Personen, die 66 Jahre und älter sind, im Haushalt

ef133 Anzahl der Personen, die 65 Jahre und älter sind, im Haushalt
ef134 Anzahl der ledigen Schüler/Studenten unter 18 Jahre im Haushalt
ef135 Alter der ältesten Person im Haushalt
ef136 Haushaltstyp
ef138 Religionszugehörigkeit des verheirateten Partners der Bezugsperson des Haushalts
ef144 Erwerbstätigkeit des weiblichen Partners
ef146 Kennzeichen für Auszubildende des weiblichen Partners, falls weiblicher Partner Auszubildende ist
ef147 Sozio-ökonomische Gliederung (Deutsch)
ef152 Pendlertyp
ef153 Wie viele Personen im Haushalt, sind verheiratet ohne Partner im gleichen Haushalt sowie nicht verheiratet, verwitwet oder geschieden?
ef156 Geburtsjahr des weiblichen Partners
ef157 Alter des weiblichen Partners
ef158 Religionszugehörigkeit des weiblichen Partners

V. Beschluss

Die unter IV. beschriebenen Anonymisierungsmaßnahmen führen in Verbindung mit dem Alter der Daten zu einem Mikrodatenfile, bei dem eine De-Anonymisierung einzelner Merkmalsträger ausgeschlossen ist. Der Datensatz ist damit absolut anonym und kann in dieser Form als Public-Use-File veröffentlicht werden.