Kurzbeschreibung der Anonymisierung der Lohn- und Einkommensteuerstatistik 1998

1 Vorbemerkung

Der erstellte Scientific-Use-File der Lohn- und Einkommensteuerstatistik basiert auf der 10%-Stichprobe, die dem Bundesfinanzministerium für Zusatzaufbereitungen zur Abschätzung finanzieller und organisatorischer Auswirkungen der Änderungen von Regelungen im Rahmen der Fortentwicklung des Steuer- und Transfersystems vom Statistischen Bundesamt zur Verfügung gestellt wird. Die Stichprobe wurde nicht zu Anonymisierungszwecken gezogen und somit nicht für den größtmöglichen Schutz der Daten optimiert. Stattdessen stand die Analysefähigkeit der Daten bei der Stichprobenziehung im Vordergrund, daher wurde die Stichprobe auf der Grundlage des "Prinzips der vergleichbaren Präzision für gegliederte Ergebnisse" so gezogen, dass der Stichprobenfehler für den GdE minimiert wird.[1] Dennoch entfaltet die Stichprobenziehung bei den Merkmalsträgern mit geringen und mittlerem Einkommen eine Anonymisierungs- und damit Schutzwirkung.

Bei der Anonymisierung der Lohn- und Einkommensteuerstatistik 1998 wurde auf so genannte datenverändernde Verfahren verzichtet. Statt dessen wurden Verfahren verwendet, die bereits seit längerer Zeit erfolgreich in der amtlichen Statistik eingesetzt werden. Beispiele hierfür sind die Scientific-Use-Files, die mit den Daten des Mikrozensus erstellt wurden und werden.

2 Anonymisierungsbereiche

Die Anonymisierung eines Merkmalsträgers steht immer im Verhältnis zu dessen Re-Identifikationsrisiko. Bezieher niedriger und mittlerer Einkommen sind prinzipiell weniger gefährdet als Bezieher höherer Einkommen. Daher wurden die Merkmalsträger mit niedrigeren Einkommen weniger stark anonymisiert als diejenigen mit höheren Einkommen. Aus diesem Grund wurden fünf Einkommensklassen mit positiven Einkommen und drei Klassen mit negativen Einkommen gebildet. Innerhalb der Klassen (im Folgenden Anonymisierungsbereiche) wurden jeweils spezielle auf die Gefahrenpotenziale abgestimmte Anonymisierungen durchgeführt.
Eine Sondergruppe stellen die Abgeordneten dar. Da die Diäten der Abgeordneten direkt aus den Daten erkennbar sind, handelt es sich um eine gut abgrenzbare kleine Gruppe mit sehr genauen Einkommensinformationen. Daher bedürfen die Abgeordneten einen besonderen Schutz gegen De-Anonymisierung. Um diesen zu gewährleisten, wurden die Abgeordneten, unabhängig von deren tatsächlichen Einkommenshöhe auf gleiche Weise anonymisiert, wie die höchsten Einkommensbezieher.

Zur Einteilung der Merkmalsträger in Anonymisierungsbereiche wurde das GDE verwendet. Bei den so genannten manuellen Fällen, bei denen kein GDE vorliegt, wurde auf das Merkmal „Bruttoarbeitslohn“ zurückgegriffen. Bei den negativen Einkommensbeziehern wurde, sofern kein GDE vorlag, das Merkmal „Einkommen“ verwendet. Die Klasseneinteilung ist der Tabelle zu entnehmen.

[1] Zum Prinzip der Stichprobenziehung vgl. Zwick, M.: Einzeldatenmaterial und Stichproben innerhalb der Steuerstatistik, in: Wirtschaft und Statistik, Heft 7, 1998, S. 566-573, der aktuelle Schichtungsplan ist dem Anhang zu entnehmen.

