C105 – Forschungsdatenzentrum Wiesbaden, Juli 2013

Konzept zur Anonymisierung der Mikrozensen 1973, 1982 und 1987 BRD zur Verwendung als Public-Use-Files (PUF)

I. Vorbemerkungen

Das vorliegende Konzept befasst sich mit der absoluten Anonymisierung der Mikrozensen der Bundesrepublik Deutschlands (BRD) von 1973, 1982 und 1987. 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 1973, 1982 und 1987 vorgenommen, Anonymisierungsmaßnahmen beschrieben, die zur absoluten Anonymität der Einzeldaten führen.

II. Ausgangsdatenfiles

Der Mikrozensus ist eine jährlich erhobene Haushaltsbefragung über die Bevölkerung und den Arbeitsmarkt, an der 1% aller Haushalte in Deutschland beteiligt ist. Insgesamt nahmen an den Mikrozensen der BRD

- 1973 rund 220.000 Haushalte mit 640.000 Personen
- 1982 rund 230.000 Haushalte mit 630.000 Personen
- 1987 rund 260.000 Haushalte mit 670.000 Personen teil.

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 gebildet. Mit Hilfe der Bautätigkeitsstatistik wird die Auswahl aktualisiert. Jährlich wird ein Viertel der Auswahlbezirke ausgetauscht. Seit 1977 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 seit 1974 in vierjährigem Rhythmus Zusatzprogramme, die teilweise von der Auskunftspflicht befreit sind[1]. 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 1973, 1982 und 1987 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.

III. Anonymisierungsmaßnahmen

Die Anonymisierungsmaßnahmen zur Erstellung der Mikrozensen 1973, 1982 und 1987 als PUF werden in Anlehnung an die Empfehlungen für die absolute Anonymisierung von Einzeldaten von Südfeld (1987)[2] 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
- 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

[1] Vor 1974 konnten Zusatzprogramme per Verordnung, auch separat, durchgeführt werden

[2] 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.

Diese Empfehlungen dienen als Leitfaden für das Anonymisierungskonzept. Der Leitfaden muss an die Mikrozensen und speziell an die Erhebungsjahre 1973, 1982 und 1987 angepasst werden. In den folgenden Kapiteln wird die Ausgestaltung der Empfehlungen für die Anonymisierung der Mikrozensen 1973, 1982 und 1987 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 1973, 1982 und 1987 bereits ein hohes Alter 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 1973, 1982 und 1987 bereits zu einem früheren Zeitpunkt der Datenproduktion entfernt und sind in den Ausgangsdatenfiles nicht enthalten.

3. Löschung von Variablen

Der für die PUF der Mikrozensen 1973, 1982 und 1987 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, Gebietseinheiten der Arbeitsstätte (nur 1973) bzw. Bundesland des früheren Wohnsitzes (nur 1987) – s. hierzu Abschnitt III.4) werden aus den Ausgangsdatenfiles nicht in die PUF übernommen: Gemeindeschlüssel, Regierungsbezirk, Kreis, Gemeinde, Gemeindeteil und Regionallisten-Nummer. Um eine eindeutige Unterscheidung der Wohnungen, 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.

Der Merkmalskatalog der Ausgangsdatenfiles wird für die Mikrozensen 1973, 1982 und 1987 PUF gegenüber den SUF nochmals stark reduziert. Folgende Übersicht zeigt die Anzahl der Variablen in den jeweiligen Datenfiles:

Erhebungsjahr	Ausgangs-datenfile	SUF	PUF
MZ 1973	164	141	105
MZ 1982	165	141	105
MZ 1987	260	192	145

Die gegenüber den SUF nicht in die PUF aufgenommenen Variablen betreffen vor allem die Bereiche Erwerbsbeteiligung, Versicherung, Einkommen und Unterhalt sowie Anzahl der Kinder im Heimatland (nur 1987).
Der ausgewählte Merkmalskranz für die absolut anonymisierten Datenfiles findet sich in den Schlüsselverzeichnissen der Jahre 1973, 1982 und 1987 (s. Anhang).

