

## **Konzept zur Anonymisierung der Volkszählung (VZ) von 1971 der Deutschen Demokratischen Republik (DDR) zur Verwendung als Public-Use-File (PUF)**

### **I. Vorbemerkungen**

Die Volkszählungen in der ehemaligen 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 so genannte 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 2000 vom Statistischen Bundesamt an das Bundesarchiv abgegeben. Die 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 1971 zur Erstellung eines absolut anonymisierten Public Use File.

### **II. Basismaterial**

Das Basismaterial der Volkszählung 1971 umfasst zwei Datenfiles: die Personendaten und die Wohnungsdaten. 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 16,4 Mio. Personen, 6,2 Mio. Haushalte und 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 Bundesland, Regierungsbezirk, Kreis, Gemeinde, Stützpunkt, Zählbereich, Zählabschnitt, Gebäude und Wohnung zusammengeführt.

Nach der Zusammenführung der Daten findet eine Plausibilitätsprüfung mithilfe von in beiden Datenfiles vorhandenen Variablen gleichen Inhaltes statt. Diese sind „Anzahl der Hauptbewohner im 1. Haushalt“ sowie „Anzahl der Kinder unter 17. Jahren im 1. Haushalt“.

Datensätze von leerstehenden, als Nebenwohnung 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 1971:

##### 1. Alter der Daten

Die Volkszählung von 1971 liegt mittlerweile 36 Jahre zurück. Es kann also 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 ein erhebliches Anonymitätskriterium dar.

##### 2. Stichprobenziehung

Als erster Schritt der verfahrensseitigen Datenanonymisierung wird eine systematische 25% Haushaltsstichprobe, auf Basis des Schlussziffernverfahrens gezogen. Personen in Gemeinschaftsunterkünften zählen jeweils als ein Haushalt. Durch die Stichprobenziehung wird sichergestellt, dass ein potenzieller Datenangreifer nicht sicher sein kann, ob die gesuchte Person oder der gesuchte Haushalt sich in der Stichprobe befinden.

Zunächst wird das Originalmaterial nach Land, Regierungsbezirk, Anzahl der Personen im Haushalt, Kreis, Gemeinde, Stützpunkt, lfd. Nummer der GE oder des Zählbereichs und Zählabschnitts für PH, lfd. Nummer des Gebäudes im Zählabschnitt, Nummer der Person in GE oder lfd. Nr. der Wohnung im Gebäude für PH und Nummer des Haushaltes in der Wohnung sortiert und anschließend die Wohnungen mit einer laufenden Wohnungsnummer über den gesamten Datenfile versehen. Bei der Ziehung 25% Stichprobe werden die letzten zwei Endziffern der Wohnungsnummer verwendet ( $=X_i$ ). Die Auswahlwahrscheinlichkeit beträgt 25 aus 100 oder 1 aus 4. Zunächst wird im Intervall zwischen 0 und 3 eine Zahl Z zufällig ausgewählt. Ausgehend von diesem zufällig ausgewählten Startwert Z werden 25 Werte  $X_i$  im Intervall von 0 bis 99 nach der Formel:

$$X_i = Z + i * 4, \text{ mit } i=0,1,\dots,24.$$

ermittelt. Alle Wohnungen mit den Endziffernkombinationen  $X_i$  (von 0 bis 99) werden in die Stichprobe aufgenommen.

### 3. Löschung von regionalen Informationen

Als weitere Anonymisierungsmaßnahme werden alle regionalen Informationen bis auf Bundesland aus dem Datenmaterial gelöscht. Im Einzelnen handelt es sich dabei um die Variablen: Regierungsbezirk, Kreis, Gemeinde, Stützpunkt, Zählbereich und Zählabschnitt sowie die Gemeindefnummer des Pendlerziels.

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 (siehe hierzu den Abschnitt „systemfreie Sortierung“).

Die Gemeindegrößenklasse wird als zusätzliches Merkmal (vp60) generiert. Sie teilt die Gemeinden der Länder Brandenburg, Mecklenburg-Vorpommern, Sachsen, Sachsen-Anhalt und Thüringen in folgende 4 Größenklassen ein:

1:	unter	2 000 Einwohner
2:	2 000 bis unter	10 000 Einwohner
3:	10 000 bis unter	50 000 Einwohner
4:	50 000 Einwohner und mehr.	

