Prof. Dr. Joachim Merz, Dr. Daniel Vorgrimler, Dipl.-Volkswirt Markus Zwick[1])

Faktisch anonymisiertes Mikrodatenfile der Lohn- und Einkommensteuerstatistik 1998

Mit den Daten der faktisch anonymisierten Lohn- und Einkommensteuerstatistik 1998 (FAST 98) veröffentlicht die deutsche amtliche Statistik erstmals Mikrodaten aus dem Bereich der Steuerstatistik. Mit diesen Daten kann die Wissenschaft, unter den Prämissen des § 16 Abs. 6 des Gesetzes über die Statistik für Bundeszwecke, auf Grundlage „echter" Veranlagungsdaten politisch relevante Fragestellungen zum Steuer- und Transfersystem am eigenen Arbeitsplatz analysieren.

Eine Weitergabe von Einzeldaten an die Wissenschaft ist nur in faktisch anonymisierter Form möglich. In dieser Form können wissenschaftliche Analysemöglichkeiten beeinträchtigt sein. Damit anonymisierte Daten dennoch 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 in bestmöglicher Weise erhalten. Um die richtige Balance der beiden Ziele zu erreichen, wurden von den statistischen Ämtern im Rahmen eines Forschungsprojektes potenzielle wissenschaftliche Nutzer in die Anonymisierungsarbeiten integriert.

In dem Beitrag „Faktisch anonymisiertes Mikrodatenfile der Lohn- und Einkommensteuerstatistik 1998" werden neben der Anonymisierungskonzeption die Rahmenbedingungen des Projektes erläutert und die Analysemöglichkeiten der Lohn- und Einkommensteuerstatistik aufgezeigt.

1 Einführung

1.1 Faktische Anonymität von Mikrodaten

Mikrodaten sind der „Rohstoff" des Statistikers. Die persönlichen oder sachlichen Informationen über einzelne Merkmalsträger, seien es Personen, Haushalte oder Unternehmen, sind die Ausgangsinformationen, die im statistischen Produktionsprozess verdichtet werden und zum Beispiel in Form von Tabellen eine übersichtliche Darstellung von Massenerscheinungen erlauben. War es bis weit in die sechziger Jahre des vorigen Jahrhunderts in der Regel nur den statistischen Ämtern möglich, diese Massendaten zu verarbeiten, so ist es heute durch die rasante Entwicklung der Datenverarbeitung nahezu jedem bzw. jeder Studierenden möglich, große Datenmengen auszuwerten. Da Mikrodaten vielschichtige Analysen erlauben, ist der Wunsch der Wissenschaft[2]), diese Daten in ihrer Urform als Einzeldaten zu analysieren, über die Zeit ständig gewachsen. Der Gesetzgeber reagierte Ende der 1970er-Jahre auf den wachsenden Bedarf an amtlichen Einzeldaten und schuf mit dem § 11 des Gesetzes über die Statistik für Bundeszwecke (Bundesstatistikgesetz – BStatG) vom 14. März 1980 die rechtliche Grundlage, Einzeldaten an Nutzer außerhalb der statistischen Ämter zu übermitteln. Hiernach waren Einzeldaten für eine Übermittlung geeignet, wenn sie so anonymisiert wurden, dass eine Identifizierung der Merkmalsträger mit absoluter Sicherheit ausgeschlossen werden konnte. Die Forderung nach einer absoluten Anonymisierung der

1) Prof. Dr. Joachim Merz, Universität Lüneburg, Fachbereich Wirtschafts- und Sozialwissenschaften, Forschungsinstitut Freie Berufe (FFB); Dr. Daniel Vorgrimler, Dipl.-Volkswirt Markus Zwick, Statistisches Bundesamt.

2) Wenn in diesem Beitrag von der „Wissenschaft" oder den „Wissenschaftlern" gesprochen wird, sind die außerhalb der statistischen Ämter arbeitenden Wissenschaftler angesprochen.

Einzeldaten, so zeigte in der Folge die Praxis, führte dazu, dass nahezu keine Einzeldatenbestände an die Wissenschaft ausgeliefert wurden. Wenn es trotz der hohen Anforderungen ermöglicht wurde, Datenbestände für die Wissenschaft zu erschließen, hatten diese meist einen so geringen Informationsgehalt, dass sie für wissenschaftliche Fragestellungen oftmals nicht mehr ausreichend waren.

Augrund dieser Erfahrungen wurde das BStatG bei der darauffolgenden Novellierung im Jahr 1987 um eine Vorschrift erweitert, nach der Einzeldaten an die Wissenschaft weitergegeben werden dürfen, „wenn die Einzelangaben nur mit unverhältnismäßig großem Aufwand an Zeit, Kosten und Arbeitskraft zugeordnet werden können“ (§ 16 Abs. 6 BStatG). Dieses „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 so genannte faktische Anonymität 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])

Mit den Arbeiten von Müller, Blien, Knoche und Wirth[5]) wurde Anfang der 1990er-Jahre der Begriff der faktischen Anonymität operationalisiert und erste faktisch anonymisierte Einzeldatenbestände aus dem Mikrozensus und in der Folge aus der Einkommens- und Verbrauchsstichprobe an die Wissenschaft übermittelt.

1.2 Die Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder

Die Diskussion um den Datenzugang und insbesondere um den Zugang zu amtlichen Einzeldaten hielt dessen ungeachtet weiter an. Mit dem Memorandum der Professoren Richard Hauser, Gerd Wagner und Klaus F. Zimmermann im Allgemeinen Statistischen Archiv[6]) erhielt die Thematik im Jahr 1998 eine neue Dynamik, die letztlich zur Einrichtung der „Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik“ führte. Das im Jahr 2001 veröffentlichte Gutachten dieser Kommission enthielt vielfältige Empfehlungen zur Verbesserung der informationellen Infrastruktur.[7]) Neben dem Vorschlag, dauerhaft einen Rat für Sozial- und Wirtschaftsdaten zu etablieren[8]), sollte insbesondere die Empfehlung zur Einrichtung von Forschungsdatenzentren bei den Datenproduzenten nachhaltig zu einem verbesserten Datenzugang für die Wissenschaft führen.

Mit der Gründung der beiden Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder in den Jahren 2001 und 2002 reagierte die amtliche Statistik unmittelbar auf die Kommissionsempfehlungen. Das wesentliche Ziel der beiden Forschungsdatenzentren besteht darin, den Zugang der Wissenschaft zu den Mikrodaten der amtlichen Statistik durch den Ausbau bestehender und die Einrichtung neuer Zugangswege zu erleichtern. Hierzu haben die Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder, neben der Einrichtung von Gastwissenschaftlerarbeitsplätzen in den geschützten Räumen der amtlichen Statistik und dem kontrollierten Fernrechnen, mit der weiteren Erstellung von Scientific-Use-Files einen entscheidenden Schritt getan.[9]) In Zusammenarbeit der beiden Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder mit den späteren wissenschaftlichen Nutzern, die als wissenschaftlicher Beirat die Arbeiten begleiteten, ist das „Faktisch anonymisierte Mikrodatenfile der Einkommensteuerstatistik 1998 (FAST 98)“ entstanden. FAST 98 ist über die Forschungsdatenzentren (*www.forschungsdatenzentren.de*) zu einem Preis von 65,– Euro für die Wissenschaft erhältlich.

1.3 Der wissenschaftliche Beirat zum Projekt FAST 98

Anonymisierungsmaßnahmen bedeuten immer eine Informationsreduktion der vorhandenen Einzeldaten, sei es durch Vergröbern, Löschen oder Verfälschen. Die resultierenden anonymisierten Daten verfügen daher über ein – im Vergleich zum Ausgangsdatenmaterial – eingeschränktes Analysepotenzial. Welche Informationen im Anonymisierungsprozess unterdrückt werden, ist zumindest teilweise variabel zu entscheiden. Es ist also in einem bestimmten Rahmen möglich, Informationen zu ersetzen.

Da es diese Substitutionsmöglichkeiten gibt, war zu entscheiden, welche Informationen zur Sicherung der Geheimhaltung der Merkmalsträger in den Daten unterdrückt werden sollen. Bei der Erstellung von FAST 98 wurde diese Entscheidung gemeinsam mit den späteren Nutzern innerhalb eines wissenschaftlichen Beirates diskutiert und getroffen. Die Daten der Lohn- und Einkommensteuerstatistik stehen für wissenschaftliche Fragestellungen (insbesondere im Rahmen der Politikberatung) schon seit geraumer Zeit in den statistischen Ämtern zur Verfügung, sodass bereits zu Projektbeginn ein großer Kreis von späteren Nutzern bekannt war. Diese wurden angeschrieben und zur Mitarbeit an FAST 98 gebeten. Die wissenschaftliche Leitung, die nach Auffassung der beiden Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder von

