Statistisches Bundesamt, IB-FDZ, Brandt
Hessisches Statistisches Landesamt, Hafner

Wiesbaden, den 14.02.2007

Leitfaden zur Anonymisierung für die Erstellung eines Campus-Files aus den Einzeldaten der zweiten europäischen Erhebung zur beruflichen Weiterbildung (CVTS 2)

1. Vorbemerkungen

Im Jahr 1987 wurde mit § 16 Abs. 6 des Bundesstatistikgesetzes[1] 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 großem Aufwand reidentifiziert werden können. „Unverhältnismäßig" bedeutet hier, dass der Aufwand einer Reidentifikation deren Nutzen übersteigt (faktische Anonymität). Die Deanonymisierung von Einzelangaben in einem faktisch anonymen Datensatz kann nicht mit absoluter Sicherheit ausgeschlossen werden. Die Schutzmaßnahmen für einen im Rahmen der Wissenschaft und Lehre frei zugänglichen Datensatz, dem so genannten Campus-File, müssen weitaus höher angesetzt werden. In diesem Fall ist es nur zulässig, absolut anonymisierte Daten weiterzugeben, d.h. die eindeutige Identifikation von Fällen ist ausgeschlossen. Der vorliegende Leitfaden behandelt absolut anonymisierte Datensätze für die Wissenschaft und Lehre, generiert aus den Daten der CVTS 2-Erhebung aus dem Jahre 2000 mit Berichtsjahr 1999.

2. Basismaterial

In der Erhebung liegen Angaben von 3184 Unternehmen zur Teilnahme von Beschäftigten an Maßnahmen zur beruflichen Weiterbildung im Jahr 1999 vor. Besonders im Dienstleistungsbereich ist der Anteil der Unternehmen, die geantwortet haben, sehr niedrig; für manche Schichten liegt er im Verhältnis zur Grundgesamtheit bei unter 1 : 600. In anderen Wirtschaftsbereichen war zwar für manche Schichten eine Vollerfassung geplant; da die Teilnahme an der CVTS freiwillig ist, wurde diese aber nicht erreicht. Ein Auswahlsatz von über 50% wird nur in manchen Schichten der Bereiche *Mit dem Kredit- und Versicherungsgewerbe verbundene Tätigkeiten* und *Bergbau und Gewinnung von Steinen und Erden* erreicht. Somit besteht schon ein gewisser Schutz für den Datensatz darin, dass ein potentieller Angreifer keine Gewissheit hat, ob das von ihm gesuchte Unternehmen tatsächlich an der Erhebung teilgenommen hat.

Eine vollständige Liste der im Campus-File enthaltenen Merkmale findet sich im Anhang.

3. Anonymisierungsmaßnahmen

Regionale Gliederung

Es wird keine Regionalinformation weitergegeben. Regionale Angaben sind besonders geeignet für Reidentifikationen. Ein Erhalt solcher Merkmale in einem Campus-File stellt daher für die Anonymisierung ein schwieriges Unterfangen dar. Hinzu kommt, dass selbst bei einer Reduktion der

Regionalinformation auf Ost/West aufgrund der geringen Fallzahl in den neuen Bundesländern (585 Unternehmen) für die meisten Fragestellungen keine belastbaren Ergebnisse getrennt nach alten und neuen Bundesländern zu erzielen wären, so dass der Wegfall des Regionalmerkmals keine wesentliche Einschränkung des Analysepotentials bedeutet.

Wirtschaftszweigklassifikation

Ausgangspunkt sind die 30 Wirtschaftsbereiche, nach denen die Stichprobenauswahl erfolgte.

Auf Basis der Hochrechnungsfaktoren für ganz Deutschland haben sich einige Wirtschaftsbereiche als besonders gefährdet herausgestellt. Diese mussten daher mit anderen Wirtschaftsbereichen zusammengelegt werden. Um eine absolute Anonymisierung zu erreichen, wurden die Wirtschaftszweige stark vergröbert.
Die NACE-Kategorien mit den 30 Wirtschaftsbereichen wurden zu 9 Kategorien zusammengefasst:

Für die zusammengelegten Wirtschaftsbereiche wurden die Hochrechnungsfaktoren für jede Beschäftigtengrößenklasse neu berechnet als Quotient aus der Anzahl der Unternehmen des zusammengefassten Bereichs in der Grundgesamtheit und der Anzahl der Unternehmen des zusammengefassten Bereichs in der Erhebung.

