

Markus Zwick

CAMPUS-Files - Kostenfreie Public Use Files für die Lehre

1. Einleitung

Mikrodaten sind mittlerweile Grundlage vielfältiger wissenschaftlicher Analysen. In der empirischen Sozial- und Wirtschaftsforschung, wie auch aus der daraus abgeleiteten Politikberatung, ist heute die Verwendung der originären statistischen Information, die Angaben über die einzelnen Merkmalsträger, nahezu selbstverständlich. Auf Empfehlung der Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik (KVI)¹ haben heute die wichtigsten Datenproduzenten Forschungsdatenzentren eingerichtet, die den Zugang zu amtlichen Mikrodaten ermöglichen.² Kaum ein politisches Reformvorhaben erlangt heute Gesetzeskraft, ohne das vorher die Auswirkungen empirisch auf der Grundlage amtlicher Mikrodaten quantifiziert worden wären. Seien es Steuerreformvorhaben, die Gesundheitsreform oder die sogenannten Hartz IV Gesetze, zu all diesen Themen liegen zahlreiche empirische Arbeiten vor. Darüber hinaus wächst die Anzahl der Diplom- und Magisterarbeiten, der Dissertationen und Habilitationen auf der Grundlage amtlicher Mikrodaten, wie die Nutzeranträge der Forschungsdatenzentren dies deutlich zeigen.

Diese neuen Möglichkeiten wie Anforderungen an die empirische Sozial- und Wirtschaftsforschung führen auch zu Forderungen im Bereich der akademischen Lehre. Schon das KVI-Gutachten hat hier Mikrodaten für die wissenschaftliche Ausbildung angemahnt. Die Statistischen Ämter des Bundes und der Länder haben auf diese Datennachfrage reagiert und bieten mit der Reihe CAMPUS-Files spezielle Mikrodatensätze für die Lehre an. CAMPUS-Files stehen seit 2004 kostenfrei, mittlerweile für die Statistiken Mikrozensus, Einkommensteuer, Kostenstrukturerhebung und Sozialhilfestatistik, unter www.forschungsdatenzentrum.de zur Verfügung. Die Daten werden jeweils in den Formaten SAS, SPSS und STATA sowie als ASCII CSV angeboten.

Diese als Public Use Files konzipierten Datenbestände werden in den folgenden Kapiteln vorgestellt. Neben den Daten werden insbesondere auch die Anonymisierungsmethoden dargelegt. Zu Beginn der Ausführungen widmet sich das nächste Kapitel einführend den verschiedenen Arten von anonymisierten Einzeldaten und Formen des Mikrodatenzugangs, um dann im Weiteren die spezielle Form der CAMPUS-Files in Produktportfolio der amtlichen Statistik zu erläutern.

2. Anonymisierte Mikrodaten und Formen des Mikrodatenzugangs

Als Mikro- oder Einzeldaten werden die in Datensätzen erfassten Merkmalsausprägungen einzelner Merkmalsträger verstanden. Üblicherweise stellt ein Datensatz einen Merkmalsträger, sei es Unternehmen, Haushalt oder Person, mit seinen innerhalb einer Statistik erfassten Eigenschaften, in der Form der numerischen Ausprägung eines beobachteten oder gemessenen Merkmals, dar. Nach der Erfassung, Aufbereitung und Plausibilisierung aller Angaben einer Statistik, liegt die maximale Menge an Information einer Erhebung vor. Jeder weiterer Schritt, sei es Ergebnisaufbereitung in Form der Tabellierung oder Anonymisierung für eine weitere Übermittlung bedeutet eine Informationsreduktion. Die Informationsreduktion durch die Ergebnisdarstellung der Statistischen Ämter führte ja gerade zur Nachfrage der Wissenschaft nach Mikrodaten.

Die statistische Ergebnisaufbereitung erfolgt in den statistischen Ämtern schon in einer ersten anonymisierten und damit informationsreduzierten Form. Gemäß § 12 des Gesetzes über die Statistik für Bundeszwecke (Bundesstatistikgesetz, BStatG) sind Hilfsmerkmale einer Statistik zum frühestmöglichen Zeitpunkt von den sogenannten Erhebungsmerkmalen zu trennen. Hilfsmerkmale sind Angaben die zur

¹ siehe KVI (2001).

² siehe hierzu z.B. Zwick (2006).

technischen Durchführung einer Statistik benötigt werden.³ Hierzu zählen insbesondere die persönlichen Identifikatoren der Merkmalsträger einer Erhebung, also z.B. Name und Anschrift einer Person oder eines Unternehmens. Einzeldatenbestände ohne direkte Identifikatoren werden als formal anonymisiert bzw. als vertrauliche Einzeldaten bezeichnet.

