Faktische Anonymisierung der Steuerstatistik (FAST)
- Lohn- und Einkommensteuer 2020 -

Anna Schrahe und Moritz Wittmaack

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 2020 ist die 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“.[2] Das Unverhältnismäßigkeitsgebot impliziert, dass eine Verletzung der Anonymität von Merkmalsträgern nur bei nutzbringenden Zuordnungen gegeben ist.[3] 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.[4]

2 Hinweise zum Vergleich mit FAST 2017

Im Vergleich zu FAST 2017 wurde das Anonymisierungskonzept sprachlich angepasst: Die Bezeichnungen „Steuerfall A“ und „Steuerfall B“ treten an die Stelle von „Mann“ und „Frau“, da in FAST nicht zwischen zusammenveranlagten Ehepaaren unterschiedlichen Geschlechts, gleichgeschlechtlichen Ehepaaren und eingetragenen Lebenspartnerschaften unterschieden werden kann. Daher können bei Zusammenveranlagungen hinter den Angaben von Steuerfall A und Steuerfall B jeweils sowohl Männer als auch Frauen stehen.[5]

---

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

[2] Für Hochschulen oder sonstige Einrichtungen mit der Aufgabe unabhängiger wissenschaftlicher Forschung gibt es innerhalb speziell abgesicherter Bereiche im Statistischen Bundesamt und in den statistischen Ämtern der Länder die Möglichkeit, Zugang zu formal anonymisierten Einzeldaten zu erhalten (§ 16 Abs. 6 S. 1 Nr. 2).

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

[4] 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.

[5] Darüber hinaus wird das Geschlecht bei der Einkommensteuerveranlagung nicht explizit festgestellt. Für die Lohn- und Einkommensteuerstatistik wird das Geschlechtsmerkmal technisch abgeleitet. In Folge dessen ist die Qualität des Geschlechtsmerkmals eingeschränkt. Für einzeln oder getrennt Veranlagte wird das Geschlecht aus dem Anredeschlüssel bei der Finanzverwaltung ermittelt. Die Angaben der als Frauen abgeleiteten einzeln oder getrennt veranlagten Steuerpflichtigen werden im Zuge der Datenaufbereitung auf die B-Kennzahlen umgesetzt, während die Angaben der als Männer abgeleiteten einzeln oder getrennt veranlagten Steuerpflichtige auf den A-

Auch die Datenaufbereitung wurde angepasst: Zur Ermittlung der Kinderanzahl wurden neben den in den Vorjahren verwendeten Geburtsdaten der Kinder weitere Merkmale herangezogen. Somit ist dieses Merkmal nun auch für nichtveranlagte Fälle befüllt. Des Weiteren wurde die maximale Höhe des Entlastungsbetrags für Alleinerziehende erstmals auf vier Kinder begrenzt. Zur Begrenzung des Kindergelds sowie den Kinderfreibeträgen für die Riester-Rente auf maximal vier Kinder wurden die 2020 gültigen Beträge genutzt.