Für das Land Berlin werden lediglich zwei Gemeindegrößenklassen ausgewiesen:

5:	unter 100 000 Einwohner
6:	100 000 Einwohner und mehr.

### 4. Löschung von weiteren Variablen

Die Ausprägungen der Variablen „Fläche der 3. Küche in der Wohnung“ bzw. „Fläche der 4. Küche in der Wohnung“ sind nicht genügend besetzt. Aus diesem Grund werden die zwei Variablen ebenfalls aus dem Datenmaterial gelöscht.

### 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. Vergrößerung von Merkmalsausprägungen

Für alle Variablen des Public Use File der VZ 1971 gilt, dass jede ausgewiesene Merkmalsausprägung mindestens 3 Fälle umfassen muss. Um diese Voraussetzung zu erfüllen wird eine sachgerechte Vergrößerung der betroffenen Merkmalsausprägungen vorgenommen.

Bei folgenden Variablen werden Vergrößerungen der Ausprägungen vorgenommen:

- vp21 - Alter:  
Ausprägungen von „100“ bis „105“ zusammengefasst zu „100 und älter“
- vw23 - Fläche der Wohnräume in der Wohnung in 1/10 qm:  
Ausprägungen von „1700“ bis „5816“ zusammengefasst zu „1700 und größer“
- vw24 - Fläche der Küchen in der Wohnung in 1/10 qm  
Ausprägungen von „448“ bis „449“ zusammengefasst zu „449“  
Ausprägungen von „450“ bis „799“ zusammengefasst zu „450 und größer“

- vw25 - Fläche der 1. Küche in der Wohnung in 1/10 qm  
Ausprägungen von „360“ bis „497“ zusammengefasst zu „360 und größer“
- vw26 - Fläche der 2. Küche in der Wohnung in 1/10 qm  
Ausprägungen von „230“ bis „497“ zusammengefasst zu „230 und größer“
- vw41 - Anzahl der Nebenbewohner im 1. HH  
Ausprägungen von „12“ bis „24“ zusammengefasst zu „12 und mehr“
- vw43 - Fläche der Wohnräume des 1. HH 1/10 qm  
Ausprägungen von „1599“ bis „4599“ zusammengefasst zu „1600 und größer“
- vw48 - Anzahl der Wohnräume des 2. HH  
Ausprägungen von „7“ bis „9“ zusammengefasst zu „7 und mehr“
- vw47 - Anzahl der Nebenbewohner im 2. HH, 7 und mehr\*/  
Ausprägungen von „7“ bis „8“ zusammengefasst zu „7 und mehr“
- vw49 - Fläche der Wohnräume des 2. HH in 1/10 qm  
Ausprägungen von „800“ bis „1800“ zusammengefasst zu „800 und größer“

Das Top-Coding der Variablen mit Flächenangaben wurde derart vorgenommen, dass die kleinste Ausprägung mit einer Besetzungszahl kleiner als 3 als Untergrenze definiert wurde und alle weiteren größeren Ausprägungen in diese Klasse zusammengefasst wurden.

## 7. Löschen von Fällen

Nach langer Recherche waren folgende für die Interpretation der Daten notwendigen Klassifikationen nicht aufzufinden:

- Schlüssel Facharbeiter und Meisterabschlüsse
- Schlüssel Fachschulgrundstudieneinrichtungen
- Schlüssel Hochschulgrundstudieneinrichtungen
- Schlüssel Wirtschaftsbereich, -sektor, -zweig
- Schlüssel Eigentumsform der Arbeitstätte.

Die Variable vp38 „Abgeschlossener Meisterberuf“, deren Interpretation auf den oben erwähnten Schlüsseln basiert, weist zwei Ausprägungen auf (262, 264), die sowohl in der Grundgesamtheit als auch in der Stichprobe eine Besetzungszahl kleiner als drei haben. Aufgrund der fehlenden Klassifikationen ist eine sinnvolle Zusammenfassung der zu gering besetzten Ausprägungen nicht möglich. Aus diesem Grunde werden die 2 Personen mit ihren Haushalten aus dem Datenmaterial gelöscht.

Ebenfalls werden ein Haushalt mit 15 Einkommensbeziehern sowie ein Haushalt mit 15 zu Unterstützenden je Einkommensbezieher als einzelne Fälle aus dem Datenmaterial gelöscht.

## 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.