C105 – Forschungsdatenzentrum Wiesbaden, März 2013

Konzept zur Anonymisierung der Mikrozensen 1991 und 2001 zur Verwendung als Public-Use-Files (PUF)

I. Vorbemerkungen

Das vorliegende Konzept befasst sich mit der absoluten Anonymisierung der Mikrozensen 1991 und 2001. Absolut anonymisierte Mikrodaten fallen unter §16 Abs. 1 Pkt. 4 des BStatG und sind vom Geheimhaltungsgebot ausgenommen. Vor einer Weitergabe der Daten als Public-Use-File (PUF) in absolut anonymisierter Form muss sichergestellt werden, dass eine Zuordnung der Einzelangaben zu den Merkmalsträgern nicht mehr möglich ist. Im Folgenden werden die, in den Mikrozensen 1991 und 2001 vorgenommen, Anonymisierungsmaßnahmen beschrieben, die zur absoluten Anonymität der Einzeldaten führen.

II. Ausgangsdatenfile

Der Mikrozensus ist eine jährlich erhobene Haushaltsbefragung über die Bevölkerung und den Arbeitsmarkt, an der 1% aller Haushalte in Deutschland beteiligt ist. Insgesamt nahmen rund 737.000 Personen in 320.000 Haushalten im Jahr 1991 und rund 720.000 Personen in 332.000 Haushalten im Jahr 2001 am Mikrozensus teil. Darunter waren 1991 etwa 168.000 Personen in 72.000 Haushalten und 2001 etwa 149.000 Personen in 71.000 Haushalten aus den neuen Bundesländern.

Alle Haushalte haben beim Mikrozensus die gleiche Wahrscheinlichkeit, in die Auswahl zu gelangen. Es wird eine einstufige geschichtete Flächenstichprobe durchgeführt, das heißt, aus dem Bundesgebiet werden Flächen (Auswahlbezirke) ausgewählt, in denen alle Haushalte und Personen befragt werden. Die Auswahlbezirke werden aus den Daten der Volkszählung 1987 gebildet. Für die neuen Bundesländer wurde auf der Basis des "Bevölkerungsregister Statistik" eine vergleichbare Auswahlgrundlage erstellt. Mit Hilfe der Bautätigkeitsstatistik wird die Auswahl aktualisiert. Jährlich wird ein Viertel der Auswahlbezirke ausgetauscht. Folglich bleibt jeder Auswahlbezirk vier Jahre in der Stichprobe (Verfahren der partiellen Rotation).

Das Frageprogramm des Mikrozensus besteht aus einem festen Grundprogramm mit jährlich wiederkehrenden Tatbeständen, die überwiegend mit Auskunftspflicht belegt sind. Darüber hin-

aus gibt es in vierjährigem Rhythmus Zusatzprogramme, die teilweise von der Auskunftspflicht befreit sind. Der Mikrozensus dient der Bereitstellung statistischer Informationen über die wirtschaftliche und soziale Lage der Bevölkerung sowie über die Erwerbstätigkeit, den Arbeitsmarkt und die Ausbildung (Mehrzweckstichprobe). Er schreibt die Ergebnisse der Volkszählung fort. Die Stichprobenerhebung über Arbeitskräfte in der Europäischen Union (Arbeitskräftestichprobe der Europäischen Union) ist seit 1968 in den Mikrozensus integriert.
Die plausibilisierten Einzeldaten der Mikrozensen 1991 und 2001 dienen als Datenbasis bei der Erstellung der PUF. In den Daten sind keine direkten Identifikationsmerkmale mehr enthalten, da diese bereits zu einem früheren Zeitpunkt entfernt wurden. Der aufbereitete Datensatz des Mikrozensus 1991 umfasst 312 Variablen, der des Mikrozensus 2001 umfasst 661 Variablen. Die höhere Anzahl der Variablen im Jahr 2001 ist jedoch nicht auf eine umfassende Ausweitung des Erhebungsprogramms zurückzuführen, sondern darauf, dass insbesondere seit dem Mikrozensus 1996 eine Vielzahl neuer abgeleiteter Variablen generiert werden.

