Faktische Anonymisierung der Steuerstatistik (FAST)
- Lohn- und Einkommensteuer 1998 -

Dr. Daniel Vorgrimler, Markus Zwick

1 Einführung

Die Möglichkeit der Weitergabe von Einzeldaten aus der amtlichen Statistik an die Wissenschaft ist in § 16 Abs. 6 des Gesetzes über Statistik für Bundeszwecke (Bundesstatistikgesetz, BStatG) geregelt. Danach dürfen Einzeldaten an die Wissenschaft dann weitergegeben werden, „wenn die Einzelangaben nur mit unverhältnismäßig großem Aufwand an Zeit, Kosten und Arbeitskraft zugeordnet werden können". Das Unverhältnismäßigkeitsgebot impliziert, dass eine Verletzung der Anonymität von Merkmalsträgern nur bei nutzbringenden Zuordnungen gegeben ist.[1] Damit wird vom Gesetzgeber keine absolute Anonymität mehr vorausgesetzt, sondern eine faktische wird als ausreichend erachtet. Da dies nur für „Hochschulen oder sonstige Einrichtungen mit der Aufgabe unabhängiger wissenschaftlicher Forschung" gilt, wird diese Regelung auch als „Wissenschaftsprivileg" bezeichnet.[2]

Damit anonymisierte Daten von der Wissenschaft angenommen werden, muss eine Anonymisierung zwei gleichrangigen Herausforderungen gerecht werden: Sie muss einerseits einen ausreichenden Schutz der Einzelangaben gewährleisten und andererseits die Analysemöglichkeiten der anonymisierten Daten bestmöglichst erhalten.

Das Projekt „Faktische Anonymisierung der Lohn- und Einkommensteuerstatistik" stellte sich diesen Herausforderungen. In Zusammenarbeit mit potenziellen Datennutzern wurde ein Anonymisierungskonzept entwickelt, mit dem Ziel, den beiden Herausforderungen gerecht zu werden. Ein vorläufiges Konzept wurde auf der zweiten Sitzung des wissenschaftlichen Beraterkreises im Januar 2004 vorgestellt und diskutiert. Hierbei kam von Seiten der Wissenschaft klar zum Ausdruck, dass der Erhalt der (quasi)stetigen Merkmale dem vollständigen Nachweis diskret ausgeprägter sozioökonomischer Merkmale vorzuziehen ist. Diesem Anliegen wurde bei der Weiterentwicklung des Anonymisierungskonzeptes Rechnung getragen. Das aus diesen Arbeiten entstandene Konzept wurde auf der dritten Sitzung des wissenschaftlichen Beraterkreises am 25. März 2004 verabschiedet und wird im Folgenden vorgestellt.

2 Anonymisierung

2.1 Das Prinzip der Tannenbaumanonymisierung

Mit einer Anonymisierung von Merkmalsträgern ist immer ein Informationsverlust der dazugehörigen Daten verbunden. Um diesen Verlust so gering wie möglich zu halten und somit die obige zweite Bedingung bestmöglichst zu erfüllen, sollten diejenigen Merkmalsträger schwächer ano-

[1] Vgl. Höhne, Sturm, Vorgrimler, Konzept zur Schutzwirkung faktischer Anonymisierung, in Wirtschaft und Statistik, 4/2003, S. 287.

[2] Zur Anonymisierung in der Bundesstatistik vgl. Köhler, S., Anonymisierung von Mikrodaten in der Bundesstatistik und ihre Nutzung – Ein Überblick, in: Forum der Bundesstatistik Band 31, 1999, S 133-150.

nymisiert werden, die einem geringeren Risiko ausgesetzt sind. D.h. weitergehende Anonymisierungsmaßnahmen sind auf diejenigen zu beschränken, die diesen auch tatsächlich bedürfen. Analysen zum Schutzbedürfnis haben gezeigt, dass das Risiko mit steigendem Einkommen zunimmt. Aus diesem Grund werden daher Merkmalsträger mit höherem Einkommen stärker anonymisiert als Steuerpflichtige, die ein geringeres Einkommen beziehen. Für die Anonymisierung in Abhängigkeit zur Einkommenshöhe wurden die Daten in unterschiedliche Einkommensbereiche untergliedert und jeweils auf diese Bereiche abgestimmte Anonymisierungen durchgeführt (Tannenbaumanonymisierung).