Übersicht 1: Informationsgehalt und Anonymisierungsgrad

Einzeldaten	Informationsgehalt	Zugangsform
nicht anonym	voller Informationsgehalt	kein Zugang
formal anonym	keine direkten Identifikatoren wie z.B. Name und Anschrift vollständiger Analysegehalt	kontrollierte Datenfernverarbeitung
faktisch anonym	Informationsreduktion bis Aufwand der Deanonymisierung größer als der Nutzen mittlerer Analysegehalt	Gastwissenschaftsarbetsplatz Scientific Use Files
absolut anonym	Informationsreduktion bis Angaben den Merkmalsträgern nach menschlichem Ermessen nicht mehr zugeordnet werden können geringerer Analysegehalt	Public Use Files CAMPUS-Files

Im Jahr 1987 wurde mit dem § 16 Abs. 6 BStatG der Wissenschaft ein privilegierter Zugang zu Mikrodaten der amtlichen Statistik eingeräumt. Hiernach ist die Übermittlung von Einzeldaten an die Wissenschaft erlaubt, sofern diese nur mit unverhältnismäßig hohem Aufwand reidentifiziert werden können (faktische Anonymität). „Unverhältnismäßig“ bedeutet hier, dass der Aufwand einer Reidentifikation deren Nutzen übersteigt. Dies impliziert, dass eine Deanonymisierung von Einzelangaben in einem faktisch anonymen Datensatz nicht mit absoluter Sicherheit ausgeschlossen wird, es für einen potentiellen Datenangreifer aber unattraktiv wäre, eine Deanonymisierung zu versuchen. Datenbestände die in faktisch anonymisierter Form an die Wissenschaft übermittelt werden sind daher so weit informationsreduziert, dass eine Zuordnung von Angaben zu einzelnen Merkmalsträgern nur noch mit einem unverhältnismäßig hohem Aufwand an Zeit, Kosten und Arbeitskraft möglich ist.

Für Nutzer außerhalb der unabhängigen Wissenschaft stehen Mikrodaten nur in einer absolut anonymisierten Form zur Verfügung. Hierbei enthalten die Mikrodaten nur noch so viel Information, dass eine Zuordnung von Angaben nach menschlichem Ermessen ausgeschlossen ist. Die Informationsreduktion ist dabei aber in der Regel so groß, dass das verbleibende Analysepotential für fundierte Analysen in der Regel nicht mehr ausreicht.

Standardisierte Public und Scientific Use Files können außerhalb der Statistischen Ämter genutzt werden (Off-Site-Nutzung). Hierzu werden die Mikrodatenbestände mit einem festem Anonymisierungskonzept bearbeitet und stehen dann für eine Übermittlung an die Nutzer zur Verfügung.

Daneben bieten die Forschungsdatenzentren der Statistischen Ämter mit den Arbeitsplätzen für Gastwissenschaftler und der kontrollierten Datenfernverarbeitung auch einen speziell auf den jeweiligen Bedarf zugeschnittenen Datenzugang an. Hier können weniger stark anonymisierte Mikrodaten genutzt

³ vgl. § 10 BStatG.

werden, die dafür nur in den abgeschotteten Bereichen der Statistischen Ämter bereitgestellt werden (On-Site-Nutzung).

Gemäß den Vorgaben des Bundesstatistikgesetzes (BStatG) bestehen weitere Unterschiede im Personenkreis, dem die Daten zugänglich gemacht werden dürfen. Während Public Use Files und Datenfernverarbeitung von allen interessierten Personen und Einrichtungen genutzt werden können, sind Scientific Use Files und Arbeitsplätze für Gastwissenschaftler der Nutzung durch unabhängige wissenschaftliche Einrichtungen vorbehalten.⁴

3. CAMPUS-Files als Spezialform eines Public Use File

Die bisher von der amtlichen Statistik angebotenen Public Use Files stießen nur auf ein geringes Interesse. So führt zum einen die durch das Anonymisierungsverfahren doch deutlich eingeschränkte Informationsmenge dieser Daten zu einer zurückhaltenden Nachfrage. Auf der anderen Seite waren diese Daten bisher oftmals teurer als Scientific Use Files, da für die Erstellung keine Förderung vorlag und daher die Kosten der Erstellung in der Regel vollständig an die Nutzer weitergegeben wurden. Die Kombination aus einem gegenüber dem Scientific Use File weiter eingeschränktem Analysepotential bei einem höheren Preis bewirkte doch eine deutliche Zurückhaltung seitens der Nutzer.

Die Reihe der CAMPUS-Files vollzieht hier einen vollständig neuen Ansatz. Auf der einen Seite ist es nicht der primäre Anspruch dieses Datenangebotes, einen möglichst breiten Erhalt des Analysepotentials der Daten zu erreichen und auf der anderen Seite werden diese Daten kostenfrei unter www.forschungsdatenzentren.de angeboten.

Zielrichtung des Angebots von CAMPUS-Files durch die Statistischen Ämter des Bundes und der Länder, ist der Einsatz dieser Daten in der Lehre an den Hochschulen. Die Daten sollen zum einen die statistische Methodenausbildung anreichern sowie im Bereich der Sozial- und Wirtschaftsstatistik auch die Vermittlung der anwendungsbezogenen Statistik erleichtern. Das empirische Arbeiten mit ‚echten‘ Daten ist aufwendig und fehleranfällig. Der Zeitraum der Studierenden für empirisch ausgerichtete Diplom- oder Masterarbeiten zur Verfügung steht, ist oftmals nicht ausreichend, um die Details und Tücken eines Datenmaterials kennen zu lernen, zusätzlich sich ein Statistikprogramm auf der Syntaxebene anzueignen um dann beides für das eigentliche Thema zu verwenden. Die CAMPUS-Files bieten die Möglichkeit wesentliche Ausbildungsschritte in den Studienverlauf zu verlagern. Komplexe Datenmaterialien wie der Mikrozensus oder die Einkommensteuerstatistik können vor der eigentlichen Arbeit studiert werden. Hierbei besteht auch die Notwendigkeit sich mit einem Statistikprogramm auf der Syntaxebene auseinander zu setzen. Datenmaterial sowie die möglichen Fragestellung sind in der Regel so ausgestaltet, dass das Klicken unter SPSS sowie das Arbeiten unter Excel nicht ausreicht.