3) Siehe Höhne, J./Sturm, R./Vorgrimler, D.: „Konzept zur Beurteilung der Schutzwirkung von faktischer Anonymisierung“ in WiSta 4/2003, S. 287 ff.
4) Zu diesen Entwicklungen siehe Krupp, H.-J.: „Mikroanalysen und amtliche Statistik – gestern, heute, morgen“ in Merz, J./Zwick, M. (Hrsg.): „MIKAS – Mikroanalysen und amtliche Statistik“ in Statistik und Wissenschaft, Band 1, Wiesbaden 2004, S. 27 ff.
5) Müller, W./Blien, U./Knoche, P./Wirth, H.: „Die faktische Anonymität von Mikrodaten“, Band 19 der Schriftenreihe „Forum der Bundesstatistik“, Wiesbaden 1991.
6) Hauser, R./Wagner, G./Zimmermann, K.: „Erfolgsbedingungen empirischer Wirtschaftsforschung und empirisch gestützter wirtschafts- und sozialpolitischer Beratung: Ein Memorandum“, Allgemeines Statistisches Archiv, Band 82, S. 369 ff.
7) Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik (Hrsg.): „Wege zu einer besseren informationellen Infrastruktur“, Baden-Baden 2001.
8) Siehe hierzu *http://www.RatSWD.de*.
9) Siehe hierzu Zühlke, S./Zwick, M./Scharnhorst, S./Wende, T.: „Die Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder“ in WiSta 10/2003, S. 906 ff. sowie *www.forschungsdatenzentrum.de*.

einer Persönlichkeit außerhalb der amtlichen Statistik wahrgenommen werden sollte, wurde ausgeschrieben. Der wissenschaftliche Beirat setzte sich schließlich aus folgenden Personen zusammen:

Herr Prof. Dr. Joachim Merz
Universität Lüneburg (wissenschaftlicher Leiter)

Herr Prof. Dr. Dr. Giacomo Corneo
Freie Universität Berlin

Herr Dr. Markus Eltges
Bundesamt für Bauwesen und Raumordnung

Herr Prof. Dr. Heinz Galler
Martin-Luther-Universität Halle-Wittenberg

Herr Hans-Joachim Georg
Bayerisches Landesamt für Statistik und Datenverarbeitung

Herr Joachim Goletz
Landesamt für Datenverarbeitung und Statistik
Nordrhein-Westfalen

Herr Volker Kordsmeyer
Gruppe „Steuern“ des Statistischen Bundesamtes

Herr Dr. Hermann Quinke
Fraunhofer-Institut für Angewandte Informationstechnik

Herr Dr. Claus Schäfer
Hans-Böckler-Stiftung

Herr Prof. Dr. Viktor Steiner
Deutsches Institut für Wirtschaftsforschung

Herr Dr. Stefan Weil
Statistisches Landesamt Rheinland-Pfalz

Frau Dr. Heike Wirth
Zentrum für Umfragen, Methoden und Analysen

Frau Dr. Sylvia Zühlke
Forschungsdatenzentrum der Statistischen Landesämter

Herr Markus Zwick
Forschungsdatenzentrum des Statistischen Bundesamtes

Auf der ersten Sitzung des wissenschaftlichen Beirats wurde die grundsätzliche Vorgehensweise anhand eines von den statistischen Ämtern vorgelegten Anonymisierungskonzeptes diskutiert. Auf der Grundlage der Ergebnisse dieser Sitzung wurde dann ein erstes faktisch anonymisiertes Mikrodatenfile der Einkommensteuerstatistik 1998 erstellt und umfangreich auf ausreichende Anonymität der Merkmalsträger getestet. Das zur zweiten Sitzung vorgelegte weiterentwickelte Konzept zur Anonymisierung wurde wiederum intensiv diskutiert. Dies führte neuerlich zu umfangreichen Erstellungs- und Prüfarbeiten. Auf der dritten Sitzung im April 2004 wurden die Konzepte zur Anonymisierung der Lohn- und Einkommensteuerstatistik 1998 vom wissenschaftlichen Beirat angenommen und auf der Grundlage der vorgelegten Ergebnisse der Geheimhaltungsüberprüfung wurde das Scientific-Use-File von den statistischen Ämtern und den beteiligten Juristen als faktisch anonym eingestuft.

Abschließend kam der wissenschaftliche Beirat zu folgendem Resümee und zu folgenden Empfehlungen:[10])

> „Mit der Erstellung eines Scientific-Use-File der Lohn- und Einkommensteuerstatistik 1998 wird die informationelle Infrastruktur in Deutschland nachhaltig verbessert. Die Lohn- und Einkommensteuerstatistik ist hinsichtlich der Differenziertheit der Einkommensangaben, ihrer Qualität als amtliche Vollerhebung sowie ihrer Möglichkeit, auch höchste Einkommen zu beschreiben, für die Wissenschaft von hohem Interesse.
>
> FAST ist ein dynamisches Produkt. Die praktischen Erfahrungen der damit arbeitenden wissenschaftlichen Nutzer werden gesammelt und in das nächste zu entwickelnde Scientific-Use-File der Lohn- und Einkommensteuerstatistik 2001 mit einfließen, sodass eine methodische Weiterentwicklung gewährleistet ist. Dies bedeutet auch eine permanente Überprüfung des gefundenen Anonymisierungsgrades.
>
> Der Beirat spricht sich dafür aus, auf Grundlage der gesammelten Erfahrungen auch ein FAST-Regionalfile zu entwickeln. So könnte FAST zukünftig zum Beispiel mit Raumordnungsmerkmalen ergänzt werden.“

Der vorliegende Beitrag beschreibt die Vorgehensweise und begründet die Entscheidungen, die zum faktisch anonymen File geführt haben. Während Kapitel 2 ausführlich die Einkommensteuerstatistik und die daraus gezogene 10%-Stichprobe erläutert, findet sich in Kapitel 3 das Konzept zur Anonymisierung der Lohn- und Einkommensteuerstatistik 1998. In Kapitel 4 folgt die Beschreibung der umfangreichen Tests auf Datensicherheit. Ein Ausblick schließt diesen Aufsatz ab.

2 Einkommensteuerstatistik

2.1 Methodische Grundlagen und Struktur der Einzeldaten der Lohn- und Einkommensteuerstatistik 1998[11])

Im § 2 Abs. 2 des Gesetzes über Steuerstatistiken (StStatG)[12]) ist geregelt, dass die Lohn- und Einkommensteuerstatistik alle drei Jahre erhoben wird. Darüber hinaus sind die Merkmale benannt, die erhoben werden. Dazu zählen neben den Merkmalen des Besteuerungsprozesses auch sozioökonomische Merkmale, wie zum Beispiel das Alter oder das Geschlecht der Steuerpflichtigen.

10) Siehe Stellungnahme des wissenschaftlichen Beirats zum Projekt FAST 98 im Anhang zu diesem Beitrag auf S. 1090 f.

11) Zu den Ausführungen in diesem Abschnitt siehe Zwick, M.: „Einzeldatenmaterial und Stichproben innerhalb der Steuerstatistiken“ in WiSta 7/1998, S. 566 ff., sowie Rosinus, W.: „Die steuerliche Einkommensverteilung“ in WiSta 6/2000, S. 456 ff.

12) Artikel 35 des Jahressteuergesetzes 1996 vom 11. Oktober 1995 (BGBl. I S. 1250), zuletzt geändert durch Artikel 56 des Gesetzes vom 23. Dezember 2003 (BGBl. I S. 2848).

Die Lohn- und Einkommensteuerstatistik ist eine dezentral durchgeführte Sekundärstatistik, das heißt die Angaben werden nicht für den statistischen Zweck erhoben, sondern fallen in einem anderem Zusammenhang an, hier im Besteuerungsprozess, und werden in einer zweiten Stufe statistisch genutzt. Die Finanzverwaltungen liefern hierzu die jeweiligen Angaben der Steuerpflichtigen zu vorgegebenen Terminen an die Statistischen Ämter der Länder. Diese stellen die jeweiligen Landesergebnisse zusammen und übermitteln sie an das Statistische Bundesamt. Das Statistische Bundesamt führt dann die Landesergebnisse im nächsten Schritt zum Bundesergebnis zusammen. Durch die Novellierung des StStatG im Rahmen des Jahressteuergesetzes 1996 können nunmehr auch die Einzelangaben von den Statistischen Landesämtern u. a. für Zusatzaufbereitungen an das Statistische Bundesamt übermittelt werden. Durch diese zentrale Verfügbarkeit der Einzeldaten ergeben sich im Rahmen dieser Statistik weit reichende Analysemöglichkeiten.

Als Sekundärstatistik ist die Lohn- und Einkommensteuerstatistik abhängig von den von den Finanzverwaltungen durchgeführten Einkommensteuerveranlagungen. Aufgrund der den Steuerpflichtigen zugestandenen Fristen zur Einreichung ihrer Einkommensteuererklärung vergehen 2¾ Jahre, bis die letzten Daten den jeweiligen Statistischen Landesämtern zur Verfügung stehen. Die Erstellung von statistischen Ergebnissen erfährt bereits hierdurch einen beträchtlichen „time lag“. Die Möglichkeit, die Veröffentlichung der Statistik über eine Hochrechnung erster Ergebnisse zu beschleunigen, wird dadurch erschwert, dass die komplizierten und gewichtigen Fälle in der Regel erst gegen Ende dieses Zeitraumes von der Finanzverwaltung verarbeitet werden. Die dreijährliche Periodizität der Statistik und die Fristen zur Einkommensteuerveranlagung haben zur Folge, dass erst im vierten Jahr nach Ende des betreffenden Veranlagungsjahres Ergebnisse vorliegen und diese zum Teil bis in das siebte Jahr die aktuellsten Ergebnisse bleiben. So sind derzeit im Jahr 2004 die Daten zum Veranlagungsjahr 1998 die aktuellsten Ergebnisse der Lohn- und Einkommensteuerstatistik.

