Forschungsdatenzentrum Düsseldorf Düsseldorf, 24. November 2014

Konzept zur Anonymisierung des Mikrozensus 2010 zur Verwendung als CAMPUS-File

I. Vorbemerkungen

Das CAMPUS-File ist ein vollständig anonymisiertes Public-Use-File (PUF), welches eigens 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 2010. Ausgehend von der Anonymisierungsmethodik für das Scientific-Use-File (SUF) des Mikrozensus 2010, der bereits die faktische Anonymität gewährleistet, werden hier zusätzliche Anonymisierungsmaßnahmen durchgeführt, die u.a. auf den Maßnahmen der PUF basieren und zur absoluten Anonymität der Einzeldaten führen. Grundsätzlich werden im CAMPUS-File des Mikrozensus 2010 nur Merkmale übernommen, die auch im SUF des Mikrozensus 2010 enthalten sind.

II. Basismaterial

Das plausibilisierte Einzelmaterial des Mikrozensus 2010 dient als Ausgangsmaterial bei der Erstellung des CAMPUS-Files. Das Erhebungsprogramm des Mikrozensus 2010 umfasst insgesamt 828 Variablen und 770.344 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 2010 führen.

Südfeld empfiehlt im Einzelnen:

- 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

[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.

- 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 2010 angepasst und mit Inhalt gefüllt werden. Die letztgenannte Empfehlung zu einer Mindestbesetzung in Hinblick auf die Kombination von sensiblen und identifizierenden Merkmalen, kommt bei der Generierung des CAMPUS-Files nicht zur Anwendung. Denn durch die außerordentlich hohe Substichprobenziehung von 3,5 Prozent in Verbindung mit dem generellen Mindesbesetzungskriterium für alle Merkmale, sowie dem nochmals deutlich höheren Mindestbesetzungskriterium für das Merkmal Staatsangehörigkeit und der starken, regionalen Vergröberung – das Merkmal EF1 ‚Land der Bundesrepublik' ist nur dichotom vorhanden – wird bereits eine ausreichend hohe Schutzwirkung erreicht.
In den folgenden Kapiteln wird die Ausgestaltung der Empfehlungen für die absolute Anonymisierung des Mikrozensus CAMPUS-File 2010 dargestellt. Nach Anwendung der Maßnahmen ist eine Zuordnung von Einzelangaben zu den Merkmalsträgern im CAMPUS-File nach menschlichem Ermessen auszuschließen.

1. Mindestalter

Südfeld empfiehlt ein Mindestalter für die zu anonymisierende Einzeldaten. In der Regel sollen die Angaben durch eine neue Erhebung bereits überholt sein. Diese Forderung ist für den Mikrozensus 2010 erfüllt.

2. Entfernen der direkten Identifikationsmerkmale

Die direkten Identifikationsmerkmale wurden aus dem Mikrozensus 2010 bereits zu einem früheren Zeitpunkt der Datenproduktion entfernt und sind im Originalmaterial nicht enthalten. Folglich sind diese auch nicht im CAMPUS-File vorhanden.

3. Vergröberung der Regionalangaben

Im CAMPUS-File wird als Regionalvariable ausschließlich das Merkmal EF1 ‚Land der Bundesrepublik' dichotomisiert (*Ost/West)* weitergegeben.

4. Generelle Vergröberungen von Merkmalen

Für alle Merkmale des CAMPUS-Files gilt, dass jede ausgewiesene Merkmalsausprägung in der Grundgesamtheit, d.h. für Deutschland gesamt, hochgerechnet mindestens 5.000 Fälle umfassen muss. Um eine sachgerechte Vergröberung der Merkmalsausprägungen vorzunehmen, wurden folgende Maßnahmen durchgeführt:

Bildung von Klassen
Bspw.: EF140 Wunsch nach weniger Arbeitsstunden: Anzahl

Zusammenfassen von Ausprägungen mit verwandter Bedeutung auf Grund der Besetzungszahl
Bspw.: EF87 Erwerbsunterbrechung (Berichtsw.): Bezug von Gehalt oder soz. Unterstützung

Zusammenfassen von höchsten Ausprägungen auf Grund der Besetzungszahl (Top-Coding)
Bspw.: EF131 Normalerweise geleistete Arbeitszeit je Woche (einschließlich regelmäßig geleisteter Überstunden) (Stunden)

Zusammenfassen von niedrigsten Ausprägungen auf Grund der Besetzungszahl (Bottom-Coding)
Bspw.: EF314 Höchster berufl. Abschluss: Jahr

Gemischte Maßnahmen (Klassenbildung und zusätzlich Top- und/oder Bottom Coding)
Bspw.: EF217 Tatsächlich geleistete Arbeitszeit in der Berichtswoche (Stunden) 2. Erwerbstätigkeit

Nähere Informationen zu den Vergröberungen finden sich im Schlüsselverzeichnis.

