VID/37313100-04
16.01.2008

Absolute Anonymisierung in der Steuerstatistik
– CAMPUS-File der Einkommensteuer 2001 –

Dr. Nicole Buschle

1 Einführung

Die Möglichkeit der Weitergabe von Einzeldaten aus der amtlichen Statistik ist in § 16 des Gesetzes über die Statistik für Bundeszwecke (Bundesstatistikgesetz, BStatG) geregelt. Nach § 16 Abs. 1 Satz 2 Nr. 4 dürfen Einzeldaten beispielsweise dann weitergegeben werden, „wenn sie dem Befragten oder Betroffenen nicht zuzuordnen sind." Hier wird vom Gesetzgeber absolute Anonymität vorausgesetzt.

Nach der Veröffentlichung des CAMPUS-File für das Veranlagungsjahr 1998 bietet das Statistische Bundesamt mit dem CAMPUS-File 2001 weitere frei zugängliche Einzeldaten aus der Einkommensteuerstatistik, anhand derer das Besteuerungsverfahren nachvollzogen und Simulationen durchgeführt werden können.

2 Anonymisierung

Mit einer Anonymisierung von Merkmalsträgern ist immer ein Informationsverlust der dazugehörigen Daten verbunden. Da die Einzeldaten vor einer Zuordnung sicher geschützt sein müssen, ist der Spielraum bei der absoluten Anonymisierung (im Vergleich zur faktischen Anonymisierung[1] nach § 16 Abs. 6 BStatG) sehr begrenzt. Die eingesetzten Anonymisierungsmaßnahmen beschränken sich auf die traditionellen Anonymisierungsmethoden.[2]

Ausgehend von der 1%-Stichprobe werden mit Hilfe des Gesamtbetrags der Einkünfte (GdE) die Steuerpflichtigen in zwei Bereiche unterteilt. Der erste erstreckt sich über den Bereich vom niedrigsten (negativen) GdE bis zu einem GdE von 500 000 Euro. Der zweite umfasst Merkmalsträger mit einem 500 000 Euro übersteigenden GdE. Für die zehn Merkmalsträger mit dem höchsten bzw. geringsten GdE sind zusätzliche Anonymisierungsmaßnahmen vorgesehen (top bzw. bottom coding).

Bei den so genannten manuellen Fällen, bei denen in den Originaldaten kein GdE vorliegt, wurde das GdE ermittelt aus Bruttoarbeitslohn abzüglich Werbungskostenpauschbetrag nach § 9a EStG und die Merkmalsträger auf diese Weise den jeweiligen Bereichen zugeordnet.

[1] Vergleiche hierzu Merz, J., Vorgrimler, D., Zwick, M., Faktisch anonymisiertes Mikrodatenfile der Lohn- und Einkommensteuerstatistik 1998, in: Wirtschaft und Statistik, Heft 10, 2004, S. 1079-1091.

[2] Zu den Methoden vgl. Höhne, J., Methoden zur Anonymisierung wirtschaftsstatistischer Einzeldaten, in: Ronning, G., Gnoss, R., Anonymisierung wirtschaftsstatistischer Einzeldaten, 2003, Forum der Bundesstatistik Band 42, S. 69-94.

2.1 Allgemeine Anonymisierung

Die in diesem Abschnitt erläuterten Anonymisierungsmaßnahmen gelten für alle Steuerpflichtigen. Unterschieden wird zwischen diskreten und stetigen Merkmalen. Für die in Tabelle 1 aufgeführten diskreten Merkmale erfolgt eine Umkodierung, teilweise mit Abschneidegrenzen. Die stetigen Merkmale werden in drei Kategorien eingeteilt, die unterschiedlich behandelt werden.

Die Beschränkung der Einkommensteuerdaten auf eine 1%-Stichprobe mit etwa 270.000 Datensätzen stellt darüber hinaus eine Anonymisierungsmaßnahme dar. Während das Vergröbern oder Streichen von Merkmalen die Identifikation eines Merkmalsträgers bereits erheblich erschwert, wird durch die fehlende Kenntnis, ob der gesuchte Merkmalsträger in der Stichprobe enthalten ist, die Chance auf eine korrekte Zuordnung vollständig vereitelt.