Die Datensatzbeschreibung wurde ebenfalls überarbeitet. In den Reitern „Feste Felder" und „Datenkatalog" weist die Spalte „Verändert gegenüber Querschnitt im Zuge der FAST-Datenaufbereitung" mit der Ausprägung „Ja" alle festen Felder bzw. Kennzahlen aus, welche aufgrund von allgemeinen Anonymisierungsmaßnahmen im Vergleich zu den Querschnittsdaten verändert wurden (siehe Kapitel 3.2). Für die sonstigen, mit „Nein" ausgezeichneten festen Felder bzw. Kennzahlen gelten die von dem Anonymisierungsbereich des Steuerpflichtigen abhängigen Anonymisierungsmaßnahmen (siehe Kapitel 3.3.2). Im Reiter „Feste Felder" wurde die Spalte „Merkmalsname im Querschnitt" hinzugefügt. Sofern das Merkmal im Querschnitt eine Entsprechung besitzt, dann ist dieser Merkmalsname hier eingetragen. Im Reiter „Kennzahlen" ist für jede Kennzahl die Anzahl der Datensätze mit negativen oder positiven Ausprägungen („Beobachtungen"), die Anzahl der Datensätze ohne Ausprägung bzw. mit der Ausprägung Null („Missings (leer oder Null)"), die gewichtete Summe der Kennzahlen („Summe (gewichtet)"), ihr gewichtetes arithmetisches Mittel („Mittelwert (gewichtet)") und der gewichtete Medianwert („Median (gewichtet)") angegeben, sowohl für FAST 2020 als auch für das Vollmaterial 2020. Forschende erhalten somit einen Anhaltspunkt, ob zur finalen Beantwortung der gewählten Forschungsfrage ein Rückgriff auf das Vollmaterial empfehlenswert ist, oder ob FAST in Hinblick auf die Abdeckung der interessierenden Kennzahlen ausreicht. Die Reiter „Weggefallen" und „Neu" weisen die festen Felder und Kennzahlen aus, welche im Vergleich zu FAST 2017 weggefallen bzw. neu hinzugekommen sind.

Der Umfang der in FAST 2020 enthaltenen Kennzahlen wurde im Vergleich zu FAST 2017 stark gestrafft. Unter anderem wurden alle Kennzahlen ausgeschlossen, welche weniger als 10 Beobachtungen aufwiesen.

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 Aufdeckungsrisiko 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 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

Kennzahlen verbleiben. Bei zusammenveranlagten Ehepaaren unterschiedlichen Geschlechts wird vor dem Hintergrund der entsprechenden Befüllungsvorschrift auf dem Hauptvordruck der Steuerfall A als Ehemann und der Steuerfall B als Ehefrau angesehen. Bei gleichgeschlechtlichen Ehepaaren sowie bei eingetragenen Lebenspartnerschaften stehen weitere Merkmale zur Ableitung der Geschlechter zur Verfügung. Diese weiteren Merkmale gehören jedoch aufgrund der angestrebten Minimierung des Aufdeckungsrisikos einzelner Steuerpflichtiger nicht zum Merkmalskranz von FAST.

Einkommensbereiche untergliedert und jeweils auf diese Bereiche abgestimmte Anonymisierungen durchgeführt (Tannenbaumanonymisierung). Die eingesetzten Anonymisierungsmaßnahmen beschränken sich auf die traditionellen Anonymisierungsmethoden.[6]

Mit Hilfe des Gesamtbetrags der Einkünfte (GdE) wurden die Daten bei den positiven Einkünften in fünf Bereiche unterteilt (vgl. Tabelle 1 und rechter Teil Abbildung 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 die höchsten GdE aufweisen, abdeckt. Den fünften Bereich bilden die 1 000 Steuerpflichtigen mit den höchsten GdE. Die zwanzig Steuerpflichtigen mit den höchsten Einkünften, getrennt nach Steuerfall A und Steuerfall B, stellen eine Untergruppe in diesem Bereich dar: Im Unterschied zu den übrigen Fällen im fünften Bereich werden die Ausprägungen der individuellen stetigen Merkmale dieser jeweils zehn Merkmalsträger durch die arithmetischen Mittel der Ausprägungen ersetzt (Anonymisierungsbereich 6).