Bei dieser Zielsetzung ist das vorhandene Analysepotential der CAMPUS-Files von nach geordnetem Interesse. Primäres Ziel ist das Training mit größeren, komplexen Statistikdaten. Nach dieser Grundausbildung stehen dann für die akademischen Abschlussarbeiten, die deutlich umfangreicheren Scientific Use Files zur Verfügung. Hier kann aber nun der Hauptaugenmerk der wissenschaftlichen Fragestellung gelten, da die Kenntnis der Werkzeuge, Daten sowie Statistikprogramm, vorhanden sein sollten.

Die CAMPUS-Files weisen gegenüber ihren großen Brüdern den Scientific Use Files wesentlich weniger Datensätze und Merkmale aus. So umfasst das CAMPUS-File ‚Mikrozensus‘ nur rund 12 000 Haushalte im Gegensatz zum Scientific Use File mit rund 230 000 Haushalten. Aber im Vergleich mit anderen Haushaltsstichproben wie z.B. der Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften (ALLBUS) oder dem Sozi-oekonomischen Panel (SOEP) ist auch das CAMPUS-File groß. Aus diesem Grund werden zurzeit die CAMPUS-Files optimiert. Primäres Ziel ist zwar nicht der wissenschaftliche Einsatz,

⁴ Zu den Zugangsformen im Detail siehe Zwick (2006) und Zühlke et al (2003) sowie www.forschungsdatenzentrum.de.

trotzdem werden Merkmalsauswahl und Stichprobenkonzept so gewählt, dass bestimmte Themenbereiche auch schon in der Lehre mit ausreichender Präzision bearbeitet werden können.

4. Die CAMPUS-Files der Statistischen Ämter des Bundes und der Länder

CAMPUS-Files, als eine spezielle Form des Public Use File, sind nicht für die wissenschaftliche Anwendung konzipiert. Aufgrund ihrer Größe und der laufenden, an Nutzeranforderungen ausgerichtete Weiterentwicklungen, sind mit diesen Daten zwar Forschungsanwendungen möglich, aber das vorrangige Ziel der CAMPUS-Files ist die Ausbildung an ‚Echtdaten‘ der Statistischen Ämter des Bundes und der Länder. Dieser Überlegung folgt auch die Anonymisierungsstrategie bei diesen Daten. Bei üblichen Scientific sowie Public Use File hat das Anonymisierungsverfahren zwei konkurrierende Ziele zu berücksichtigen. Als Ergebnis soll ein Material verfügbar sein, dass ein möglichst großes Analysepotential der Daten, bei gleichzeitiger Erfüllung der Datenschutzaufgaben, erhält. Diesen Anspruch haben die CAMPUS-Files nicht. Das verbleibende Analysepotential war bei der Aufstellung der Anonymisierungskonzepte sekundär. Durch das Ziel ein Angebot zu schaffen, dass kostenfrei und unbeschränkt im Internet verfügbar ist und dies weltweit, geht die Informationsreduktion der CAMPUS-Files deutlich über die Reduktion bei sonstigen Public Use Files hinaus. Durch das primäre Ziel der CAMPUS-Files, die empirische Ausbildung in der Lehre anzureichern ist die starke Informationsreduktion keine Einschränkung. Die Daten machen es möglich Statistiken wie Verfahren in der akademischen Ausbildung kennen zu lernen, für wissenschaftliche Abschlussarbeiten stehen dann Scientific Use Files zur Verfügung.

Die im Folgenden vorgestellten CAMPUS-Files des Mikrozensus, der Einkommensteuerstatistik und der Kostenstrukturerhebung ähneln sich daher in der Anonymisierungskonzeption. Zum einen handelt es sich jeweils um deutlich kleinere Stichproben als bei den jeweiligen Scientific Use Files. Weiter ist die Anzahl der Merkmale deutlich reduziert, diese liegen darüber hinaus in der Regel nur grob klassiert vor. Darüber hinaus wurden ältere Jahrgänge der jeweiligen Statistiken ausgewählt und somit die Zeit als Anonymisierungsmaßnahme genutzt.⁵

Auf das CAMPUS-File der Sozialhilfestatistik 1998 wird in der Folge nicht eingegangen. Durch das ‚Vierte Gesetz für moderne Dienstleistungen am Arbeitsmarkt‘ (Hartz IV Gesetz) und dem damit einhergehenden Übergang der laufenden Sozialhilfe, für Empfänger die grundsätzlich erwerbsfähig sind, in die Grundsicherung für Arbeitssuchende, liegt derzeit keine Statistik vor, die in analoger Form wie das Datenmaterial von 1998 die Sozialleistungen abbilden könnte. Ein Teil der Statistiken werden zurzeit von den Statistischen Ämtern betrieben, ein weiterer Teil von der Bundesagentur für Arbeit. Es fehlt hier aktuell an einem einheitlichen Datenbestand, so dass die Reihe CAMPUS-Files der Sozialhilfestatistik vorläufig nicht fortgesetzt werden kann.

