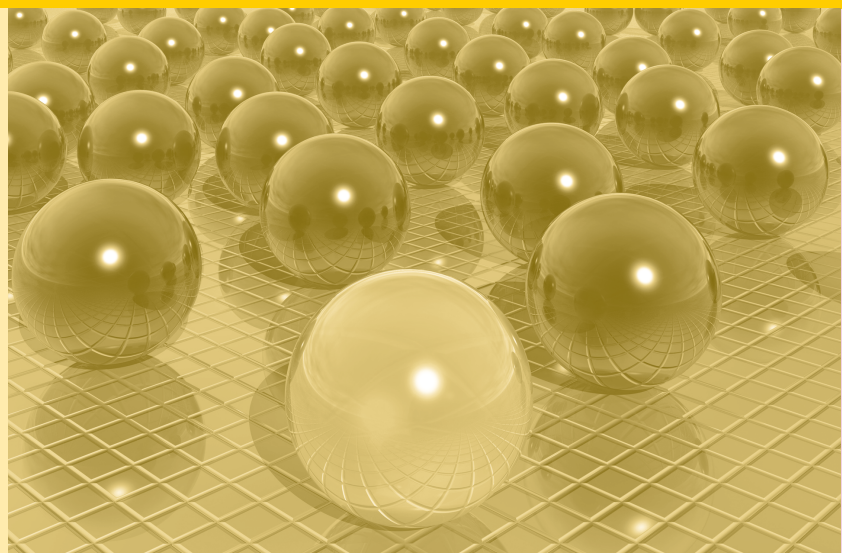
FDZ-Arbeitspapier
Nr. 40



Definition von nutzerseitigen Kriterien für Datenstrukturfiles

Julia Höninger, Martin Rosemann, Ramona Voshage

2011

STATISTISCHE ÄMTER
DES BUNDES UND DER LÄNDER
FORSCHUNGSDATENZENTREN

Impressum

Herausgeber: Statistische Ämter des Bundes und der Länder
Herstellung: Information und Technik Nordrhein-Westfalen
Telefon 0211 9449-01 • Telefax 0211 442006
Internet: www.forschungsdatenzentrum.de
E-Mail: forschungsdatenzentrum@it.nrw.de

Fachliche Informationen
zu dieser Veröffentlichung:

Forschungsdatenzentrum der
Statistischen Landesämter
– Geschäftsstelle –
Tel.: 0211 9449-2873
Fax: 0211 9449-8087
forschungsdatenzentrum@it.nrw.de

Informationen zum Datenangebot:

Statistisches Bundesamt
Forschungsdatenzentrum

Tel.: 0611 75-4220
Fax: 0611 72-3915
forschungsdatenzentrum@destatis.de

Forschungsdatenzentrum der
Statistischen Landesämter
– Geschäftsstelle –
Tel.: 0211 9449-2876
Fax: 0211 9449-8087
forschungsdatenzentrum@it.nrw.de

Erscheinungsfolge: unregelmäßig
Erschienen im Juni 2011

Diese Publikation wird kostenlos als **PDF-Datei** zum Download unter www.forschungsdatenzentrum.de angeboten.

Bei den enthaltenen statistischen Angaben handelt es sich um eigene Arbeitsergebnisse der genannten Autoren im Zusammenhang mit der Nutzung der bereitgestellten Daten der Forschungsdatenzentren. Es handelt sich hierbei ausdrücklich nicht um Ergebnisse der Statistischen Ämter des Bundes und der Länder.

FDZ-Arbeitspapier
Nr. 40

Definition von nutzerseitigen Kriterien für Datenstrukturfiles

Julia Höninger, Martin Rosemann, Ramona Voshage

2011

STATISTISCHE ÄMTER
DES BUNDES UND DER LÄNDER
FORSCHUNGSDATENZENTREN

Definition von nutzerseitigen Kriterien für Datenstrukturfiles

Julia Höninger[1], Martin Rosemann[2] und Ramona Voshage[3]

Zusammenfassung. Datenstrukturfiles werden Nutzern der Forschungsdatenzentren zur Verfügung gestellt, wenn sie per kontrollierter Datenfernverarbeitung Mikrodaten auswerten wollen. Anhand der Datenstrukturfiles können Wissenschaftler Programmcodes entwickeln, die sie per E-Mail an das Forschungsdatenzentrum senden, um dann die statistischen Ergebnisse zurück zu erhalten. Damit Auswertungsprogramme möglichst gut entwickelt werden können, sollten Datenstrukturfiles die Struktur der Originaldaten semantisch und syntaktisch korrekt abbilden. Dieser Artikel erläutert, warum dies aus Nutzersicht notwendig ist und was dies konkret bedeutet. Es werden sowohl die Anforderungen an Inhalt und Form von Datenstrukturfiles sowie an Bedienbarkeit und Verfügbarkeit von Datenstrukturfiles dargelegt.