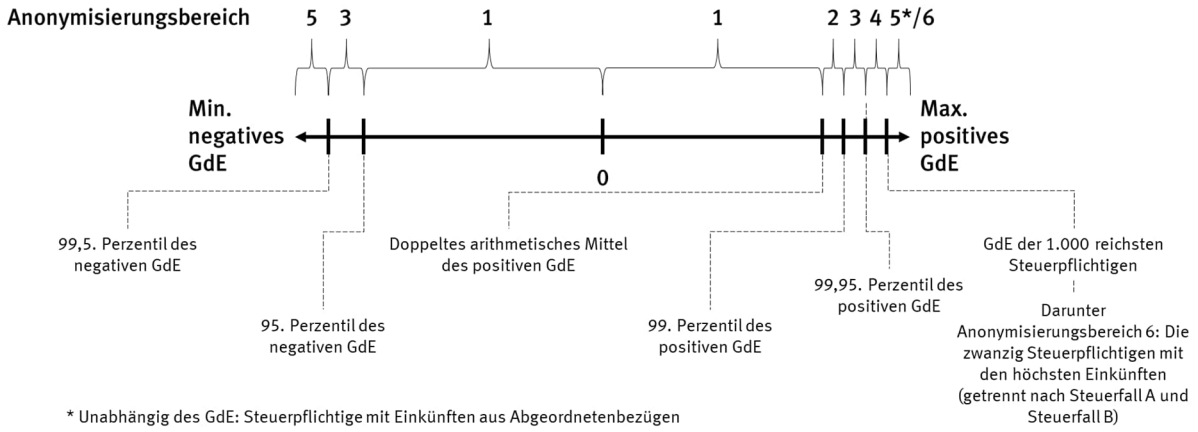
Tabelle 1: Einteilung der Anonymisierungsbereiche auf Basis des GdE

Anonymisierungs-bereich	Positiver GdE in €	Negativer GdE in €
1	0 bis unter 87 762 (zweimal der durchschnittliche GdE)	-1 bis -47 670 (95. Perzentil)
2	87 762 bis unter 243 874 (99. Perzentil)	–
3	243 874 bis unter 1 209 899 (99,95. Perzentil)	-47 670 bis -425 050 (99,5. Perzentil)
4	1 209 899 bis unter 8 636 178 (bis zu den 1.000 Reichsten)	–
5	8 636 178 oder mehr Darunter als Anonymisierungsbereich 6: 10 Steuerpflichtige mit höchsten Einkünften Steuerfall A und 10 Steuerpflichtige mit höchsten Einkünften Steuerfall B	Weniger als -425 050

Bei den Steuerpflichtigen mit negativem GdE wurden zur Anonymisierung drei Bereiche gebildet (vgl. Tabelle 1 und linker Teil Abbildung 1). Der erste Bereich enthält die Merkmalsträger, deren negativer GdE zwischen dem 1. und 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 dem absolut höchsten negativem GdE fallen. Die Anonymisierungsmethoden in diesen Bereichen sind mit den Methoden in den Bereichen eins, drei und fünf der Merkmalsträger mit positivem GdE identisch.

[6] 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.

Abbildung 1: Einteilung der Anonymisierungsbereiche auf Basis des GdE



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 rund 4,2 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.[7] Aus ebendiesem Grund sind kleinere homogene Gruppen von Merkmalsträgern, sogenannte Ränder der Stichprobe, als Vollerhebung enthalten. Für diese besitzt ein Datenangreifer somit weiterhin Teilnahmekenntnis, so dass das oben genannte 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.

[7] 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[8]

Eingabefeld	Merkmal(e)	Maßnahme
EF1	Merker	Umkodierung der elf Ausprägungen in: 01 = veranlagte Fälle 02 = nicht veranlagte Fälle
EF13, EF14	Religion (jeweils getrennt für Steuerfall A und Steuerfall B)	Umkodierung der zwölf Ausprägungen in: 01 = evangelisch 02 = katholisch 03 = sonstige 04 = konfessionslos
EF19	Grund-/Splittingtabelle	Umkodierung der zwei Ausprägungen in: 1 = Grundtabelle 2 = Splittingtabelle
EF64, EF68	Alter (jeweils getrennt für Steuerfall A und Steuerfall B)	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.
EF73, EF74, EF75	Alter Kinder (erstes, zweites, drittes Kind)	Einführung einer Obergrenze. Bei allen Kindern ab 27 Jahren wird das Alter auf 27 Jahre festgesetzt.
EF80, EF81	Anzahl Riester-Verträge (jeweils getrennt für Steuerfall A und Steuerfall B)	Merkmal, das angibt, ob einer oder mehrere Riester-Verträge vorhanden sind. Diese werden als Merkmale der dritten Kategorie behandelt.
c22191, c22192, c22391, c22392	Einnahmen aus nebenberuflichen Tätigkeiten	Die Angaben zu Einnahmen aus der 1. und 2. nebenberuflichen Tätigkeit wurden zusammengefasst.
c25850, c25851	Überschüsse aus Objekten	Zusammenfassung der Überschüsse des 4. und aller weiteren Objekte (getrennt nach Steuerfall A und B).
c36102-c36403, c65399, c65432, c65748, c65879, c65880, c65881, c66998, c65897, c65898, c65899, c65319	Kennzahlen zu den Kindern sowie Kennzahlen, welche von Kinderzahl abhängen	Die Merkmale der ersten vier Kinder wurden auf die Verwandtschaftsverhältnisse zu Steuerfall A und B reduziert. Die Anzahl und Angaben zum Alter der Kinder sind in den Daten enthalten. Bei mehr als vier Kindern wurde die Anzahl auf vier gesetzt. Kindergeld, Kinderfreibeträge für die Riester-Rente, Schulgeld, Kinderbetreuungskosten und der Entlastungsbetrag für Alleinerziehende 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 unter anderem 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.

