I B – Forschungsdatenzentrum
Anja Crößmann

Wiesbaden, November 2010

Konzept zur Anonymisierung der Volkszählung der Bundesrepublik Deutschland im Jahre 1970 zur Verwendung als Public-Use-File

I. Vorbemerkung

Die Volkszählung der Bundesrepublik Deutschland (BRD) im Jahre 1970 diente als Mehrzweckerhebung zur Bevölkerungs-, Berufs- sowie zur Arbeitsstättenzählung. 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- und Arbeitsstättenzählung (Volkszählungsgesetz 1970) vom 14. April 1969 (BGBl. I S. 292) in Verbindung mit dem Gesetz über die Statistik für Bundeszwecke (Bundesstatistikgesetz – BStatG) vom 3. September 1953 (BGBl. I S. 1314).
Während die Arbeitsstättenzählung in einer Totalerhebung durchgeführt wurde, war die Volks- und Berufszählung in zwei Teile gegliedert. In der Totalerhebung wurden nur die grundlegenden Merkmale erfasst, die für regional und fachlich tief gegliederte Auswertungen nötig waren. Alle übrigen Merkmale wurden mit Ausnahme des bevölkerungsschwachen Saarlands für weitergehende demographische, wirtschaftliche und soziale Strukturanalysen in einer 10-Prozent-Stichprobe der Bevölkerung zusammengefasst.
Dem Forschungsdatenzentrum des Statistischen Bundesamtes wurden die Volkszählungsdaten der BRD 1970 von den Statistischen Landesämtern zum Zwecke der Aufbereitung und Anonymisierung zur Verfügung gestellt. 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.
Das vorliegende Konzept beschreibt die Vorgehensweise des Forschungsdatenzentrums des Statistischen Bundesamtes bei der Aufbereitung und Anonymisierung der Daten der Volkszählung der BRD 1970 zur Erstellung eines absolut anonymen Mikrodatenfiles, eines sogenannten Public-Use-File (PUF).

II. Basismaterial

Die Volkszählung der BRD von 1970 umfasst ca. 63,9 Mio. Personen in 26,7 Mio. Haushalte sowie 205 Merkmale. Die darin enthaltenen Personenangaben geben Auskunft über demografische Informationen und enthalten Angaben zu den Quellen des Lebensunterhalts, der Bildung sowie der Erwerbstätigkeit. Weiterhin in den Daten enthalten sind haushalts- und familienbezogene Angaben, die Auskunft über die jeweilige Zusammensetzung geben.

Da die Originaldaten der Totalerhebung der Volkszählung von 1970 der BRD in den statistischen Ämtern nicht mehr vorliegen, wird zur Erstellung des PUF auf die 10-Prozent-Stichprobe des faktisch anonymen, sogenannten Scientific-Use-File (SUF) zurückgegriffen. Dieses enthält knapp 6,2 Mio. Personen in 2,3 Mio. Haushalten und wurde in Zusammenarbeit der statistischen Landes-

ämter mit dem Zentrum für Umfragen, Methoden und Analysen (ZUMA, dem heutigen GESIS – Leibniz-Institut für Sozialwissenschaften) erstellt.[1]

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 ersetzt, falls diese aus dem Material ableitbar waren.

IV. Anonymisierungsmaßnahmen

Folgendes Bündel an Anonymisierungsmaßnahmen baut auf denen des SUF auf[1] und führt zur absoluten Anonymität der Volkszählungsdaten der BRD von 1970:

1. Alter der Daten

Da die Volkszählung der BRD von 1970 mittlerweile vierzig Jahre 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

Zur absoluten Anonymisierung der Volkszählung der BRD 1970 wird aus dem 10-Prozent-Material des SUF eine systematische, zufallsbedingte 50-Prozent-Unterstichprobe auf der Haushaltsebene mit Hilfe des Schlussziffernverfahrens gezogen. Insgesamt steht somit eine 5-Prozent-Stichprobe der BRD Bevölkerung von 1970 im PUF zur Verfügung. Durch die Stichprobenziehung kann ein potenzieller Datenangreifer nicht sicher sein, ob die gesuchte Person oder der gesuchte Haushalt sich in der Stichprobe befindet. In Kombination mit den weiteren getroffenen Maßnahmen, führt die 10-Prozent-Stichprobe zur faktischen und die 5-Prozent-Unterstichprobe zur absoluten Anonymität des jeweiligen Datenfiles.
Als Vorbedingung der Stichprobenziehung wird das Basismaterial nach einer neu generierten Haushaltsnummer[2] sortiert.

---

[1] Informationen und Dokumentationen des SUF der Volkszählung der BRD 1970 befinden sich auf den Internetseiten des Metadatensystems der Forschungsdatenzentren der statistischen Ämter des Bundes und der Länder: http://dok.fdz-metadaten.de/1/12/121/12111A/erheb/197000/index.html sowie bei GESIS: http://www.gesis.org/dienstleistungen/daten/amtliche-mikrodaten/volks-und-berufszaehlung-1970 .

[2] Die Generierung der Haushaltsnummern erfolgt nach der von Bernhard Schimpl-Neimanns/ Hansjörg Frenzel entwickelten Vorgehensweise. Diese ist beschrieben in Schimpl-Neimanns/ Frenzel: 1-Prozent Stichprobe der Volks- und Berufszählung 1970. Datei mit Haushalts- und Familiennummern und revidierter Teilstichprobe für West-Berlin. Dokumentation der Datenaufbereitung. (ZUMA-Technischer Bericht 95/06). Für eine Hochrechnung auf die Gesamtbevölkerung wurde die 10%-Stichprobe des SUF im Bundesland Berlin (West) durch Doppeln und Streichen von Haushalten an die Verteilungen der Grundgesamtheit angepasst. Bei der Zuteilung der Haushaltsnummern im PUF wird daher die Variable „doppl“ generiert, die anzeigt, welche Haushalte gedoppelt wurden (vgl. Schlüsselverzeichnis des PUF BRD 1970).

