STATISTISCHE ÄMTER
DES BUNDES UND DER LÄNDER
FORSCHUNGSDATENZENTREN

29.08.2006

CAMPUS File Mikrozensus 1998 – Beschreibung der Anonymisierungsmethodik

Vorbemerkungen

Das CAMPUS File ist ein Public Use File (PUF), das speziell für Lehrende und Studierende erstellt wird. Die Idee hinter dem CAMPUS File ist es, die praktische Statistikausbildung mit amtlichen Einzeldaten anzureichern und damit den Hochschulen ein effektives Werkzeug für eine qualitativ hochwertige Lehre zu liefern.

Originalmaterial in Form von Einzeldaten ist momentan – mit Ausnahme des ALLBUS und des SOEP – an Universitäten praktisch nicht zu finden. Als Übungsdatensätze dienen oft selbst erstellte, kleine Datensätze, die die Anforderungen, die z.B. die Mikrosimulation mit einer großen Datenmenge stellt, nahezu nicht zu erfüllen vermögen. Hier schaffen nun die Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder mit Unterstützung des Bundesministeriums für Bildung und Forschung Abhilfe.

Das CAMPUS File ist durch verschiedene Anonymisierungsmaßnahmen so verändert, dass der Datenschutz einzelner Personen in der Grundgesamtheit voll gewährleistet ist. Trotzdem ermöglicht es – neben der Methodenausbildung - auch modellhaft Studien zu vielen, insbesondere in den Sozialwissenschaften, relevanten Fragestellungen durchzuführen. Mit Hilfe der vielfältig enthaltenen sozioökonomischen Merkmale zur Familienstruktur oder zum Erwerbsverhalten sind unter anderem Analysen zu „Formen des Zusammenlebens, Haushalts- und Familientypen“ oder zur „Erwerbsbeteiligung in Deutschland“ möglich. Natürlich muss man hierbei berücksichtigen, dass die kleine Stichprobe aus dem Originalmaterial, die das CAMPUS File darstellt, einen hohen Stichprobenfehler impliziert, weshalb bestimmte Verteilungen von den in der Grundgesamtheit erwarteten Verteilungen abweichen können.

Es ist aber ohne weiteres möglich, z.B. eine praktisch orientierte Hausarbeit mit Hilfe eines CAMPUS Files zu erstellen und eine darauf aufbauende Diplomarbeit dann mit den in den Forschungsdatenzentren (FDZ) bereitgestellten Scientific Use Files, oder mit anderen in den FDZ angebotenen Analysemöglichkeiten, wie dem Fernrechnen oder einem Gastwissenschaftlerarbeitsplatz, anzufertigen.

I. Basismaterial
   Ausgangspunkt für dieses CAMPUS File war das Originalmaterial des Mikrozensus 1998. Der Mikrozensus ist eine repräsentative 1-%-Bevölkerungsstichprobe.

II. Endprodukt
   Das CAMPUS File ist eine 3,5%-Wohnungsstichprobe des Mikrozensus 1998 gezogen aus dem Originalmaterial. Darin enthalten sind Angaben zu 25.410 Personen aus 11.771 Haushalten und 11.668 Wohnungen. Insgesamt gingen 199 Variablen des Originalmaterials und des Scientific Use File[1] (SUF) in das absolut anonyme Grunddatenfile ein. Drei neue Variablen, die angepassten Hochrechungsfaktoren für Personen, Haushalte und Wohnungen wurden erzeugt.

[1] Der Scientific Use File (SUF) des Mikrozensus 1998 ist eine 70%-Stichprobe des Originalmaterials und umfasst Informationen zu bevölkerungs- und arbeitsmarktstatistischen Strukturdaten von ca. 511.000 Personen in 231.000 Haushalten. Der SUF enthält 332 der ursprünglichen 757 Eingabefelder.

Anonymisierungsmaßnahmen

Der CAMPUS File ist eine absolut anonymisierte Stichprobe des Mikrozensus Originaldatenfiles. Die Maßnahmen zur Erreichung der absoluten Anonymität des CAMPUS Files bauen auf Anonymisierungsmaßnahmen zur Erreichung der faktischen Anonymität des SUF auf. Über die beim SUF angewandten Maßnahmen hinaus werden Maßnahmen wie die Ziehung einer im Vergleich zum SUF kleineren Stichprobe, die weitere Vergröberung von Merkmalen und die zusätzliche Löschung von Variablen durchgeführt, um die absolute Anonymität zu erreichen.