Schlüsselwörter: Datenangebot, Paneldaten, kontrollierte Datenfernverarbeitung, Forschungsdatenzentrum

JEL: C10, C50, C81

[1] Amt für Statistik Berlin-Brandenburg, Forschungsdatenzentrum, Alt-Friedrichsfelde 60, 10315 Berlin. E-Mail: Julia.Hoeninger@statistik-bbb.de
[2] Institut für Angewandte Wirtschaftsforschung (IAW), Ob dem Himmelreich 2, 72072 Tübingen. E-Mail: Martin.Rosemann@iaw.edu.
[3] Amt für Statistik Berlin-Brandenburg, Forschungsdatenzentrum, Alt-Friedrichsfelde 60, 10315 Berlin. E-Mail: Ramona.Voshage@statistik-bbb.de

Das Forschungsprojekt „infinitE – Eine informationelle Infrastruktur für das E-Science Age“[4] beschäftigt sich mit zwei Teilmodulen, um sich dem fernen Ziel eines echten Remote Access zu Mikrodaten der amtlichen Statistik in Deutschland zu nähern. Der Datenzugang soll durch bessere Datenstrukturfiles und eine automatisierte Ergebniskontrolle verbessert werden. Im Mittelpunkt des Projektes stehen Wirtschaftsstatistiken. Dieser Beitrag ist die überarbeitete Version eines Abschnitts aus dem Zwischenbericht des Forschungsprojektes „infinitE“.

1 Hintergrund

Seit 2002 wurden bei den großen amtlichen Datenproduzenten Forschungsdatenzentren (FDZ) eingerichtet um der Wissenschaft den Zugang zu Mikrodaten zu ermöglichen. So gibt es inzwischen unter anderem Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder (Zühlke et al. 2004, Zühlke et al. 2007), ein FDZ der Deutschen Rentenversicherung (Rehfeld 2009, Stegmann 2009) und ein FDZ der Bundesagentur für Arbeit im Institut für Arbeitsmarkt- und Berufsforschung (Allmendinger und Kohlmann 2005). Alle Forschungsdatenzentren bieten verschiedene Wege des Datenzugangs an. Die meisten Forschungsdatenzentren geben anonymisierte Datensätze, so genannte Scientific Use Files, an die Wissenschaft heraus. Hier ist durch die Anonymisierungsmaßnahmen jedoch der Informationsgehalt reduziert. Alternativ können Wissenschaftler[5] an speziell ausgestatteten Gastwissenschaftlerarbeitsplätzen in den Räumen der Datenproduzenten arbeiten und sich die auf Geheimhaltung geprüften statistischen Ergebnisse exportieren lassen. Eine dritte Möglichkeit des Datenzugangs ist die kontrollierte Datenfernverarbeitung. Bei dieser Art des Datenzugangs senden Wissenschaftler Auswertungsprogramme an das FDZ, die FDZ-Mitarbeiter führen diese auf den Originaldaten aus und senden die auf Geheimhaltung geprüften Ergebnisse an die Wissenschaftler zurück.

Wollen Wissenschaftler Mikrodaten über die kontrollierte Datenfernverarbeitung auswerten, werden ihnen vom FDZ so genannte Datenstrukturfiles zur Verfügung gestellt. Anhand dieser Datenstrukturfiles können Wissenschaftler ihren Programmcode entwickeln, den sie per E-Mail an das Forschungsdatenzentrum senden um dann die statistischen Ergebnisse zurück zu erhalten. Damit Auswertungsprogramme möglichst gut entwickelt werden können, sollten Datenstrukturfiles die Struktur der Originaldaten semantisch und syntaktisch korrekt abbilden.

[4] Das Forschungsprojekt „infinitE“ wird vom Bundesministerium für Bildung und Forschung (BMBF) gefördert.
[5] Zur sprachlichen Vereinfachung wird in diesem Text nur die männliche Form des Wortes „Wissenschaftler“ und seinen Synonymen verwendet, wobei die Forschungsdatenzentren natürlich weibliche und männliche Wissenschaftler betreuen.

Bisher besteht ein Datenstrukturfile (DSF) in der Regel aus einer Stichprobe der Originaldaten, auf welche weitere Anonymisierungsmaßnahmen angewendet werden oder aus zufällig generierten Werten im Wertebereich des Originaldatensatzes. Bei beiden Vorgehensweisen bleiben die Merkmale erhalten, ihre Ausprägungen und die inneren Abhängigkeitsstrukturen zu anderen Merkmalen werden dabei in der Regel zerstört. Der Wissenschaftler kann prüfen, ob sein Programm lauffähig ist, er bekommt aber keine Hinweise, ob er seine inhaltliche Fragestellung adäquat umgesetzt hat. Multivariate Analysen sind mit den bisherigen Datenstrukturfiles häufig nicht möglich, da die Korrelationen durch die Anonymisierungsmaßnahmen oft vollkommen zerstört werden. Die Herausforderung, auch die Abhängigkeitsstrukturen zwischen den Merkmalen annähernd zu erhalten, verschärft sich bei Paneldaten, da hier zusätzlich die zeitliche Dimension zu berücksichtigen ist. Viele Programme können nicht auf ihre Lauffähigkeit geprüft werden, da inhaltliche Strukturen der Datensätze nicht erhalten bleiben. So können Variablen für multivariate Auswertungen oder Regressionen bisher aufgrund fehlender Merkmalskonstellationen nicht generiert und somit die Auswertungen nicht ausgeführt werden (z.B. Wachstumsraten in Paneldatensätzen). Auch eine Abschätzung der voraussichtlichen Laufzeit ist meist nicht möglich, da die Datenstrukturfiles wesentlich kleiner als die Originaldatensätze sind.