Es wird darauf hingewiesen, dass die Stichprobe nicht zur Anonymisierung gezogen wurde, sondern mit dem Ziel, „handhabbare" Datenmengen mit höchstmöglicher Repräsentativität zu erhalten.[3] Aus diesem Grund sind kleinere heterogene Gruppen von Merkmalsträgern als Vollerhebung enthalten. Dies gilt besonders für die Bezieher hoher Einkommen (hier: ab 500 000 Euro), denen deshalb in dem Anonymisierungsbereich 2 stärkere Anonymisierungsmaßnahmen zugeordnet sind. Die 1%-Stichprobe entfaltet ihre Anonymisierungswirkung nur als „Nebenprodukt", und dies im Bereich der niedrigen und mittleren Einkommen. Für diese ergibt sich durch die Stichprobenziehung ein entscheidender Beitrag zur Erreichung der absoluten Anonymität.

Merkmalskategorien

Für die Anonymisierung mittels Vergröbern oder Streichen von Merkmalen wird zwischen den stetigen und diskreten (qualitativen) Merkmalen unterschieden. Die stetigen Merkmale sind nach ihrer Bedeutung in drei Kategorien eingeteilt. In der ersten sind die Merkmale enthalten, die unverändert übernommen werden. Die zweite Kategorie enthält Merkmale, die modifiziert werden. Hier werden Dummy-Variablen erzeugt, die den Wert 1 annehmen, wenn ein Merkmal vorhanden ist. Alle weiteren stetigen Merkmale zählen zur dritten Kategorie und werden gelöscht.

Merkmale der ersten Kategorie:

- Summe der Einkünfte (getrennt nach A und B[4])
- Gesamtbetrag der Einkünfte
- Einkommen
- zu versteuerndes Einkommen
- tarifliche Einkommensteuer
- festzusetzende Einkommensteuer

Merkmale der zweiten Kategorie:

- die sieben Einkunftsarten (getrennt nach A und B)
- Sonderausgaben, die nicht Vorsorgeaufwendungen sind
- Sonderausgaben: Vorsorgeaufwendungen
- Außergewöhnliche Belastungen, abzugfähig – bei getrennter Veranlagung –A –
- Außergewöhnliche Belastungen, abzugfähig – bei getrennter Veranlagung –B –
- Förderung des Wohneigentums: Steuerbegünstigungen insgesamt

[3] Zur Funktion der Stichprobe vgl. Zwick, M., Einzeldatenmaterial und Stichproben innerhalb der Steuerstatistik, in: Wirtschaft und Statistik, Heft 7, 1998, S. 566-572.

[4] A bezeichnet männliche Steuerpflichtige, B weibliche.

Die qualitativen Merkmale betreffend wird bei den Steuerpflichtigen zwischen veranlagten und manuellen Fällen unterschieden, bei der Art der Veranlagung wird zwischen Grund- und Splittingtabelle differenziert. Die Variable Kinder nimmt den Wert 1 an, wenn Kinder vorhanden sind. Das Merkmal Alter ist in Klassen mit einer Breite von 10 Jahren eingeteilt. Die Regionen sind mit West (alte Bundesländer) und Ost (neue Bundesländer und Berlin) beschrieben. Die Merkmale Geschlecht, soziale Gliederung, Steuerklassen/-kombination sowie der Hochrechnungsfaktor werden unverändert übernommen. Als neue Merkmale sind der Anonymisierungsbereich und eine zufällig erzeugte Zeilennummer enthalten. In der Tabelle 1 sind die allgemeinen Anonymisierungsmaßnahmen zusammengefasst.

Tabelle 1: Allgemeine Anonymisierungsmaßnahmen

Merkmal		Ausprägung
Merker		Umkodierung der 8 Ausprägungen in: 1 = veranlagte Fälle 2= manuelle Fälle
Veranlagungsart		Umkodierung der 8 Ausprägungen in: 1 = Grundtabelle 2= Splittingtabelle
Kinder		Umkodierung in 0 = keine Kinder 1 = 1 oder mehr Kinder
Alter		Klassifizierung der Altersangaben in: 0 = keine Angabe 1 = unter 20 2 = 20 bis unter 30 3 = 30 bis unter 40 4 = 40 bis unter 50 5 = 50 bis unter 60 6 = 60 bis unter 70 7 = 70 und älter
Region		Umkodierung der Ausprägungen in: 1 = West 2 = Ost
stetige Merkmale	Kategorie 1	Werte in Euro
	Kategorie 2	Ersetzen durch Dummies: 0 = kein Wert vorhanden 1 = Wert vorhanden
	Kategorie 3	gelöscht