Stichprobenziehung

Als erster Schritt der Anonymisierung wurde eine systematische 3,5% Wohnungsstichprobe mit einer Zufallskomponente, auf Basis des Schlussziffernverfahrens gezogen. Zunächst wurde 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 wurde lediglich die letzte Endziffer der laufenden Haushaltsnummer benötigt und die Datensätze mit den Endziffern 2, 5 sowie 9 gelöscht.

Im Gegensatz zur 70% Stichprobe wurden zur Erzeugung der 3,5% Stichprobe die letzten drei Endziffern verwendet. Die Auswahlwahrscheinlichkeit betrug 35 aus 1000 oder 1 aus 28,6. Zunächst wurde im Intervall zwischen 0 und 28 eine Zahl zufällig ausgewählt (Zufallszahl=15). Ausgehend von diesem zufällig ausgewählten Startwert =15 wurden 35 Folgewerte X_i im Intervall von 0 bis 999 nach der Formel:

$$X_i = 15 + ganzzahl\left(i * \frac{1000}{35} \right)$$, mit i = 0 bis 34.

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

Vergröberung

Neben der Stichprobenziehung wurden kritische Merkmale, deren Häufigkeiten im Originalmaterial eine zu geringe Besetzungszahl aufwiesen, vergröbert. Beispielhaft ist hier ein an den SUF angepasstes Top- und Bottom Coding der Variablen Alter, Einkommen, Staatsangehörigkeit usw. zu nennen. Die Ländergliederung nach Bundesland (ef1) wurde zu einer regionalen Gliederung nach Ost (neue Bundesländer und Ost-Berlin) und West (alte Bundesländer und West-Berlin) vergröbert. Nähere Informationen zu den Vergröberungen finden sich im Schlüsselverzeichnis (schluessel_campus_1998.pdf).

Systemfreie Sortierung

Als eine zusätzliche Anonymisierungsmaßnahme wurde der Datensatz 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.

Löschung

Eine dritte beim CAMPUS File angewandte Maßnahme zur absoluten Anonymisierung war die Löschung von 370 Variablen aus dem Originalmaterial.

Eine detaillierte Übersicht über die im CAMPUS File enthaltenen Variablen findet sich in der Datei „fdz_mikrozensus_cf_schluesselverzeichnis_98.pdf“.

Die Anpassung der Hochrechnungsfaktoren

Die Hochrechungsfaktoren für Personen (ef750), Haushalte (ef751) sowie Wohnungen (ef757) wurden an die Stichprobe des CAMPUS File nach der Methode der gebundenen Hochrechnung angepasst. Die Erzeugung der gebundenen Hochrechnungsfaktoren ef750g, ef751g, ef757g geschah nach Anpassungsklassen. Die Anpassungsklassen entstanden durch die Bildung von Schichten nach Bundesland (ef1), Staatsangehörigkeit (ef43 - Ausprägungen zu zwei Klassen - Deutsch und nicht Deutsch - zusammengefasst) und Geschlecht (ef32). Die Kombination der Variablen ergab insgesamt 52 Schichten. In 10 von 16 Bundesländern wurde nach Deutsch/nicht Deutsch getrennt. In den restlichen 6 Bundesländern (Saarland und neue Bundesländer) wurde nicht nach Deutsch/nicht Deutsch getrennt, da im CAMPUS File die Ausprägung "nicht Deutsch" nicht genügend besetzt war (nicht zweistellig). Durch die Kombination der Variablen Bundesland und Staatsangehörigkeit entstanden 26 Schichten. Nach weiterer Unterteilung der Schichten nach Geschlecht entstanden die 52 Schichten.

Die Hochrechnungsfaktoren wurden 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 52 verschiedene Gewichte.
Die gebundenen Hochrechnungsfaktoren (ef750g, ef751g und ef757g) der Schicht i berechnete man durch Multiplikation der Hochrechnungsfaktoren (ef750, ef751 und ef757) mit dem Gewicht der Schicht i.

Materialien

Das CAMPUS File liegt in verschiedenen Formaten zum Download bereit:

CSV-Rohdatenmaterial :	„fzd_mikrozensus_cf_1998_ascii-csv.zip“
SPSS Version 13 mit Labels:	„fzd_mikrozensus_cf_1998_spss.zip“
STATA Version 8.2 :	„fzd_mikrozensus_cf_1998_stata.zip“
SAS Version 8 für Windows:	„fzd_mikrozensus_cf_1998_sas.zip“