STATISTISCHE ÄMTER
DES BUNDES UND DER LÄNDER
FORSCHUNGSDATENZENTREN

Leitfaden zur Anonymisierung für die Erstellung eines Campus-Files aus den Einzeldaten der Dritten Europäischen Erhebung zur beruflichen Weiterbildung (CVTS 3)

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 3-Erhebung aus dem Jahre 2006 mit Berichtsjahr 2005.

2. Basismaterial

In der Erhebung liegen Angaben von 2857 Unternehmen zur Teilnahme von Beschäftigten an Maßnahmen zur beruflichen Weiterbildung im Jahr 2005 vor. Besonders im Dienstleistungsbereich und im Baugewerbe (Unternehmen mit bis unter 50 Beschäftigte) ist der Anteil der Unternehmen, die geantwortet haben, sehr niedrig; für manche Schichten liegt er im Verhältnis zur Grundgesamtheit bei unter 1:400. In anderen Wirtschaftsbereichen war zwar für einige Schichten eine Vollerfassung geplant; da die Teilnahme an der CVTS freiwillig ist, wurde diese in keiner Schicht erreicht. Eine Teilnahmequote von über einem Drittel wird nur in einer Schicht des Textil-, Bekleidungs- und Ledergewerbe, bei Unternehmen mit 500 und mehr Beschäftigten im Bergbau und Gewinnung von Steinen und Erden sowie für die oberste Beschäftigtengrößenklasse der Bereiche Holzgewerbe, Herstellung von Möbeln, Schmuck, Musikinstrumente, Sportgeräte, Spielwaren und sonstigen Erzeugnissen und Recycling; Papiergewerbe; Kultur, Sport, Unterhaltung; Architektur-, Ingenieurbüros; technische, physikalische, chemische Untersuchung und Werbung erreicht. Somit besteht schon ein gewisser

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).

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. Analyse des Gefährdungspotentials

Der Datensatz enthält keine besonders sensiblen Daten, die kommerziell verwertbar sind.

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 CVTS3 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 1).

Tabelle 1: Unternehmen der CVTS 3 nach Wirtschaftszweig (WZ) und Beschäftigtengrößenklasse

	Beschäftigtengrößenklasse							
WZ	1	2	3	4	5	6	56	Summe
1	61	71	59	34	27	16	0	268
2	153	172	148	131	88	96	0	788
3	17	19	21	12	10	16	0	95
4	28	15	24	15	20	10	0	112
5	56	62	66	49	36	36	0	305
6	21	37	29	23	0	0	18	128
7	51	48	45	32	18	21	0	215
8	28	24	35	24	15	20	0	146
9	51	55	25	26	32	59	0	248
Summe	466	503	452	346	246	274	18	2305

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.

4. 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 aufgrund der geringen Fallzahl in den neuen Bundesländern ohne Berlin (468 Unternehmen) für die meisten Fragestellungen keine belastbaren Ergebnisse getrennt nach 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 Teilnahmequoten für ganz Deutschland haben sich einige Wirtschaftsbereiche als besonders gefährdet herausgestellt. Diese müssen 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 2005), gezogen. Damit enthält der Campus-File Daten von 2305 Unternehmen.

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

Tabelle 1: Verteilung der CVTS 3-Unternehmen auf Wirtschaftsbereiche

	Wirtschaftsgliederung	NACE 30-Angabe	Anzahl	Grundgesamtheit
1	Ernährungsgewerbe u. Tabakverarbeitung; Textil- u. Bekleidungsgewerbe; Ledergew.	02 03	268	12993
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	788	53752
3	Energie- und Wasserversorgung	13	95	1568
4	Baugewerbe	14	112	33929
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	305	57220
6	Gastgewerbe	18	128	12051
7	Landverkehr; Transport in Rohrfernleitungen; Schiff- u. Luftfahrt; Tätigk. Für den Verkehr; Verkehrsvermittlung; Nachrichtenübermittlung	19 und 20	215	16377
8	Kreditgewerbe; Versicherungsgewerbe; mit dem Kredit- u. Versicherungsgew. Verb. Tätigkeiten	21 22 und 23	146	3790
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	248	59113

Beschäftigte des Unternehmens

Die absolute Anzahl der Beschäftigten wird nicht ausgewiesen, sondern nur jeweils 6 Beschäftigtengrößenklassen für die Jahre 2004 und 2005 sowie die Anteile für männliche und weibliche Beschäftigte an den Gesamtbeschäftigten im Jahr 2005.
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 nur für sehr wenige Unternehmen Angaben vorliegen, werden aus dem Datensatz entfernt. Im einzelnen betrifft dies die folgenden Merkmale:

Umlagen oder Beiträge an Fonds für Weiterbildung in Euro / Einnahmen aus Fonds oder sonstige Zuschüsse für Weiterbildung in Euro: Diese Merkmale fallen weg, da es hier nur 45 bzw. 30 Unternehmen mit entsprechenden Beiträgen bzw. Einnahmen gibt und deshalb die Gefahr der Reidentifikation eines Unternehmens besonders hoch ist. Aus diesem Grund werden auch die Gesamtkosten für Weiterbildung in Euro (c7tot) nicht ausgewiesen, sondern nur die Zwischensumme der direkten Kosten(c7sub) ohne Einbeziehung der Beiträge bzw. Einnahmen und die Personalausfallkosten.

