
Statistisches Bundesamt
Vorgrimler / Sturm IA

Wiesbaden, den 15.09.2003

Daniel Vorgrimler / Roland Sturm

Konzeption des Public Use Files KSE-KMU 1999

Inhaltsverzeichnis

1	Allgemeines	1
2	Einschränkung des Datenmaterials	1
3	Anonymisierungsmaßnahmen	3
4	Schutzaspekte	3
5	Analyseaspekte	5
6	Fazit	7
7	Ansprechpartner	7

1 **Allgemeines**

Einzelangaben über persönliche und sachliche Verhältnisse, die für eine Bundesstatistik gemacht werden, sind von den statistischen Ämtern grundsätzlich geheim zu halten. Das Bundesstatistikgesetz (BStatG) erlaubt die Weitergabe von Einzeldaten nur in bestimmten Fällen, z.B. dann, wenn der Datennutzer keine Möglichkeit hat, einzelne Datensätze zu re-identifizieren, d.h. die Angaben den Befragten oder Betroffenen zuzuordnen. Eine Lockerung dieser Vorgabe gilt für die Weitergabe von Einzeldaten für Forschungszwecke an die Wissenschaft (in § 16 (6) des BStatG geregelt). Hier ist die Weitergabe erlaubt, wenn der Wissenschaftler einen Datensatz nur mit unverhältnismäßig hohem Aufwand re-identifizieren kann. Für die aus den Regelungen resultierenden zwei Formen der weitergabefähigen Daten sind die Bezeichnungen Public-Use-File (für die Allgemeinheit) und Scientific-Use-File (für die Wissenschaft) gebräuchlich.

Auf dieser CD werden Mikrodaten der Kostenstrukturerhebung im Bergbau und Verarbeitenden Gewerbe (KSE) der Allgemeinheit zur Verfügung gestellt. Es handelt es sich um einen Public-Use-File, dessen Nutzung z.B. in der wissenschaftlichen Lehre gesehen werden kann. Da der Kreis der Nutzer nicht eingeschränkt wird, müssen an die Daten wesentlich härtere Anonymisierungsbedingungen gestellt werden als an einen Scientific-Use-File. Die Maßnahmen werden im Folgenden beschrieben. Die Datennutzer können daraus ersehen, welche Möglichkeiten und die Grenzen für die Verwendung des Mikrodatensatzes bestehen. Durch die Beschreibung der Anonymisierungen wird auch deutlich, dass das Statistische Bundesamt die Vertraulichkeit der Informationen gewährleistet.

Die Erstellung des Public Use Files kleiner und mittlerer Unternehmen(KMU) der KSE erfolgte auf der Grundlage von Erkenntnissen des Projektes "Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten" das die statistischen Ämter gemeinsam mit der Wissenschaft durchführen.

2 **Einschränkung des Datenmaterials**

• **Beschränkung auf Unternehmen mit max. 250 Beschäftigten**

Bisherige Re-Identifikationsversuche an formal anonymisierten Dateien¹ haben gezeigt, dass Unternehmen mit weniger als 250 Beschäftigten nur sehr schwer richtig zu re-identifizieren sind.² Daher sind Unternehmen dieser Größe grundsätzlich für einen Public-Use-File geeignet. Die Mög-

1 Formale Anonymisierung bedeutet die Abtrennung von direkten Identifikatoren wie dem Unternehmensnamen und der Anschrift.

2 Zu den Möglichkeiten von Re-Identifikationen bei der KSE vgl. Vorgrimler, D.: Re-Identifikationsmöglichkeiten am Beispiel eines konkreten Angriffsszenarios, in: Ronning, G., Gnos, R.: Anonymisierung wirtschaftsstatistischer Einzeldaten, 2003.

lichkeit einer Re-Identifikation steigen mit der Unternehmensgröße; daher wurden Unternehmen mit mehr als 250 Beschäftigten nicht für diesen Public-Use-File verwendet.³

- **Verzicht auf WZ-Abteilungen, die der Geheimhaltung unterliegen**

In den Veröffentlichungen der statistischen Ämter werden aufgrund von Geheimhaltungsaspekten die Ergebnisse einiger Abteilungen der Klassifikation der Wirtschaftszweige, Ausgabe 1993 (WZ93) nicht veröffentlicht. Unternehmen der KSE, die in diesen Abteilungen klassifiziert sind, sind für den Public-Use-File nicht berücksichtigt worden. Es handelt sich dabei um Unternehmen der Abteilungen 10, 11, 14, 16, 23, 30, 32, 35 und 37 der WZ93.

