

Faktische Anonymisierung der Steuerstatistik (FAST) ***- Lohn- und Einkommensteuer 2007-***

Wolfram Schwabbacher

1 Einführung

Die Einzeldaten der Lohn- und Einkommensteuerstatistik wurden der Wissenschaft zum ersten Mal für das Veranlagungsjahr 1998 und danach dreijährlich als faktisch anonymisierte Datei zur Verfügung gestellt. Im Rahmen des Projekts „Faktische Anonymisierung der Lohn- und Einkommensteuerstatistik“ wurde ein Anonymisierungskonzept erarbeitet, das einerseits einen ausreichenden Schutz der Einzelangaben gewährleistet und andererseits die Analysemöglichkeiten der anonymisierten Daten bestmöglich erhält. Die faktische Anonymisierung der Einzeldaten des Veranlagungsjahres 2007 ist die konsequente Weiterführung des Projektes und wird nachfolgend beschrieben.¹

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.² 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.³

2 Hinweise zum Vergleich mit früheren Jahren

In der Bundesstatistik für das Veranlagungsjahr 2004 waren erstmals auch Daten der von den Arbeitgebern direkt an die Finanzverwaltung elektronisch übermittelten Lohnsteuerbescheinigungen enthalten. Im Veranlagungszeitraum 2007 war die elektronische Übermittlung für nahezu alle Arbeitgeber verpflichtend, so dass die nichtveranlagten Fälle nun nahezu vollständig nachgewiesen sind. Dies unterscheidet die Statistiken 2004 und 2007 grundlegend von denen früherer Veranlagungsjahre, in denen nichtveranlagte Fälle nur erfasst werden konnten, wenn Arbeitgeber oder Arbeitnehmer die Lohnsteuerkarte der jeweiligen Gemeinde oder der Finanzverwaltung zur Verfügung stellten. Die somit erheblich erhöhte Anzahl der nachgewiesenen nichtveranlagten Steuerpflichtigen und die damit einhergehende Änderung der Einkommensstruktur führt zur Einschränkung der Vergleichbarkeit der Ergebnisse der Statistiken 2004 und 2007 mit denen

¹ Vgl. Merz, Vorgrimler, Zwick, Faktisch anonymisiertes Mikrodatenfile der Lohn- und Einkommensteuerstatistik 1998, in *Wirtschaft und Statistik* 10/2004, S. 1079-1090.

² Vgl. Höhne, Sturm, Vorgrimler, Konzept zur Schutzwirkung faktischer Anonymisierung, in *Wirtschaft und Statistik*, 4/2003, S. 287.

³ 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.

vorangegangener Jahre⁴. Für die veranlagten Steuerpflichtigen (EF1 = 1) ist ein Vergleich mit den Statistiken früherer Veranlagungsjahre hingegen ohne Einschränkungen dieser Art möglich.

Darüber hinaus wurden bei der Erstellung des vorliegenden Materials Nullen anders behandelt als bei den Statistiken zu früheren Veranlagungsjahren. Aufgrund des unterschiedlichen Umgangs mit Rechennullen in den Finanzverwaltungen der einzelnen Länder können im Originalmaterial echte Rechennullen und fehlende Werte bei einem Merkmal dieselbe Bedeutung haben. Daher wurden alle Nullen im Material durch fehlende Werte ersetzt, lediglich beim Merkmal Gesamtbetrag der Einkünfte wurden fehlende Werte durch Nullen ersetzt. Dieses Vorgehen sorgt dafür, dass das unsystematische Setzen von Rechennullen nicht zu Verzerrungen bei Auswertungen führen kann und gewährleistet zudem die Vergleichbarkeit mit den amtlichen Veröffentlichungen zur Lohn- und Einkommensteuerstatistik.