4.1 Das CAMPUS-File Mikrozensus 1998

Der Mikrozensus ist eine amtliche Repräsentativstatistik über die Bevölkerung und den Arbeitsmarkt, an der jährlich 1% aller Haushalte in Deutschland mit Auskunftspflicht beteiligt sind (laufende Haushaltsstichprobe). Insgesamt nehmen rund 370 000 Haushalte mit 820 000 Personen am Mikrozensus teil (2005); darunter etwa 160 000 Personen in rund 70 000 Haushalten in den neuen Bundesländern und Berlin-Ost.⁶ Alle Haushalte haben beim Mikrozensus die gleiche Auswahlwahrscheinlichkeit. Es wird eine einstufige geschichtete Flächenstichprobe durchgeführt, das heißt, aus dem Bundesgebiet werden Flächen (Auswahlbezirke) ausgewählt, in denen alle Haushalte und Personen befragt werden. Die Auswahlbezirke werden aus dem Material der Volkszählung 1987 gebildet;

⁵ Zu den Verfahren der Anonymisierung siehe Müller et al (1991) sowie Ronning et al (2005).

⁶ Zum Mikrozensus siehe z.B. Wirth/Müller (2006) sowie http://www.destatis.de/themen/d/thm_mikrozen.php, vergleiche auch <http://www.gesis.org/Dauerbeobachtung/GML/Daten/MZ/index.htm>.

für die neuen Bundesländer wurde auf der Basis des "Bevölkerungsregister Statistik" eine vergleichbare Auswahlgrundlage erstellt

Das Scientific Use File ist eine faktisch anonymisierte 70%-Substichprobe der Haushalte (bei Mikrozensus mit dem Zusatzprogramm zur Wohnsituation: eine 70%-Substichprobe der Wohnungen) des Mikrozensus.⁷ Für die faktisch anonymisierte Substichprobe gilt, dass sie als systematische Zufallsauswahl aus dem Originalmaterial gezogen wird. Das Mikrozensus SUF unterscheidet sich über die Substichprobe hinaus vom Originalfile des Mikrozensus dadurch, dass bestimmte Variablen im Scientific Use File, bedingt durch die Anonymisierung, nur in klassierter und vergrößerter Form verfügbar sind.

Das CAMPUS-File ist eine 3,5%-Substichprobe der Wohnungen des Mikrozensus 1998, gezogen aus dem Originalmaterial. Darin enthalten sind Angaben zu 25.410 Personen aus 11.771 Haushalten und 11.668 Wohnungen. Insgesamt gingen 199 Variablen des Originalmaterials und des Scientific Use File in das absolut anonyme Grunddatenfile ein. Drei neue Variablen, angepasste Hochrechnungsfaktoren für Personen, Haushalte und Wohnungen wurden erzeugt. Die Maßnahmen zur Erreichung der absoluten Anonymität des CAMPUS-Files bauen auf Anonymisierungsmaßnahmen zur Erreichung der faktischen Anonymität des Scientific Use File auf. Über die beim Scientific Use File angewandten Maßnahmen hinaus wurden Maßnahmen wie die Ziehung einer im Vergleich zum SUF deutlich kleineren Stichprobe sowie weitere Vergrößerung von Merkmalen und die zusätzliche Löschung von Variablen durchgeführt. Insbesondere kritische Merkmale, deren Häufigkeiten im Originalmaterial eine geringe Besetzungszahl aufwiesen wurden weiter vergrößert. Beispielhaft sei hier ein an den Scientific Use File angepasstes Top- und Bottom-Coding der Variablen Alter, Einkommen, Staatsangehörigkeit usw. zu nennen. Die Ländergliederung nach Bundesland wurde zu einer regionalen Gliederung nach Ost (neue Bundesländer und Ost-Berlin) und West (alte Bundesländer und West-Berlin) vergrößert. Eine dritte beim CAMPUS-File angewandte Maßnahme zur absoluten Anonymisierung war die Löschung von 370 Variablen aus dem Originalmaterial.

Übersicht 2: CAMPUS-File Mikrozensus 1998

Stichprobengröße	25.410 Personen aus 11.771 Haushalten
Datenerhebung	Ausgangspunkt ist das Originalmaterial des Mikrozensus 1998. Der Mikrozensus ist eine repräsentative 1% Bevölkerungsstichprobe. Der CAMPUS-File ist eine 3,5 % Stichprobe des Ausgangsdatenmaterials.
Inhalt	Neben Angaben zur Bezugsperson enthält dieses Campus-File u.a. Angaben zu Erwerbstätigkeit, Arbeitssuche, Unterhalt, Einkommen und Familienkonzepten. Insgesamt stehen 199 Merkmale des Mikrozensus zur Verfügung.
Datenzugang	http://www.forschungsdatenzentrum.de/bestand/mikrozensus/cf/1998/index.asp .