- **Ersatz der administrativen Gebietsschlüssel durch den siedlungsstrukturellen Kreistyp BBR9**

Der in der originalen Kostenstrukturerhebung enthaltene administrative Gebietsschlüssel differenziert die Unternehmen so stark, dass es nicht möglich erscheint, absolut anonyme Daten zu erstellen, in denen dieses Merkmal enthalten ist. Die Auswertung nach Regionen stellt jedoch einen wichtigen Analysebereich dar. Um diesen nicht zu verlieren, wurde der administrative Gebietsschlüssel durch den siedlungsstrukturellen Kreistyp BBR9 ersetzt. Dieser dient ebenfalls dem Vergleich zwischen unterschiedlichen Regionen. Es wird aber nicht nach der administrativen Lage unterschieden, sondern die Regionen werden nach ihrer Siedlungsstruktur bewertet und in 9 Klassen eingeteilt. Es wird dabei nach "Kernstädten" und sonstigen Kreisen bzw. Kreisregionen unterschieden. Als Kernstädte werden kreisfreie Städte mit mehr als 100.000 Einwohnern ausgewiesen. Kreisfreie Städte unterhalb dieser Größe werden mit ihrem Umland zu Kreisregionen zusammengefasst. Die Typisierung der Kreise und Kreisregionen erfolgt außerhalb der Kernstädte nach der Bevölkerungsdichte. Um den großräumigen Kontext zu berücksichtigen, wird dann weiter nach der Lage im siedlungsstrukturellen Regionstyp differenziert. Mit dieser Einordnung wird der Überlegung Rechnung getragen, dass die Lebensbedingungen in den Kreisen sowie ihre Entwicklung wesentlich auch von der Entwicklung und der Struktur der jeweiligen Region bzw. des Regionstyps abhängig ist.

Durch den Ersatz des administrativen Gebietsschlüssels durch den BBR9 wird einerseits das Re-Identifikationsrisiko erheblich reduziert und andererseits weiterhin die Möglichkeit geboten, mit den Daten sinnvolle regionsbezogene Analysen durchzuführen.

- **Streichen von Unternehmen, die in der Grundgesamtheit einmalige Ausprägungskombinationen aufweisen**

Unter den rund 13.000 Unternehmen der KSE mit maximal 250 Beschäftigten haben neun eine einmalige Ausprägungskombination in den diskreten Merkmalen BBR9 und zweistelliger WZ93. Diese Unternehmen sind für den Public Use File nicht berücksichtigt worden, so dass durch die Kombination der beiden Merkmale keine einmaligen Fälle mehr identifiziert werden können.

- **Beschränkung auf 500 Unternehmen**

Von den rund 13.000 möglichen Unternehmen sind nur 500 durch Zufallsziehung ausgewählte in den Public-Use-File aufgenommen worden. Dies entspricht einem Auswahlsatz von 3,8%.

3 Die Möglichkeiten, Daten solcher Unternehmen für den beschränkten Nutzerkreis der Wissenschaft im Rahmen eines Scientific-Use-Files zugänglich zu machen, werden zur Zeit untersucht..

3 *Spezifische Anonymisierungsmaßnahmen für die 500 Datensätze*

Neben der Beschränkung der Grundgesamtheit und der Stichprobenziehung, die beide bereits Anonymisierungsmaßnahmen darstellen, wurden noch folgende Maßnahmen getroffen:⁴

- Die Wirtschaftsklassifikation (WZ93) wurde von vier bzw. fünf auf zwei Stellen gekürzt.
- Es wurde eine 10prozentige Fehlerüberlagerung des siedlungsstrukturellen Kreistyps mit dem Verfahren PRAM durchgeführt, wobei der Nachbarschaftsbereich auf zwei Ausprägungen beschränkt ist.⁵ Dadurch stimmt die veröffentlichte Ausprägung des BBR9 im Public-Use-File nur noch mit einer Wahrscheinlichkeit von 90% mit der Ausprägung in den Originaldaten überein.
- Jedes stetige Merkmal wurde jeweils einzeln mikroaggregiert (Mikro33g).⁶ Bei der Mikroaggregation werden immer mindestens drei Merkmalsträger (Unternehmen) gesucht, die in dem zu anonymisierenden Merkmal die größte Ähnlichkeit aufweisen. Aus den Merkmalsausprägungen der drei Merkmalsträger wird ein Durchschnittswert ermittelt. Die „wahren“ Werte werden in den veröffentlichten Daten mit diesen Durchschnitt ersetzt. Durch dieses Vorgehen ist jede Merkmalsausprägung bei mindestens drei Merkmalsträgern vorhanden, d.h. jede Ausprägung kommt in den Daten mindestens dreimal vor. Das Verfahren wurde auf die 500 in der Stichprobe enthaltenen Unternehmen angewandt, und zwar nachdem die Stichprobe gezogen war. Dies bewirkt einen höheren Schutz als eine Mikroaggregation vor der Stichprobenziehung, da die Stichprobendatei sehr viel kleiner ist und dadurch die Merkmale stärker verändert werden als bei einer Mikroaggregation, die sämtliche 13.000 in Frage kommenden Unternehmen beinhaltet.