Mit der Etablierung der Forschungsdatenzentren als wichtige Bestandteile des Forschungsstandortes Deutschlands fragen die Wissenschaftler zunehmend komplexere Datenbestände, vor allem Paneldatensätze, nach. Die FDZ bedienen diese Nachfrage durch eine Reihe innovativer Projekte, so unter anderem die Projekte „AFiD – Amtliche Firmendaten für Deutschland“ (Malchin und Voshage 2009, Malchin et al. 2011) und „KombiFiD – Kombinierte Firmendaten für Deutschland“ (Gürke et al. 2011). Auch im FDZ der BA im IAB wurde aufgrund der häufig nachgefragten Paneldaten ein Verfahren zur Erzeugung von Datenstrukturfiles entwickelt, das insbesondere die Panelstruktur bei der Anonymisierung aufrecht erhält (Jacobebbinghaus et al. 2010).

An den oben dargestellten Kritikpunkten zu einigen heutigen Datenstrukturfiles sollte aus Sicht der Datennutzer angesetzt werden, um einen effizienteren Ablauf der kontrollierten Datenfernverarbeitung zu ermöglichen. Ziel dieses Papiers ist es daher, Kriterien für semantische Datenstrukturfiles aus Sicht der Datennutzer zu skizzieren. An diesen Kriterien können die entwickelten Anonymisierungs-verfahren zur Erstellung von Datenstrukturfiles gemessen werden.

2 Anforderungen an Inhalt und Form der Datenstrukturfiles

2.1 Allgemeine Anforderungen an Inhalt und Form der Datenstrukturfiles

Allgemein sollen die Datennutzer anhand von Datenstrukturfiles ihre Analyseprogramme so spezifizieren können, dass diese anschließend ohne Änderungen mit den Originaldaten funktionieren und sinnvolle Ergebnisse erzielen. Außerdem wäre es wünschenswert, wenn möglichst wenige Auswertungsläufe nötig sind, um zu publikationsreifen Ergebnissen zu gelangen. Der Wissenschaftler möchte möglichst schnell seine Endergebnisse erhalten, aber auch für das FDZ bedeuten viele Auswertungsläufe für Zwischenergebnisse einen erheblichen Aufwand, der so weit wie möglich reduziert werden sollte.

Damit muss angestrebt werden, dass (1) alle Analysen, die mit den Originaldaten möglich sind, auch mit den Datenstrukturfiles durchgeführt werden können und zudem (2) sollten die mit den Datenstrukturfiles erzielten Ergebnisse im Wesentlichen die inhaltlichen Aussagen reproduzieren, die auch mit den Originaldaten getroffen werden würden. Im Unterschied zu einem Scientific Use File müssen die Ergebnisse aber nicht für Publikationen oder für die weitere Forschung geeignet sein. Wissenschaftler müssen sich bewusst sein, dass die gefundenen Ergebnisse häufig nicht mit den Originalergebnissen übereinstimmen und somit für sich alleine nicht belastbar sind.

Ausgehend von diesen beiden Zielen wurden im Rahmen des Projekts „infinitE“ operationalisierbare Kriterien für Datenstrukturfiles definiert. Dabei ist leicht einsichtig, dass die Definition von Kriterien, die für das Ziel (1) relevant sind, deutlich einfacher ist, als die Definition von Kriterien für das Ziel (2). Für das zweite Ziel kann jedoch auf die Vorarbeiten in den Projekten „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten“ (Ronning et al. 2005) und „Wirtschaftsstatistische Paneldaten und faktische Anonymisierung“ (Höhne 2008, Brandt et al. 2008, Höhne 2010) zurückgegriffen werden.

Im Folgenden wird genauer auf die Kriterien für die beiden genannten Ziele eingegangen. Beide Ziele sind wichtig, um den personellen Aufwand in den FDZ zu verringern: Getestete Programme, die ohne Fehlermeldung funktionieren, ersparen die oft zeitintensiven manuellen Anpassungen oder ein ebenso zeitintensives Hin- und Herschicken der Programme zwischen FDZ und Wissenschaftlern. Grobe Kenntnisse von Analysetendenzen ermöglichen gezieltere Auswertungen der Forschungsfrage und reduzieren die Anzahl der Anfragen.

