Konzept zur Anonymisierung des Mikrozensus 2002 zur Verwendung als CAMPUS File (CF)

I. Vorbemerkungen

Das CAMPUS File (CF) ist ein absolut anonymes Public-Use-File (PUF), das speziell für Lehrende und Studierende erstellt wird. Seine Funktion besteht darin, die praktische Statistikausbildung mit amtlichen Einzeldaten anzureichern und damit den Hochschulen ein effektives Werkzeug für eine qualitativ hochwertige Lehre zu liefern.

Das vorliegende Konzept befasst sich mit der absoluten Anonymisierung des Mikrozensus 2002. Ausgehend von der Anonymisierungsmethodik für den Scientific-Use-File (SUF) des Mikrozensus 2002 (MZ 2002), der bereits die faktische Anonymität gewährleistet, werden hier zusätzliche Anonymisierungsmaßnahmen vorgesehen, die aufbauend auf die Maßnahmen des SUF zur absoluten Anonymität der Einzeldaten führen.

II. Basismaterial

Das Einzelmaterial des Mikrozensus 2002 dient als Ausgangsmaterial bei der Erstellung des CF. Das Erhebungsprogramm des Mikrozensus 2002 umfasst insgesammt 672 Variablen und 788.049 Datensätze.

III. Anonymisierungsmaßnahmen

In Anlehnung an die Empfehlungen von Südfeld (1987)[1] für die absolute Anonymisierung von Einzeldaten werden in diesem Kapitel Maßnahmen beschrieben, die durchgeführt am Originalmaterial, zur absoluten Anonymität des Mikrozensus 2002 führen.

Die Empfehlungen von Südfeld sind als Forderungen formuliert, die ein absolut anonymisiertes Datenmaterial mindestens erfüllen sollte. Diese sind im Einzelnen:

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

- Das absolut anonyme Material ist nur eine Stichprobe aus dem Originalmaterial
- Das Datenmaterial weist ein bestimmtes Mindestalter auf
- Die Datensätze sind systemfrei angeordnet
- Direkte Identifikatoren sind im Datenbestand nicht enthalten
- Regionalangaben werden nur als Typisierungsangaben weitergegeben
- Jede Ausprägung eines einzelnen Merkmals weist eine Mindestbesetzungszahl auf
- Sensible Merkmale werden klassifiziert übermittelt
- Identifizierende Merkmale, über die sehr einfach Zusatzinformationen zu gewinnen sind, werden nur klassifiziert übermittelt
- Die Kombination sensibler sowie identifizierender Merkmale weist eine Mindestbesetzungszahl auf

Diese Empfehlungen dienen als Leitfaden für das Anonymisierungskonzept. Der Leitfaden muss an den Mikrozensus und speziell an das Erhebungsjahr 2002 angepasst und mit Inhalt gefüllt werden. Die letztgenannte Empfehlung einer Mindestbesetzung in Hinblick auf die Kombination von sensiblen sowie identifizierenden Merkmale kommt bei der Generierung des Campusfiles nicht zur Anwendung. Denn, – wie im Folgenden dargestellt – wird bereits durch die ausserordentlich hohe Substichprobenziehung von 3,5 Prozent in Verbindung mit dem generellen Mindestbesetzungskriterium für alle Merkmale und dem nochmals deutlich höheren Mindestbesetzungskriterium für das Merkmal Staatsangehörigkeit, bereits eine ausreichend hohe Schutzwirkung erreicht wird.
In den folgenden Kapiteln wird die Ausgestaltung der Empfehlungen für die Anonymisierung des MZ 2002 dargestellt.