Aufgrund ihrer Datenvielfalt bietet die Lohn- und Einkommensteuerstatistik vielfältige Analysemöglichkeiten. Hierbei können neben rein steuerlichen Betrachtungen auch Untersuchungen über die Einkommensverteilung durchgeführt werden. Besonders die Bezieher hoher und höchster Einkommen sind in keiner anderen statistischen Quelle so genau erfasst wie in der Lohn- und Einkommensteuerstatistik. Dies macht diese Statistik besonders für die Betrachtung dieser gesellschaftlichen Gruppe außerordentlich wertvoll.[13])

Bei Analysen muss allerdings beachtet werden, dass die Einkommensbegriffe der Lohn- und Einkommensteuerstatistik auf dem Steuerrecht basieren. Daher sind die Merkmale nicht ohne weiteres mit denen aus den Volkswirtschaftlichen Gesamtrechnungen (VGR) vergleichbar. Dem Einkommensbegriff der VGR am nächsten kommt der Begriff „Gesamtbetrag der Einkünfte“ (GDE). Aber auch dieser berücksichtigt zum Beispiel nur zum Teil Umverteilungen und orientiert sich eher an dem primären Markteinkommen der Steuerpflichtigen. Das tatsächlich verfügbare Einkommen der Haushalte wird aber durch staatliche Umverteilungen – wie zum Beispiel durch den progressiven Einkommensteuertarif oder die Transfereinkommen, die nur teilweise in der Einkommensteuerstatistik abgebildet werden – beeinflusst. Besonders bei Verteilungsanalysen müssen diese Restriktionen beachtet werden. Dies hat bei der Verwendung der Daten der Lohn- und Einkommensteuerstatistik für Verteilungsanalysen dazu geführt, dass aus den Angaben der Einkommensteuerstatistik in einem ersten Schritt ökonomische Einkommen berechnet wurden.[14])

Die knapp 30 Mill. Einzeldatensätze der Lohn- und Einkommensteuerstatistik 1998 umfassen je Steuerpflichtigen gut 500 Merkmale, die jeweils in unterschiedlicher Anzahl besetzt sind. Die Merkmale weisen den Besteuerungsprozess angefangen von den Einkünften bis hin zur tatsächlichen Steuerschuld für jeden Steuerpflichtigen nach.

Bei den Überschusseinkünften[15]) kann deren Entstehung betrachtet werden (z. B. Bruttolohn minus Werbungskosten). Dies ist bei den Gewinneinkünften[16]) derzeit nicht möglich, da keine Angaben über die Betriebseinnahmen oder -ausgaben in den Daten enthalten sind. Die Entstehung dieser Einkunftsarten bleibt daher außer über die qualitativ eingeschränkten Angaben der Anlage für statistische Zwecke[17]) nicht nachvollziehbar.

Ein Datensatz repräsentiert einen Steuerpflichtigen. Bei einer gemeinsamen Veranlagung von Ehepaaren im Splittingfall besteht ein Steuerpflichtiger aus zwei Personen bzw. zwei Steuerfällen. Daher umfassen die knapp 30 Mill. Einzeldatensätze Angaben von über 42 Mill. Steuerfällen. Bis zum Merkmal „Summe der Einkünfte“ werden dabei die jeweiligen Merkmale für die Ehepartner getrennt ausgewiesen. Im weiteren Besteuerungsverlauf ist dies nicht mehr möglich bzw. nicht mehr sinnvoll. Als Folge der Unterscheidung zwischen Steuerpflichtigen und Steuerfällen ist die steuerliche Einkommensverteilung basierend auf der Verteilung des „Gesamtbetrags der Einkünfte“ keine Verteilung der Individualeinkommen. Sie bildet aber auch nicht exakt die Verteilung der Haushaltseinkommen ab, da innerhalb eines Haushalts mehrere Steuerpflichtige leben können (z. B. bei den Eltern lebende einzeln veranlagte Kinder). Dennoch wird in Analysen in der Regel der Steuerpflichtige

13) So z.B im Rahmen des Armuts- und Reichtumsberichts der Bundesregierung; siehe hierzu Merz, J.: „Hohe Einkommen, ihre Struktur und Verteilung – Mikroanalysen auf der Basis der Einkommensteuerstatistik“ in Lebenslagen in Deutschland – Der erste Armuts- und Reichtumsbericht der Bundesregierung, Bundesministerium für Arbeit und Sozialordnung, Berlin 2001.

14) Siehe Bach, S./Bartholmai, B.: „Möglichkeiten zur Modellierung hoher Einkommen auf Grundlage der Einkommenssteuerstatistik“, DIW-Diskussionspapiere Nr. 212, Berlin 2000.

15) Zu den Überschusseinkünften, die sich aus den Einnahmen abzüglich den Werbungskosten ergeben, zählen Einkünfte aus nichtselbstständiger Arbeit, aus Vermietung und Verpachtung, aus Kapitalvermögen sowie aus sonstigen Einkünften.

16) Zu den Gewinneinkünften, die sich aus den Betriebseinnahmen abzüglich den Betriebsausgaben ergeben, zählen Einkünfte aus Land- und Forstwirtschaft, aus Gewerbebetrieb sowie aus selbstständiger Arbeit.

17) Die Anlage St gehört zwar alle drei Jahre zu den Pflichtangaben dieser Steuerpflichtigen, da aber die Angaben zum Besteuerungsprozess nicht benötigt werden, ist seitens der Finanzverwaltungen die Motivation, einen vollständigen Rücklauf der Anlage St bei den Steuerpflichtigen anzumahnen, gering, von einer Qualitäts- und Plausibilitätsprüfung ganz zu schweigen.

als Approximation des Haushalts verwendet und die Einkommensverteilung auf Haushaltsbasis berechnet.[18])

Wie zu Beginn dieses Abschnitts beschrieben, weisen die Datensätze neben den quantitativen Merkmalen des Besteuerungsprozesses auch sozioökonomische Merkmale aus, die eine gezielte Analyse einzelner gesellschaftlicher Gruppen ermöglichen. Zu diesen Merkmalen zählen u. a. das Geschlecht, die regionale Gliederung, die Religion, das Alter und bei Gewerbetreibenden der Wirtschaftszweig (Gewerbekennzahl, GKZ93).

2.2 Die Stichprobe aus der Lohn- und Einkommensteuerstatistik

Die Stichprobe aus der Lohn- und Einkommensteuerstatistik ist in § 7 Abs. 4 StStatG gesetzlich verankert und u. a. als 10%-Stichprobe vorgeschrieben. Sie dient zur Durchführung von Zusatzaufbereitungen zur Abschätzung finanzieller und organisatorischer Auswirkungen der Änderungen von Regelungen im Rahmen der Fortentwicklung des Steuer- und Transfersystems.[19]) Die Stichprobenpläne werden hierzu zentral im Statistischen Bundesamt ausgearbeitet. Gezogen werden die Stichproben aus den von den Statistischen Landesämtern übermittelten Einzeldaten des Gesamtmaterials. Angelegt ist die Stichprobe 1998 wie in früheren Jahren als geschichtete Zufallsstichprobe. Dabei diente als Auswahlkriterium eine hohe Genauigkeitsanforderung insbesondere an den Nachweis des Gesamtbetrags der Einkünfte.

Das StStatG fordert eine „bundesweit repräsentative" Stichprobe. Daher wurde bei den Stichproben für die Veranlagungsjahre 1992 und 1995 auf das Bundesland als Schichtungsmerkmal verzichtet und lediglich eine Schichtung nach alten und neuen Bundesländern vorgenommen.[20]) Um auch möglichst genaue Länderanalysen zu ermöglichen, wurde in die Stichprobe für das Veranlagungsjahr 1998 das Bundesland als Schichtungsmerkmal aufgenommen. Übersicht 1 vergleicht die jeweiligen Schichtungsmerkmale der für die Veranlagungsjahre 1992, 1995 und 1998 gezogenen Stichproben.

Zwar blieben die Schichtungsmerkmale im Prinzip erhalten, die jeweilige Anzahl der Kategorien veränderte sich jedoch. Dies gilt insbesondere für die aktuellste Stichprobe des Jahres 1998. Zu den 2 016 Schichten, die sich hier durch die vollständige Kombination der Merkmalsausprägungen ergeben, kommen noch weitere 32 Schichten hinzu. Diese bestehen aus den so genannten manuellen Fällen (Fällen mit Lohnsteuerkarten, die nicht veranlagt werden). Bei diesen wurde lediglich nach den 16 Bundesländern sowie nach zwei Einkommensklassen geschichtet. Insgesamt bleibt die Anzahl der Schichten 1998 mit 2 048 hinter der Anzahl von 1995 (2 704 Schichten, einschl. „Sonderschichten") zurück. 1992 wurden lediglich 1 344 Schichten gebildet.[21])

Gering besetzte Schichten sind in der Stichprobe in der Regel als Vollerhebung, das heißt mit allen Merkmalsträgern, enthalten. Des Weiteren wurden alle Merkmalsträger mit einem Gesamtbetrag der Einkünfte oberhalb von 102 257 Euro (200 000 DM) aufgrund ihrer Heterogenität vollständig in die Stichprobe übernommen.