Die Hochrechnungsfaktoren für Personen, Haushalte sowie Wohnungen wurden im CAMPUS-File nach der Methode der gebundenen Hochrechnung angepasst. Die Erzeugung der gebundenen Hochrechnungsfaktoren geschah nach Anpassungsklassen. Durch die Erzeugung von gebundenen Hochrechnungsfaktoren nach Anpassungsklassen ist eine nahezu verzerrungsfreie Hochrechnung der Werte aus dem CAMPUS-File auf die Gesamtbevölkerung möglich.

⁷ Zum Scientific Use File des Mikrozensus siehe Müller et al (1991) sowie Schimpl-Neimanns (2004); siehe auch die Hinweise in Fußnote 6.

Seitens des Zentrums für Umfragen, Methoden und Analysen (ZUMA) wurde das CAMPUS-File des Mikrozensus von Anfang an genutzt. Aus diesen Erfahrungen heraus konnte das CAMPUS-File des Mikrozensus weiterentwickelt werden.⁸ So konnten Probleme bei der Stichprobe überwunden werden sowie Probleme mit den Hochrechnungsfaktoren korrigiert werden. Aus der Verwendung des CAMPUS-Files innerhalb der Workshops zum Mikrozensus, die ZUMA regelmäßig anbietet, resultierte auch die Anforderung den ausgewählten Merkmalskatalog zu überarbeiten. Das CAMPUS-File Mikrozensus 2002 das im Frühjahr 2007 veröffentlicht wird, hat ein in dieser Hinsicht optimiertes Merkmalsprogramm.

Nach den Erfahrungen, die ZUMA in der Ausbildung (Workshops) gemacht hat, erscheint es notwendig, für die Lehre einen CAMPUS-File einer älteren Mikrozensushebung verfügbar zu haben. Nur so ist es z.B. möglich, in der Lehre den Nutzer auf die Schwierigkeiten bei vergleichenden Analysen vorzubereiten. Idealerweise sollte von den zur Verfügung stehenden älteren Scientific Use Files der Mikrozensus 1976 verwendet werden, da für diesen Mikrozensus z.B. ein höherer Merkmalsumfang als beim Mikrozensus 1973 (insbesondere Bildungsvariablen) zur Verfügung steht und sich im Vergleich zu den neueren Daten die meisten Unterschiede ergeben (Stichprobenplan, Erhebungsprogramm, Variablenbrüche etc.). Dieses Datenmaterial soll in einer Zusammenarbeit des Forschungsdatenzentrums der Statistischen Ämter der Länder mit ZUMA entstehen.

Weitere Informationen zum CAMPUS-File Mikrozensus 1998 sowie die Daten finden sich unter <http://www.forschungsdatenzentrum.de/bestand/mikrozensus/cf/1998/index.asp>.

4.2 Das CAMPUS-File der Lohn- und Einkommensteuer 1998

Die Lohn- und Einkommensteuerstatistik ist eine alle drei Jahre durchgeführte dezentrale Sekundärstatistik. D.h. die Angaben werden nicht für den statistischen Zweck erhoben, sondern fallen in 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 Landesämter. Diese generieren die jeweiligen Landesergebnisse und übermitteln die sich ergebenden Tabellen an das Statistische Bundesamt. Das Statistische Bundesamt führt die Länderergebnisse dann in einem nächsten Schritt zum Bundesergebnis zusammen. Seit dem Jahressteuergesetz 1996⁹ werden neben den Tabellendaten auch die Einzelangaben von den Statistischen Landesämtern an das Statistische Bundesamt übermittelt. Durch die Möglichkeit der Steuerpflichtigen, ihre Einkünfte je nach Fall bis zu drei Jahren nach ihrer Entstehung zu erklären, liegt die Lohn- und Einkommenssteuerstatistik immer erst mit einer Zeitverzögerung vor.

Die Lohn- und Einkommensteuerstatistik weist für rund 29 Mio. Steuerpflichtige mit über 1000 Merkmalen die primäre Einkommensentstehung detailliert nach. Von den sieben Einkunftsarten, über die verschiedenen Einkommensgrößen der Einkommensteuer, bis zur festgesetzten Einkommensteuer werden vielfältige Angaben erfasst.¹⁰ Darüber hinaus stehen im begrenzten Umfang sozioökonomische Merkmale zur Verfügung. Ergebnisse auf der Grundlage der Einzeldaten der Lohn- und Einkommensteuerstatistik werden in Politik und Wissenschaft intensiv genutzt.¹¹

Seit 2004 liegt zur Lohn- und Einkommensteuerstatistik 1998 ein Scientific Use File vor (FAST98).¹² Das Scientific Use File beruht auf einer disproportional geschichteten 10%-Stichprobe mit rund 3 Mio. Datensätzen.¹³

⁸ Siehe zu diesen ersten Erfahrungen Wirth/Schimml-Neimanns (2004).

⁹ Novellierung des „Gesetzes über Steuerstatistiken“ (StStatG) mit dem 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).

¹⁰ siehe hierzu Kordsmeyer (2004).

¹¹ Zu Anwendungsbeispielen siehe Zwick/Merz (2007).

¹² Vgl. Merz et al (2006) sowie Vorgrimler (2006).