2.2 Kriterien, um die Durchführbarkeit der gleichen Analysen sicherzustellen

Im Folgenden werden zunächst gemäß Ziel (1) Kriterien für Datenstrukturfiles vorgestellt, mit denen sichergestellt wird, dass alle Analysen, die mit den Originaldaten möglich sind, auch mit den

Datenstrukturfiles durchgeführt werden können, wobei auf übliche und sinnvolle Analysen abgestellt wird:

1. Alle (belastbaren, sinnvoll verwendbaren) Merkmale aus den Originaldaten sollen auch im Datenstrukturfile enthalten sein. In manchen Datenbeständen sind noch Merkmale enthalten, deren Qualität als unzureichend eingeschätzt wird und bei denen empfohlen wird von einer Auswertung abzusehen. Da diese Variablen meist wenig befüllt sind, stellen sie bei manchen Anonymisierungsverfahren ein Problem da, wenn z.B. Korrelationsmatrizen berechnet werden. Daher sollen solche Variablen, bei denen von einer Auswertung aus inhaltlichen Gründen abgeraten wird, gar nicht erst in einem Datenstrukturfile enthalten sein.
2. Zuordnungen sollten beibehalten werden, d.h. bspw. bei einem Panel sollten die systemfreien Nummern eines Betriebes über die Jahre identisch sein; generell sollten Datensätze über die systemfreien Betriebs- und Unternehmensnummern verknüpfbar bleiben.
3. Bei metrischen Merkmalen sind die Wertebereiche zu erhalten. Merkmale, die nur positive Werte annehmen können, sollten auch im Datenstrukturfile auf positive Werte beschränkt bleiben. Es sollten z.B. keine negativen Beschäftigtenzahlen im Datenstrukturfile enthalten sein.
4. Bei kategorialen Merkmalen sind alle – zumindest alle für Auswertungen ausreichend gefüllte – Kategorien zu erhalten. Existiert im Datensatz in einer Variable eine Kategorie, die so schwach besetzt ist, dass eine Darstellung dieser Kategorie schon aus Datenschutzgründen nicht zulässig ist, könnte als Kompromiss bereits vor der Anonymisierung diese Kategorie mit einer anderen zusammengelegt werden, wenn im Originaldatensatz ebenso verfahren wird.
5. Die Dimension (Größenordnung) des Datensatzes sollte annähernd erhalten bleiben. Nur wenn ungefähr eine ähnliche Anzahl an Variablen und Beobachtungen im Datenstrukturfile enthalten ist, kann der Nutzer erkennen welche Anforderungen an Speicherplatz und welche Rechenzeit für die gewünschten Auswertungen benötigt werden. So kann der Nutzer bereits selbst abschätzen, dass Ergebnisse von Auswertungsprogrammen mit sehr langen Rechenzeiten (mehrere Stunden oder Tage) ihm erst später als Ergebnisse von kurzen Programmen zur Verfügung gestellt werden können. Alternativ hat der Nutzer mit seinem Datenstrukturfile die Möglichkeit die Programmierung effizienter zu gestalten.
6. Die Häufigkeiten in den einzelnen Kategorien einer diskreten Variablen sollten annähernd erhalten bleiben, damit bei Modellspezifikationen abgeschätzt werden kann, welche Differenzierungen dieser Variablen jeweils sinnvoll sind.

2.3 Kriterien, um die Ähnlichkeit der Ergebnisse sicherzustellen

Forscher beginnen bei Auswertungen eines Datensatzes üblicherweise mit deskriptiven Analysen. Um sich mit einem Datensatz vertraut zu machen, werden ein- und zweidimensionale Häufigkeitstabellen, Perzentile, Mittelwerte und Streuungsmaße einzelner Variablen berechnet. Neue Variablen werden gebildet und für diese ebenfalls deskriptive Kennzahlen ausgegeben. Erst danach werden gängigerweise ökonometrische Modelle geschätzt. Nichtlineare und strukturelle Schätzer folgen erst am Ende, wenn der Wissenschaftler bereits mit der Struktur der Daten vertraut ist.

Im Folgenden werden Kriterien für Datenstrukturfiles aufgelistet, die sicherstellen sollen, dass ähnliche Analyseergebnisse wie mit den Originaldaten erzielt werden. Die Reihenfolge orientiert sich dabei an der üblichen Auswertungsstrategie der Wissenschaftler:

1. Bei metrischen Merkmalen sollte die Spannweite ungefähr erhalten werden.
2. Logische Zusammenhänge sind zu erhalten. So sollte beispielweise der Gesamtumsatz eines Betriebes oder Unternehmens der Summe aus Inlands- und Auslandsumsatz entsprechen.
3. Strukturelle Nullen sind zu erhalten. Betreibt ein Unternehmen keinen Handel, was durch eine Null bei der Dummyvariable Handelstätigkeit (1=ja, 0=nein) angezeigt wird, so muss der Handelsumsatz im Datensatz ebenfalls den Wert Null annehmen. Die verschiedenen Teilnahmemuster bei Paneldatensätzen müssen auch im Datenstrukturfile enthalten sein. Aus Geheimhaltungsgründen kann das Teilnahmemuster je Beobachtungseinheit aber auch anonymisiert oder synthetisiert werden.
4. Die Größenordnung der Durchschnitte (arithmetische Mittel und Mediane) in den einzelnen Kategorien ist ungefähr abzubilden. Möglicherweise müssen hierfür Abweichungsschwellen gesetzt werden.
5. Die Korrelationen sollten annähernd erhalten werden. Insbesondere sollten die Vorzeichen bei signifikanten Korrelationen in der Regel repliziert werden.
6. Schätzergebnisse auf der Basis der Datenstrukturfiles sollten die gleiche Tendenz aufweisen wie Schätzungen mit den Originaldaten.
    a. Gleiche Einflussfaktoren sollen als signifikant und insignifikant geschätzt werden.
    b. Es sollten keine Vorzeichenwechsel bei signifikanten Einflüssen auftreten.