Umlagen oder Beiträge an Fonds für Erstausbildung in Euro / Einnahmen aus Fonds oder sonstige Zuschüsse für Erstausbildung in Euro: Diese Merkmale fallen weg, da es hier nur 100 bzw. 109 Unternehmen mit entsprechenden Beiträgen bzw. Einnahmen gibt und deshalb die Gefahr der Reidentifikation eines Unternehmens besonders hoch ist. Aus diesem Grund werden auch die Gesamtkosten für Erstausbildung in Euro (f2tot) nicht ausgewiesen, sondern nur die Ausbildungsvergütungen (f2a) und die sonstigen Kosten (f2b) ohne Einbeziehung der Beiträge bzw. Einnahmen.

Die Merkmale C4tot und C7sflag wurden entfernt. Das Merkmal C4tot entspricht dem Merkmal C3tot und ist deshalb im Datenmaterial bereits enthalten. C7sflag ist lediglich ein Steuerungsmerkmal.

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 2005) bzw. auf Werte pro Teilnehmer umgerechnet.

Weitere entfernte Merkmale

Beschäftigte am 31.12.2005 Männer (a2m05)
Beschäftigte am 31.12.2005 Frauen (a2f05)
Beschäftigte am 31.12.2004 insgesamt (a2tot04)
Beschäftigte am 31.12.2005 insgesamt (a2tot05)
Index Beschäftigte 31.12.2004 im Vergleich zum 31.12.2005 (a2_idx)
Beschäftigte unter 25 Jahren am 31.12.2005 (a3a)
Beschäftigte 25 bis unter 55 Jahre am 31.12.2005 (a3b)
Beschäftigte 55 Jahre und älter am 31.12.2005 (a3c)
Geleistete Arbeitsstunden im Jahr 2005 (a4)
Geleistete Arbeitsstunden im Jahr 2005 – Männer (a4m)
Geleistete Arbeitsstunden im Jahr 2005 – Frauen (a4f)
Personalaufwendungen in Euro (A5)
Anzahl der Teilnehmer Weiterbildung am Arbeitsplatz (b2a)
Anzahl der Teilnehmer Jobrotation, Austauschprogramme, Abordnungen, Studienbesuche (b2b)
Anzahl der Teilnehmer Lern- und Qualitätszirkel (b2c)
Anzahl der Teilnehmer Selbstgesteuertes Lernen (b2d)
Anzahl der Teilnehmer Kongresse, Informationsveranstaltungen u. ä. (b2e)
Teilnehmer an Lehrveranstaltungen (c1tot)
Teilnehmer an Lehrveranstaltungen – Männer (c1m)
Teilnehmer an Lehrveranstaltungen - Frauen (c1f)
Teilnehmer an Lehrveranstaltungen – unter 25 Jahre (c2a)
Teilnehmer an Lehrveranstaltungen – 25 bis unter 55 Jahre (c2b)
Teilnehmer an Lehrveranstaltungen – 55 Jahre und älter (c2c)
Teilnahmestunden Lehrveranstaltungen Insgesamt (c3tot)
Teilnahmestunden interne Lehrveranstaltungen (c3i)
Teilnahmestunden externe Lehrveranstaltungen (c3e)
Teilnahmestunden Lehrveranstaltungen - Männer (c4m)
Teilnahmestunden Lehrveranstaltungen – Frauen (c4f)
Zahlungen und Gebühren für externe Lehrveranstaltungen und Kosten für externes Personal
in internen Lehrveranstaltungen in Euro (c7a)
Reisekosten, Spesen und Tagegeld in Euro (c7b)
Personalaufwendungen für internes Weiterbildungspersonal in Euro (c7c)
Kosten für Räume, Ausstattung und Unterrichtsmaterialien für Weiterbildung in Euro (c7d)
Kosten für Lehrveranstaltungen in Euro - Zwischensumme (c7sub)
Personalausfallkosten (pac)
Umlagen oder Beiträge an Fonds für Weiterbildungsprogramme (c8aflag)
Auszubildende im Lauf des Jahres 2005 (f1tot05)
Auszubildende im Lauf des Jahres 2005 – Männer (f1m05)
Auszubildende im Lauf des Jahres 2005 – Frauen (f1f05)
Ausbildungsvergütungen in Euro (f2a)
Sonstige Kosten für Erstausbildung in Euro (f2b)
Personalaufwendungen für Ausbildungspersonal in Euro (f2c)