In der zu ziehenden Unterstichprobe ergibt sich eine Auswahlwahrscheinlichkeit von 50 aus 100 oder 5 aus 10. Zur Stichprobenziehung wird die letzte Ziffer der Haushaltsnummer verwendet. Hierzu werden in einem Intervall zwischen 0 und 9 fünf Werte X_i zufällig gewählt.

$$X_i = Z + ganzzahl\left(i * \frac{10}{5} \right)$$, mit i = 0 bis 9

Alle Haushalte mit der Endziffernkombination X_i (d.h. 5 aus 10) werden in die Stichprobe aufgenommen.

3. Löschung von regionalen Informationen

Als weitere Anonymisierungsmaßnahme werden alle regionalen Informationen bis auf Wohnsitzgemeinde/Land aus dem Datenmaterial gelöscht (s. hierzu Punkt 4 „Löschung von Variablen“). Um dennoch eine eindeutige Identifizierung der Haushalte zu ermöglichen wird jeder Haushalt mit einer laufenden Nummer versehen (s. hierzu Punkt 5 „Systemfreie Sortierung“).

Die Variable Gemeindegrößenklasse des Wohnortes (v78) wird – angelehnt an das Anonymisierungskonzept zur Erstellung des Mikrozensus als SUF – neu generiert. Aufgrund des Alters der Daten wird jedoch die Höhe der Mindestfallzahlen nur auf ein Zehntel der im Mikrozensus SUF angesetzten Fallzahlen gesetzt. Keine Gemeinde ist hiernach in der Grundgesamtheit mit weniger als 50 000 Fällen identifizierbar und in jeder Gemeindegrößenklasse eines Bundeslands sind mindestens 40 000 Fälle in der Grundgesamtheit vertreten. Die Variable der Gemeindegrößenklasse des Wohnortes teilt nun die Gemeinden in den Bundesländern in folgende Größenklassen ein:

01:		unter	200	Einwohner
02:	200	bis unter	500	Einwohner
03 :	500	bis unter	1 000	Einwohner
04 :	1 000	bis unter	2 000	Einwohner
05 :	2 000	bis unter	5 000	Einwohner
06 :	5 000	bis unter	10 000	Einwohner
07 :	10 000	bis unter	20 000	Einwohner
08 :	20 000	bis unter	50 000	Einwohner
09 :	50 000	bis unter	100 000	Einwohner
10 :	100 000	bis unter	200 000	Einwohner
11 :	200 000	bis unter	500 000	Einwohner
12 :	500 000 Einwohner und mehr.			

Davon abweichend werden zusätzlich folgende Größenklassen ausgewiesen:

für Schleswig-Holstein, Nordrhein-Westfalen und Baden-Württemberg:

13 :	bis unter	500	Einwohner

für Saarland:
14 : bis unter 1 000 Einwohner
für Hamburg:
15 : bis unter 5 000 Einwohner

4. Löschung von Variablen

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

Variablen, die im SUF nicht enthalten sind:
v2 Geburtsdatum: Tag
v95 Sozioökonomische Gliederung (Ausländer)
v118 Familientyp
v152 Geburtsjahr des Ehepartners
v165 Geburtsjahr des Familienvorstandes
v186 Hausnummer-Zusatz
v187 Prioritätsmerkmal
v188 Belegart
v189 Hausnummer-Zusatz
v190 Prioritätsmerkmal
v191 Belegart
v192 Korrekturnummer
v193 Bündelnummer
v194 Zählerlistennummer
v195 Anstaltslisten-Nummer
v196 Anstaltslisten-Nummer
v197 Zählbezirksnummer
v198 Haushaltsnummer
v199 Paginiernummer
v200 Volkszählungs-Kenn-Nummer
v201 Wohnsitzgemeinde/Kreis
v202 Wohngemeinde/Gemeinde
v203 Zielgemeinde/Land
v204 Zielgemeinde/Gemeinde
v205 Auspendler über Landesgrenzen

Zusätzlich im PUF entnommene Variablen:
v3 Geburtsdatum: Monat
v69 Geburtsjahr des 6. Kindes
v70 Geburtsjahr des 7. Kindes
v71 Geburtsjahr des 8. Kindes
v72 Geburtsjahr des 9. Kindes
v73 Geburtsjahr des 10. Kindes
v74 Geburtsjahr des 11. Kindes
v75 Geburtsjahr des 12. Kindes

v80 10%-Kennzeichen alle
v81 Anstaltskennzeichen
v82 10%-Kennzeichen alle
v90 Differenzierung Ernährer/Ernährte
v148 Anzahl Ernährte Männlich
v149 Anzahl Ernährte Weiblich
v150 Anzahl Ernährte Männlich
v151 Anzahl Ernährte Weiblich

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 Variable laufende Haushaltsnummer mit einer eindeutig systemfreien Nummerierung erstellt.

6. Vergröberung von Merkmalsausprägungen

Für alle Variablen des Public-Use-Files der Volkszählung der BRD von 1970 gilt, dass jede ausgewiesene Merkmalsausprägung in der univariaten Verteilung der Grundgesamtheit mindestens 1 000 Fälle umfassen muss. Ausgenommen hiervon sind die Variablen der Staatsangehörigkeit (mindestens 10 000 Fälle) sowie die Gemeindegrößenklasse der Wohnsitzgemeinde (mindestens 50 000 Fälle, s. hierzu Punkt 3 „Löschung von regionalen Informationen“). 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 BRD von 1970 zu entnehmen (s. Anhang).

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.