3 Anonymisierung

3.1 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 bestmöglich zu erfüllen, werden diejenigen Merkmalsträger schwächer anonymisiert, die einem geringeren Risiko ausgesetzt sind. Weitergehende Anonymisierungsmaßnahmen werden somit auf diejenigen beschränkt, die dieser 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.⁵

⁴ Im Gesamtmaterial der Statistik für das Veranlagungsjahr 2007 sind 12,3 nichtveranlagte Steuerpflichtige nachgewiesen, 2004 betrug diese Zahl 9,3 Millionen und im Jahr 2001 nur 1,9 Millionen. Da diese Steuerpflichtigen überwiegend vergleichsweise niedrige Einkommen aus nichtselbständiger Arbeit beziehen, ist mit Auswirkungen auf die Einkommensverteilung und die Zusammensetzung des erzielten Gesamteinkommens aus den einzelnen Einkunftsarten zu rechnen.

⁵ Zu den Methoden vgl. Höhne J., Methoden zur Anonymisierung wirtschaftsstatistischer Einzeldaten in: Ronning, G., Gnos, R., Anonymisierung wirtschaftsstatistischer Einzeldaten, 2003, Forum der Bundesstatistik Band 42, S. 69-94.

Tabelle 1: Einteilung der Anonymisierungsbereiche

Anonymisierungsbereich	Positives GdE in €	Negatives GdE / Einkommen in €
1	0 bis 63.331 (zweimal das durchschnittliche GdE)	-1 bis -54.272 (95. Perzentil)
2	63.331 bis 171.800 (99. Perzentil)	–
3	171.800 bis 919.297 (99,95. Perzentil)	-54.272 bis -426.007 (99,5. Perzentil)
4	919.297 bis 7.191.333 (bis zu den 1.000 Reichsten)	–
5	> 7.191.333	< -426.007

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. Perzentil der Einkommensverteilung. Der dritte Bereich umfasst das Intervall vom 99. Perzentil bis zum 99,95. Perzentil, 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 Steuerpflichtigen mit dem höchsten GdE. Einen Sonderbereich stellen die drei Merkmalsträger mit den höchsten Einkünften, getrennt nach Mann und Frau, dar: Im Unterschied zum fünften Bereich werden die Ausprägungen der individuellen stetigen Merkmale dieser jeweils drei Merkmalsträger durch die arithmetischen Mittel der Ausprägungen ersetzt. Bei den nicht veranlagten Fällen, bei denen kein GdE vorliegt, wurde ersatzweise der „gesamte Bruttolohn“ als Teilungsmerkmal verwendet.

Bei den Steuerpflichtigen mit negativen Einkommen wurde in den Fällen, in denen der GdE nicht besetzt war, das Merkmal „Einkommen“ zur Einteilung verwendet. Zur Anonymisierung dieser Merkmalsträger wurden drei Bereiche gebildet (vgl. Tabelle 1). Der erste Bereich enthält die Merkmalsträger, deren negativen Einkommen zwischen 1. und dem 95. Perzentil der absoluten negativen Einkommensverteilung liegen. Der zweite Bereich erstreckt sich von dieser Grenze bis zu dem 99,5. Perzentil, während in den dritten Bereich alle restlichen Merkmalsträger mit den absolut höchsten negativen Einkommen fallen. Die Anonymisierungsmethoden in diesen Bereichen sind mit den Methoden in den Bereichen eins, drei und fünf der Merkmalsträger mit positiven Einkommen identisch.

3.2 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 etwa 3,9 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.⁶ Aus diesem Grund sind kleinere homogene Gruppen von Merkmalsträgern, so genannte Ränder der Stichprobe, als Vollerhebung enthalten. Dies gilt besonders für die veranlagten Bezieher hoher Einkommen (ab 75.000 bzw. 150.000 Euro⁷). Für diese besitzt ein Datenangreifer somit weiterhin Teilnahmekennntnis, 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.

Tabelle 2: Allgemeine Anonymisierungsmaßnahmen⁸

Eingabefeld	Merkmal(e)	Maßnahme
EF1	Merker	Umkodierung der acht Ausprägungen in: 1 = veranlagte Fälle 2 = nicht veranlagte Fälle
EF7	Amtlicher Gemeindegemeinschaftsschlüssel	Verkürzung auf 16 Bundesländer
EF13 + EF14	Religion (jeweils getrennt für 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 für 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.
c36116 - c37727, c46016, c65748, c65879, c65880, c65881, c66998	Anzahl der Kinder	Die Merkmale der Kinder wurden entfernt. Lediglich die Anzahl und Angaben zum Alter der Kinder sind in den Daten enthalten. Fünf und mehr Kinder wurden der Ausprägung ≥ 4 Kinder zugewiesen. Kindergeld und Kinderfreibeträge wurden entsprechend begrenzt.
c65500	Einkommen	Das Merkmal Einkommen wurde entfernt.