1. Stichprobenziehung

Als erster Schritt der Anonymisierung wird eine systematische 3,5% Wohnungsstichprobe, auf Basis des Schlussziffernverfahrens gezogen. Zunächst wird das Originalmaterial nach Bundesland, Regierungsbezirk, Gemeindegrößenklasse, Zahl der Personen in Privathaushalten, Auswahlbezirksnummer und laufende Nummer der Wohnung im Auswahlbezirk sortiert und anschließend die Wohnungen mit einer laufenden Wohnungsnummer über den gesamten Datenfile versehen. Bei der Ziehung der üblichen 70% Stichprobe beim SUF wird lediglich die letzte Endziffer der laufenden Haushaltsnummer benötigt und die Datensätze mit den Endziffern 2, 5 sowie 9 gelöscht.
Im Gegensatz zur 70% Stichprobe werden zur Erzeugung der 3,5% Stichprobe die letzten drei Endziffern verwendet. Die Auswahlwahrscheinlichkeit beträgt 35 aus 1000 oder 1 aus 1000/35. Zunächst wird im Intervall zwischen 0 und 1000/35 eine Zahl Z zufällig ausgewählt. Ausgehend von diesem zufällig ausgewählten Startwert Z werden 35 Werte X_i im Intervall von 0 bis 999 nach der Formel:

$$X_i = runden\left(Z + i * \frac{1000}{35} \right) \text{, mit i=0,1,...,34.}$$

ermittelt. Alle Wohnungen mit den Endziffernkombinationen X_i (d.h. 35 aus 1000) werden in die Stichprobe aufgenommen.

2. Mindestalter

Südfeld empfiehlt ein Mindesalter für die zu anonymisierende Einzeldaten. „In der Regel sollen die Angaben durch eine neue Erhebung bereits überholt sein." Der Mikrozensus 2002 erfüllt die Forderung nach dem Mindestalter zur Zeit seiner Anonymisierung in 2007.

3. 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 Auswahlbezirk, Gebäude, Haushalt, Wohnung sowie Person mit einer eindeutigen systemfreien Nummerierung versehen.

4. Entfernen der direkten Identifikationsmerkmale

Die direkten Identifikationsmerkmale wurden aus dem Mikrozensus 2002 bereits zu einem früheren Zeitpunkt der Datenproduktion entfernt und sind im Originalmaterial nicht enthalten.

5. Regionalangaben

Im CF werden das Bundesland und die Gemeindegrößenklasse als Regionalvariablen weitergegeben.
Beim Merkmal Gemeindegrößenklasse darf analog zum SUF keine einzelne Gemeinde mit 500.000 Einwohnern identifizierbar sein. Bei mehreren Gemeinden in einer Klasse müssen diese insgesammt in jedem Bundesland mindestens 400.000 Einwohner erfassen.

6. Hinreichende Besetzung der Merkmalsausprägungen

Für alle Variablen des CF gilt, dass jede ausgewiesene Merkmalsausprägung in der Grundgesamtheit mindestens 5.000 Fälle umfassen muss. Um diese Voraussetzung zu erfüllen ist eine sachgerechte Vergröberung der betroffenen Merkmalsausprägungen vorzunehmen. Diese Vergröberungen entsprechen den Maßnahmen der faktischen Anonymisierung des SUF, die ebenfalls eine Mindestbesetzung von hochgerechnet 5.000 Fällen vorsieht.
Bei der Anonymisierung des SUF werden folgende Methoden der Vergröberung von Ausprägungen angewandt:

- Bildung von Klassen

Bspw. bei Einkommen und Gemeindegrößenklasse

- Zusammenfassen von Ausprägungen mit verwandter Bedeutung auf Grundlage der Besetzungszahl
  Bspw. bei Merkmalen zur Staatsangehörigkeit, Beruf, Wirtschaftszweig, Miete, Wohnungsgröße, Wochenarbeitszeit, Stellung im Beruf u.a.
- Zusammenfassen von höchsten Ausprägungen auf Grundlage der Besetzungszahl (Top-Coding)
  Bspw. bei Merkmalen zur Hauhalts- bzw. Wohnungsgröße, Anzahl der Kinder, Zahl der Erwerbstätigen sowie der Erwerbslosen, Miete, Wohnungsgröße, Wochenarbeitszeit, Stellung im Beruf u.a.
- Zusammenfassen von niedrigsten Ausprägungen auf Grundlage der Besetzungszahl (Bottom-Coding)
  Bspw. bei Eheschliessungsjahr, Beginn der Erwerbstätigkeit u.a.
- Gemischte Maßnahmen (Klassenbildung und zusätzlich Top- und/oder Bottom Coding)
  Bspw. bei Merkmalen wie Miete und Alter

7. Sensible und identifizierende Merkmale

Der Forderung von Südfeld (1987) zur allenfalls klassifizierten Darstellung von sensiblen Merkmalen wie Gesundheit, Einkommen und Vermögen wird bereits in der Erhebungsmethode des Mikrozensus Rechnung getragen. Diese Merkmale werden in der Erhebung soweit klassifiziert erfasst, dass bei allen die Mindestanzahl von 50.000 Fällen bereits im Originalmaterial überschritten wird.