Die eingesetzten Anonymisierungsmaßnahmen beschränken sich auf die traditionellen Anonymisierungsmethoden.[3]

Tabelle 1: Einteilung der Anonymisierungsbereiche

Anonymisierungsbereich	Positives GDE in € (DM)	Negatives GDE / Einkommen in € (DM)
1	0 – 64.106 (0 – 125.381) (zweimal das durchschnittliche GDE)	0 – - 102.258 (0 – -200.000)
2	64.107 – 137.532 (125.382 – 268.990) (99 % Percentil)	–
3	137.533 – 970.202 (268.991 – 1.897.552) (99,95 % Percentil)	- 102.259 - -511.292 (-200.001 – -1.000.000)
4	970.203 - 7.354.714 (1.897.553 – 14.384.571) (bis zu den 1.000 reichsten)	–
5	> 7.354.714 (> 14.384.572)	< - 511.292 (< - 1.000.000)

Mit Hilfe des Gesamtbetrags der Einkünfte (GDE) wurden die Daten bei den positiven Einkünften in fünf Bereiche unterteilt (vgl. Tabelle 1). Der erste erstreckt sich von einem GDE von Null bis zu dem doppelten des mittleren GDE. Der zweite Bereich geht von diesem bis zum 99% Percentil der Einkommensverteilung. Der dritte Bereich umfasst das Intervall vom 99% Percentil bis zum 99,95% Percentil, während der vierte Bereich diese Grenze bis zu den 1.000 Merkmalsträgern, die das höchste GDE aufweisen, abdeckt. Den fünften Bereich bilden die 1.000 Personen mit dem höchsten GDE. Einen Sonderbereich stellen die drei Merkmalsträger mit den höchsten Einkommen dar. Im Unterschied zum fünften Bereich werden die Ausprägungen der individuellen stetigen Merkmale dieser drei durch die arithmetischen Mittel der jeweils drei Ausprägungen ersetzt. Bei den so genannten manuellen Fällen, bei denen kein GDE vorliegt, wurde ersatzweise der „gesamte Bruttolohn“ als Teilungsmerkmal verwendet und die Merkmalsträger wurden auf diese Weise den jeweiligen Bereichen hinzugespielt.

Bei den Steuerpflichtigen mit negativen Einkommen wurde in den Fällen in denen das GDE nicht besetzt war das Merkmal „Einkommen“ zur Einteilung verwendet. Zur Anonymisierung dieser Merkmalsträger wurden drei Bereiche gebildet (vgl. Tabelle 1). Es ist vorgesehen, dass der erste Bereich diejenigen Merkmalsträger beinhaltet, deren negativen Einkommen zwischen – 1 und dem 95 % Percentil der absoluten negativen Einkommensverteilung liegen. Der zweite Bereich

[3] 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.

soll sich von dieser Grenze bis zu dem 99,5 % Percentil erstrecken, während in den dritten Bereich alle restlichen Merkmalsträger mit den absolut höchsten negativen Einkommen fallen. In der derzeitigen Version sind anstelle dieser relativen Grenzen absolute verwendet worden. Der erste deckt dabei den Bereich von einem GDE von weniger als Null bis zu einem negativen Einkommen von 102.258 € (200.000 DM) ab. Der zweite geht von diesem Einkommen bis zu einem negativen Einkommen von 511.292 € (1.000.000 DM). Im dritten Bereich sind alle Merkmalsträger mit „höheren" negativen Einkommen vertreten.[4] Die Anonymisierungsmethoden in diesen Bereichen sind mit den Methoden in den Bereichen eins, drei und fünf der Merkmalsträger mit positiven Einkommen identisch.