3 Anonymisierungskonzept für die Lohn- und Einkommensteuerstatistik 1998

Als Datenbasis für die Anonymisierung der Lohn- und Einkommensteuerstatistik 1998 diente die im Abschnitt 2.2 beschriebene 10%-Stichprobe. Das in der Einführung beschriebene Unverhältnismäßigkeitsgebot ist nur eine notwendige Bedingung für ein Scientific-Use-File. Diese Bedingung gewährleistet die faktische Anonymität von Daten, nicht aber die Verwendbarkeit der Daten für wissenschaftliche Analysen. Faktisch anonyme Daten sollten sinnvollerweise nur dann von den statistischen Ämtern als Scientific-Use-File der Wissenschaft zur Verfügung gestellt werden, wenn sie ausreichende wissenschaftliche Analysemöglichkeiten bieten. Da eine Anonymisierung von Merkmalsträgern immer eine Reduktion von Information impliziert, folgt daraus, dass eine Anonymisierung auf das Notwendigste zu beschränken ist. Um dies zu erreichen, sind bei FAST 98 die Merkmalsträger in Abhängigkeit von ihrem Reidentifikati-

Übersicht 1: Schichtungsmerkmale der Stichproben der Lohn- und Einkommensteuerstatistik

1992		1995		1998	
Merkmal	Kategorien	Merkmal	Kategorien	Merkmal	Kategorien
Bundesland neu/alt	2	Bundesland neu/alt	2	Bundesland	16
Veranlagungsart	4	Veranlagungsart	4	Veranlagungsart	2
Kinderfreibetragsschritte	4	Kinderfreibeträge	4	Kinder	3
Überwiegende Einkunftsart	7	Überwiegende Einkunftsart	7	Überwiegende Einkunftsart	3
Gesamtbetrag der Einkünfte	6	Gesamtbetrag der Einkünfte	12	Gesamtbetrag der Einkünfte	7

18) Z. B. siehe Bach, S./Haan, P./Rudolph, H.-J./Steiner, V.: „Reformkonzepte zur Einkommens- und Ertragsbesteuerung: Erhebliche Aufkommens- und Verteilungswirkungen, aber relativ geringe Effekte auf das Arbeitsangebot" in DIW-Wochenbericht 16/2004, sowie Rosinus, W.: „Die steuerliche Einkommensverteilung" in WiSta 6/2000, S. 456 ff.

19) § 7 Abs. 4 StStatG.

20) Siehe Zwick, M.: „Einzeldatenmaterial und Stichproben innerhalb der Steuerstatistiken" in WiSta 7/1998, S. 570.

21) Siehe Fußnote 20.

onsrisiko anonymisiert worden. Diejenigen mit einem geringeren Reidentifikationsrisiko sind entsprechend schwächer anonymisiert als diejenigen mit erhöhten Risiken.[22]) Des Weiteren wurde auf die Verwendung datenverändernder Anonymisierungsverfahren, wie sie vor allem bei wirtschaftsstatistischen Einzeldaten getestet werden[23]), verzichtet. Es kamen nur Verfahren zum Einsatz, die bereits seit längerem bei anderen personenbezogenen Einzeldaten innerhalb der statistischen Ämter Verwendung finden.[24])[25])

3.1 Das Prinzip Tannenbaumanonymisierung

Nicht jeder der rund 2,8 Mill. Merkmalsträger der Stichprobe konnte individuell auf sein Reidentifikationsrisiko hin überprüft werden. Vielmehr wurde angenommen, dass das Risiko der Reidentifikation mit der Einkommenshöhe zunimmt. Auf Grundlage dieser Annahme wurden die Merkmalsträger in verschiedene Einkommensbereiche eingeteilt und erhielten so einen Indikator für ihr Risiko. Innerhalb der Anonymisierungsbereiche wurden speziell auf das Risiko abgestimmte Anonymisierungsmaßnahmen durchgeführt (siehe Übersicht 2). Analog zum Tannenbaum, der mit steigender Stammhöhe weniger Grün aufweist, weisen die Daten mit steigendem Einkommen aufgrund der Anonymisierungsmaßnahmen weniger Informationen auf (so genannte Tannenbaumanonymisierung).

Übersicht 2: Einteilung der Anonymisierungsbereiche

Anonymisierungsbereich	Positiver Gesamtbetrag der Einkünfte in EUR (DM)	Negativer Gesamtbetrag der Einkünfte/Einkommen in EUR (DM)
1	0 bis 64 106 (0 bis 125 381) (zweimal der durchschnittliche Gesamtbetrag der Einkünfte)	0 bis – 102 258 (0 bis – 200 000)
2	64 107 bis 137 532 (125 382 bis 268 990) (99%-Percentil)	–
3	137 533 bis 970 202 (268 991 bis 1 897 552) (99,95%-Percentil)	– 102 259 bis – 511 292 (– 200 001 bis – 1 000 000)
4	970 203 bis 7 354 714 (1 897 553 bis 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 (siehe Übersicht 2). Der erste erstreckt sich von einem GDE von Null bis zum Doppelten des durchschnittlichen 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 den höchsten GDE aufweisen, abdeckt. Den fünften Bereich bilden die 1 000 Personen mit dem höchsten GDE. In den Fällen, bei denen kein GDE vorlag, wurde der „gesamte Bruttolohn“ als Einteilungsmerkmal 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 (siehe Übersicht 2). Der erste Bereich umfasst diejenigen Merkmalsträger, deren negatives Einkommen zwischen – 1 und dem 95%-Percentil der absoluten negativen Einkommensverteilung liegt. Der zweite Bereich erstreckt sich von dieser Grenze bis zu dem 99,5%-Percentil, während in den dritten Bereich alle restlichen Merkmalsträger mit den absolut höchsten negativen Einkommen fallen. In der derzeitigen Version sind anstelle dieser relativen Grenzen noch absolute verwendet worden (siehe Übersicht 2). Die Anonymisierungsmethoden sind mit den Methoden in den Bereichen 1, 3 und 5 der Merkmalsträger mit positiven Einkommen identisch.

Tabelle 1 gibt die Bedeutung der Anonymisierungsbereiche nach den Kriterien Steuerpflichtige, Gesamtbetrag der Einkünfte und festgesetzte Einkommensteuer wieder. Sie zeigt, dass dem Anonymisierungsbereich 1 die mit Abstand größte Bedeutung hinsichtlich der Steuerpflichtigen zukommt. Diese geht allerdings bei den Wertebetrachtungen etwas zurück, was mit der Schiefe der Einkommensverteilung zusammenhängt.

Tabelle 1: Anteile der Anonymisierungsbereiche
Prozent

Anonymisierungsbereich	Steuerpflichtige		Gesamtbetrag der Einkünfte		Einkommensteuer	
	Anteil	kumuliert	Anteil	kumuliert	Anteil	kumuliert
1	92,2	92,2	68,8	68,8	51,7	51,7
2	6,6	98,8	17,1	85,9	22,2	73,9
3	1,2	99,99	8,6	94,5	16,2	90,1
4	0,05	99,99	3,2	97,7	6,2	96,3
5	0,02	100	2,3	100	3,7	100

3.2 Allgemeine Anonymisierung

Neben der auf die Einkommenshöhe abgestimmten Anonymisierung wurden weitere Maßnahmen ergriffen, mit denen

22) Ausführlich zur Anonymisierungskonzeption siehe Vorgrimler, D./Zwick, M.: „Faktische Anonymisierung der Steuerstatistik (FAST) – Lohn- und Einkommensteuer 1998“, erscheint in der Reihe FDZ-Arbeitspapiere, *www.forschungsdatenzentrum.de.*

23) Zu den gegenüber der traditionellen Anonymisierung verstärkten Auswirkungen der datenverändernden Verfahren siehe Rosemann, M./Vorgrimler, D./Lenz, R.: „Erste Ergebnisse faktischer Anonymisierung wirtschaftsstatistischer Einzeldaten“ in Allgemeines Statistisches Archiv, Heft 1, 2004, S. 73 ff.

24) Eine allgemeine Übersicht über Anonymisierungsmethoden findet sich in Höhne, J.: „Methoden zur Anonymisierung wirtschaftsstatistischer Einzeldaten“ in Gnoss, R./Ronning, G. (Hrsg.): „Anonymisierung wirtschaftsstatistischer Einzeldaten“, Band 42 der Schriftenreihe „Forum der Bundesstatistik“, Wiesbaden 2003, S. 69 ff., sowie zur Anonymisierung in der Bundesstatistik siehe Köhler, S.: „Anonymisierung von Mikrodaten in der Bundesstatistik und ihre Nutzung – Ein Überblick“ in Band 31 der Schriftenreihe „Forum der Bundesstatistik“, Wiesbaden 1999, S. 133 ff.

25) Eine Ausnahme bilden die drei Steuerpflichtigen mit den höchsten Einkommen, deren Merkmale mittels der Mikroaggregation zusätzlich anonymisiert wurden (siehe Abschnitt 3.3).