Unter identifizierenden Merkmalen werden diejenigen Angaben verstanden, über die sehr einfach Zusatzinformationen zu gewinnen sind. Diese definieren wir für den MZ 2002 als: Staatsangehörigkeit, Alter, Zahl der Kinder, Beruf und Wirtschaftszweig.

Im Einzelnen handelt es sich um die Variablen:

<u>Staatsangehörigkeit</u>
ef44 Staatsangehörigkeit
ef52 2. Staatsangehörigkeit

<u>Alter</u>
ef30 Alter
ef558 Alter der Haushaltsbezugsperson
ef593 Alter der Bezugsperson in der Familie

<u>Zahl der Kinder</u>

Haushalt:
ef528 bis ef531 Zahl der Kinder unter 15 nach 4 Altersstufen
ef532 bis ef534 Zahl der Kinder über 15 nach 3 Altersstufen
ef535 bis ef536 Zahl der Kinder über 15 nach 2 Altersstufen, die Schüler sind

Familie:
ef576 bis 579 Zahl der Kinder unter 15 nach 4 Altersstufen
ef580 bis 582 Zahl der Kinder über 15 nach 3 Altersstufen
ef583 bis 584 Zahl der Kinder über 15 nach 2 Altersstufen, die Schüler sind

Lebensgemeinschaft:
ef631 bis 634 Zahl der Kinder unter 15 nach 4 Altersstufen
ef635 bis 637 Zahl der Kinder über 15 nach 3 Altersstufen
ef638 bis 639 Zahl der Kinder über 15 nach 2 Altersstufen, die Schüler sind

Beruf
ef114 - Beruf in letzter Tätigkeit (= im Originalfile als ef114UG2)
ef128 - Beruf (gegenwärtige Tätigkeit) (= im Originalfile als ef128UG2)
ef190 - Beruf 2. Erwerbstätigkeit (= im Originalfile als ef190UG2)
ef563 - Ausgeübter Beruf der Haushaltsbezugsperson(= im Originalfile als ef563UG2)
ef683 - ISCO-Beruf (1. Erwerbstätigkeit) (= im Originalfile als ef683UG1)
ef688 - ISCO-Beruf frühere Erwerbstätigkeit (= im Originalfile als ef688UG1)

Wirtschaftszweig
ef115 - Wirtschaftszweig in der letzten Tätigkeit (= im Originalfile als ef115UG1)
ef129 - Wirtschaftszweig (gegenwärtige Tätigkeit) (= im Originalfile als ef129UG1)
ef191 - Wirtschaftszweig 2. Erwerbstätigkeit (= im Originalfile als ef191UG1)
ef562 - Wirtschaftszweig der Haushaltsbezugsperson (= im Originalfile als ef562UG1)
ef382 - Wirtschaftszweig Ende April 2001 (= im Originalfile als ef382UG1)
ef598 - Wirtschaftszweig der Bezugsperson in der Familie (= im Originalfile als ef598UG1)
ef614 - Wirtschaftszweig der Ehefrau der Bezugsperson in der Familie (= im Originalfile als ef614UG1)
ef662 - Wirtschaftszweig des Lebenspartners der Bezugsperson im Haushalt (= im Originalfile als ef662UG1)

Jede ausgewiesene Merkmalsausprägung dieser Variablen muss in ihrer univariaten Verteilung in der Grundgesamtheit mindestens hochgerechnet 10.000 Fälle umfassen. Die Ausprägungen abgeleiteter Variablen werden analog zu denen der Ursprungsvariablen zusammengefasst.
Die Ausprägungen der Variable Staatsangehörigkeit werden soweit zusammengefasst, dass jede Ausprägung eine Häufigkeit von hochgerechnet mindestens 1 Million Fälle aufweist.

Im Folgenden werden die Anonymisierungsmaßnahmen beschreiben, die eine hinreichende Vergöberung der identifizierenden Merkmale im CF sicherstellen, sodass die Möglichkeit einer Zuordnung von Merkmalen zu Merkmalsträgern ausgeschlossen werden kann.

