Anonymisierung der PHF-Daten (Private Haushalte und ihre Finanzen)

Martin Eisele (Deutsche Bundesbank)

Für die Bereitstellung der PHF-Daten als Scientific Use File (SUF) wurden verschiedene Anonymisierungsmaßnahmen getroffen mit dem Ziel, einen Datensatz zur Nutzung durch die wissenschaftliche Forschung zu erstellen. Dabei wurden unter Beachtung der rechtlichen Erfordernisse verschiedene in den Rohdaten vorhandene Variablen weggelassen oder vergröbert. Die diesbezüglichen Datenschutzmaßnahmen orientieren sich an der gegenwärtigen Praxis der faktischen Anonymisierung von Befragungsdaten bei gleichzeitig größtmöglicher Erhaltung des Analysepotenzials. Zudem ist bei der Anonymisierung des SUF zu beachten gewesen, dass es sich beim PHF um eine Panel-Studie handelt, und daher Änderungen in den teilnehmenden Haushalten über die Zeit anhand der Daten nachvollziehbar sein werden. Diese Änderungen, insbesondere in der Haushaltszusammensetzung, bergen ein zusätzliches Reidentifikationsrisiko, dem bereits proaktiv in der ersten Befragungswelle durch stärkere Anonymisierungsmaßnahmen begegnet wurde als dies für eine Querschnittstudie ohne Panel-Komponente nötig gewesen wäre.

<u>Designinformationen:</u>

Die Indikatoren für die drei Schichten der Stichprobenziehung auf der ersten Stufe (vermögende und sonstige kleinere Gemeinden, Großstädte) sind im Datensatz vorhanden; ebenso die beiden Schichten der zweiten Stufe für die Ziehung in Großstädten (vermögende und sonstige Straßenabschnitte).

Die Point-Nummer, die den Clustern des Stichprobenmodells entspricht, ist ebenfalls im Datensatz enthalten. Haushalte mit derselben Point-Nummer liegen geographisch nahe beieinander; ansonsten enthält die Point-Nummer keine regionale Information.

Das SUF enthält die finalen Gewichte, die aus Design-Gewichtung, Nonresponse-Adjustierung und Kalibrierung hervorgingen.

Die folgenden Maßnahmen betreffen einzelne Variablen:

Regionale Merkmale:

Die Bundesländer wurden zu vier Großregionen Nord, Süd, West, Ost zusammengefasst. Die Region Nord umfasst Bremen, Hamburg, Niedersachsen und Schleswig-Holstein. Die Region Süd umfasst Baden-Württemberg, Bayern und Hessen. Die Region West umfasst Nordrhein-Westfalen, Rheinland-Pfalz und das Saarland. Die Region Ost umfasst Berlin, Brandenburg, Mecklenburg-Vorpommern, Sachsen, Sachsen-Anhalt und Thüringen.

Die politische Gemeindengrößenklasse wurde von sieben auf fünf Kategorien reduziert (Klassen 1 und 2 sowie 4 und 5 zusammengelegt). Bei den vermögenden kleinen Gemeinden in Ostdeutschland wurde die politische Gemeindengrößenklasse gelöscht.

Die BIK-Regionsgrößenklasse wurde von zehn auf fünf Kategorien reduziert (Klassen 5 und 7, 6 und 8, sowie 1-4 zusammengelegt). Bei den vermögenden kleinen Gemeinden in Ostdeutschland wurde die BIK-Regionsgrößenklasse gelöscht.

In Hinblick auf die Vergröberung der Regionalinformationen sind die Anonymisierungsmaßnahmen im Vergleich zu den bei Scientific Use Files üblichen Standards weitergehend. Die damit potenziell einhergehende Einschränkung des Analysepotenzials zugunsten des Datenschutzes der Befragten ist beabsichtigt. Auf diese Weise konnte das Analysepotenzial bei den im Zentrum des Forschungsinteresses stehenden Merkmalen zu Vermögen und Finanzen weitgehend erhalten werden.

Demographische Merkmale:

Die Beziehungen der Haushaltsmitglieder wurden vergröbert in folgende sechs Kategorien: Ehemann/-frau, Partner, (Schwieger-/Adoptiv-/Stief-)Eltern, (Adoptiv-/Stief-)Kinder, sonstige Verwandte, nicht verwandt.

Das Geburtsland und die Nationalität wurden in die Kategorien Deutschland, Euro-Länder ohne Deutschland, EU ohne Euro-Länder, Staaten der ehem. UdSSR ohne Baltikum, Rest-Europa inklusive Türkei, Amerika und Rest der Welt vergröbert. Die dritte Staatsbürgerschaft wurde entfernt.

Dem Alter der über 70-jährigen wurde ein stochastischer Fehlerterm hinzugefügt, so dass die Altersangabe im SUF um bis zu +/- 2 Jahre um das wirkliche Alter streut. Die Verteilung des Alters bei den über 70-jährigen in der Stichprobe wurde dadurch nur sehr geringfügig verändert. Zudem wurde ein Top-Coding des Alters bei 90 vorgenommen.

Alle Variablen, die in direktem logischen Zusammenhang mit dem Alter einer Person stehen (z.B. das Geburtsjahr) wurden zur Erhaltung der Datenkonsistenz entsprechend der Anonymisierungsmaßnahme für das Alter angepasst.

Beschäftigung / Unternehmen:

ISCO-Codes (dem ISCO 88-Standard entsprechend) sind in der Regel als Zweisteller ausgewiesen. Bei einigen Personen war nur eine einstellige Codierung möglich, und in wenigen Fällen ist als Anonymisierungsmaßnahme nur eine Stelle angegeben.

NACE-Codes werden gemäß NACE Rev. 2 mit den Buchstaben A bis U ausgewiesen; im Datensatz sind diese mit den Zahlen von 1 bis 21 codiert.

Die Anzahl der Beschäftigten in den eigenen Unternehmen (Variablen hd0501, hd0502, hd0503) wurde in vier Kategorien vergröbert (1,2-3,4-9,10+).

Die Rechtsform eigener Unternehmen wurde auf vier Kategorien vergröbert (Einzelunternehmen, Personengesellschaft, Kapitalgesellschaft, Sonstige).

Das Gründungsjahr der eigenen Unternehmen wurde auf Zehnjahresintervalle vergröbert.

Sonstiges:

Die meisten stetigen Variablen wurden auf zwei signifikante Stellen gerundet (durch ein sogenanntes Random Rounding, das die negativen statistischen Effekte des kaufmännischen Rundens ausgleichen soll). Diese Anonymisierungsmaßnahme entspricht zudem dem natürlichen Antwortverhalten vieler der befragten Haushalte. Etwa 80% aller stetigen Werte wurden bereits von den Befragten mit nur einer oder zwei signifikanten Stellen angegeben.

Über die bis hier beschriebenen Maßnahmen hinausgehend wurden vereinzelt bei besonders auffälligen Merkmalsausprägungen oder Merkmalskombinationen zu Anonymisierungszwecken Vergröberungen oder Veränderungen vorgenommen.

Sämtliche von den Interviewern gesammelten Paradaten sind im Datensatz nicht enthalten. Einzige Ausnahme: die Gebäudecharakteristik (Variable dsc0100).

Sämtliche Angaben und Antworten, die in Textform erfasst wurden, sind entweder in numerische Angaben umcodiert worden (z.B. NACE, ISCO) oder aus dem Datensatz entfernt worden (z.B. Interviewer-Kommentare).