Übersicht 3: Allgemeine Anonymisierungsmaßnahmen

Eingabefeld	Merkmal(e)	Maßnahme
EF1	Veranlagungs-grund	Umkodierung der acht Ausprägungen in: 1 = veranlagte Fälle 2 = manuelle Fälle
EF13 + EF14	Religionen (jeweils getrennt für Männer und Frauen)	Umkodierung der zwölf Ausprägungen in: 1 = evangelisch 2 = katholisch 3 = sonstige 4 = konfessionslos
EF19	Veranlagungs-art	Umkodierung der acht Ausprägungen in: 1 = Grundtabelle 2 = Splittingtabelle
EF64 + EF67	Alter (jeweils getrennt für Männer und Frauen)	Einführung einer Unter- (15 Jahre) und einer 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 Zahl und Angaben zum Alter der Kinder sind in den Daten enthalten. 5 und mehr Kinder wurden der Ausprägung ∅4 Kinder zugewiesen.

alle Merkmalsträger mindestens anonymisiert wurden (allgemeine Anonymisierung). Übersicht 3 gibt über diese Anonymisierungsmaßnahmen Auskunft.

Die Beschränkung der Einkommensteuerdaten auf eine 10%-Stichprobe stellt darüber hinaus eine allgemeine Anonymisierungsmaßnahme dar, da ein potenzieller Datenangreifer aufgrund der Stichprobe keine Kenntnis darüber hat, ob der gesuchte Merkmalsträger überhaupt in den Daten enthalten ist.[26]) Dies macht eine erfolgreiche Reidentifikation unwahrscheinlicher und unsicherer. Allerdings wurde die Stichprobe nicht zur Anonymisierung gezogen, sondern mit dem Ziel, „handhabbare" Datenmengen mit höchstmöglicher Repräsentativität zu erhalten.[27]) Aus diesem Grund sind kleinere heterogene Gruppen von Merkmalsträgern als Totalschichten enthalten. Dies gilt wie gesehen besonders für die Bezieher hoher Einkommen. Für diese besitzt ein potenzieller Datenangreifer somit weiterhin Teilnahmekenntnis, sodass das o.g. Argument nicht gilt. Die Stichprobe entfaltet daher ihre Anonymisierungswirkung im Bereich der niedrigen und mittleren Einkommen.

Das Alter der Daten wirkt in zweierlei Hinsicht als Anonymisierungsmaßnahme. Zum einen ist es für einen potenziellen Datenangreifer umso schwieriger, relevantes Zusatzwissen für einen Merkmalsträger zu generieren, je älter die Daten sind. Zum anderen ist der Nutzen einer Information aktualitätsabhängig. Daher sinkt der Nutzen einer Identifikation mit zunehmendem Alter der Daten. Dieses Argument gilt allerdings nur, wenn die Datenaktualität ein positives Element der Nutzenfunktion des potenziellen Datenangreifers ist.

3.3 Spezifische Anonymisierung

3.3.1 Merkmalskategorien

In den beschriebenen Anonymisierungsbereichen wurden Merkmale unterschiedlich 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 eingeschränkt werden.

Merkmale der ersten Kategorie:

- Summe der Einkünfte (nach weiblichen und männlichen Steuerpflichtigen getrennt)
- 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 selbstständiger Arbeit (A+B)
- Einkünfte aus nichtselbstständiger 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 rund 300 stetigen Merkmale zählen zur dritten Kategorie.

Informationen, die zur Anonymisierung entweder nur vergröbert oder verfälscht wie auch solche, die überhaupt

26) Zur Teilnahmekenntnis siehe auch Lenz, R./Sturm, R./Vorgrimler, D.: „Maße für die faktische Anonymität von Mikrodaten" in WiSta 6/2004, S. 623 ff.
27) Zur Funktion der Stichprobe siehe Zwick, M., a.a.O. (Fußnote 20), S. 556 ff.

nicht mehr in den Zieldaten enthalten sind, weisen für einen Datenangreifer einen geringeren Wert auf als die Originalinformationen.[28]) Anonymisierung wirkt sich daher nicht nur auf die Kosten eines Datenangreifers aus, sondern auch sein Nutzen wird negativ beeinflusst. Bei der Lohn- und Einkommensteuerstatistik gilt dieser Aspekt besonders bei den stetigen Merkmalen. Diese sind eventuell schwierig als Überschneidungsmerkmale einsetzbar, wodurch eine Veränderung ihrer Werte keinen zusätzlichen Schutz der Merkmalsträger darstellen würde, jedoch dürften die stetigen Merkmale einem Datenangreifer den höchsten Nutzen stiften. Werden daher stetige Merkmale aus den Daten gelöscht oder vergröbert, so hat dies vor allem auf die Nutzenseite des Datenangriffs eine Wirkung. Dies ist ein wesentlicher Aspekt zur Erreichung einer faktischen Anonymität, bei der die Unverhältnismäßigkeit eines Datenangriffs mitberücksichtigt wird.

3.3.2 Spezielle Anonymisierungsmaßnahmen

Übersicht 4 fasst die getroffenen speziellen Anonymisierungsmaßnahmen für die unterschiedlichen Teilbereiche zusammen.

Die an der Diskussion über das Anonymisierungskonzept beteiligten Wissenschaftler bevorzugen den Erhalt der stetigen Merkmale gegenüber dem der sozioökonomisch diskreten Merkmale. Dies spiegelt sich in der speziellen Anonymisierung dahingehend wider, dass zunächst die diskreten Merkmale vergröbert bzw. gelöscht wurden, bevor die stetigen Merkmale zur Anonymisierung herangezogen wurden. So ist das Merkmal „Religion“ nur in den ersten beiden Einkommensbereichen mit vier Ausprägungen vertreten und das Merkmal „Region“ nur in diesen Bereichen mit dem Bundesland als Ausprägung. Das Alter ist ab dem zweiten Bereich klassiert und im ersten wurde eine obere und untere Grenze gezogen. Im fünften Bereich ist die Anzahl der Kinder nicht mehr angegeben. Dagegen ist im ersten Bereich nicht nur die Anzahl, sondern auch das Alter der ersten drei Kinder in den Daten enthalten.

Aufgrund dieses stärkeren Eingriffs in die sozioökonomischen Merkmale konnten die stetigen Merkmale bis einschließlich des dritten Bereichs unverändert in den Daten verbleiben. Im vierten Bereich sind die Merkmale der dritten Kategorie noch als Dummy-Variablen enthalten, während sie im fünften Bereich gelöscht sind. Die Transformation der Angaben in eine Dummy-Variable bedeutet in diesem Zusammenhang, dass das neue (Dummy-)Merkmal eine 1 als Ausprägung annimmt, wenn das ursprüngliche Merkmal mit positiver Ausprägung vorlag, eine Null aufweist, wenn das ursprüngliche Merkmal beim Steuerpflichtigen nicht vorhanden war und eine –1 bei negativen Ursprungswerten. Die Merkmale der zweiten Kategorie werden im vierten Bereich weiterhin ausgewiesen, allerdings ohne geschlechterspezifische Trennung. Im fünften Bereich sind diese Merkmale nur noch als Dummy-Variablen enthalten. Bei den Merkmalen der ersten Kategorie gibt es lediglich eine Einschränkung: Die Werte der drei Merkmalsträger mit den jeweils höchsten Ausprägungen wurden ersetzt durch die Durchschnittswerte ihrer jeweiligen Ausprägungen [Mikroaggregation[29])]. So entsprechen die Maxima der Merkmale der ersten Kategorie nicht mehr den Originalwerten, sondern stellen die arithmetischen Mittel der drei höchsten Werte dar, erhalten geblieben ist jedoch die volle Wertsumme.

Steuerpflichtige mit negativen Einkommen werden innerhalb der Bereiche 1, 2 und 3 anonymisiert.

3.3.3 Zusatzmerkmale der anonymisierten Datei

Neben der Informationsreduktion durch die Anonymisierung sind in die Datei für die Wissenschaft zusätzlich generierte Informationen aufgenommen worden.

Übersicht 4: Spezielle Anonymisierungsmaßnahmen in den Einkommensbereichen

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	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
Gewerbekennzahl		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 männlich weiblich als Summe	Dummy (+ Bedeutungsmerkmale)
	3	Ja	Ja	Ja	Dummy	Nein

1) Bei positiven Einkommen: 1 = von 0 bis zu einem Gesamtbetrag der Einkünfte von 64 106 EUR; 2 = 64 107 bis 137 532 EUR (99%-Percentil); 3 = 137 533 bis 970 202 EUR (99,95%-Percentil); 4 = 970 203 EUR bis zu den 1 000 höchsten Gesamtbeträgen der Einkünfte; 5 = die 1 000 höchsten Gesamtbeträge der Einkünfte + Abgeordnete. Bei negativen Einkommen: 1 = von 0 bis zu einem negativen Einkommen von 102 258 EUR (95%-Percentil); 3 = von 102 259 bis zu einem negativen Einkommen von 511 292 EUR (99,5%-Percentil) EUR; 5 = bei einem negativen Einkommen von über 511 292 EUR.

28) Siehe hierzu Höhne, J./Sturm, R./Vorgrimler, D., a.a.O. (Fußnote 3) sowie Lenz, R./Sturm, R./Vorgrimler, D., a.a.O. (Fußnote 26), S. 621 ff.

29) Zur Mikroaggregation siehe Domingo-Ferrer, J./Mateo-Sanz, J. M.: “Practical data-oriented microaggregation for statistical disclosure control”, IEEE Transactions on Knowledge and Data Engineering, Vol. 14(1), 2002, S. 189 ff.