Um die Unsicherheit eines potentiellen Datenangreifers, ob das Unternehmen an der Befragung teilgenommen hat, weiter zu erhöhen, wurde eine 80% –Stichprobe, geschichtet nach dem Wirtschaftszweig (9 Kategorien) und der Beschäftigtengrößenklasse (6 Kategorien, gebildet aus den Beschäftigtenangaben für das Jahr 1999), gezogen. Damit besteht der Campus-File aus einem Datensatz mit 2569 Unternehmen.

Eine Übersicht der in der anonymisierten Datei zusammengelegten Wirtschaftsbereiche samt Fallzahlen ist in nachfolgender Tabelle gegeben:

1 Gesetz über die Statistik für Bundeszwecke (Bundesstatistikgesetz – BStatG) vom 22. Januar 1987 (BGBl. I S. 462, 565), zuletzt geändert durch Artikel 2 des Gesetzes vom 9. Juni 2005 (BGBl. I S. 1534).

Tabelle 1: Verteilung der CVTS 2-Unternehmen auf Wirtschaftsbereiche

	Wirtschaftsgliederung	NACE 30-Angabe	Anzahl	Grundgesamtheit
1	Ernährungsgewerbe u. Tabakverarbeitung; Textil- u. Bekleidungsgewerbe; Ledergew.	02 03	304	20435
2	Verlags- u. Druckgewerbe;Vervielfältigung von besp. Ton-, Bild- und Datenträgern; H. v. Gummi- u. Kunststoffwaren; Glasgewerbe, Keramik, Ver. V. Steinen u. Erden; Metallerzg. U. –bearbeitung; H. v. Metallerzeugnissen; Maschinenbau; H. v. Büromasch., Dv-Gerät. U. einr.; Elekt.-technik; Feinmech. Und Optik; Fahrzeugbau; Bergbau, Gew. V. Steinen u. Erden; Kokerei, Mineralölver., H. u. Ver. V. Spalt- u. Brutstoffen, Chem. Industrie; Holzgew., H. v. Möbeln, Schmuck, Musikinstr., Sportger., Spielw. U. sonst. Erzeugnissen; Recycling; Papiergew.;	06 08 09 10 11 12 01 und 07 04 und 05	1046	65420
3	Energie- und Wasserversorgung	13	185	1363
4	Baugewerbe	14	164	61083
5	Handelsvermittlung und Großhandel; KFZ-handel; Instandhaltung u. Reparatur v. KFZ; Tankstellen; Einzelhandel (ohne Handel mit Kraftfahrzeugen und ohne Tankstellen); Reparatur von Gebrauchsgütern	16 15 und 17	271	69070
6	Gastgewerbe	18	103	16543
7	Landverkehr; Transport in Rohrfernleitungen; Schiff- u. Luftfahrt; Tätigk. Für den Verkehr; Verkehrsvermittlung; Nachrichtenübermittlung	19 und 20	165	16138
8	Kreditgewerbe; Versicherungsgewerbe; mit dem Kredit- u. Versicherungsgew. Verb. Tätigkeiten	21 22 und 23	202	3693
9	Grundstücks- und Wohnungswesen; Vermietung bewegl. Sachen; F&E; Datenverarb. Und Datenbanken; Rechts-, Steuer- u. Unternehmensber.; Markt- u. Meinungsforschung; Beteiligungsgesellschaften; Architektur- und Ingenieurbüros; Technische, physikalische und chemische Untersuchung; Werbung; Gewerbsmäßige Vermittlung und Überlassung von Arbeitskräften; Detekteien und Schutzdienste; Reinigung von Gebäuden, Inventar und Verkehrsmitteln; Erbringung von sonstigen Dienstleistungen überwiegend für Unternehmen; Kultur, Sport und Unterhaltung; Abwasser- u. Abfallbes., sonst.. Entsor-gung; Interessenvertretungen; kirchl. U. sonst. relig. Einr.; Erbringung v. sonst. Dienstleistungen	24 25 26 27 28 29 30	129	89189