2.2 Allgemeine Anonymisierung

Neben den auf spezifische Einkommensbereiche abgestimmten Anonymisierungsmaßnahmen, die weiter unten erläutert werden, wurden allgemeine Anonymisierungsmaßnahmen durchgeführt, mit denen die Merkmale bei allen Merkmalsträgern mindestens verändert wurden. Dies schließt nicht aus, dass das gleiche Merkmal im Zuge der spezifischen Bereichsanonymisierung weitergehenden Maßnahmen unterworfen wurde. Tabelle 2 gibt über die allgemeinen Anonymisierungsmaßnahmen Auskunft.

Die Beschränkung der Einkommensteuerdaten auf eine 10% Stichprobe mit knapp 3 Mio. Datensätzen, stellt darüber hinaus eine Anonymisierungsmaßnahme dar. Ein Datenangreifer verliert bei einer versuchten Identifikation durch die Stichprobenziehung die Kenntnis, ob der gesuchte Merkmalsträger in der Stichprobe enthalten ist. Ordnet er einen Datensatz aus dem Zusatzwissen einem Merkmalsträger zu, so trägt er das Risiko, dass diese Zuordnung nur deshalb zustande kam, weil der richtige Merkmalsträger nicht in der Stichprobe enthalten ist. Die Zuordnung ist für ihn wertlos. Es sei aber darauf hingewiesen, dass die Stichprobe nicht zur Anonymisierung gezogen wurde, sondern mit dem Ziel, „handhabbare" Datenmengen mit höchstmöglicher Repräsentativität zu erhalten.[5] Aus diesem Grund sind kleinere homogene Gruppen von Merkmalsträgern, so genannte Ränder der Stichprobe, als Vollerhebung enthalten. Dies gilt besonders für die Bezieher hoher Einkommen (ab 200.000 DM). Für diese besitzt ein Datenangreifer somit weiterhin Teilnahmekenntnis, so dass das o.g. Argument nicht gilt. Die Stichprobe entfaltet ihre Wirkung als Anonymisierung 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 faktischen Anonymität.

[4] Die relativen Grenzen sollen in einer späteren Überarbeitung der Daten eingebaut werden.

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

Tabelle 2: Allgemeine 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 Jahre). Ober- bzw. unterhalb der Grenzen wurde das Alter als Durchschnitt derjenigen, die ober- bzw. unterhalb der Grenzen liegen, angegeben.
c36010 - c37066	Anzahl der Kinder	Die Merkmale der Kinder wurden entfernt. Lediglich die Anzahl und Angaben zum Alter der Kinder ist in den Daten enthalten. Fünf und mehr Kinder wurden der Ausprägung ≥ 4 Kinder zugewiesen

Das Alter der Daten wirkt ebenfalls als eine allgemeine Anonymisierungsmaßnahme und zwar in zweierlei Hinsicht. Zum einen ist es für einen Datenangreifer umso schwieriger, relevantes Zusatzwissen für einen Merkmalsträger zu generieren, je älter die Daten sind. Aus diesem Grund steigen die Kosten eines Identifikationsversuchs. Zum anderen ist der Nutzen einer Information u.a. von der Aktualität der selben abhängig. Daher sinkt der Nutzen einer Identifikation mit zunehmenden Alter der Daten. Das Nutzenargument gilt allerdings nur, wenn die Aktualität der Daten auch eine Rolle für den Datenangreifer spielt.

2.3 Spezifische Anonymisierung

2.3.1 Merkmalskategorien

In den fünf unter Abschnitt 2.1 beschriebenen Anonymisierungsbereichen wurden unterschiedliche Merkmale vergröbert oder gestrichen. Hierzu wurden die stetigen Merkmale nach Ihrer Bedeutung in drei Kategorien eingeteilt. In der ersten sind die Merkmale enthalten, die auch bei den Merkmalsträgern mit den höchsten Einkommen noch ausgewiesen werden. Die zweite Kategorie enthält Merkmale, die nur bei den höchsten Einkommen behandelt werden, während die Merkmale der dritten Kategorie als erstes zur Anonymisierung der Merkmalsträger eingeschränkt werden.

Merkmale der ersten Kategorie:

- Summe der Einkünfte (A+B[6])
- Gesamtbetrag der Einkünfte
- Einkommen
- zu versteuerndes Einkommen
- tarifliche Einkommensteuer
- festzusetzende Einkommensteuer

Merkmale der zweiten Kategorie:

[6] A sind männliche Steuerpflichtige, B weibliche.