¹³ Zur Stichprobe siehe Zwick (1998)

Das CAMPUS-File Einkommensteuer 1998 ist eine knapp 1%-Stichprobe mit 234 510 Steuerpflichtigen. Die vollständige, alle Merkmale umfassende Stichprobe, die analog zum Stichprobenplan der 10%-Stichprobe gezogen wurde, ist Grundlage des Steuersimulationsmodells des Bundesministeriums der Finanzen.¹⁴ Aufbauend auf den Anonymisierungsverfahren von FAST98 wurde der CAMPUS-File Einkommensteuer 1998 entwickelt.

Übersicht 3: CAMPUS-File Einkommensteuer 1998

Stichprobengröße	234 510 Steuerpflichtige
Datenerhebung	Ausgangspunkt ist die Einkommensteuerstatistik 1998 mit rund 29 Mio. Steuerpflichtigen. Der CAMPUS-File ist eine geschichtete rund 1 %-Stichprobe der unbeschränkt Steuerpflichtigen.
Inhalt	Aus den Einkommensteuererklärungen wurden für das CAMPUS-File 30 Merkmale des Besteuerungsverfahrens übernommen. Insgesamt stehen 38 Merkmale zur Verfügung.
Datenzugang	http://www.forschungsdatenzentrum.de/bestand/lest/cf/1998/index.asp

In einem ersten Schritt wurde hierzu das Datenmaterial nach drei Merkmalskategorien strukturiert.

Merkmale der 1. Kategorie:

- Summe der Einkünfte (im Splittingfall Männer (M) und Frauen(F))
- Gesamtbetrag der Einkünfte
- Einkommen
- zu versteuerndes Einkommen
- tarifliche Einkommensteuer
- festzusetzende Einkommensteuer

Merkmale der 2. Kategorie:

- die sieben Einkunftsarten, getrennt nach M und F
- Sonderausgaben, die nicht Vorsorgeaufwendungen sind
- Sonderausgaben: Vorsorgeaufwendungen
- Außergewöhnliche Belastungen, abzugfähig – bei getrennter Veranlagung –M –
- Außergewöhnliche Belastungen, abzugfähig – bei getrennter Veranlagung –F –
- Förderung des Wohneigentums: Steuerbegünstigungen insgesamt

Merkmale der 3. Kategorie:

alle weiteren stetigen Merkmale

Die Merkmale der ersten Kategorie stehen vollständig zur Verfügung. Die Merkmale der Kategorie 2 sind im CAMPUS-File nur als Dummy-Variablen vorhanden. D.h. wenn das Merkmal für einen Steuerpflichtigen bzw. Steuerfall vorliegt weist die Merkmalsausprägung den Wert ‚Eins‘ aus, ansonsten den Wert ‚Null‘. Die Merkmale der Kategorie 3 sind im CAMPUS-File nicht enthalten. Bei den zehn Steuerpflichtigen mit den höchsten Einkommensangaben, gemessen am Gesamtbetrag der Einkünfte, wurden die stetigen Merkmalswerte der Kategorie 1 durch das jeweilige arithmetische Mittel dieser zehn Steuerpflichtigen ersetzt. Analog wurde bei den zehn Steuerpflichtigen mit den niedrigsten Einkommensangaben verfahren. Aufgrund vorhandener zum Teil sehr hoher Verlustfälle in der Einkommensteuerstatistik ein notwendiges Verfahren, da diese Angaben, wie auch bei hohen Einkommen, stärker gefährdet sind.

¹⁴ Vgl. Lietmeyer (2007).

Weitere Metadaten zum CAMPUS-File sowie Daten und Einleseroutinen für die Programme SAS, SPSS und Stata findet sich unter <http://www.forschungsdatenzentrum.de/bestand/lest/cf/1998/index.asp>.

4.3 Das CAMPUS-File der Kostenstrukturerhebung 1999

Im Grunde waren die anonymisierten Mikrodaten der Kostenstrukturerhebung (KSE) das erste CAMPUS-File, nur gab es bei der Erstellung dieses Datensatzes noch nicht den markanten Namen. Im Rahmen des Projektes ‚Anonymisierung wirtschaftsstatistischer Einzeldaten‘ entstand in einer ersten Phase ein Public Use File der KSE. Hier war es das Ziel, einen ersten Datensatz im Bereich der wirtschaftsstatistischen Daten zu konzipieren. Um dieses Ziel zu erreichen wurde das Informationspotential der Daten der KSE stark eingeschränkt. Darüber hinaus wurden die Ausprägungen der verschiedenen Merkmale mit datenverändernde Verfahren behandelt. Der daraus resultierte Datensatz ist für fundierte wissenschaftliche Analysen nur sehr eingeschränkt zu verwenden. Aber mit diesem File lag erstmals ein anonymisierter Unternehmensdatensatz aus der amtlichen Statistik vor.¹⁵ Im weiteren Verlauf des Projektes wurde dann auch ein Scientific Use File, mit deutlich höherem Analysegehalt, entwickelt.¹⁶

Übersicht 4: CAMPUS-File Kostenstrukturerhebung 1999

Stichprobengröße	500 Unternehmen mit maximal 250 Beschäftigte
Datenerhebung	Die Kostenstrukturerhebung im Verarbeitenden Gewerbe erfasst als geschichtete Zufallsstichprobe jährlich rund 13 000 Unternehmen mit 20 Beschäftigten und mehr.
Inhalt	Das Erhebungsprogramm umfasst die tätigen Personen, den Gesamtumsatz, die Kosten nach Kostenarten sowie zusätzlich die Investitionen.
Datenzugang	http://www.forschungsdatenzentrum.de/bestand/kse/cf/1999/index.asp