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 derselben abhängig. Daher sinkt der Nutzen einer Identifikation mit zunehmendem Alter der Daten. Das Nutzenargument gilt allerdings nur, wenn die Aktualität der Daten auch eine Rolle für den Datenangreifer spielt.

⁶ Zur Funktion der Stichprobe vgl. Zwick, M. Einzeldatenmaterial und Stichproben innerhalb der Steuerstatistik, in: Wirtschaft und Statistik, Heft 7, 1998, Seite 566-572.

⁷ Nach der Grundtabelle Besteuerte sind als Vollerhebung in der Stichprobe enthalten, wenn der Gesamtbetrag der Einkünfte mindestens 75.000 Euro beträgt, nach der Splittingtabelle Besteuerte ab einem Gesamtbetrag der Einkünfte von mindestens 150.000.

⁸ Neben den hier genannten Maßnahmen wurden weitere Merkmale entfernt. Dies betrifft vor allem sensible Angaben wie Vertragsdaten und Merkmale mit sehr niedrigen Besetzungszahlen.

3.3 Spezifische Anonymisierung

3.3.1 Merkmalskategorien

In den fünf unter Abschnitt 2.1 beschriebenen Anonymisierungsbereichen wurden unterschiedliche Merkmale vergrößert 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 und B)⁹
- Gesamtbetrag der Einkünfte
- zu versteuerndes Einkommen
- tarifliche Einkommensteuer
- festzusetzende Einkommensteuer

Merkmale der zweiten Kategorie:

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

*Alle weiteren stetigen Merkmale zählen zur dritten Kategorie.*¹⁰

3.3.2 Anonymisierungsmaßnahmen in den spezifischen Bereichen

Erster Bereich (von 0 bis zu zum doppelten durchschnittlichen GdE):

Zusätzlich zur allgemeinen Anonymisierung wurden beim Alter „15 Jahre“ und „70 Jahre“ 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. Ferner sind in den Daten Altersangaben zu den ersten drei Kindern enthalten. Die Gewerkekennzahl (GKZ) wurde auf eine Stelle reduziert. Die stetigen Merkmale wurden in diesem Bereich nicht verändert.

Zweiter Bereich (vom doppelten durchschnittlichen GdE bis zum 99. Perzentil):

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

⁹ A bezeichnet dabei männliche und B weibliche Steuerfälle.

¹⁰ Vgl. Höhne, Sturm, Vorgrimler, Konzept zur Schutzwirkung faktischer Anonymisierung, in *Wirtschaft und Statistik*, 4/2003, S. 287-292.

sind und 0 bei jüngeren Kindern. Alle weiteren Merkmale wurden wie im ersten Bereich behandelt.

Dritter Bereich (vom 99. Perzentil bis zum 99,95. Perzentil):

Im dritten Bereich ist das Merkmal Alter 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 (vom 99,95. Perzentil 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 und -1 negative Werte darstellen. Die Merkmale der Kategorien 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, welche das Vorhandensein eines Merkmals anzeigen. Die Merkmale der dritten Kategorie wurden gelöscht. Die Ausprägungen der drei Merkmalsträger mit den höchsten Einkünften 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. Hiervon betroffen sind jeweils die drei Steuerpflichtigen mit den höchsten Einkünften des Mannes und denen der Frau.

Bei den diskreten Merkmalen wurde die GKZ gelöscht und die Freiberufler nur noch als Dummy angegeben (zu den Freiberuflern vgl. Abschnitt 4). Die Anzahl der Kinder wird nicht mehr angegeben, sondern nur noch, ob Kinder vorhanden sind.