- 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.

Zu beachten ist, dass Informationen, die zur Anonymisierung entweder nur vergröbert, verfälscht oder überhaupt nicht mehr in den Zieldaten enthalten sind, für einen Datenangreifer einen geringeren Wert aufweisen als die Originalinformationen.[7] Anonymisierungsmaßnahmen wirken sich daher nicht nur auf die Kosten eines Datenangreifers bei einem Datenangriff aus (indem eine Identifikation schwieriger wird), sondern darüber hinaus wird sein Nutzen negativ beeinflusst. Bei der Einkommensteuerstatistik gilt dieser Aspekt besonders bei den stetigen Merkmalen. Diese sind evtl. schwierig als Überschneidungsmerkmale einsetzbar, wodurch eine Veränderung ihrer Werte kein zusätzlicher Schutz der Merkmalsträger darstellen würde (da sie zum „Angriff" sowieso nicht verwendet werden). Jedoch dürften die stetigen Merkmale – zu denen z.B. sämtliche Einkommensinformationen gehören – einem Datenangreifer den höchsten Nutzen stiften. Werden daher stetige Merkmale aus den Daten gelöscht oder vergröbert, so hat dies weniger Auswirkungen auf die Kostenseite eines Angriffes als vielmehr auf die Nutzenseite. Dies ist ein wesentlicher Aspekt zur Erreichung der faktischen Anonymität.

2.3.2 Anonymisierungsmaßnahmen in den spezifischen Bereichen

Erster Bereich (von 0 bis zu einem GDE von 64.106 € (125.381 DM)):

Zusätzlich zur allgemeinen Anonymisierung sind in diesem Bereich bei den diskreten Merkmalen die Regionskennung auf Bundeslandebene gekürzt. Beim Alter wurden „15 Jahren" und „70 Jahren" als Unter- und Obergrenzen eingeführt. Das Alter derjenigen, die ober- bzw. unterhalb der Grenzen liegen, ist als Durchschnitt derjenigen, die ober- bzw. unterhalb der Grenzen liegen, angegeben. Zusätzlich sind in den Daten die Altersangaben der ersten drei Kinder enthalten. Die Gewerbekennzahl (GKZ) wurde auf eine Stelle reduziert. Die stetigen Merkmale sind in diesem Bereich unbehandelt geblieben.

[7] Vgl. hierzu Höhne, Sturm, Vorgrimler, Konzept zur Schutzwirkung faktischer Anonymisierung, in Wirtschaft und Statistik, 4/2003, S. 287-292.

Zweiter Bereich (GDE von 64.107 € (125.382 DM) bis 137.532 € (268.990 DM), 99% Percentil):

Im Unterschied zum ersten Bereich ist in diesem das Alter klassifiziert. Als Klassenbreite wurde fünf Jahre gewählt. Das Alter der ersten drei Kinder wird jeweils mit einer Dummy-Variable beschrieben. Diese Variablen nehmen den Wert 1 an, wenn die Kinder mindestens 15 Jahre alt sind und 0 bei jüngeren Kindern. Alle weiteren Merkmale wurden wie im ersten Bereich behandelt.

Dritter Bereich (GDE von 137.533 € (268.991 DM) – 970.202 € (1.897.551 DM), 99,95% Percentil):

Im dritten Bereich ist das Merkmal Alter in Klassen mit einer Klassenbreite von 10 Jahren klassifiziert. Die Regionen sind nur noch mit West (alte Bundesländer) und Ost (neue Bundesländer und Berlin) beschrieben. Des Weiteren sind die Ausprägungen des Merkmals Religion gelöscht. Die stetigen Merkmale blieben weiterhin unbehandelt.