Inhaltlich liefern die Kostenstrukturerhebungen im Verarbeitenden Gewerbe umfassende Informationen zu Unternehmen im Bereich des Produzierenden Gewerbes. Sie dienen als Ausgangspunkt für vielfältige Strukturuntersuchungen in Politik und Verwaltung sowie in der Wirtschaft und hier insbesondere in den Verbänden. Die Informationen der KSE bilden darüber hinaus eine Datengrundlage für die Volkswirtschaftlichen Gesamtrechnungen. Hier werden die Ergebnisse vor allem für die Berechnung der Wertschöpfung und ihrer Komponenten nach Wirtschaftsbereichen im Rahmen der Entstehungsrechnung herangezogen; schließlich liefern sie auch wichtige Informationen für die Input-Output-Rechnungen

Die Kostenstrukturerhebung im Verarbeitenden Gewerbe des Jahres 1999 – reduziert auf Unternehmen mit 20 bis einschließlich 249 Beschäftigten (kleinere und mittlere Unternehmen) – erfasst als Stichprobe etwa 13 000 Unternehmen. Die Befragung erfolgt zentral durch das Statistische Bundesamt im Wege der Selbstausfüllung durch die Unternehmen. Die in der Stichprobe gewonnenen Ergebnisse werden auf die Gesamtheit der Unternehmen zwischen 20 und 249 Beschäftigten hochgerechnet. Die Stichprobe wird in der Regel alle 4 Jahre neu gezogen, so dass kleinere und mittlere Unternehmen durch Rotation entlastet werden können.

Die Daten der KSE bilden im Kern den Wertschöpfungsprozess ab:

¹⁵ Siehe zu diesen ersten Projektergebnissen Ronning/Gnoss (2003)

¹⁶ Zu den Ergebnissen des Projektes ‚Anonymisierung wirtschaftsstatistischer Einzeldaten‘ siehe Ronning et al (2005)

	Gesamtumsatz
+	Bestandsveränderungen an unfertigen und fertigen Erzeugnissen aus eigener Produktion
+	Selbsterstellte Anlagen
=	Bruttoproduktionswert (Gesamtleistung)
-	Materialverbrauch, Einsatz an Handelsware, Kosten für Lohnarbeiten
=	Nettoproduktionswert
-	Sonstige Vorleistungen
=	Bruttowertschöpfung
-	indirekte Steuern (ohne Umsatzsteuer) abz. Subventionen
=	Bruttowertschöpfung zu Faktorkosten
-	Abschreibungen
=	Nettowertschöpfung zu Faktorkosten

Auch mit dem CAMPUS-File KSE lässt sich die Nettowertschöpfung für die enthaltenen Unternehmen nachvollziehen. Neben dieser Möglichkeit bietet dieser CAMPUS-File die Möglichkeit, die Wirkungsweise alternativer Anonymisierungsverfahren zu vermitteln. Bei diesem Datensatz wurden neben den klassischen informationsreduzierenden Verfahren wie oben schon benannt auch datenverändernde Verfahren eingesetzt.¹⁷ Hierbei zeigt sich, dass die Verfahren der Zufallsüberlagerung von Merkmalsausprägungen je nach Datenkonzeption schwerwiegende Nachteile aufweisen. Im CAMPUS-File der KSE hat der Einsatz dieser Verfahren dazu geführt, dass der definitorische Zusammenhang vom Gesamtumsatz bis zur Nettowertschöpfung durch das Anonymisierungsverfahren verloren gegangen ist. Gerade diese Problematik ermöglicht es, in der Lehre Vorteile und Probleme verschiedener Anonymisierungsverfahren mit ‚echten‘ Daten zu vermitteln.

Mikro- wie Metadaten der Kostenstrukturerhebung im Verarbeitenden Gewerbe finden sich unter <http://www.forschungsdatenzentrum.de/bestand/kse/cf/1999/index.asp>

5. Einsatz und Ausblick

Zurzeit ist leider nur eingeschränkt die Verbreitung der CAMPUS-Files bekannt. Sucht man unter diesem Begriff im Internet stößt man aber mittlerweile schon auf verschiedene Lehrangebote. An der Johann Wolfgang Goethe Universität in Frankfurt werden die CAMPUS-Files in den Veranstaltungen ‚Wirtschaftsstatistik‘ sowie ‚Multivariate Statistik‘ durch den Autor verwandt. Darüber hinaus ist, neben der Verwendung bei ZUMA, bisher nur wenig über die Nutzung der CAMPUS-Files bekannt. Dies wird sich durch zwei Maßnahmen verändern.

Eine technische Maßnahme wird im Laufe des Jahres dazu führen, dass die Statistischen Ämter des Bundes und der Länder durch die Downloads der CAMPUS-Files etwas mehr über die Nutzer erfahren. Der Datendownload der CAMPUS-Files wird weiterhin kostenfrei sein aber es ist geplant, vor der Möglichkeit des Datenzugangs eine Registrierung vorzusehen. Hierbei kann der Nutzer auch wählen, ob er zukünftig über Neuerungen oder Veranstaltungen informiert werden möchte.