III. Anonymisierungsmaßnahmen

Die Anonymisierungsmaßnahmen zur Erstellung der Mikrozensen 1991 und 2001 als PUF werden in Anlehnung an die Empfehlungen für die absolute Anonymisierung von Einzeldaten von Südfeld (1987)[1] an den, aus den plausibilisierten Einzeldaten erstellten Datenfiles durchgeführt.
Die Empfehlungen von Südfeld, die ein absolut anonymisiertes Datenfile mindestens erfüllen sollte, betreffen folgende Maßnahmen:

- Das absolut anonyme Datenfile soll nur eine Stichprobe aus dem Originaldatenfile sein
- Das Datenfile soll ein bestimmtes Mindestalter aufweisen und durch eine neuere Erhebung bereits überholt sein
- Die Datensätze sollen systemfrei angeordnet werden
- Direkte Identifikatoren sollen im Datenbestand nicht enthalten sein
- Regionalangaben sollen nur als Typisierungsangaben weitergegeben werden
- Jede Ausprägung eines einzelnen Merkmals soll eine Mindestbesetzungszahl aufweisen
- Sensible Merkmale sollen nur klassifiziert übermittelt werden

[1] Südfeld, E.: Anonymisierungsstandards und generelle Abwicklungsregelungen für Anforderungen nach anonymisierten Einzelangaben im Statistischen Bundesamt. In: „Nutzung von anonymisierten Einzelangaben aus Daten der amtlichen Statistik – Bedingungen und Möglichkeiten“, Schriftenreihe Forum der Bundesstatistik, Band 5, S.146 – 156, Wiesbaden 1987.

- Identifizierende Merkmale, über die sehr einfach Zusatzinformationen zu gewinnen sind, sollen nur klassifiziert übermittelt werden
- Die Kombination sensibler sowie identifizierender Merkmale soll eine Mindestbesetzungszahl aufweisen

Diese Empfehlungen dienen als Leitfaden für das Anonymisierungskonzept. Der Leitfaden muss an die Mikrozensen und speziell an die Erhebungsjahre 1991 und 2001 angepasst werden.
In den folgenden Kapiteln wird die Ausgestaltung der Empfehlungen für die Anonymisierung der Mikrozensen 1991 und 2001 dargestellt. Nach Anwendung der Maßnahmen ist eine Zuordnung von Einzelangaben zu Merkmalsträgern in den PUF nach menschlichem Ermessen auszuschließen.

1. Alter der Daten

Da die Datenbestände der Mikrozensen 1991 und 2001 bereits ein bestimmtes Mindestalter aufweisen und durch neuere Erhebungen überholt sind, kann angenommen werden, dass Zusatzinformationen nur in eingeschränktem Umfang verfügbar oder 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. Entfernen der direkten Identifikationsmerkmale

Die direkten Identifikationsmerkmale wurden aus den Mikrozensen 1991 und 2001 bereits zu einem früheren Zeitpunkt der Datenproduktion entfernt und sind im Originaldatenfile nicht enthalten.

3. Löschung von Variablen

Der für die PUF der Mikrozensen 1991 und 2001 ausgewählte Merkmalskranz richtet sich weitestgehend nach den Merkmalen, die in den SUF enthalten sind. Um die Anonymität der Einzelangaben zu gewährleisten, werden in die absolut anonymen Datenfiles bspw. folgende hoch sensible Variablen nicht in die PUF übernommen:

<u>Regionalangaben</u>

Regionale Informationen (bis auf Gebietseinheiten, Gemeindegrößenklassen und Gebietseinheiten der Arbeitsstätte – s. hierzu Abschnitt III.4) werden aus dem Originaldatenfile nicht in die PUF übernommen: Gemeindeschlüssel, Regierungsbezirk, Kreis, Gemeinde, Gemeinde-

teil und Regionallisten-Nummer. Um eine eindeutige Unterscheidung der Haushalte sowie Personen zu ermöglichen, werden diese mit einer eindeutigen, systemfreien laufenden Nummer versehen (s. Abschnitt III.7).

Auswahlbezirksnummer

Um keine Rückschlüsse auf Haushalte zuzulassen, die in direkter Nachbarschaft leben, wird die Auswahlbezirksnummer nicht in die PUF aufgenommen.

Geburtsmonat und Geburtsjahr

Die Merkmale Geburtsmonat und Geburtsjahr werden als hoch sensible Merkmale definiert, und gehen daher nicht in die PUF ein.