Für Steuerpflichtige, die Einkünfte aus freien Berufen erzielen, wurde aus der in der ursprünglichen Einkommensteuerstatistik vorliegenden Gewerbekennzahl das Merkmal „Freiberufler“ mit folgenden Ausprägungen in den ersten vier Anonymisierungsbereichen generiert:

Technische Beratung, Forschung, Architekten, Ingenieure; Rechtsanwälte, Notare; Wirtschaftsprüfer, -berater; Ärzte; Sonstige Gesundheitsberufe; Werbung, Foto, Kunst und Kultur; Schriftberufe; Schulen und Sonstige.

Zusätzlich enthalten die Daten eine Dummy-Variable, die angibt, ob der Steuerpflichtige freiberuflich tätig ist. Damit wird einer langen Tradition der Steuerstatistiken gefolgt, in der die Gruppe der Freien Berufe in dieser Typisierung ausgewertet und analysiert wird.

Im Anonymisierungsbereich 5 sind die Merkmale der zweiten Kategorie nur noch als Dummy-Variablen 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 nichtselbstständiger Tätigkeit und sonstige Überschusseinkünfte). Für jede dieser Kategorien existiert ein Bedeutungsmerkmal. 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. Entsprechend weist dieses Merkmal die Ausprägung 2 für eine mittlere Bedeutung aus. Entstehen keine Einkünfte aus der Kategorie, wird das Merkmal auf 0 gesetzt. Als Beispiel ist in Tabelle 2 die Häufigkeitsverteilung der Merkmale für die 1 000 Steuerpflichtigen mit den höchsten Einkommen angegeben. Sie zeigt demnach, welche Einkommenskategorien zur Erzielung der höchsten Einkommen am meisten beitragen.

Tabelle 2: Struktur der Einkünfte bei den höchsten Einkommen
Häufigkeitsverteilung für die 1 000 Merkmalsträger mit den höchsten Einkommen

Bedeutung	Gewinneinkünfte	Einkünfte aus nichtselbstständiger Arbeit	Sonstige Überschusseinkünfte
Hoch	910	10	80
Mittel	49	318	616
Niedrig	30	275	283
Keine	11	397	21

Als weitere Zusatzinformation ist in der anonymisierten Datei die Anonymisierungsstärke jedes Merkmalsträgers angegeben. Die Ausprägungen 1 bis 5 geben hierbei die Anonymisierungsbereiche wieder. Zusätzlich wurde eine Ausprägung 6 eingeführt, die diejenigen Merkmalsträger bezeichnet, deren stetige Merkmale punktuell mikroaggregiert wurden.

4 Test der Datensicherheit

Der Test der Datensicherheit setzt bei den in Kapitel 3 beschriebenen anonymisierten Daten an. Naturgemäß sind die Möglichkeiten der Reidentifikation bei den Originaldaten größer. Auf diese Darstellung wird hier verzichtet, da diese Daten aufgrund der Ex-ante-Abschätzungen des Deanonymisierungsrisikos für ein Scientific-Use-File nicht in Frage kamen.

4.1 Ansatzpunkte möglicher Reidentifikationsversuche

Bevor eine anonymisierte Datei als faktisch anonym im Sinne des § 16 Abs. 6 BStatG gelten kann, muss diese auf einen ausreichenden Datenschutz überprüft werden.[30]) Eine Möglichkeit hierzu bieten Simulationen von Reidentifikationsversuchen. Diese lassen sich in zwei Arten unterteilen: in so genannte Massenfischzüge mit dem Ziel, mit Hilfe externer Datenbanken als Zusatzwissen so viele Merkmalsträger wie möglich zu reidentifizieren, und in Einzelangriffe, bei denen versucht wird, gezielt einen bestimmten Merkmalsträger in den anonymisierten Daten zu finden. Mit beiden Verfahren sind in der Vergangenheit anonymisierte Daten auf ihre Sicherheit überprüft worden.[31])

Als erster Schritt des Tests auf ausreichende Anonymisierung ist zu überlegen, welches der beiden Verfahren grundsätzlich zum Sicherheitstest bei der Lohn- und Einkommensteuerstatistik 1998 in Betracht kommt.

Um Merkmalsträger einer anonymisierten Datei erfolgreich reidentifizieren zu können, sind folgende Grundannahmen für einen Datenangreifer nötig:[32])

- Externes Zusatzwissen über die gesuchten Merkmalsträger (etwa in Form einer Datenbank)
- Kenntnis über die Teilnahme des gesuchten Merkmalsträgers an der Erhebung
- Merkmale, welche sowohl in externen als auch in Zieldaten enthalten sind (Überschneidungsmerkmale).

Diese Bedingungen schränken die Möglichkeiten von Reidentifikationsversuchen erheblich ein. Massenfischzüge erscheinen faktisch ausgeschlossen, da das hierzu benötigte Zusatzwissen weder die nötigen Überschneidungsmerkmale aufweist noch in geeigneter und hinreichender Form vorliegt.[33]) Für Einzelangriffe bei bestimmten Personengruppen ist jedoch genügend Zusatzwissen vorhanden, auch wenn das Zusatzwissen aus verschiedenen Quellen zusammengefügt werden muss. Das Hauptaugenmerk

30) Zum Test der Datensicherheit der anonymisierten Einkommensteuerstatistik 1998 siehe ausführlich Scharnhorst, S./Zühlke, S./Stegenwaller, L.: „Beiträge zum Projekt „Faktische Anonymisierung der Lohn- und Einkommensteuerstatistik 1998“, erscheint in der Reihe FDZ-Arbeitspapiere, *www.forschungsdatenzentrum.de*.

31) So z.B. für die Kostenstrukturerhebung im Verarbeitenden Gewerbe siehe Lenz, R.: “Disclosure of confidential information by means of multi-objective optimisation”, Proceedings of the Comparative Analysis of (micro) Enterprise Data Conference (CAED), London 2003 (*http://www.statistics.gov.uk/events/caed/abstracts/lenz.asp*) und Vorgrimler, D.: „Re-Identifikationsmöglichkeiten am Beispiel eines konkreten Angriffsszenarios“ in Gnoss, R./Ronning, G. (Hrsg.): „Anonymisierung wirtschaftsstatistischer Einzeldaten“, Band 42 der Schriftenreihe „Forum der Bundesstatistik“, Wiesbaden 2003, S. 40 ff. Für die Umsatzsteuerstatistik siehe Lenz, R./Vorgrimler, D.: „Geheimhaltungsmethoden auf dem Prüfstand – eine Analyse anhand der Umsatzsteuerstatistik“ in WiSta 6/2004, S. 639 ff.

32) Siehe Brand, R./Bender, S./Kohaut, S.: “Possibilities for the creation of a scientific-use-file for the IAB-Establishment-Panel”, Proceedings of the Joint Eurostat/UN-ECE Work Session on Statistical Data Protection, Thessaloniki 1999, S. 57 ff.

33) Siehe Scharnhorst, S./Zühlke, S./Stegenwaller, L., a. a. O. (Fußnote 31).

muss dabei auf diejenigen Personengruppen gerichtet werden, die aufgrund ihrer besonderen Stellung als Totalschichten in der Stichprobe enthalten sind. Nur für diese Gruppen besitzt ein Datenangreifer Teilnahmekenntnis. Da die Bezieher mittlerer und niedrigerer Einkommen in der Regel nur als Stichprobe enthalten sind, können diese ex ante als faktisch anonym angesehen werden, auch wenn über sie aufgrund der schwächeren Anonymisierung mehr Informationen vorliegen.[34]) Merkmalsträger, die einzeln in die Stichprobe eingegangen sind, können daran erkannt werden, dass ihr Hochrechnungsfaktor den Wert 1 annimmt. Aufgrund dieser Argumente konzentriert sich die Sicherheitsanalyse auf folgende Gruppen:

– Prominente, Manager,

– Personen mit freiberuflicher Tätigkeit,

– Abgeordnete,

– persönliches Umfeld.

4.2 Überschneidungsmerkmale

In Anlehnung an Elliot/Dale[35]) lassen sich die Überschneidungsmerkmale aus dem Zusatzwissen in folgende vier Kategorien einteilen:

1. Leicht zugängliches Zusatzwissen von hoher Qualität (prime keys)

2. Leicht zugängliches Zusatzwissen von niedrigerer Qualität (background keys)

3. Schwer zugängliches Zusatzwissen von hoher Qualität (critical keys)

4. Schwer zugängliches Zusatzwissen von niedriger Qualität (inefficient keys)

Hierbei hängt die Qualität des Zusatzwissens davon ab, wie stark es die Daten differenziert, wie stabil das Zusatzwissen über die Zeit ist und wie hoch die Wahrscheinlichkeit von Messfehlern ist.

Als prime keys in der Einkommensteuerstatistik sind das Alter der Steuerpflichtigen und die Anzahl der Kinder anzusehen. Für bestimmte Personengruppen, wie die Freiberufler, ist auch die Information über ihre berufliche Gruppenzugehörigkeit ein prime key. Als background keys können die Angaben über den Wohnort und die Geschlechtszugehörigkeit verwendet werden (leicht ermittelbar, aber mit nur geringer Differenzierungswirkung). Zu den critical keys zählen die Angaben zum Alter der ersten drei Kinder, die Religionszugehörigkeit, die Spendentätigkeit und die Unterhaltspflichten.[36])

4.3 Ergebnisse der Reidentifikationsversuche

Prominente und Manager

Es wurden sowohl Prominente aus dem Sport- als auch aus dem Medienbereich betrachtet. Vorab wurde das Risiko einer Reidentifikation besonders bei Sportlern als hoch angesehen, da bei ihnen die relativ seltene Merkmalskombination hohes Einkommen/geringes Alter häufiger auftritt. Dennoch konnte in keinem der insgesamt zwölf Einzelangriffe auf Prominente eine eindeutige Zuordnung erzielt werden. Auch das Hinzuziehen weiterer Überschneidungsmerkmale – wenn sie denn vorhanden gewesen wären – lässt eine Reidentifikation nicht wahrscheinlicher werden.