5. Vergröberungen sensibler und identifizierender Merkmale

Sensibel und identifizierend sind solche Merkmale, über die sehr einfach Zusatzinformationen zu gewinnen sind (hier: *Staatsangehörigkeit/Migration, Wirtschaftszweig, Beruf/Erwerbstätigkeit*). Diese wurden entweder nicht aus dem SUF übernommen, oder aber so vergröbert, dass nach menschlichem Ermessen eine Reidentifikation einzelner Personen nicht mehr möglich ist. Jede Merkmalsausprägung umfasst hochgerechnet mindestens 10.000 Fälle in der Grundgesamtheit, d.h. in Deutschland gesamt.

Die Merkmale zur *Staatsangehörigkeit* sowie zur *Migration* wurden dichotomisiert und umfassen hochgerechnet mindestens 1 Mio. Fälle in der Grundgesamtheit. Die Merkmale zu den *Wirtschaftszweigen* werden nur auf Zweisteller-Ebene weitergegeben. Die Merkmale zum *Alter* wurden so vergröbert, dass es eine Kategorie für Hochbetagte ab dem 95. Lebensjahr gibt. Darüber hinaus wurden weitere sensible und identifizierende Merkmale gelöscht. Enthalten sind im CAMPUS-File die nachfolgenden, vergröberten sensiblen und identifizierenden Merkmale.

Staatsangehörigkeit/Migration

EF367	Zuzugsjahr (dichotomisiert)
EF368	Deutsche Staatsangehörigkeit vorhanden (dichotomisiert)

Alter

EF44	Alter
EF47	Geburtsjahr

Wirtschaftszweig

EF105	Wirtschaftszweig in der letzten Tätigkeit
EF137	Wirtschaftszweig (gegenwärtige Tätigkeit)
EF214	Wirtschaftszweig weitere Erwerbstätigkeit
EF445	Wirtschaftszweig vor 12 Monaten (freiwillige Beantwortung)

Beruf/Erwerbstätigkeit

EF136	Beruf (1. Erwerbstätigkeit) - ISCO-88 (COM) (3-Steller)
EF517	Beruf (1. Erwerbstätigkeit) - ISCO-88 (COM) (3-Steller)

Nähere Informationen zu den Vergröberungen finden sich im Schlüsselverzeichnis.

6. Stichprobenziehung

Auf Basis des Schlussziffernverfahrens wird eine systematische 3,5% Wohnungsstichprobe gezogen. Zunächst wird das Originalmaterial nach *Bundesland*, *Regierungsbezirk*, *Gemeindegrößenklasse*, *Zahl der Personen in der Wohnung*, *Auswahlbezirksnummer*, *lfd. Nr. des Gebäudes* und *lfd. Nr. der Wohnung im Gebäude* sortiert und anschließend die Wohnungen mit einer laufenden Wohnungsnummer über das gesamte Datenfile versehen.
Bei der Ziehung der 3,5-Prozent-Stichprobe werden 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:

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

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

7. 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. Im Anschluss daran, werden die Merkmale *Auswahlbezirksnummer*, *lfd. Nr. des Haushalts*, sowie *lfd. Nr. der Person im Haushalt* mit einer eindeutigen systemfreien Nummerierung versehen.

IV. Anpassung der Hochrechnungsfaktoren an die geringe Stichprobengröße

Der Standardhochrechnungsfaktor Jahr (EF952g), sowie der Hochrechnungsfaktor Jahr für die Wohnsituation der Haushalte (EF960g) werden an die Stichprobe des CAMPUS-File nach der Methode der gebundenen Hochrechnung angepasst. Die Erzeugung der gebundenen Hochrechnungsfaktoren EF952g und EF960g geschieht nach Anpassungsklassen. Die Anpassungsklassen entstehen durch die Bildung von Schichten nach *Bundesland* (EF1), *Staatsangehörigkeit* (EF368 – Ausprägungen zu zwei Klassen – Deutsch und Ausländer – zusammengefasst) und *Geschlecht* (EF46). Die Kombination der Variablen ergibt insgesamt 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 Hochrechnungsfaktoren in derselben Schicht i im CAMPUS-File ist das Gewicht der Schicht i. Es entstehen also 64 verschiedene Gewichte.
Die gebundenen Hochrechnungsfaktoren (EF952g, EF960g) der Schicht i berechnet man durch Multiplikation der Hochrechnungsfaktoren (EF952, EF960) 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.

V. Endprodukt

Das CAMPUS-File ist eine 3,5%-Stichprobe des Mikrozensus 2010. Darin enthalten sind Angaben zu 23.374 Personen aus 11.494 Haushalten. Insgesamt gingen 427 der im Mikrozensus enthaltenen 828 Variablen in das absolut anonyme Grunddatenfile ein.

VI. Fazit

Die in dem vorliegenden Konzept beschriebenen Anonymisierungsmaßnahmen führen zur absoluten Anonymität des Mikrozensus CAMPUS-File 2010.