3 Anforderungen an Verfügbarkeit und Bedienbarkeit von Datenstrukturfiles

Um eine hohe Akzeptanz bei den Wissenschaftlern zu erreichen und vor allem neuen Nutzern den Einstieg zu eigenen Analysen mit Daten der FDZ zu erleichtern, soll der Gebrauch von Datenstrukturfiles einfach und selbsterklärend sein. Datenstrukturfiles sollen den Ablauf des kontrollierten Fernrechnens optimieren und für beide Seiten beschleunigen.

Datenstrukturfiles müssen damit spätestens nach Vertragsabschluss zwischen Datennutzer und FDZ zum Testen und Entwickeln der Auswertungsprogramme zur Verfügung stehen. Wenn die Übergabe des Datenstrukturfiles erst nach Abschluss des Vertrages erfolgt, könnten auch Datenstrukturfiles übergeben werden, die nur faktisch anonym sind.

Es hätte jedoch auch Vorteile, wenn Datenstrukturfiles bereits vor einem Vertragsabschluss zum Test von Auswertungsideen sowie zum Kennenlernen von Datensätzen zur Verfügung stehen würden. Beispielsweise könnten Datenstrukturfiles im Internet zum Download angeboten werden. Dann könnten Wissenschaftler bereits bevor sie einen Nutzungsantrag stellen prüfen, ob ihre Forschungsfrage mit einem bestimmten Datensatz überhaupt sinnvoll zu bearbeiten ist, welche Merkmale sie ggf. benötigen und bereits beginnen ihre Programme zu spezifizieren. Wenn der Zugang zum Datenstrukturfile ohne die bürokratischen Hürden des Nutzungsantrags und des Nutzungsvertrags erfolgen könnte, würde dies für die FDZ eine Arbeitserleichterung bedeuten. Diese Form der Bereitstellung von Datenstruktufiles hätte zudem den Vorteil, dass auch ausländische Wissenschaftler oder sich im Ausland aufhaltende Wissenschaftler auf die Datensätze im Internet zugreifen könnten. Online downloadbare Datenstrukturfiles würden damit einen Schritt in Richtung Internationalisierung des Datenangebots darstellen. Allerdings wäre es dann erforderlich, dass die Datenstrukturfiles die Kriterien der absoluten Anonymität erfüllen. Nach dem Bundestatistikgesetz (BStatG) haben nur Wissenschaftler das Privileg, Zugang zu faktisch anonymen Daten zu erhalten. Sobald Datensätze im Internet verfügbar sind, kann potenziell jeder darauf zu greifen. Damit müssen allerdings auch die strengeren Maßstäbe der absoluten Anonymität zwingend erfüllt werden.

4 Schlussfolgerungen

Das Ziel (1), die Durchführbarkeit der gleichen Analysen sicherzustellen, ist bei der Erstellung von Datenstrukturfiles in jedem Fall einzuhalten. Ein Datenstrukturfile kann als solches nur den oben beschriebenen Nutzen für die Wissenschaft und damit auch für die FDZ erfüllen, wenn alle Analysen, die mit den Originaldaten möglich sind, auch mit den anonymisierten Daten durchgeführt werden können und somit alle Auswertungsprogramme, die mit dem Datenstrukturfile fehlerfrei funktionieren auch mit dem Originaldatensatz fehlerfrei laufen – und umgekehrt.

Dem gegenüber ergibt sich bei Ziel (2) – dem weitestgehenden Erhalt der originalen Analyseergebnisse – zwangsläufig ein Interessenskonflikt zwischen der Sicherstellung der Anonymität einerseits und der bestmöglichen Erhaltung der Analyseergebnisse andererseits, der letztlich im Rahmen eines Kompromisses gelöst werden muss. Es sollte derjenige Weg der Anonymisierung beschritten werden, der gleichzeitig die georderte Anonymität sicherstellt und dabei das Analysepotenzial bestmöglich erhält.

Damit stellt sich die Frage, ob ein Datenstrukturfile in absolut oder faktisch anonymisierter Form bereitgestellt werden sollte. Wie in Abschnitt 3 ausgeführt, hätte die Bereitstellung von Datenstrukturfiles als für jedermann zugängliche absolut anonymisierte Datensätze viele Vorteile für die potenziellen Nutzer und die FDZ. Allerdings sind die Anforderungen der absoluten Anonymität an die Datensicherheit weitaus höher als die der faktischen Anonymität, so dass weitaus stärkere Anonymisierungsverfahren zum Einsatz kommen müssen, was zu einer stärkeren Veränderung von Analyseergebnissen führen und damit eine Beeinträchtigung der Kriterien für das Ziel (2) darstellen würde. Gleichzeitig sind jedoch auch die Anforderungen an das Analysepotenzial von Datenstrukturfiles im Rahmen des Ziels (2) weniger anspruchsvoll als diejenigen an echte Scientific Use Files, bei denen das Ziel die möglichst echte Replikation der originalen Ergebnisse für wissenschaftliche Publikationen ist. Aus diesen Gründen wird im weiteren Verlauf des Projekts “infinitE“ das ehrgeizige Ziel verfolgt, absolut anonymisierte Datenstrukturfiles zu erzeugen, die auch hinsichtlich der Zielkategorie (2) ein akzeptables Analysepotenzial aufweisen. Sollte dieses Ziel nicht gelingen, so muss auf das Konzept von faktisch anonymisierten Datenstrukturfiles ausgewichen werden.