[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 3.1 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 und B)
- 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 selbstständiger Arbeit (A und B)
- Einkünfte aus nichtselbstständiger 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.[9]

3.3.2 Anonymisierungsmaßnahmen in den spezifischen Bereichen

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

Beim Alter „15 Jahre“ und „70 Jahre“ wurde eine Unter- und eine Obergrenze 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 entsprechend der allgemeinen Anonymisierung bis zu 26 Jahren einzeln und ab 27 Jahren und älter zusammengefasst enthalten. Zusätzlich zur allgemeinen Anonymisierung wurde die Gewerbekennzahl (GKZ) auf den Wirtschaftszweig-Abschnitt reduziert. Die stetigen Merkmale sind in diesem Bereich unverändert.

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 wurde jeweils mit einer Dummy-Variablen

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

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

Im dritten Bereich wurde das Merkmal Alter mit einer Klassenbreite von 10 Jahren klassifiziert. Die Regionen sind nur noch mit West (alte Bundesländer) und Ost (neue Bundesländer inklusive Berlin) beschrieben. Des Weiteren wurden die Ausprägungen des Merkmals Religion gelöscht. Die stetigen Merkmale bleiben 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 kategoriale Variablen umkodiert, wobei 0 keine Werte, 1 positive und -1 negative Werte darstellen. Die Merkmale der Kategorien eins und zwei bleiben unbehandelt, mit Ausnahme der sieben Einkunftsarten. Diese sind nur als Summen der Angaben von Steuerfall A und Steuerfall B angegeben. Bei den diskreten Merkmalen ist die Dummy-Variable für das Alter der Kinder nicht mehr enthalten. Die weiteren diskreten Merkmale wurden 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 kategoriale Variablen ersetzt, wobei 0 keine Werte, 1 positive und -1 negative Werte darstellen. Die Merkmale der dritten Kategorie wurden gelöscht. Die Ausprägungen der zwanzig Merkmalsträger mit den höchsten Einkünften wurden durch die Durchschnittswerte ihrer jeweiligen Ausprägungen, getrennt nach Steuerfall A und Steuerfall B, ersetzt. So entsprechen die Maxima der Merkmale der ersten Kategorie nicht mehr den Originalwerten, sondern stellen die arithmetischen Mittel der höchsten Werte dar. Hiervon betroffen sind jeweils die zehn Steuerpflichtigen mit den höchsten Einkünften von Steuerfall A und Steuerfall B. Des Weiteren wurde für diese zwanzig Steuerpflichtigen ab 2010 eine sehr grobe Altersklassifikation eingeführt und Informationen zur Region und Freiberuflern geleert.