Vierter Bereich (GDE von 970.203 € (1.897.552 DM) bis zu den 1.000 Merkmalsträgern mit den höchsten GDE):

Zusätzlich zu den bereits angewandten Anonymisierungsmethoden wurden in diesem Bereich die stetigen Merkmale der Kategorie drei in Dummy-Variablen umkodiert. Wobei 1 positive Werte und -1 negative Werte darstellen. Die Merkmale der Kategorie eins und zwei bleiben unbehandelt, mit Ausnahme der sieben Einkunftsarten, die nur als Summe der beiden Untergruppen A+B angegeben sind. Bei den diskreten Merkmalen ist die Dummy-Variable für das Alter der Kinder nicht mehr enthalten. Die weiteren diskreten Merkmale sind analog zum dritten Bereich behandelt.

Fünfter Bereich (1.000 Merkmalsträger mit den höchsten GDE):

Der fünfte Bereich enthält die stetigen Merkmale der ersten Kategorie. Die Ausprägungen der Merkmale der zweiten Kategorie sind durch Dummy-Variablen ersetzt (vgl. Kapitel 3) und die der dritten gelöscht. Die Ausprägungen der drei Merkmalsträger mit dem höchsten GdE wurden ersetzt durch die Durchschnittswerte ihrer jeweiligen Ausprägungen. So entsprechen die Maxima der Merkmale der ersten Kategorie nicht mehr den Originalwerten, sondern stellen die arithmetischen Mittel der drei höchsten Werte dar.

Bei den diskreten Merkmalen wurde die GKZ gelöscht und die Freiberufler nur noch als Dummy angegeben (zu den Freiberuflern vgl. Kapitel 3). Die Anzahl der Kinder wird nicht mehr angegeben, sondern nur noch, ob Kinder vorhanden sind.

Sonderbereich „negative Einkommen“:

Steuerpflichtige mit negativen Einkommen werden innerhalb dreier Bereiche anonymisiert. Diese entsprechen den Bereichen 1, 3 und 5 bei den positiven Einkommen.

Sonderbereich "Abgeordnete"

Abgeordnetendiäten werden in der Einkommensteuererklärung als 'Sonstige Einkünfte als Abgeordneter' erfasst. Für diese kleine Gruppe liegen somit sehr spezifische Angaben in den Daten vor. Darüber hinaus lassen sich im Internet, insbesondere auf den Seiten der Bundes- und Landesparlamente, detaillierte Informationen über diese Gruppe gewinnen. Ein erster Schritt zur Anonymisierung dieser Gruppe von Steuerpflichtigen war, die Angaben mit weiteren Sonstigen Einkünften zusammenzufassen. Dies erwies sich als nicht ausreichend. Aus diesem

Grund wurden sämtliche Steuerpflichtigen die 'Sonstige Einkünfte als Abgeordneter' bezogen, in den Anonymisierungsbereich 5 aufgenommen und die Merkmale entsprechend behandelt.

In der Tabelle 3 sind die getroffenen Anonymisierungsmaßnahmen für die unterschiedlichen Teilbereiche zusammengefasst.

Tabelle 3: Anonymisierungsmaßnahmen in den speziellen Bereichen

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[8]); 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 €.

3 Zusatzinformationen in FAST

Neben der Reduktion von Informationen durch die Anonymisierung wurden in die FAST-Datei zusätzlich generierte Informationen aufgenommen, die der Wissenschaft das Arbeiten mit den Daten erleichtern sollen. Diese aus den originären Daten erstellten Zusatzinformationen werden in diesem Abschnitt kurz vorgestellt.

Aus der GKZ wurden neun Kategorien für die freien Berufe gebildet und in allen Bereichen außer dem fünften aufgenommen (vgl. Tabelle 3). Dabei handelt es sich um folgende Kategorien:

1 Technische Beratung, Forschung, Architekten, Ingenieur
2 Rechtsanwälte, Notar
3 Wirtschaftsprüfer, -berater
4 Ärzte
5 Sonstige Gesundheitsberufe
6 Werbung, Foto, Kunst und Kultur
7 Schriftberufe
8 Schulen

[8] In späteren Versionen vorgesehene relative Grenzen

9 Sonstige

Zusätzlich wurde eine Dummy-Variable eingeführt, die angibt ob der Merkmalsträger freiberuflich tätig ist. Diese ist in allen Bereichen enthalten. Damit folgen wir einer langen Tradition der Steuerstatistiken, in der die Gruppe der Freien Berufe schon über lange Jahre in dieser Typisierung ausgewertet und analysiert werden.