Tabelle 1: Einteilung der Anonymisierungsbereiche

Anonymisierungsbereich	Positives GDE in € (DM)	Negatives GDE / Einkommen in € (DM)
1	0 – 64.106 (0 – 125.381) (zweimal durchschnittliche GDE)	0 – - 102.258 (0 – -200.000) (95% Percentil[a)])
2	64.107 – 137.532 (125.382 – 268.990) (99 % Percentil)	–
3	137.533 – 970.202 (268.991 – 1.897.551) (99,95 % Percentil)	- 102.259 - -511.292 (-200.001 – -1.000.000) (99,5% Percentil [a)])
4	970.203 - 7.354.714 (1.897.552 – 14.384.570) (bis zu den 1.000 reichsten)	–
5[b)]	> 7.354.714 + Abgeordnete (> 14.384.572)	< - 511.292 (< - 1.000.000)

a) in späteren Versionen vorgesehene relative Grenzen
b) Als Anonymisierungsbereich 6 wurden bei den 3 reichsten männlichen und weiblichen Merkmalsträgern zusätzliche Maßnahmen durchgeführt.

3 Allgemeine Anonymisierung

Allgemeine Anonymisierung bedeutet in diesem Fall, dass ein bestimmtes Merkmal mit einer bestimmten Methode bei allen Merkmalsträgern behandelt wurde, unabhängig von dem jeweiligen Einkommen. Aufgrund der Einkommenshöhe kann das Merkmal jedoch mit weitergehenden Maßnahmen behandelt worden sein. Eine Zusammenfassung der allgemeinen Anonymisierung liefert Tabelle 2.

Tabelle 2: Übersicht über die allgemeinen Anonymisierungsmaßnahmen

Eingabefeld	Merkmal(e)	Maßnahme
EF1	Merker	Umkodierung der acht Ausprägungen in: 1 = veranlagte Fälle 2 = manuelle Fälle
EF13 + EF14	Religionen (jeweils getrennt nach Männer und Frauen)	Umkodierung der zwölf Ausprägungen in: 1 = evangelisch 2 = katholisch 3 = sonstige 4 = konfessionslos
EF19	Veranlagungsart	Umkodierung der acht Ausprägungen in: 1 = Grundtabelle 2 = Splittingtabelle
EF64 + EF67	Alter (jeweils getrennt nach Männer und Frauen)	Einführung einer Unter- (15 Jahre) und Obergrenze (70 Jahren). Ober- bzw. unterhalb der Grenzen wurde das Alter als Durchschnitt derjenigen, die ober- bzw. unterhalb der Grenzen liegen, angegeben.
c36010 - c37066	Anzahl und Alter der Kinder	Die Merkmale der Kinder wurden entfernt. Lediglich die Anzahl und Angaben zum Alter der Kinder ist in den Daten enthalten (EF70-EF73). Fünf und mehr Kinder wurden der Ausprägung ≥ 4 Kinder zugewiesen

4 Spezielle Anonymisierung

Die für die Bereiche speziell durchgeführten Anonymisierungen sind in Tabelle 3 aufgelistet. Die stetigen Merkmale wurden nach Ihrer Bedeutung in unterschiedliche Kategorien eingeteilt. In der ersten sind Merkmale enthalten, die auch bei den Merkmalsträgern mit den höchsten Einkommen ausgewiesen werden, die zweite enthält Merkmale, die nur bei den höchsten Einkommen behandelt werden, während die Merkmale der dritten Kategorie als erstes anonymisiert wurden. Die genaue Einteilung der stetigen Merkmale in die drei Kategorien ist im Anhang enthalten. Die

Ausprägungen des Merkmals „Freiberufler“ kann der Datensatzbeschreibung entnommen werden.

Zusätzlich zu den in der Tabelle beschriebenen Anonymisierungen, wurden die Merkmalsausprägungen der drei „reichsten“ Merkmalsträger durch die durchschnittliche Merkmalsausprägung der drei Reichsten ersetzt (mikroaggregiert). Dies betrifft nur die Merkmale der ersten Kategorie. Insgesamt sind davon sechs Merkmalsträger betroffen. Bei den drei reichsten weiblichen Merkmalsträgern ist das Merkmal „Summe der Einkünfte – weiblich“ mikroaggregiert, die restlichen dagegen unbehandelt. Bei den drei reichsten männlichen sind alle weiteren Merkmale der ersten Kategorie mikroaggregiert, da diese Personen bei allen Merkmalen (außer dem SdE – weiblich) jeweils die drei höchsten Werte aufweisen.