Die Datennutzer sollen mit Hilfe von Datenstrukturfiles in die Lage versetzt werden, Analyseprogramme und Modellspezifikationen zu testen. Dabei sollen die Analyseprogramme exakt denjenigen entsprechen, die später auf die Echtdaten angewendet werden. Zusätzliche Programmcodes, mit deren Hilfe durch Anonymisierungsverfahren verursachte Verzerrungen korrigiert werden (Korrekturverfahren wie z.B. in Büttner und Rässler (2008, S.14) oder Krug (2010, S.32) für die multiple Imputation oder in Biewen und Ronning (2008) oder Ronning et al. (2010) für die stochastische Überlagerung dargestellt), sollten bei Datenstrukturfiles daher nicht erforderlich sein, um die Spezifikation einer Analyse zu testen. Deshalb werden nur Anonymisierungsverfahren ohne ihre eventuell existierenden Korrekturverfahren bei der Erstellung von Datenstrukturfiles betrachtet. Dabei werden solche Verfahren gesucht, bei denen Ergebnisse zwar ggf. verzerrt werden, sich die Verzerrung aber in einem für die Kriterien der Datenstrukturfiles vertretbaren Rahmen bewegt.

Im Hinblick auf die Zielkategorie (2) können die Teilziele 1 bis 3 dadurch berücksichtigt werden, dass diese unmittelbar als Vorgaben in der Konzeption der Anonymisierungsverfahren berücksichtigt werden. Dem gegenüber muss die möglichst weitgehende Einhaltung der Kriterien 4 bis 6 anhand

von empirischen Analysen und Verfahrensvergleichen untersucht werden. Hierzu kann auf Ergebnisse der Forschungsprojekte „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten“ (Ronning et al. 2005) und „Wirtschaftsstatistische Paneldaten und faktische Anonymisierung“ (Höhne 2008, Brandt et al. 2008, Höhne 2010) zurückgegriffen werden.

Bei Gastwissenschaftlerarbeitsplätzen wurde die Erfahrung gemacht, dass wissenschaftliche Analysen meist einen Schwerpunkt legen und die Daten in nur einer Dimension tiefer gegliedert benötigt werden. Analysen werden beispielsweise für disaggregierte Wirtschaftszweige oder Regionalräume durchgeführt. Bei Wirtschaftsstatistiken ist jedoch bereits die Sicherstellung faktischer Anonymität bei Erhaltung tiefer Gliederungen bei allen kategorialen Variablen schwierig (insbesondere Branchen- und Regionalkennung). Da jedoch die im Rahmen der Analysen tiefer gegliederte Dimension je nach Datennutzer und Forschungsfrage unterschiedlich sein kann, allerdings für einen Datenbestand – zumindest nach dem Konzept der absoluten Anonymität – ein einheitlicher Datenstrukturfile erzeugt werden muss, ist die Herausforderung bei der Anonymisierung der kategorialen Merkmale wirtschaftsstatistischer Daten besonders anspruchsvoll.

Referenzen

Allmendinger, Jutta und Annette Kohlmann (2005): Datenverfügbarkeit und Datenzugang am Forschungsdatenzentrum der Bundesagentur für Arbeit im Institut für Arbeitsmarkt- und Berufsforschung. Allgemeines Statistisches Archiv 88, S. 159-182.

Biewen, Elena und Gerd Ronning (2008): Estimation of Linear Models with Anonymised Panel Data. AStA - Advances in Statistical Analysis, 92(4), S. 423-438.

Brandt, Maurice und Markus Zwick (2009): infinitE – Eine informationelle Infrastruktur für das E-Science Age; Verbesserung des Mikrodatenzugangs durch „Remote-Access“. Wirtschaft und Statistik 7/2009, Statistisches Bundesamt, Wiesbaden.

Brandt, Maurice, Stefan Dittrich und Michael Konold (2008): Wirtschaftsstatistische Längsschnittdaten für die Wissenschaft. Wirtschaft und Statistik 3/2008, Statistisches Bundesamt, S. 217-224.

Büttner, Thomas und Susanne Rässler (2008): Multiple Imputation of Right-Censored Wages in the German IAB Employment Sample Considering Heteroscedasticity. IAB Discussion Paper 44/2008.

Höhne, Jörg (2010): Verfahren zur Anonymisierung von Einzeldaten. Statistik und Wissenschaft, Band 16. Statistisches Bundesamt, Wiesbaden.