Sonderbereich „negative Einkommen“:

Steuerpflichtige mit negativen Einkommen wurden innerhalb dreier Bereiche anonymisiert. Die durchgeführten Anonymisierungsmaßnahmen entsprechen denen in 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 Tabelle 3 sind die getroffenen Anonymisierungsmaßnahmen für die unterschiedlichen Teilbereiche zusammengefasst.

Tabelle 3: Anonymisierungsmaßnahmen in den speziellen Bereichen

Merkmal	Anonymisierungsbereich 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	Dummy 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
	2	Ja	Ja	Ja	Ja, aber A + B als Summe
	3	Ja	Ja	Ja	Dummy
					Dummy (+ Bedeutungsmerkmale)
					Nein

¹ Zur Zuordnung der Steuerpflichtigen zu den Bereichen siehe Tabelle 1.

4 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.

Es wurde eine in allen Anonymisierungsbereichen enthaltene Dummy-Variable eingeführt, die angibt, ob der Merkmalsträger freiberuflich tätig ist. Ein Steuerfall wird als Freiberufler (EF59 für Mann / EF61 für Frau) klassifiziert, wenn „Einkünfte aus freiberuflicher Tätigkeit“, „Einkünfte aus Beteiligungen“ oder „Einkünfte laut gesonderter Feststellung aus selbständiger Arbeit“ vorhanden sind.

Zusätzlich wurde in allen außer dem fünften Anonymisierungsbereich (vgl. Tabelle 3) der Steuerfall anhand der GKZ einer von neun Kategorien für die freien Berufe zugeordnet. 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
- 9 Sonstige

Hat ein Steuerpflichtiger keine „Einkünfte aus freiberuflicher Tätigkeit“ sondern nur „Einkünfte aus Beteiligungen“ oder „Einkünfte laut gesonderter Feststellung aus selbständiger Arbeit“ und lässt er sich Anhand der Gewerbekennzahl nicht in die ersten acht Kategorien der Klassifizierung (1-8) einordnen, wird er nicht als Freiberufler ausgewiesen. Dieses Vorgehen entspricht dem bei der Erstellung von Standardveröffentlichung und sichert somit die Vergleichbarkeit der mit FAST 2007 ermittelten Ergebnisse mit denen der amtlichen Statistik.

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 Kategorien 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 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).

5 Fazit

Mit der vorgelegten Datei „Fast 2007“ ist es gelungen, die Reihe der faktisch anonymisierte Daten aus der Lohn- und Einkommensteuerstatistik, die der Wissenschaft zu Analysezwecken zur Verfügung gestellt werden können, fortzusetzen. Auch wenn bei der Anonymisierung größten Wert auf den Erhalt des Analysepotenzials gelegt wurde, sind nicht alle Fragestellungen der Wissenschaft exakt mit den Daten analysierbar¹¹. 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 dieses Datenangebot einen großen Mehrwert für die Wissenschaft dar. Das umfangreiche Material ist für die Wissenschaft zu geringen Kosten zugänglich. Darüber hinaus stellt FAST trotz der Anonymisierung bei den höchsten Einkommen konsequent detaillierte Informationen über die Bezieher der höchsten Einkommen bereit. Gerade dies ist ein Bereich, der in anderen Datenquellen entweder nicht oder nur sehr unzureichend abgebildet ist.

Wiesbaden, April 2013

¹¹ Insbesondere Merkmale mit niedrigen Besetzungszahlen und bzw. oder sehr schiefen Verteilungen können sowohl durch die Stichprobenziehung als auch durch die angewendeten Anonymisierungsmaßnahmen größeren Verfälschungen unterworfen sein. Ein Vergleich der hochgerechneten Ergebnisse kritischer Variablen mit der der DVD beigefügten Randauszählung des Gesamtmaterials kann hierzu Hinweise liefern.

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.

Merz, J.; Vorgrimler, D.; Zwick, M.: Faktisch anonymisiertes Mikrodatenfile der Lohn- und Einkommensteuerstatistik 1998, in: *Wirtschaft und Statistik*, Heft 10, 2004, S. 1079-1090.

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