Im Anonymisierungsbereich 5 sind die Merkmale der zweiten Kategorie nur noch als Dummy-Variable enthalten. Damit die Datennutzer die Struktur der Einkünfte auch im höchsten Einkommensbereich nachbilden können, wurden die sieben Einkunftsarten in drei Kategorien eingeteilt (Gewinneinkünfte, Einkünfte aus nichtselbständiger Tätigkeit und Sonstige Überschusseinkünfte). Für jede dieser Kategorie wurde ein Merkmal gebildet. Dieses nimmt den Wert 1 an, wenn in dieser Einkunftsart die höchsten Einkünfte erzielt werden und 3 wenn die geringsten Einkünfte aus dieser Kategorie stammen. Entstehen keine Einkünfte aus der Kategorie, wird das Merkmal auf 0 gesetzt. Die Merkmale wurden für alle Anonymisierungsbereiche gebildet. Als Beispiel ist in Tabelle 4 die Häufigkeitsverteilung der Merkmale für die 1.000 Merkmalsträger mit den höchsten Einkommen angegeben. Sie zeigt demnach welche Einkommenskategorien zur Erzielung der höchsten Einkommen am meisten beitragen.

Tabelle 4: Bedeutung der Einkommensklassen bei den höchsten Einkommen

Bedeutung	Gewinneinkünfte	Einkünfte aus nichtselbständiger Arbeit	Sonstige Überschusseinkünfte
hoch	910	10	80
mittel	49	318	616
niedrig	30	275	283
keine	11	397	21

Quelle: eigene Berechnungen

Als weitere Zusatzinformation wurde ein Merkmal eingeführt, welches die Anonymisierungsstärke des jeweiligen Merkmalsträger angibt. Die Ausprägungen 1-5 geben hierbei die Anonymisierungsbereiche wieder. Zusätzlich wurde die Ausprägung 6 eingeführt, die diejenigen Merkmalsträger aus dem Anonymisierungsbereich 5 angibt, deren stetige Merkmale mikroaggregiert wurden. Die Bezieher negativer Einkünfte wurden ebenso entsprechend ihrer Stärke gekennzeichnet (mit 1, 3 oder 5) wie die Abgeordneten (mit 5).

4 Fazit

Mit der vorgelegten Anonymisierungskonzeption ist es gelungen, erstmalig faktisch anonymisierte Daten aus der Lohn- und Einkommensteuerstatistik zu erstellen, die der Wissenschaft zu Analysezwecken zur Verfügung gestellt werden können. Auch wenn bei der Anonymisierung größten Wert auf den Erhalt des Analysepotenzials gelegt wurde, sind nicht alle Fragestellungen der Wissenschaft exakt mit den Daten analysierbar. Für diese Fälle sei auf die alternativen Zugangswege zu Mikrodaten, die von den Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder angeboten werden, verwiesen.

Trotz dieser Einschränkung stellt die angebotene Datei einen großen Fortschritt in der steuerstatistischen Datenbasis für die Wissenschaft dar. Erstmalig ist ein so umfangreiches Material für steuerstatistische Analysen für die Wissenschaft mit geringen Kosten allgemein zugänglich. Dar-

über hinaus stellt FAST trotz der Anonymisierung bei den höchsten Einkommen erstmalig überhaupt detaillierte 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.

Wiesbaden, im April 2004

Literatur

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

Höhne, J.; Sturm, R.; Vorgrimler, D.: Konzept zur Schutzwirkung faktischer Anonymisierung, in Wirtschaft und Statistik, 4/2003, S. 287-292.

Köhler, S.: Anonymisierung von Mikrodaten in der Bundesstatistik und ihre Nutzung – Ein Überblick, in: Forum der Bundesstatistik Band 31, 1999, S 133-150.

Zwick, M.: Einzeldatenmaterial und Stichproben innerhalb der Steuerstatistik, in: Wirtschaft und Statistik, Heft 7, 1998, Seite 566-572.