2.2 Anonymisierungsmaßnahmen in den spezifischen Bereichen

Da ab einem GdE von 500 000 Euro alle Steuerpflichtigen in die 1%-Stichprobe eingegangen sind, wird aus dem Anonymisierungsbereich 2 zusätzlich zu den oben beschriebenen Anonymisierungsmaßnahmen eine Unterstichprobe mit einem Umfang von 33 Prozent gezogen. Danach werden die Ausprägungen der verbleibenden zehn Merkmalsträger mit dem höchsten GdE durch die Durchschnittswerte ihrer jeweiligen Ausprägungen ersetzt. Für die 10 Merkmalsträger aus Anonymisierungsbereich 1 mit den geringsten GdE wird ebenso verfahren. So entsprechen die Maxima und Minima der Merkmale der ersten Kategorie nicht mehr den Originalwerten, sondern stellen die arithmetischen Mittel der zehn höchsten bzw. geringsten Werte dar. Tabelle 2 fasst die Maßnahmen zusammen.

Tabelle 2: Spezielle Anonymisierungsmaßnahmen

Anonymisierungs-bereich	GdE in € (DM)	spezielle Anonymisierungsmaßnahmen
1	bis 500 000 (bis 977 915)	Ersetzen der Ausprägungen der Merkmale der Kategorie 1 bei den unteren 10 Merkmalsträgern durch ihre Mittelwerte
2	über 500 000 (über 977 915)	Ziehen einer 33%-Stichprobe und Ersetzen der Ausprägungen der Merkmale der Kategorie 1 bei den oberen 10 Merkmalsträgern durch ihre Mittelwerte

3 Kontrolle der Anonymisierung

Zur Kontrolle, ob die Anonymisierungsmaßnahmen hinreichend greifen und damit ein Steuerpflichtiger absolut sicher vor einer Identifikation ist, wurde sowohl das Gesamtmaterial als auch das Campus-File nach Randsummen ausgewertet. Für beide Dateien wurde eine mehrdimensionale Kreuztabelle aus den Merkmalen ef8, k0_tabelle, k0_alter_a, k0_alter_b, k0_kinder und k0_region gebildet. Es ist festzuhalten, dass Merkmalskombinationen, die im Gesamtmaterial nur ein- oder zweimal vorhanden sind, im Campus-File nicht vertreten sind.

Selbst ein Merkmalsträger mit einer eindeutigen Merkmalskombination und einem Hochrechnungsfaktor von eins kann aus mehreren Gründen nicht eindeutig identifiziert werden: Erstens ist einem Angreifer nicht bekannt, wie viele Steuerpflichtige im Gesamtmaterial eine solche Merkmalskombination aufweisen. Zweitens sind die Schichten der Stichprobe anhand des Campus-Files nicht rekonstruierbar, da nicht alle Schichtmerkmale zur Verfügung stehen, so dass drittens aufgrund der vermeintlich eindeutigen Merkmalskombination und dem Hochrechnungsfaktor von eins nicht geschlossen werden kann, dass kein weiterer Steuerpflichtiger mit dieser Merkmalskombination in einer anderen Schicht existiert und der Stichprobenziehung „zum Opfer fiel“. Darüber hinaus gibt es viertens in dem relevanten Bereich mehrere Steuerpflichtige ohne Altersangabe, die prinzipiell in die Kategorie mit dem eindeutigen Fall gehören könnten. Schlussendlich herrscht fünftens keine Gewissheit, wer eine Einkommensteuererklärung (oder wenigstens die Lohnsteuerkarte) abgegeben hat und damit überhaupt in die Datenbasis eingegangen ist.

4 Fazit

Mit der vorgelegten Anonymisierungskonzeption ist es erneut gelungen, absolut anonymisierte Daten aus der Einkommensteuerstatistik zu erstellen, die einem breiten Nutzerkreis zu Analysezwecken zur Verfügung gestellt werden können. Die Datei umfasst mehr als 270 000 Steuerpflichtige mit je 38 Merkmalen. Da das Analysepotenzial durch die Anonymisierung eingeschränkt ist, sind nicht alle Fragestellungen mit den Daten analysierbar. Für die Wissenschaft sei deshalb auf die alternativen Zugangswege zu Mikrodaten verwiesen, die von den Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder angeboten werden (http://www.forschungsdatenzentren.de).

Trotz dieser Einschränkung stellt die angebotene Datei einen großen Fortschritt für die steuerstatistische Datenbasis dar. Zum zweiten Mal ist ein so umfangreiches Material für steuerstatistische Analysen allgemein und kostenlos zugänglich. Darüber hinaus stellt das CAMPUS-File trotz der Anonymisierung Informationen über die Bezieher der höchsten Einkommen bereit. Gerade dies ist ein Bereich, der in anderen Datenbasen entweder nicht oder nur sehr unzureichend abgebildet ist.