Eine weitere Maßnahme zur Verbesserung der Kommunikation mit den Nutzern werden gezielte Workshops zu den CAMPUS-Files sein. Hier finden sich erste Angebote zum CAMPUS-File des Mikrozensus bei ZUMA. Zum CAMPUS-File der Einkommensteuer wird das Statistische Bundesamt voraussichtlich im Frühjahr 2008 einen ersten Workshop anbieten. Neben der Vermittlung der Inhalte sollen diese Veranstaltungen durch den Dialog mit den Anwendern auch dazu genutzt werden, die CAMPUS-Files inhaltlich weiterzuentwickeln.

¹⁷ Zu klassischen und datenverändernden Verfahren der Anonymisierung siehe u.a. Rosemann (2006)

Der Dialog soll auch dazu führen, den Bedarf an weiteren CAMPUS-Files zu bestimmen. Zurzeit zeigt sich, dass insbesondere der Wunsch geäußert wird, dass die Statistischen Ämter auch ein CAMPUS-File der Einkommens- und Verbrauchsstichprobe zur Verfügung stellen. Dieses Projekt soll daher in diesem Jahr noch angegangen werden.

Weiter soll die Idee des CAMPUS-Files in die Richtung eines SCHOOL-Files weitergedacht werden. Hier laufen zum einen Gespräche, ob und wie CAMPUS-Files mit Lernprogrammen, wie z.B. dem Statistiklabor, kombiniert werden können. Eine Überlegung in diesem Zusammenhang ist es, aus CAMPUS-Files kleinere SCHOOL-Files zu konzipieren, die dann problemlos in Lernprogramme integrierbar und auch mit Hilfe von Excel zu bearbeiten sind.

Literatur:

Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und amtlicher Statistik (Hrsg. 2001), Wege zu einer besseren informationellen Infrastruktur, Nomos Verlag, Baden-Baden.

Kordsmeyer, V. (2004), Die Einkommensteuerstatistik als Mikrodatenfile, in: Merz, J., Zwick, M. (Hrsg.); Mikroanalysen und amtliche Statistik – MIKAS; Reihe ‚Statistik und Wissenschaft‘ des Statistisches Bundesamt, Band 1, Wiesbaden.

Lietmeyer, V. (2007), Neue Wege in der mikroanalytischen Steuerschätzung, in: Zwick, M.; Merz, J. (2007).

Merz, J., Vorgrimler, D., Zwick, M. (2006), De facto anonymised microdata file on income tax statistics 1998, Zeitschrift für Wirtschafts- und Sozialwissenschaften - Schmollers Jahrbuch, 126. Jahrgang, Heft 2, S. 313 - 327.

Müller, W., Blien, U., Knoche, P., Wirth, H. u.a. (1991); Die faktische Anonymität von Mikrodaten, Schriftenreihe Forum der Bundesstatistik, Band 19, Metzler-Poeschel, Stuttgart.

Ronning, G., Gnoss, R. (2003), Anonymisierung wirtschaftsstatistischer Einzeldaten, Forum der Bundesstatistik, Statistisches Bundesamt, Band 42.

Ronning, G., Sturm, R., Höhne, J., Lenz, R., Rosemann, M., Scheffler, M., Vorgrimler, D. (2005), Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten, Statistisches Bundesamt, Statistik und Wissenschaft, Band 4.

Rosemann, R. (2006), Auswirkungen datenverändernder Anonymisierungsverfahren auf die Analyse von Mikrodaten, IAW-Forschungsberichte Nr. 66, Tübingen.

Schimpl-Neimanns, B. (2004), Anwendungen und Erfahrungen mit dem Scientific Use File des Mikrozensus, in: Merz, J., Zwick, M. (Hrsg.); Mikroanalysen und amtliche Statistik – MIKAS; Reihe ‚Statistik und Wissenschaft‘ des Statistisches Bundesamt, Band 1, Wiesbaden.

Vorgrimler, D. (2006), Anonymisierte Daten der amtlichen Steuerstatistik, FDZ-Arbeitspapiere Nr. 13.

Wirth, H., Müller, W. (2006), Mikrodaten der amtliche Statistik – Ihr Potential in der empirischen Sozialforschung; in: Diekmann, A. (Hrsg.), Methoden der Sozialforschung, Kölner Zeitschrift für Soziologie und Sozialpsychologie, Sonderheft 44.

Wirth, H., Schimpl-Neimanns, B. (2004), Anregungen zur konzeptionellen Überarbeitung des Campus File Mikrozensus 1998, German Microdata Lab, Arbeitspapier 03.

Zühlke, S., Zwick, M., Scharnhorst, S., Wende, T (2003), Die Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder, Wirtschaft und Statistik 10.

Zwick, M., Merz, J. (Hrsg. 2007); Mikroanalysen und Steuerpolitik, Statistik und Wissenschaft, Band 7, Wiesbaden

Zwick, M. (2006); Forschungsdatenzentren, Nutzen und Kosten einer informationellen Infrastruktur für Wissenschaft, Politik und Datenproduzenten, *Wirtschaft und Statistik* 12.

Zwick, M. (1998), Einzeldatenmaterial und Stichproben innerhalb der Steuerstatistiken, *Wirtschaft und Statistik* 7.