4. Vergröberung von Regionalangaben

Als Regionalangaben werden das Bundesland und die Gemeindegrößenklasse weitergegeben. Die Merkmale werden wie folgt vergröbert:

4.1 Univariate Vergröberung

Bundesland

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

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

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

4.2 Bivariate Vergröberung

Gemeindegrößenklasse

Die Variable Gemeindegrößenklasse der Wohnsitzgemeinde wird – angelehnt an die Anonymisierungskonzepte zur Erstellung der Mikrozensen 1973, 1982 und 1987 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) 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.

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. Variablen, deren Ausprägungen aus anderen Variablen erschlossen werden können, werden einheitlich so vergröbert, dass ein Vergleich der Variablen zu keiner Aufdeckung zusätzlicher Informationen führen kann.

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:

Alter

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
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 für den Mikrozensus 1987 PUF zusammengefasst und als "Zahl der Kinder von 18 Jahren und älter" herausgeben. Für die Jahrgänge 1973 und 1982 lag diese Zusammenfassung bereits in den Ausgangsdaten vor.

Beruf

Die Ausprägungen des Merkmals Beruf werden entsprechend der Klassifizierung der Berufe[3] 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[4] zweistellig – also auf der Ebene der Wirtschaftsabteilungen – in den PUF aufgenommen; die dritte Stelle bleibt leer. Zusätzlich werden Zusammenfassungen verwandter Aus-

[3] Mikrozensus 1973: KldB70, Mikrozensen 1982 und 1987: KldB75
[4] Mikrozensen 1973 und 1982: WZ71, Mikrozensus 1987: WZ79

prägungen 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
2) Ausland oder Staatenlos

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 1973, 1982 und 1987 PUF (s. Anhang).

6. Substichprobenziehung

Zur Erstellung des PUF der Mikrozensen 1973, 1982 und 1987 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 Ausgangsdatenfiles aufweist, wird dieses zunächst im Fall der Privathaushalte nach Bundesland, Regierungsbezirk, Gemeindegrößenklasse, Auswahlbezirksnummer und laufender Nummer des Haushalts im Auswahlbezirk, im Fall der Gemeinschaftseinrichtungen nach Bundesland, Regierungsbezirk, Auswahlbezirksnummer, laufender Nummer des Haushalts im Auswahlbezirk und laufender Nummer der Person im Haushalt

sortiert. Nach der Sortierung werden alle Haushalte mit einer fortlaufenden Haushaltsnummer versehen, wobei Personen in Gemeinschaftseinrichtungen wie Haushalte behandelt und entsprechend mit einer Haushaltsnummer versehen werden. Zur Anpassung der Mikrozensus Stichprobe an die Gesamtbevölkerung wird das Ausgangsdatenmaterial in Anpassungsschichten unterteilt. Aus dem Abgleich der Häufigkeitsverteilungen in den Anpassungsschichten mit denen der realen Bevölkerung ergeben sich Über- bzw. Unterrepräsentationen. In den Mikrozensen 1973, 1982 und 1987 werden diese Fehlbestände ausgeglichen, indem einzelne Personendatensätze gedoppelt oder gestrichen werden. Die aus diesem Verfahren entstehenden Dopplungsfälle erhalten im Rahmen der Substichprobenziehung die gleiche Haushaltsnummer wie die dazugehörigen Originalfälle. Auf diese Weise ist sichergestellt, dass keine Dopplungsfälle ohne Originalfall in die Substichprobe gelangen können.
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 Hochrechnung

Da es sich bei den PUF der Mikrozensen 1973, 1982 und 1987 jeweils um eine 50% Substichprobe der Ausgangsdatenfiles handelt, müssen die PUF, um wieder auf die ursprüngliche Stichprobengröße hochrechnen zu können, mit 100/50 multipliziert werden. Dies kann erfolgen, in-

dem eine Konstante mit dem Wert 2 generiert wird, die dann wie ein herkömmlicher Gewichtungsfaktor verwendet wird.
Um auf die Gesamtbevölkerung hochzurechnen, muss die jeweilige Substichprobe 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 1973, 1982 und 1987 sind damit absolut anonym und können in dieser Form als Public-Use-Files veröffentlicht werden.