2. Staatsangehörigkeit

Die Angaben zur zweiten Staatsangehörigkeit werden als hoch sensibel betrachtet und daher nicht in die PUF übernommen.

Der SUF des Mikrozensus 1991 enthält 169 von 312 Variablen des Originaldatenfiles. Für den PUF 1991 wird der Merkmalskatalog des SUF nochmals reduziert und umfasst 129 Variablen. Im Rahmen der Anonymisierung werden also insgesamt 184 Variablen des Originalfiles nicht in den PUF übernommen[2]. Die gegenüber dem SUF nicht in den PUF aufgenommenen Variablen betreffen vor allem die Bereiche Erwerbsbeteiligung, Versicherung, Schicht- und Feiertagsarbeit, Einkommen und Unterhalt sowie Anzahl der Kinder im Heimatland.

Der SUF des Mikrozensus 2001 enthält 319 von 661 Variablen des Originalfiles. Auch hier wird für die Erstellung des PUF 2001 der Merkmalskatalog des SUF reduziert, nämlich auf 196 Variablen[2]. Im Rahmen der Anonymisierung werden demnach insgesamt 466 Variablen des Originaldatenfiles nicht in den PUF übernommen. Die nicht in den PUF aufgenommenen Variablen betreffen vor allem die Bereiche Erwerbsbeteiligung in der zweiten Erwerbstätigkeit, Schicht- und Feiertagsarbeit, Arbeitssuche, Fort- und Weiterbildung sowie Unterhalt und Einkommen.

Der ausgewählte Merkmalskranz findet sich in den Schlüsselverzeichnissen der Mikrozensen 1991 und 2001 PUF (s. Anhang).

[2] Die Anzahl der Variablen im PUF und die Anzahl der gelöschten Variablen addieren sich nicht genau auf die Anzahl der Variablen im Originaldatenfile, da die Variable lfd_nr (Laufende Nummer der Person) für die PUF neu generiert wurde und nicht im Originaldatenfile enthalten ist.

4. Vergröberung von Regionalangaben

Als Regionalangaben werden das Bundesland, die Gemeindegrößenklasse sowie das Bundesland der Arbeitsstätte weitergegeben. Die Merkmale werden wie folgt vergröbert:

4.1 Univariate Vergröberung

Bundesland

Das Merkmal Bundesland wird in drei Gebietseinheiten Deutschlands mit folgenden Bundesländern unterteilt:

Nord: Schleswig-Holstein, Hamburg, Niedersachsen, Bremen und Nordrhein-Westfalen

Süd: Hessen, Rheinland-Pfalz, Baden-Württemberg, Bayern und Saarland

Ost: Berlin, Brandenburg, Mecklenburg-Vorpommern, Sachsen, Sachsen-Anhalt und Thüringen

4.2 Bivariate Vergröberung

Gemeindegrößenklasse

Die Variable Gemeindegrößenklasse der Wohnsitzgemeinde wird – angelehnt an die Anonymisierungskonzepte zur Erstellung der Mikrozensen 1991 und 2001 SUF – neu aggregiert. Keine Gemeinde ist hiernach in der Grundgesamtheit mit weniger als 500.000 Einwohnern belegt. Zusätzlich hierzu sind in jeder Gemeindegrößenklasse einer Gebietseinheit (Nord, Süd, Ost) mindestens 400.000 Einwohner in der Grundgesamtheit vertreten. Die Variable der Gemeindegrößenklasse der Wohnsitzgemeinde teilt nun die Gemeinden in den Gebietseinheiten 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.

Bundesland der Arbeitsstätte

Das Bundesland der Arbeitsstätte wird ebenfalls in den drei Gebietseinheiten Nord, Süd und Ost ausgegeben (s. Bundesland).

5. Vergröberung von Merkmalsausprägungen

5.1 Bivariate Vergröberung

Neben der Vergröberung von Regionalangaben (s. Abschnitt III.4) werden alle Merkmalsausprägungen, deren Häufigkeiten in den Gebietseinheiten eine Besetzungszahl von unter 10.000 Fälle in der Grundgesamtheit aufweisen, vergröbert. Bei den Vergröberungen kann es durch die jeweilige Fallzahlbesetzung zu Unterschieden zwischen den Gebietseinheiten kommen[3]. Variablen, deren Ausprägungen aus anderen Variablen erschlossen werden können, werden nicht nach Fallzahlen vergröbert, sondern nur dann, wenn eine Einheitlichkeit mit der Ausgangsvariablen hergestellt werden muss.