Höhne, Jörg (2008): Anonymisierungsverfahren für Paneldaten, AStA Wirtschafts- und Sozialstatistisches Archiv, 2(3), S. 259-275.

Gürke, Christopher, Anja Gruhl und Tanja Hethey-Maier (2011): Verknüpfung von Unternehmensdaten verschiedener Datenproduzenten – Das Projekt „Kombinierte Firmendaten für Deutschland“. Wirtschaft und Statistik 2/2011, Statistisches Bundesamt, Wiesbaden, S. 91-97.

Jacobebbinghaus, Peter; Dana Müller und Agnes Orban (2010): How to use data swapping to create useful dummy data for panel datasets. FDZ Methodenreport 03/2010. Bundesagentur für Arbeit, Nürnberg.

Krug, Gerhard (2010): Fehlende Daten bei der Verknüpfung von Prozess- und Befragungsdaten - Ein empirischer Vergleich ausgewählter Missing Data Verfahren. Methoden - Daten - Analysen 2010, Jg. 4, Heft 1, S. 27-57.

Malchin, Anja und Ramona, Voshage (2009): Official Firm Data for Germany, Schmollers Jahrbuch / Journal of Applied Social Science Studies 129 (2009), S. 501-513, Berlin.

Malchin, Anja, Ramona Voshage und Joachim Wagner (Hrsg. 2011): Empirical Studies with New German Firm Level Data from Official Statistics. Jahrbücher für Nationalökonomie und Statistik. Band 231 (3), Lucius & Lucius, Stuttgart.

Rehfeld, Uwe G. (2009): Das Forschungsdatenzentrum der Rentenversicherung in stetiger Fortentwicklung, DRV-Schriften Band 55/2009, S. 17-26.

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

Ronning, Gerd, Martin Rosemann und Elena Biewen (2010): IV-Schätzung eines linearen Panelmodells mit anonymisierten Betrieb- und Unternehmensdaten. Schmollers Jahrbuch - Journal of Applied Social Science Studies, Jg. 130, S. 357-380.

Stegmann, Michael (2009): Das aktuelle Datenangebot und Neuentwicklungen im FDZ-RV, DRV-Schriften Band 55/2009, S. 27-36.

Zühlke, Sylvia, Markus Zwick, Sebastian Scharnhorst and Tim Wende (2004): The Research Data Centres of the Federal Statistical Office and the Statistical Offices of the Länder. Schmollers Jahrbuch - Journal of Applied Social Science Studies 124, S. 567 – 578.

Zühlke, Sylvia, Helga Christians und Katharina Cramer (2007): Das Forschungsdatenzentrum der Statistischen Landesämter – eine Serviceeinrichtung für die Wissenschaft. AStA Wirtschafts- und Sozialstatistisches Archiv 3-4, S. 169-178.

Bisher sind in der Reihe folgende FDZ-Arbeitspapiere erschienen:

Arbeitspapier Nr. 39: Improvement of data access – The long way to remote data access in Germany, M. Brandt/ M. Zwick, Juni 2011

Arbeitspapier Nr. 38: Decentralised Access to Confidential Microdata in Europe, M. Brandt/ P. Eilsberger/ M. Zwick, Juni 2011

Arbeitspapier Nr. 37: Masking Micro Data with Stochastic Noise, J. Höhne/ J. Höninger, Mai 2011

Arbeitspapier Nr. 36: Enthüllungsrisiko beim Remote Access: Die Schwerpunkteigenschaft der Regressionsgerade, A. Vogel, April 2011

Arbeitspapier Nr. 35: Temporary agency work and firm performance, S. Nielen/ A. Schiersch, April 2011

Arbeitspapier Nr. 34: Harmonisation of statistical confidentiality in the Federal Republic of Germany, M. Brandt/ A. Crößmann/ C. Gürke, März 2011

Arbeitspapier Nr. 33: Remote Access. Eine Welt ohne Mikrodaten ??, G. Ronning/ P. Bleninger/ J. Drechsler/ C. Gürke, Februar 2011

Arbeitspapier Nr. 32: Compiling a Harmonized Database from Germany´s 1978 to 2003 Sample Surveys of Income and Expenditure. T. Bönke/ C. Schröder/ C. Werdt, Mai 2010

Arbeitspapier Nr. 31: The Research Potential of New Types of Enterprise Data based on Surveys from Official Statistics in Germany., J. Wagner, Oktober 2009

Arbeitspapier Nr. 30: Geschlechterspezifische Einkommensunterschiede bei Selbstständigen im Vergleich zu abhängig Beschäftigten - Ein empirischer Vergleich auf der Grundlage steuerstatistischer Mikrodaten, P. Eilsberger/ M. Zwick, Januar 2008

Arbeitspapier Nr. 29: Reichtum in Niedersachsen und anderen Bundesländern -Ergebnisse der Steuergeschäftsstatistik 2003 für Selbstständige (Freie Berufe und Unternehmer) und abhängig Beschäftigte, P. Böhm/J. Merz, November 2008

Arbeitspapier Nr. 28: Exports and Productivity in the German Business Services Sector. First Evidence from the Turnover Tax Statistics Panel, A. Vogel, Juli 2009

