I B – Forschungsdatenzentrum
Anja Crößmann

Wiesbaden, September 2010

Konzept zur Anonymisierung der Volkszählung von 1981 der Deutschen Demokratischen Republik zur Verwendung als Public-Use-File

I. Vorbemerkungen

Die Volkszählungen in der ehemaligen Deutschen Demokratischen Republik (DDR) dienten der stichtagsbezogenen Ermittlung der wichtigsten demographischen, sozialen und ökonomischen Merkmale der Einwohner und der Haushalte. Rechtsgrundlage beider Zählungen war das Gesetz über die Durchführung von Volks-, Berufs-, Wohnraum- und Gebäudezählungen in der DDR vom 1. Dezember 1967.
Die Erhebungsdaten der letzten beiden Zählungen, 1971 und 1981, liegen in elektronischer Form vor. Sie gingen von der Staatlichen Zentrale für Statistik der ehemaligen DDR nach der Wiedervereinigung in das Eigentum des Rechtsnachfolgers, des Statistischen Bundesamtes, über. Hier fand eine sogenannte Rückrechnung der Volkszählungsdaten von 1971 und 1981 statt, d.h. die Daten wurden gesichert und dokumentiert, aber auch in einigen Punkten der Systematik der Bundesdeutschen Volkszählung von 1987 angepasst, um eine Vergleichbarkeit herzustellen.
Die rückgerechneten Daten wurden im Jahre 2000 vom Statistischen Bundesamt an das Bundesarchiv abgegeben. Diese Daten sind gemäß § 2 Abs. 4 Satz 2 und § 5 Abs. 3 BArchG nach einer Sperrfrist von 60 Jahren benutzbar, eine Sondergenehmigung gemäß § 16 Abs. 6-9 BStatG für wissenschaftliche Forschungsvorhaben ist möglich.
Das Forschungsdatenzentrum (FDZ) erhielt in 2005 die im Bundesarchiv befindlichen Volkszählungsdaten zum Zwecke der Aufbereitung und Anonymisierung für eine Verwendung durch die Wissenschaft.
Vorliegendes Konzept beschreibt die Vorgehensweise des FDZ bei der Aufbereitung und Anonymisierung der Daten der Volkszählung von 1981 zur Erstellung eines absolut anonymisierten Mikrodatenfiles, eines sogenannten Public Use Files (PUF).

II. Basismaterial

Das Basismaterial der Volkszählung 1981 umfasst drei Datenfiles: Personendaten, Wohnungsdaten und Daten der Gemeinschaftseinrichtungen. Die Personendaten enthalten demografische Informationen sowie Angaben zu den Quellen des Lebensunterhalts, der Bildung, der Erwerbstätigkeit und der Haushaltszusammensetzung. Die Wohnungsdaten geben Auskunft über den baulichen Zustand der Gebäude, die Wohnungsbelegung und ihre Ausstattung (Heizung, Warmwasserversorgung, sanitäre Anlagen). Das Basismaterial umfasst 17,2 Mio. Personen in 6,5 Mio. Haushalten und 6,6 Mio. Wohnungen.

III. Plausibilisieren der Daten

Die Plausibilisierung der Daten erfolgt auf der Grundlage von Häufigkeitsauszählungen der Ausprägungen aller Variablen. Hier werden Ausreisserwerte identifiziert und fehlerhafte Angaben durch richtige, falls diese aus dem Material ableitbar sind, ersetzt.

IV. Zusammenführen der Personen- und Wohnungsdaten

Die Personen und Wohnungsdaten werden anhand der Merkmale Land, Regierungsbezirk, Kreis, Gemeinde, Ortsteil- bzw. Wohnbezirksnummer, Zählbereich im Ortsteil/Wohnbezirk, Nummer des Stützpunkts und Wohnung zusammengeführt. Da die Daten über die Gemeinschaftseinrichtungen keine gemeinsamen Identifikatoren mit den Personen- und Wohnungsangaben aufweisen, gehen sie nicht in das PUF mit ein.
Datensätze von leerstehenden oder zweckentfremdet genutzten Wohnungen werden aus dem Datenmaterial gelöscht, da für diese Wohnungen keine Personeninformationen vorhanden sind.