Zum gleichen Ergebnis kommt man im Bereich der Wirtschaftsmanager, obwohl für diese Personengruppe im Zusatzwissen mehr Einkommensangaben vorhanden waren. Auch hier konnte keine der sechs Zielpersonen trotz intensiver Recherche, insbesondere über das Internet, reidentifiziert werden.

Aufgrund der Ergebnisse bei insgesamt 18 prominenten Personen und Managern können die Einzelangriffe auf diese Personengruppen als gescheitert angesehen werden.

Personen mit freiberuflicher Tätigkeit

Wie in Abschnitt 4.2 erläutert, kann das Merkmal der freiberuflichen Tätigkeit eines Merkmalsträgers als prime key verwendet werden, da dies den Steuerpflichtigen als Teil einer relativ kleinen Untergruppe charakterisiert.[37]) In die faktisch anonymisierten Daten ist das Merkmal der freiberuflichen Tätigkeit als zusätzliche Information aufgenommen worden, wodurch die Verwendung als Überschneidungsmerkmal erleichtert wird (siehe Abschnitt 3.2.3). Aus diesem Grunde wurden die „Freiberufler“ einem separaten Test unterzogen. Allerdings bestand dieser nicht aus „realen“ Einzelangriffen, sondern es wurde in den Daten der Anteil einmaliger Ausprägungskombinationen ermittelt. Für Nordrhein-Westfalen war dieser so gering, dass keine Gefährdung der Datensicherheit zu befürchten ist. Für Mecklenburg-Vorpommern ist der Anteil zwar deutlich größer, allerdings sind so genaue Informationen über die Merkmalsträger notwendig, um in der Realität einen einmaligen Fall zu identifizieren, dass dies nur im persönlichen Umfeld eines Freiberuflers möglich erscheint. Die alleinige Tatsache, zur Freiberuflergruppe zu gehören, gefährdet dagegen die Datensicherheit des Merkmalsträgers nicht.

Abgeordnete

Die Reidentifikation von Abgeordneten wird sowohl durch besseres Zusatzwissen als auch durch bessere Angaben in den Zieldaten erleichtert. Die Höhe der von den Abgeordneten bezogenen Diäten sind öffentlich zugänglich und in den

34) Siehe Fußnote 34.

35) Siehe Elliot, M./Dale, A.: “Scenarios of attack: the data intruder’s perspective on statistical disclosure risk” in Netherlands Official Statistics, Vol. 14 (2) 1999, S. 6 ff.

36) Zur ausführlichen Begründung dieser Einteilung siehe Fußnote 34.

37) In den FAST-Daten sind rund 260 000 Datensätze als männliche und rund 100 000 Datensätze als weibliche Steuerpflichtige mit freiberuflichen Einkünften gekennzeichnet. Auf Deutschland hochgerechnet entspricht dies rund 800 000 männlichen und rund 400 000 weiblichen Steuerfällen mit freiberuflichen Einkünften.

Daten der Einkommensteuerstatistik als separates Merkmal enthalten. Im ersten Schritt der Anonymisierung war bereits ersichtlich, dass dieses Merkmal zumindest mit einem weiteren Merkmal innerhalb der gleichen Einkunftsart (sonstige Einkünfte) zusammengefasst werden muss. Auf Basis dieser Maßnahmen wurden reale Einzelangriffe durchgeführt.

Bei einer ersten Versuchswelle wurden 16 Abgeordnete in den Daten gesucht. Eindeutig und richtig gelang die Reidentifikation allerdings nur in einem Fall.

Der zweite Schritt bestand in einer Veränderung der Suchrichtung. Da einzelne Ausprägungen des Merkmals „sonstige Einkünfte aus Leistungen“ sich bei Werten häuften, die den Diäten der nordrhein-westfälischen Landtags- und Bundestagsabgeordneten entsprechen, konnte eine Gruppe von 86 Abgeordneten aus Nordrhein-Westfalen identifiziert werden. Elf dieser Personen konnten eindeutig zugeordnet werden. Darüber hinaus sind noch zwei Doppelzuordnungen vorhanden. Von diesen insgesamt 15 Zuordnungen sind zehn als korrekt anzusehen.

Da es mit einem relativ geringen Aufwand möglich erscheint, Abgeordnete trotz der bisher durchgeführten Anonymisierungsmaßnahmen zu identifizieren, musste für diese Teilpopulation die Anonymisierung verschärft werden. Hierzu mussten die identifizierenden Angaben eliminiert werden. Die Abgeordneten ließen sich über Häufigkeitsauswertungen des Merkmals „sonstige Einkünfte aus Leistungen“ nicht nur dieser Gruppe zuordnen, sondern darüber hinaus einzelnen Bundesländern. Weitere Tests zeigten, dass dies in der Regel so lange möglich ist, wie die Einkunftsart „sonstige Einkünfte“ ausgewiesen wird. Erst im Anonymisierungsbereich 5 ist dieses Merkmal nur noch als Dummy-Variable enthalten. Daher sind als zusätzliche Schutzmaßnahme alle als Abgeordnete identifizierten Merkmalsträger dem Anonymisierungsbereich 5 zugeordnet worden. Aus diesem Grunde sind im Anonymisierungsbereich 5 nicht nur die 1 000 Merkmalsträger mit den höchsten positiven Gesamtbeträgen der Einkünfte und rund 2 200 mit den höchsten negativen Gesamtbeträgen der Einkünfte enthalten, sondern zusätzlich rund 3 000 Abgeordnete. Die verschärften Anonymisierungsmaßnahmen des Anonymisierungsbereichs 5 bieten einen zusätzlichen Schutz, sodass nach dieser Überarbeitung der Anonymisierung auch die Teilpopulation der Abgeordneten als faktisch anonym anzusehen ist.

Persönliches Umfeld

Für die Untergruppe „persönliches Umfeld“ kann im Allgemeinen für einen Datenangreifer das beste verfügbare Zusatzwissen angenommen werden. Dies gilt sowohl in quantitativer (Anzahl der vorhandenen Überschneidungsmerkmale) als auch in qualitativer (Verlässlichkeit der Werte) Hinsicht.

Drei Personen aus dem persönlichen Umfeld wurden als Zielpersonen in den faktisch anonymisierten Daten gesucht. Bei den ersten beiden Personen war eine erfolgreiche Suche schon deshalb unwahrscheinlich, da die Zielpersonen aufgrund des vorliegenden Stichprobenplans Schichten mit sehr geringen Auswahlsätzen zugeordnet wurden. Der Versuch einer Reidentifikation machte daher bereits aus diesem Grund keinen Sinn.

Bei der dritten Person konnte eine eindeutige Zuordnung erzielt werden, die sich aber bei der Überprüfung als falsch herausstellte. Somit konnten auch die Reidentifikationsversuche im persönlichen Umfeld als gescheitert angesehen werden. Es ist aber darauf hinzuweisen, dass gerade das persönliche Umfeld eines Datenangreifers eine sehr subjektive Einschätzung ist. Gehört ein potenzieller Datenangreifer zu einer Gesellschaftsgruppe, die einem höheren Risiko einer Reidentifikation ausgesetzt ist, wird sein Umfeld eher aus solchen Personen bestehen als bei einem Datenangreifer, der selbst zu einer als ungefährdet einzustufenden Gesellschaftsgruppe gehört. Die Möglichkeit, einen Merkmalsträger zu reidentifizieren, ist daher direkt abhängig von den Eigenschaften des Datenangreifers, eine „objektive“ Beurteilung ist daher weniger möglich als bei den anderen Subpopulationen des Datensatzes.

4.4 Fazit der Reidentifikationsversuche

Die auf Basis von Einzelangriffen erfolgten Reidentifikationsversuche haben gezeigt, dass die Daten faktisch anonym sind. Zu diesem Ergebnis kamen die beteiligten statistischen Ämter, die hinzugezogenen Juristen und der beratende wissenschaftliche Nutzerkreis einstimmig. Die faktische Anonymität konnte allerdings bei den Daten von Abgeordneten erst durch zusätzliche Anonymisierungsmaßnahmen sichergestellt werden.

Dass ein Datenangreifer theoretisch einen Merkmalsträger reidentifizieren kann, ist mit diesem Ergebnis nicht ausgeschlossen. Doch die absolute Anonymität wird vom Gesetzgeber im § 16 Abs. 6 BStatG auch nicht gefordert, sondern der Aufwand, den ein Datenangreifer für eine erfolgreiche Reidentifikation betreiben muss, muss unverhältnismäßig hoch sein. Diese Bedingung ist erfüllt und damit sind die Daten faktisch anonym und können unter den weiteren strikten Auflagen des § 16 Abs. 6 BStatG an die Wissenschaft übermittelt werden.