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 „negativer GdE“:

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 Steuerpflichtige, 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	6
Religion		4 Ausprägungen	4 Ausprägungen	k. A.	k. A.	k. A.	k. A.
Kirchen-steuer-festset-zung		3 Ausprägungen	3 Ausprägungen	k. A.	k. A.	k. A.	k. A.
Kinder		Anzahl bis 4, Alter der ersten 3 Kinder	Anzahl bis 4, Alter der ersten 3 Kinder als Dummy	Anzahl bis 4, Alter der ersten 3 Kinder als Dummy	Anzahl bis 4	Dummy ja / nein	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	Jünger als 50 Jahre / mindestens 50 Jahre
Region		Bundesland	Bundesland	West/Ost	West/Ost	West/Ost	k. A.
GKZ		1-Steller	1-Steller	1-Steller	1-Steller	k. A.	k. A.
Frei-berufler		9 Ausprägungen	9 Ausprägungen	9 Ausprägungen	9 Ausprägungen	Dummy ja / nein	k. A.
Stetige Merkmale	1	Ja	Ja	Ja	Ja	Ja	Ja, als Durchschnitt
	2	Ja	Ja	Ja	Ja, aber A + B als Summe	Kategorial (+ Bedeutungs-merkmale)	Kategorial (+ Bedeutungs-merkmale)
	3	Ja	Ja	Ja	Dummy	Nein	Nein

[1] Zur Zuordnung der Steuerpflichtigen zu den Bereichen siehe Tabelle 1 und Abbildung 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.[10] Ein Steuerfall wird als Freiberufler (EF59 für Steuerfall A / EF61 für Steuerfall B) 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 und sechsten Anonymisierungsbereich (vgl. Tabelle 3) der Steuerfall anhand der GKZ einer von neun Kategorien für die freien Berufe zugeordnet (EF58 für Steuerfall A / EF60 für Steuerfall B). Dabei handelt es sich um folgende Kategorien:

01 Technische Beratung, Forschung, Architekten, Ingenieur

02 Rechtsanwälte, Notar

03 Wirtschaftsprüfer, -berater

04 Ärzte

05 Sonstige Gesundheitsberufe

06 Werbung, Foto, Kunst und Kultur

07 Schriftberufe

08 Schulen

09 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 GKZ 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 Standardveröffentlichungen und sichert somit die Vergleichbarkeit der mit FAST 2020 ermittelten Ergebnisse mit denen der amtlichen Statistik.

Im Anonymisierungsbereich 5 sind die Merkmale der zweiten Kategorie nur noch als kategoriale Merkmale enthalten. Damit die Datennutzenden 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 (EF77, EF78, EF79). 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.

[10] Für die zehn Steuerpflichtigen mit den höchsten Einkünften von Steuerfall A bzw. den zehn Steuerpflichtigen mit den höchsten Einkünften von Steuerfall B wurde diese Information gelöscht (siehe Kapitel 3.3.2).

Als weitere Zusatzinformation wurde ein Merkmal eingeführt, welches die Anonymisierungsstärke des jeweiligen Merkmalsträger angibt (EF79). Die Ausprägungen 1 bis 6 geben hierbei die Anonymisierungsbereiche wieder. 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 2020“ ist es gelungen, die Reihe der faktisch anonymisierten Daten aus der Lohn- und Einkommensteuerstatistik, die der Wissenschaft zu Analysezwecken zur Verfügung gestellt werden können, fortzusetzen. Auch wenn bei der Anonymisierung größter Wert auf den Erhalt des Analysepotenzials gelegt wurde, sind nicht alle Fragestellungen der Wissenschaft exakt mit den Daten analysierbar.[11] 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.[12]

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 GdE detaillierte Informationen über die Bezieher dieser Einkommen bereit. Gerade dies ist ein Bereich, der in anderen Datenquellen entweder nicht oder nur sehr unzureichend abgebildet ist.

Wiesbaden, September 2024

[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. Die neu hinzugefügten Spalten im Reiter „Feste Felder“ der Datensatzbeschreibung können Hinweise hierzu liefern.

[12] https://www.forschungsdatenzentrum.de/de/steuern/lest (abgerufen am 02.09.2024)

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.