V. Anonymisierungsmaßnahmen

Folgendes Bündel an Anonymisierungsmaßnahmen führt zur absoluten Anonymität der Volkszählungsdaten von 1981:

1. Alter der Daten

Die Volkszählung von 1981 liegt mittlerweile fast dreißig Jahre zurück. Es kann daher angenommen werden, dass Zusatzinformationen nur in eingeschränktem Umfang verfügbar 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

Zur Erstellung des PUF der Volkszählung der DDR von 1981 wird aus dem Originalmaterial eine systematische 25% Zufallsstichprobe auf Haushaltsebene mit Hilfe des Schlussziffernverfahrens gezogen. Als Vorbedingung der Stichprobenziehung wird das Originalmaterial nach den Variablen Land, Regierungsbezirk, Kreis, Gemeinde, Ortsteil- bzw. Wohnbezirksnummer, Zählbereich im Ortsteil/Wohnbezirk, Nummer des Stützpunktes und 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 zugeteilt. Jede Person in einer Gemeinschaftsunterkunft erhält hierbei eine eigene fortlaufende Haushaltsnummer.
Zur Ziehung der 25% Haushaltsstichprobe werden die letzten zwei Endziffern der Haushaltsnummer verwendet. Die Auswahlwahrscheinlichkeit beträgt 25 aus 100 oder 1 aus 4. Daher wird in einem Intervall zwischen 0 und 4/1 eine Zahl Z zufällig gewählt. Ausgehend von diesem zufällig ausgewählten Startwert Z werden 25 Werte X_i im Intervall von 00 bis 99 nach der Formel:

$$X_i = Z + ganzzahl\left(i * \frac{100}{25} \right)$$, mit i= 0 bis 99

ermittelt. Alle Haushalte mit den Endziffernkombinationen X_i (d.h. 25 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 Land aus dem Datenmaterial gelöscht (s. 4. Löschung von Variablen).
Um dennoch eine eindeutige Identifizierung der Gebäude, der Wohnungen und der Haushalte zu ermöglichen werden die Gebäude im Bundesland mit einer eindeutigen laufenden Nummer versehen (s. hierzu den Abschnitt „systemfreie Sortierung“).

4. Löschung von Variablen

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

vp1	Satzart
vp2	Gemeindeschlüssel
vp2u2	Gemeindeschlüssel: Regierungsbezirk
vp2u3	Gemeindeschlüssel: Kreis
vp2u4	Gemeindeschlüssel: Gemeinde
vp3	Ortsteil- bzw. Wohnbezirksnummer (neu)
vp4	Gemeindeschlüssel
vp4u1	Gemeindeschlüssel: Bezirk
vp4u2	Gemeindeschlüssel: Kreis
vp4u3	Gemeindeschlüssel: Gemeinde
vp5	Ortsteil- bzw. Wohnbezirksnummer (alt)
vp54	Gemeindenummer des Pendlerzieles
vp54u2	Gemeindenummer des Pendlerzieles: Regierungsbezirk
vp54u3	Gemeindenummer des Pendlerzieles: Kreis
vp54u4	Gemeindenummer des Pendlerzieles: Gemeinde
vp55	Gemeindenummer des Pendlerzieles
vp55u1	Gemeindenummer des Pendlerzieles: Bezirk
vp55u2	Gemeindenummer des Pendlerzieles: Kreis
vp55u3	Gemeindenummer des Pendlerzieles: Gemeinde
vp60	Leerfeld
vw1	Satzart
vw2	Gemeindeschlüssel
vw2u2	Gemeindeschlüssel: Regierungsbezirk
vw2u3	Gemeindeschlüssel: Kreis
vw2u4	Gemeindeschlüssel: Gemeinde
vw3	Ortsteil- bzw. Wohnbezirksnummer (neu)
vw4	Gemeindeschlüssel

vw4u1	Gemeindeschlüssel: Bezirk
vw4u2	Gemeindeschlüssel: Kreis
vw4u3	Gemeindeschlüssel: Gemeinde
vw5	Ortsteil- bzw. Wohnbezirksnummer (alt)
vw48	Leerfeld

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-, Haushalts- und Personennummer mit einer eindeutigen systemfreien Nummerierung versehen.

6. Vergröberung von Merkmalsausprägungen

Für alle Variablen des Public-Use-Files der Volkszählung der DDR von 1981 gilt, dass jede ausgewiesene Merkmalsausprägung in der univariaten Verteilung der Grundgesamtheit mindestens 3 Fälle umfassen muss. Um diese Voraussetzung zu erfüllen wird eine sachgerechte Vergröberung der betroffenen Merkmalsausprägungen vorgenommen. Die Vergröberungen der betroffen Variablen und deren Umsetzungen sind dem Schlüsselverzeichnis des Public-Use-Files der Volkszählung der DDR von 1981 zu entnehmen.

VI. Beschluss

Die unter V. 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.

2. L-FDZ
3. FDZ Umsetzung d. Anonymisierungsmaßnahmen
4. Kopie an Geschäftsstelle FDZ-Länder z.K.