7.1 Vergröberung

Die Vergröberung von Merkmalsausprägungen kann auf unterschiedlicher Art vorgenommen werden.

Weitergabe von zweistelligen Codes aus systematischen Klassifikationen

Die im SUF dreistellig ausgewiesenen Codes der Merkmale zum Wirtschaftszweig werden im Campus File allenfalls zweistellig weitergegeben. Darüber hinaus werden weitere Zusammenfassungen vorgenommen, um die geforderte Besetzungszahl von hochgerechnet 10.000 Fällen zu erreichen.

Zusammenfassen von Ausprägungen auf Grundlage der Besetzungszahl

- Zusammenfassung von Ausprägungen mit verwandter Bedeutung

Die Ausprägungen der identifizierenden Variable Staatsangehörigkeit (ef44) werden zu vier Klassen zusammengefasst: 1 = Deutsch, 2 = Eu-Ausland (Stand 2002), 3 = Türkisch und 4= Sonstige.

Um die Mindestanzahl von 10.000 Fällen je Ausprägung zu erreichen, werden einzelne Ausprägungen der Merkmale zu Beruf und Wirtschaftszweig weiter zusammengefasst.

- Zusammenfassung von Ausprägungen mit den höchsten Werten – Top Coding

Die höchsten Ausprägungen des Merkmals Alter werden in die Kategorie 95 Jahre und älter zusammengefasst.

7.2 Löschung von Variablen

In den CF werden nur diejenigen Merkmale übernommen, die auch im SUF enthalten sind. Die Variable 2. Staatsangehörigkeit wird zusätzlich aus dem Material gelöscht.

Nähere Informationen zu den Vergöberungen finden sich im Schlüsselverzeichnis (fdz_mikrozensus_cf_2002_schluesselverzeichnis.pdf).

IV. Anpassung der Hochrechnungsfaktoren an die geringe Stichprobengröße

Die Hochrechungsfaktoren für Personen (ef750), Haushalte (ef751) sowie Wohnungen (ef761) werden an die Stichprobe des CAMPUS File nach der Methode der gebundenen Hochrechnung angepasst. Die Erzeugung der gebundenen Hochrechnungsfaktoren ef750g, ef751g, ef761g geschieht nach Anpassungsklassen. Die Anpassungsklassen entstehen durch die Bildung von Schichten nach Bundesland (ef1), Staatsangehörigkeit (ef52 - Ausprägungen zu zwei Klassen - Deutsch und Ausländer - zusammengefasst) und Geschlecht (ef32). Die Kombination der Variablen ergibt insgesammt 64 Schichten. Die Hochrechnungsfaktoren werden sowohl im Originalfile als auch im CAMPUS File pro Schicht aufsummiert. Der Quotient aus der Summe der Hochrechnungsfaktoren in der Schicht i im Originalfile und der Summe der Hochrechnugnsfaktoren in derselben Schicht i im CAMPUS File ist das Gewicht der Schicht i. Es entstehen also 64 verschiedene Gewichte.
Die gebundenen Hochrechnungsfaktoren (ef750g, ef751g und ef761g) der Schicht i berechnet man durch Multiplikation der Hochrechnungsfaktoren (ef750, ef751 und ef761) mit dem Gewicht der Schicht i.
Durch die Erzeugung von gebundenen Hochrechnungsfaktoren nach Anpassungsklassen ist eine nahezu verzerrungsfreie Hochrechnung der Werte aus dem CAMPUS File auf die Gesamtbevölkerung möglich. Dies zeigt sich auch in der differenzierten regionalen Analyse (Bundesländer). Die Gesamtbevölkerung nach Bundesländern wird mit der gebundenen Hochrechnung in keinem Fall um mehr als 1,4 % im Vergleich zum Mikrozensus Originalmaterial verfehlt. Selbst die sehr differenzierten Auszählungen nach Bundesland, Geschlecht und Altersklasse zeigen nur für sehr gering besetzte Zellen größere Differenzen.

V. Fazit

Die in vorliegendem Konzept beschriebenen Anonymisierungsmaßnahmen führen zur absoluten Anonymität des CF MZ 2002.