Arbeitspapier Nr. 27: Künstler in den Daten der amtlichen Statistik, C. Haak, August 2008

Arbeitspapier Nr. 26: Union Density and Varieties of Coverage: The Anatomy of Union Wage Effects in Germany, B. Fitzenberger/ K. Kohn/ A. C. Lembcke, August 2008

Arbeitspapier Nr. 25: German engineering firms during the 1990's. How efficient are export champions?, A. Schiersch, Juli 2008

Arbeitspapier Nr. 24: Zum Einkommensreichtum Älterer in Deutschland – Neue Reichtumskennzahlen und Ergebnisse aus der Lohn- und Einkommensteuerstatistik (FAST 2001), P. Böhm/ J. Merz, Februar 2008

Arbeitspapier Nr. 23: Neue Datenangebote in den Forschungsdatenzentren. Betriebs- und Unternehmensdaten im Längsschnitt, M. Brandt/ D. Oberschachtsiek/ R. Pohl, November 2007

Arbeitspapier Nr. 22: Stichprobendaten von Versicherten der gesetzlichen Krankenversicherung - Grundlage und Struktur des Datenmaterials, P. Lugert, Dezember 2007

Arbeitspapier Nr. 21: KombiFid - Kombinierte Firmendaten für Deutschland, S. Bender/ J. Wagner/ M. Zwick, November 2007

Arbeitspapier Nr. 20: Neue Möglichkeiten zur Nutzung vertraulicher amtlicher Personen- und Firmendaten, U. Kaiser/ J. Wagner, Juni 2007

Arbeitspapier Nr. 18: Die Gehalts- und Lohnstrukturerhebung: Methodik, Datenzugang und Forschungspotential, H.-P. Hafner/ R. Lenz, Mai 2007

Arbeitspapier Nr. 17: Anonymisation of Linked Employer Employee Datasets. Theoretical Thoughts and an Application to the German Structure of Earnings Survey, H.-P. Hafner/ R. Lenz, Dezember 2006

Arbeitspapier Nr. 16: Die europäische Union - Integration von unten oder Eliteprojekt? Eine Sekundäranalyse von Mikrodaten der amtlichen Statistik, R. Nauenburg, November 2006

Arbeitspapier Nr. 15: Keeping in Touch - A Benefit of Public Holidays Using German Time Use diary Data, J. Merz/ L. Osberg, November 2006

Arbeitspapier Nr. 14: Zur Konzeption eines Taxpayer-Panels für Deutschland, D. Vorgrimler/ C. Gräb/ S. Kriete-Dodds, November 2006

Arbeitspapier Nr. 13: Anonymisierte Daten der amtlichen Steuerstatistik, D. Vorgrimler, September 2006

Arbeitspapier Nr. 12: Mikrosimulation in der Betriebswirtschaftlichen Steuerlehre, R. Maiterth, August 2006

Arbeitspapier Nr. 11: Der Anteil der freien Berufe und der Gewerbetreibenden an der Gemeindefinanzierung, M. Zwick, September 2006

Arbeitspapier Nr. 10: Konstruktion und Bewertung eines ökonomischen Einkommens aus der Faktisch Anonymisierten Lohn- und Einkommensteuerstatistik, T. Bönke/ F. Neher/ C. Schröder, August 2006

Arbeitspapier Nr. 9: Anonymising business micro data - results of a German project, R. Lenz/ M. Rosemann/ D. Vorgrimler/ R. Sturm, Juni 2006

Arbeitspapier Nr. 8: Scientific analyses using the Continuing Vocational Training Survey 2000, R. Lenz/H.-P. Hafner/ D. Schmidt, Juni 2006

Arbeitspapier Nr. 7: A standard for the release of microdata, R. Lenz/ D. Vorgrimler/ M. Scheffler, Juni 2006

Arbeitspapier Nr. 6: Measuring the disclosure protection of micro aggregated business microdata, R. Lenz, Juni 2006

Arbeitspapier Nr. 5: De facto anonymised microdata file on income tax statistics 1998, J. Merz/ D. Vorgrimler/ M. Zwick, Oktober 2005

Arbeitspapier Nr. 4: Matching German turnover tax statistics, R. Lenz/ D. Vorgrimler, Juni 2005

Arbeitspapier Nr. 3: The research data centres of the Federal Statistical Office and the statistical offices of the Länder,
S. Zühlke/ M. Zwick/ S. Scharnhorst/ T. Wende, März 2005

Arbeitspapier Nr. 2: Eine kommunale Einkommen- und Körperschaftsteuer als Alternative zur deutschen Gewerbesteuer: Eine empirische Analyse für ausgewählte Gemeinden, R. Maiterth/ M. Zwick, April 2005

Arbeitspapier Nr. 1: Ein Vergleich der Ergebnisse von Mikrosimulationen mit denen von Gruppensimulationen auf Basis der Einkommensteuerstatistik, H. Müller, März 2005

Statistische Ämter des Bundes und der Länder,
FDZ-Arbeitspapier Nr. 40, Definition von nutzerseitigen Kriterien für Datenstrukturfiles, 2011