5 Ausblick

Mit dem faktisch anonymisierten Mikrodatenfile der Lohn- und Einkommensteuerstatistik 1998 erweitert die amtliche Statistik ihr Angebot an standardisierten Scientific-Use-Files. Die Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder kommen damit dem Auftrag nach, neben einer Erweiterung der Datenzugangswege insbesondere Daten in einer Form zur Verfügung zu stellen, die es erlaubt, „vor Ort“ zu forschen.

Das Scientific-Use-File ist zum Preis von 65,– Euro über die Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder für die Wissenschaft erhältlich. Der kostengünstige Preis resultiert aus der Förderung durch das Bundesministerium für Bildung und Forschung. Durch die finanzielle Zuwendung der Bundesregierung ist es der amtlichen Statistik möglich, mit den Forschungsdatenzentren Kapazitäten zur Verfügung zu stellen, die es erlauben, Ano-

nymisierungsprojekte zu realisieren und die resultierenden Scientific-Use-Files kostengünstig anzubieten.

Mit standardisierten faktisch anonymen Mikrodaten, die den Bereich der amtlichen Statistik unter den Auflagen des § 16 Abs. 6 BStatG verlassen können, ist der Datenbedarf der Wissenschaft allerdings nicht gedeckt. Es verbleiben Forschungsbereiche, die mit FAST 98 nicht ausreichend erforscht werden können. So werden detaillierte Untersuchungen zu hohen Einkommen oder regional tief gegliederte Analysen aufgrund der Informationsreduktion durch die Anonymisierungsmaßnahmen nur beschränkt möglich sein.

Hier bieten die weiteren Zugangswege über die Forschungsdatenzentren Ansätze zur Lösung. Neben standardisierten (off-site) Scientific-Use-Files können in den Räumen der amtlichen Statistik individuelle, auf den Forschungszweck hin erstellte (on-site) Scientific-Use-Files an Gastwissenschaftlerarbeitsplätzen genutzt werden. Weiter besteht über das kontrollierte Fernrechnen die Möglichkeit, das vollständige Informationspotenzial amtlicher Einzeldaten zu nutzen.[38]) Erklärtes Ziel der Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder ist es, keine Forschungsprojekte mehr aufgrund mangelnder Zugangsmöglichkeiten zu amtlichen Einzeldaten scheitern zu lassen. Es wird nicht immer möglich sein, den komfortabelsten Weg zu den amtlichen Daten anzubieten, aber die Nutzung des Informationspotenzials der amtlichen Einzeldaten sollte immer kostengünstig für die Wissenschaft möglich sein. Kosten zum Beispiel in Form unterdrückter Information oder weiterer Wege zur Information ergeben sich jedoch aufgrund der Datenschutzauflagen. Diese sind für die amtliche Statistik ebenso zu beachten wie die Wissenschaftsfreiheit.

Das berechtigte Anliegen der Wissenschaft nach möglichst aktuellen Daten bringt es mit sich, dass die Anonymisierung der Lohn- und Einkommensteuerstatistik mit FAST 98 nicht abgeschlossen sein kann. Vielmehr gilt es nun, aufbauend auf den geleisteten Arbeiten, Scientific-Use-Files mit aktuelleren Daten anzubieten. Als ein erster Schritt hierzu sind Daten des Veranlagungsjahres 2001 faktisch zu anonymisieren, sobald diese vorliegen. Es gilt darüber hinaus zu prüfen, ob nach 2001 anonymisierte Einkommensteuerdaten sogar jährlich angeboten werden können. Als Basis hierfür könnten die jährlich nach § 2a StStatG erhobenen Daten zur Einkommensteuerstatistik dienen. Dies würde den Grad an Aktualität und Vergleichbarkeit zwischen den Veranlagungsjahren optimieren.

Mit dem nun vorgelegten Scientific-Use-File der Einkommensteuerstatistik 1998 und den begonnenen Projekten zur Anonymisierung der Gehalts- und Lohnstrukturstatistik, der Krankenhausstatistik und dem im Sommer 2005 vorliegenden Projektergebnis zur Anonymisierung wirtschaftsstatistischer Daten haben die statistischen Ämter einen entscheidenden Schritt hin zu einer verbesserten informationellen Infrastruktur getan. Damit gelingt es der amtlichen Statistik, den Zielkonflikt zwischen „Wissenschaftsfreiheit“ und „Datenschutz“ ein weiteres Stück abzumildern.

38) Siehe Fußnote 9.

Stellungnahme des wissenschaftlichen Beirates

Mit der Vorlage des Gutachtens der Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik wurde erneut der Wunsch der Wissenschaft nach weiteren faktisch anonymisierten Datenbeständen aus dem Bereich der amtlichen Statistik unterstrichen. So empfiehlt die Kommission u. a. „... die Entwicklung von Scientific-Use-Files (SUF) als wichtiges Instrument des Mikrozugangs voranzutreiben“.

Die aus einer weiteren Empfehlung der Kommission hervorgegangenen Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder haben sich dieses Themas angenommen und gemeinsam mit den Fachstatistikern und der Wissenschaft ein faktisch anonymisiertes Mikrodatenfile der Lohn- und Einkommensteuerstatistik (FAST) erstellt. Die Arbeiten zu FAST wurden von den genannten Gruppen innerhalb eines wissenschaftlichen Beirates begleitet.

Der wissenschaftliche Beirat setzte sich zusammen aus

– der wissenschaftlichen Leitung (Prof. Dr. Merz),

– der Gruppe „Steuern“ des Statistischen Bundesamtes,

– den Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder,

– den Fachbereichen Steuern der Statistischen Landesämter Nordrhein-Westfalen, Rheinland-Pfalz und Bayern und

– Vetretern des wissenschaftlichen Nutzerkreises.

Aufgabenstellung

Der Beirat hatte die Aufgabe, als gemeinsames Gremium von Datenproduzenten und Datennutzern, die Erstellung von FAST beratend und bewertend zu begleiten. Innerhalb dieses Gremiums wurden zum einen verschiedene Anonymisierungskonzepte aus Sicht des Datenschutzes und zum anderen die Analysefähigkeit der jeweils erstellten anonymisierten Datei intensiv diskutiert.

Ergebnis

Auf der Grundlage der entwickelten Konzepte zur Anonymisierung der Lohn- und Einkommensteuerstatistik 1998 sowie der überprüften Angriffsszenarien auf die anonymisierte Datei, wurde nach intensiver Diskussion innerhalb des Beirats ein Scientific-Use-File der Lohn- und Einkommensteuer 1998 entwickelt, das nach Ansicht des Beirates im Sinne des § 16 Abs. 6 BStatG faktisch anonym ist.

Der wissenschaftlichen Forschung steht mit dem Scientific-Use-File der Lohn- und Einkommensteuerstatistik 1998 damit ein Datenmaterial zur Verfügung, mit dem es möglich ist, einen großen Teil der steuerpolitischen Fragestellungen zu beantworten. Zudem ermöglicht FAST sozioökonomische Fragestellungen, wie z. B. Fragestellungen im Rahmen des zweiten Armuts- und Reichtumsberichts der Bundesregierung 2004, anzugehen.

Resümee und Empfehlung

Mit der Erstellung eines Scientific-Use-File der Lohn- und Einkommensteuerstatistik 1998 wird die informationelle Infrastruktur in Deutschland nachhaltig verbessert. Die Lohn- und Einkommensteuerstatistik ist hinsichtlich der Differenziertheit der Einkommensangaben, ihrer Qualität als amtliche Vollerhebung sowie ihrer Möglichkeit, auch höchste Einkommen zu beschreiben, für die Wissenschaft von hohem Interesse.

FAST ist ein dynamisches Produkt. Die praktischen Erfahrungen der damit arbeitenden wissenschaftlichen Nutzer werden gesammelt und in das nächste zu entwickelnde Scientific-Use-File der Lohn- und Einkommensteuerstatistik 2001 mit einfließen, sodass eine methodische Weiterentwicklung gewährleistet ist. Dies bedeutet auch eine permanente Überprüfung des gefundenen Anonymisierungsgrades.

Der Beirat spricht sich dafür aus, auf Grundlage der gesammelten Erfahrungen auch ein FAST-Regionalfile zu entwickeln. So könnte FAST zukünftig zum Beispiel mit Raumordnungsmerkmalen ergänzt werden.

Auszug aus Wirtschaft und Statistik

Herausgeber: Statistisches Bundesamt, Wiesbaden

Schriftleitung: Johann Hahlen
Präsident des Statistischen Bundesamtes
Verantwortlich für den Inhalt:
Brigitte Reimann,
65180 Wiesbaden

- Telefon: +49 (0) 6 11/75 20 86
- E-Mail: wirtschaft-und-statistik@destatis.de

Vertriebspartner: SFG Servicecenter Fachverlage
Part of the Elsevier Group
Postfach 43 43
72774 Reutlingen
Telefon: +49 (0) 70 71/93 53 50
Telefax: +49 (0) 70 71/93 53 35
E-Mail: destatis@s-f-g.com

Erscheinungsfolge: monatlich

Allgemeine Informationen über das Statistische Bundesamt und sein Datenangebot erhalten Sie:

- im Internet: www.destatis.de

oder bei unserem Informationsservice
65180 Wiesbaden

- Telefon: +49 (0) 6 11/75 24 05
- Telefax: +49 (0) 6 11/75 33 30
- E-Mail: info@destatis.de