Sensible und identifizierende Merkmale, über die sehr einfach Zusatzinformationen zu gewinnen sind (hier: Alter, Beruf, Wirtschaftszweig, Staatsangehörigkeit), werden in ihren Merkmalsausprägungen mit einer Mindestfallzahl von 50.000 bzw. bei der Staatsangehörigkeit von 100.000 Fällen in der Gebietseinheit der Grundgesamtheit ausgewiesen. Zusätzlich werden folgende Klassifikationen der sensiblen Merkmale erstellt:

<u>Alter</u>

Das Merkmal Alter in Jahren sowie alle weiteren Altersvariablen werden auf folgende Altersklassen vergröbert:

bis unter 3 Jahre
3 bis unter 6 Jahre
6 bis unter 10 Jahre
10 bis unter 15 Jahre
15 bis unter 18 Jahre
18 bis unter 20 Jahre
20 bis unter 25 Jahre
25 bis unter 30 Jahre
30 bis unter 35 Jahre
35 bis unter 40 Jahre
40 bis unter 45 Jahre
45 bis unter 50 Jahre
50 bis unter 55 Jahre
55 bis unter 60 Jahre

[3] Unter den zusätzlichen Vergröberungen für die Gebietseinheit Ost finden sich im Schlüsselverzeichnis in der Spalte "Umsetzungen" Angaben in eckigen Klammern (z.B. [= 2+3]). Diese Angaben beschreiben, welche Kategorien der Gebietseinheiten Nord und Süd zusammengefasst werden müssen, um die für die Gebietseinheit Ost ausgewiesene Kategorie zu erzeugen.

60 bis unter 63 Jahre
63 bis unter 65 Jahre
65 bis unter 70 Jahre
70 bis unter 75 Jahre
75 bis unter 80 Jahre
80 Jahre oder älter.

Die Variablen der Altersklassen "Zahl der Kinder von 18 bis unter 27 Jahren" und "Zahl der Kinder von 27 Jahren und älter" werden zusammengefasst und als "Zahl der Kinder von 18 Jahren und älter" herausgeben.

Beruf

Die Ausprägungen des Merkmals Beruf werden entsprechend der Klassifizierung der Berufe dreistellig – also auf der Ebene der Berufsordnungen – nachgewiesen. Zusätzlich werden Zusammenfassungen verwandter Ausprägungen vorgenommen, um die Mindestbesetzungszahl von 50.000 Personen der Grundgesamtheit in der Gebietseinheit zu erreichen.

Wirtschaftszweig

Die Ausprägungen des Merkmals Wirtschaftszweig werden entsprechend der Klassifizierung der Wirtschaftszweige zweistellig – also auf der Ebene der Wirtschaftsabteilungen – in den PUF aufgenommen; die dritte Stelle bleibt leer. Zusätzlich werden Zusammenfassungen verwandter Ausprägungen vorgenommen, um die Mindestbesetzungszahl von 50.000 Personen der Grundgesamtheit in der Gebietseinheit zu erreichen.

Staatsangehörigkeit

Als besonders stark identifizierend werden Variablen der Staatsangehörigkeit betrachtet, deren Merkmalsausprägungen für jede Nationalitätengruppe in der Grundgesamtheit nicht weniger als 100.000 Einwohner in der Gebietseinheit aufweisen sollen. Die Staatsangehörigkeit soll zudem nur dichotom mit den folgenden beiden Kategorien ausgegeben werden:

1) Deutschland sowie Deutschland und Ausland (MZ 1991)
   bzw. Deutscher ohne weitere Staatsangehörigkeit (MZ 2001)
2) Ausland oder Staatenlos (MZ 1991)
   bzw. Ausländer oder Staatenloser (MZ 2001)

5.2 Multivariate Vergröberung

In einem abschließenden Prozess wird überprüft, ob jede Zelle der Kreuztabellen der Merkmale Gebietseinheit × Altersklasse × dichotome Staatsangehörigkeit × (ggf. zuvor bivariat vergröbertes) Merkmal X mit mindestens drei Personen besetzt ist. Bei allen Beobachtungen, die dieses Kriterium nicht erfüllen, werden die Ausprägungen des (ggf. zuvor bivariat vergröberten) Merkmals X auf „keine Angabe“ gesetzt. Liegt diese Kategorie in einer Variablen nicht vor, so wird sie hierfür erzeugt.