Beschäftigte des Unternehmens

Die absolute Anzahl der Beschäftigten wird nicht ausgewiesen, sondern nur jeweils 6 Beschäftigtengrößenklassen für die Jahre 1998 und 1999 sowie die Anteile für männliche und weibliche Beschäftigte an den Gesamtbeschäftigten im Jahr 1999.
Die Beschäftigtengrößenklassen 5 und 6 sind für den Wirtschaftszweig „Gastgewerbe" zusammengelegt worden, da in der Größenklasse 5 nur sehr wenige Unternehmen vorhanden sind und daher ein Reidentifikationsrisiko bei einer getrennten Ausweisung dieser Größenklasse gegeben wäre.

Weitere Maßnahmen

Einige Merkmale, bei denen entweder die Qualität sehr schlecht war oder nur für sehr wenige Unternehmen Angaben vorlagen, werden aus dem Datensatz entfernt oder modifiziert. Im Einzelnen betrifft dies die folgenden Merkmale:

Anteil der indirekten Kosten: Fällt weg, da nur vereinzelt gesicherte Angaben vorliegen.

Zahl der ganz oder teilweise mit Lehrveranstaltungen beschäftigten Personen: Fällt weg wegen mangelhafter Qualität, die mitunter darauf zurückzuführen ist, dass die erfragte Information nicht in dieser Form in der Buchhaltung der Unternehmen vorlag.

Fonds, an die Beiträge gezahlt werden: Diese Merkmale fallen weg, da es bei regionalen Fonds nur 77 Ja-Antworten gibt, bei nationalen 7 und bei Sonstigen 47.

Einnahmen, Einnahmequellen, Saldo: Diese Merkmale fallen weg, da es im Datensatz nur 57 Unternehmen gibt, die Einnahmen aus Lehrveranstaltungen haben. Sollte ein potenzieller Datenangreifer Kenntnisse über diese Merkmale haben, so können in Kombination mit den anderen Überschneidungsmerkmalen (Wirtschaftsbereich und Beschäftigtenangabe) eindeutige Fälle entstehen. Es wird daher ein neues Merkmal berechnet, welches den Wert 1 annimmt, wenn ein Unternehmen Einnahmen aus Lehrveranstaltungen hat und 0 sonst.

Unterdrückung von Merkmalen

Bei den Merkmalen, die abhängig von der Anzahl der Beschäftigten sind (geleistete Arbeitsstunden, Personalaufwendungen, Kosten, Teilnahmestunden, Teilnehmer), werden die Originalwerte auf Pro-Kopf-Werte (Division durch Anzahl der Beschäftigten Ende 1999) bzw. auf Werte pro Teilnehmer umgerechnet.

Weitere entfernte Merkmale

Beschäftigte am 31.12.1999 Männer (A299m)
Beschäftigte am 31.12.1999 Frauen (A299f)
Beschäftigte am 31.12.1998 insgesamt (A298tot)
Beschäftigte am 31.12.1999 insgesamt (A299tot)
Saisonale Schwankungen der Beschäftigtenzahl (Saison)
Geleistete Arbeitsstunden (A3)

Personalaufwendungen in Euro (A4a)
Fusionen, Übernahmen oder Umstrukturierungen in 1999 (A5c)
Andere organisatorische Änderungen in 1999 (A5d)
Teilnahme an Lehrveranstaltungen – Männer (C2nm)
Teilnahme an Lehrveranstaltungen – Frauen (C2nf)
Teilnahme an Lehrveranstaltungen – Personen insgesamt (C2ntot)
Teilnahmestunden insgesamt (C2ttot)
Teilnahmestunden Männer (C2tm)
Teilnahmestunden Frauen (C2tf)
Kosten für Lehrveranstaltungen insgesamt in Euro (Kosten)
Einnahmen für Lehrveranstaltungen (C6g)