Tabelle 3: Übersicht über die speziellen Anonymisierungsmaßnahmen

Merkmal		Bereiche[1]				
		1	2	3	4	5
Religion		4 Ausprägungen	4 Ausprägungen	k. A.	k. A.	k. A.
Kinder		Anzahl bis vier Alter der ersten 3 Kinder	Anzahl bis vier Alter als Dummy	Anzahl bis vier Alter als Dummy	Anzahl bis vier	Ja/nein
Alter		Ja mit 15 / 70 Grenze	Klasse mit 5 Jahren	Klasse mit 10 Jahren	Klasse mit 10 Jahren	Klasse mit 10 Jahren
Region		Bundesland	Bundesland	West/Ost	West/Ost	West/Ost
GKZ		1-Steller	1-Steller	1-Steller	1-Steller	k. A.
Freiberufler		9 Ausprägungen	9 Ausprägungen	9 Ausprägungen	9 Ausprägungen	Dummy ja /nein
Stetige Merkmale	1	Ja	Ja	Ja	Ja	Ja
	2	Ja	Ja	Ja	Ja, aber A + B als Summe	Dummy (+ Bedeutungsmerkmale)
	3	Ja	Ja	Ja	Dummy	Nein

[1] **Bei positiven Einkommen:** 1 = von 0 bis zu einem GDE von 64.106€; 2 = 64.107 – 137.532 € (99% Percentil); 3 = 137.533 – 970.202 € (99,95% Percentil); 4 = 970.203 € bis zum 1.000 höchsten GDE; 5 = die 1.000 höchsten GDE + Abgeordnete.

Bei negativen Einkommen 1 = von 0 bis zu einem negativen Einkommen von 102.258 € (95% Percentil[2]); 3 = von 102.259 bis zu einem negativen Einkommen von 511.292 (99,5% Percentil[8]) €; 5 = bei einem negativen Einkommen von über 511.292 €.

Die sieben Einkunftsarten wurden in drei Kategorien (Gewinneinkünfte, Einkünfte aus nichtselbständiger Arbeit und Überschusseinkünfte) zusammengefasst. Für jede Kategorie wurde ein Merkmal gebildet, das die Bedeutung der Einkunftsart für den Merkmalsträger widerspiegelt. Die Merkmale nehmen den Wert 1 an, wenn in der dazugehörigen Einkunftsart die höchsten und 3 wenn die geringsten Einkünfte erzielt werden. Entstehen keine Einkünfte aus der Kategorie wird das Merkmal auf 0 gesetzt. Die Merkmale wurden für alle Anonymisierungsbereiche gebildet, sind aber speziell für den fünften Bereich von Bedeutung, da durch sie weiterhin die Entstehungsstruktur der Einkünfte für alle Merkmalsträger abgebildet werden kann. Dies wäre ansonsten nicht möglich, da die Einkunftsarten zu den Merkmalen der zweiten Kategorie gehören.

[2] In späteren Versionen vorgesehene relative Grenzen

Anhang:

Einteilung der stetigen Merkmale in Kategorien

Merkmale der ersten Kategorie:

- Summe der Einkünfte (A+B[3])
- Gesamtbetrag der Einkünfte
- Einkommen
- zu versteuerndes Einkommen
- tarifliche Einkommensteuer
- festzusetzende Einkommensteuer

Merkmale der zweiten Kategorie:

- Einkünfte aus Land- und Forstwirtschaft (A+B)
- Einkünfte aus Gewerbebetrieb (A+B)
- Einkünfte aus selbständiger Arbeit (A+B)
- Einkünfte aus nichtselbständiger Arbeit (A+B)
- Einkünfte aus Kapitalvermögen (A+B)
- Einkünfte aus Vermietung und Verpachtung (A+B)
- sonstige Einkünfte (A+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

Alle weiteren stetigen Merkmale zählen zur dritten Kategorie.

[3] A sind männliche Steuerpflichtige, B weibliche.