Nähere Informationen zu den Vergröberungen finden sich in den Schlüsselverzeichnissen der Mikrozensen 1991 und 2001 PUF (s. Anhang).

6. Substichprobenziehung

Zur Erstellung der PUF der Mikrozensen 1991 und 2001 wird aus den plausibilisierten und vergröberten Datenfiles jeweils eine systematische 50% Haushaltsstichprobe mit einer Zufallskomponente auf Basis des Schlussziffernverfahrens gezogen. Mit der Substichprobenziehung wird die Möglichkeit der Reidentifikation eines Merkmalsträgers stark eingeschränkt. Unter der Annahme, dass ein/e potentielle/r Datenangreifer/in weiß, dass eine Person oder ein Haushalt im Mikrozensus befragt wurde, kann er/sie sich nicht sicher sein, ob sich diese Person bzw. dieser Haushalt in der Substichprobe befindet.

Um zu gewährleisten, dass die Substichprobe in bestimmten Merkmalen nur geringe zufallsbedingte Abweichungen von den Verteilungen des Originaldatenfiles aufweist, wird dieses zunächst nach Bundesland, Unterstichprobenkennung, Regierungsbezirk, Gemeindegrößenklasse, Zahl der Personen im Haushalt, Gebäudegrößenklasse und Zahl der Haushalte in der Wohnung sortiert. Anschließend werden alle Haushalte mit einer fortlaufenden Haushaltsnummer versehen. Personen in Gemeinschaftseinrichtungen werden bei der Substichprobenziehung wie Haushalte behandelt. Jede Person in einer Gemeinschaftseinrichtung bekommt daher ebenso eine fortlaufende Haushaltsnummer zugeteilt.

Zur Ziehung der 50% Substichprobe wird die Endziffer der Haushaltsnummer verwendet. Die Auswahlwahrscheinlichkeit beträgt 50 aus 100 oder 5 aus 10. Daher werden aus einem Intervall zwischen 0 und 9 fünf einfache Zufallszahlen (Z_i) ausgewählt. Die für die Stichprobenziehung genutzten Endziffern berechnen sich demnach entsprechend folgender Formel:

$$X_i = Z + ganzzahl\left(i * \frac{100}{50} \right)$$, mit i= 0 bis 9

Jeder Haushalt, dessen Endziffer eine der fünf Zufallszahlen aufweist, geht in die Substichprobe ein.

7. Systemfreie Sortierung

Aus der Anordnung der Datensätze lassen sich im Originaldatenfile indirekt Regionalinformationen ableiten. Um eine Re-Identifizierung der Regionalangaben auszuschließen werden die Datensätze systemfrei, d.h. nach einem nicht nachvollziehbaren System, sortiert. Um dennoch eine eindeutige Unterscheidung der Haushalte und Personen zu ermöglichen, werden diese mit einer neuen systemfreien Nummerierung versehen.

IV. Anpassung der Hochrechnungsfaktoren

Da es sich bei den PUF der Mikrozensen 1991 und 2001 jeweils um eine Substichprobe des Originaldatenfiles handelt, wurden die Hochrechnungsfaktoren des Ausgangsdatenfiles an die Substichprobe des PUF angepasst. Die Größe der Substichproben der PUF liegt bei 50% (50/100) der Ausgangsdatenfiles. Um wieder auf die ursprüngliche Stichprobe hochrechnen zu können wurden die Hochrechnungsfaktoren daher mit 100/50 multipliziert.
Um auf die Gesamtbevölkerung hochzurechnen, müssen die Hochrechnungsfaktoren nochmals mit 100 multipliziert werden, da es sich bei den Ausgangsdatenfiles der Mikrozensen um eine 1%-Stichprobe der Gesamtbevölkerung handelt.

V. Beschluss

Die unter Abschnitt III. beschriebenen Anonymisierungsmaßnahmen führen zu Mikrodatenfiles, bei denen eine De-Anonymisierung einzelner Merkmalsträger ausgeschlossen ist. Die Datensätze der Mikrozensen 1991 und 2001 sind damit absolut anonym und können in dieser Form als Public-Use-Files veröffentlicht werden.