Informationen über geschätzte Merkmale wurden aus dem Datensatz entfernt

Zahlungen und Gebühren für die Teilnahme der Beschäftigten und externes Weiterbildungspersonal geschätzt (C6aest)
Reisekosten, Spesen, Tagegeld geschätzt (C6best)
Lohn- und Gehaltskosten für internes Weiterbildungspersonal, ausschließlich mit Lehrveranstaltungen beschäftigt geschätzt (C6cest)
Lohn- und Gehaltskosten für internes Weiterbildungspersonal, teilweise mit Lehrveranstaltungen beschäftigt geschätzt (C6dest)
Kosten für Räume und Ausstattung geschätzt (C6eest)
Beiträge an öffentliche und andere Einrichtungen geschätzt (C6fest)

4. Analyse des Gefährdungspotentials

Der Datensatz enthält keine sensiblen Daten wie z.B Angaben über Umsatz oder Steuern. Die erfragten Informationen sind i.d.R. nicht kommerziell verwertbar, da sie 7 Jahre alt sind und sich in diesem Zeitraum die Weiterbildungsphilosophie eines Unternehmens stark ändern kann, sei es durch interne Einflüsse wie etwa den Wechsel der Geschäftsführung oder durch externe Faktoren wie steigenden Konkurrenzdruck und dadurch notwendig werdende bessere Qualifizierung der Beschäftigten. Aufgrund der mangelnden Aktualität der Daten sind sie als Informationsquelle über einen Mitbewerber nur stark eingeschränkt nutzbar.

Trotz des eher zweifelhaften Nutzens einer Deanonymisierung sind insbesondere bei der Erstellung eines Campus-Files, der jedermann den Datenzugang außerhalb der geschützten Räume der amtlichen Statistik ermöglicht, weitere Schutzmaßnahmen erforderlich, um keinesfalls die Deanonymisierung eines Merkmalsträgers zu riskieren. Dies ist u. a. wichtig, um das Vertrauen der teilnehmenden Unternehmen in die amtliche Statistik zu erhalten und sie zu einer Teilnahme an zukünftigen Erhebungen motivieren zu können.

Zusatzwissen über Weiterbildungsmaßnahmen von Unternehmen liegt nicht in systematischer Form vor, so dass für die diesbezüglichen Merkmale höchstens Einzelangriffe denkbar sind. Dazu müsste man sehr genaue Kenntnis über ein Unternehmen besitzen und würde dann kaum mehr zusätzliche Informationen gewinnen.

Die Überschneidungsmerkmale, die einem potentiellen Datenangreifer für die CVTS2 aus kommerziellen Datenbanken zur Verfügung stehen könnten, sind der Wirtschaftszweig und die Beschäftigtengrößenklasse. Diese Merkmale wurden im Rahmen der Anonymisierung so kategorisiert, dass keine Rückschlüsse auf einzelne Merkmalsträger mehr möglich sind (Tabelle 2).

Tabelle 2: Unternehmen der CVTS 2 nach Wirtschaftszweig (WZ) und Beschäftigtengrößenklasse

	Beschäftigtengrößenklasse							
WZ	1	2	3	4	5	6	56	Summe
1	67	62	64	48	42	21	0	304
2	196	244	198	153	127	128	0	1046
3	28	40	38	36	15	28	0	185
4	37	39	42	19	16	11	0	164
5	60	64	45	41	27	34	0	271
6	18	34	24	16	0	0	11	103
7	58	57	20	18	6	6	0	165
8	20	38	30	34	28	52	0	202
9	45	26	20	18	9	11	0	129
Summe	529	604	481	383	270	291	11	2569

Eine eindeutige Reidentifikation von Einheiten im Campus-File ist nicht möglich. Damit ist die Weitergabe der Daten für die Nutzung in Wissenschaft und Lehre unbedenklich.