4 *Prüfung der Schutzwirkung*

Gemäß dem in den statistischen Ämtern entwickelten Schutzwirkungskonzept von faktischer Anonymisierung⁷ kann ein Datensatz dann als sicher angesehen werden, wenn für einen Datenangreifer die Wahrscheinlichkeit einen brauchbaren Wert zu enthüllen eine zu definierende Schwelle unterschreitet. Da dieses Konzept für die faktische Anonymisierung entwickelt wurde, muss für den Public-Use-File eine strengere Regel angewandt werden. Diese Prüfung der Schutzwirkung wird im folgenden erläutert.

- **Möglichkeit der korrekten Zuordnung**

Im Rahmen der bisher im Forschungsprojekt durchgeführten Re-Identifikationsexperimenten konnte bei den Originaldaten unter den 13.000 Unternehmen mit weniger als 250 Beschäftigten im Rahmen von 25 Versuchen ein Unternehmen erfolgreich re-identifiziert werden. Hinzu kommen noch zwei weitere Unternehmen, die mit Hilfe kommerzieller Datenbanken re-identifiziert

4 zur Beschreibung von Anonymisierungsmaßnahmen vgl. Höhne, J.: Methoden zur Anonymisierung wirtschaftsstatistischer Einzeldaten, in: Ronning, G., Gnos, R.: Anonymisierung wirtschaftsstatistischer Einzeldaten, 2003.

5 Nachbarschaftsbereich ist der Bereich, innerhalb dessen die Ausprägungen getauscht werden können. So wird z.B. die Originalausprägung 6 mit einer Wahrscheinlichkeit von jeweils 2,5% durch die Ausprägung 4,5,7, oder 8 ersetzt. Mit einer Wahrscheinlichkeit von 90% wird die Originalausprägung beibehalten. Vgl. Höhne, J., 2003.

6 Zur Methode der Mikroaggregation vgl. Höhne, J., 2003.

7 Zum Schutzwirkungskonzept vgl. Höhne, J., Sturm, R., Vorgrimler, D., Konzept zur Schutzwirkung faktischer Anonymisierung, in: Wirtschaft und Statistik, Heft 4, 2003, Seite 287-292.

werden konnten. Die Wiederholung der Re-Identifikationsexperimenten bei den nach dem beschriebene Schema anonymisierten Daten führte zu dem Ergebnis, dass *kein* Unternehmen re-identifiziert werden konnte. So sind auch alle drei Unternehmen im Public-Use-File nicht mehr auffindbar.

Ebenfalls im Rahmen des Forschungsprojektes wurde ein erster Massenfischzug bei Originaldaten (ohne Anonymisierungsmaßnahmen) simuliert. Als Zusatzwissen diente die Datenbank der Creditreform. Von den 500 Unternehmen des Public-Use-File wurden dabei 488 falsch zugeordnet, lediglich 9 Unternehmen wurden durch das Programm richtig zugeordnet werden, 3 Zuordnungen konnten nicht überprüft werden. Das Risiko einer Falschzuordnung liegt demnach für einen Datenangreifer in dieser Simulation bei über 98%, da er die Richtigkeit einer Zuordnung nicht überprüfen kann. Bezieht man ein, dass es sich bei diesem Experiment um einen Massenfischzug beim nicht-anonymisierten Datenmaterial handelte und nicht um den hier vorliegenden Public-Use-File, dessen Merkmale wie beschrieben geschützt wurden, so kann ein erfolgreicher Massenfischzug ausgeschlossen werden.

- **Abweichungen der veröffentlichten Werte zu den wahren Werten⁸**

Aufgrund der durchgeführten Mikroaggregation entsprechen die ausgewiesenen Werte des Public-Use-File nicht den Werten der Originalerhebung. In der Tabelle 1 ist für jedes stetige Merkmal die Anzahl der Datensätze angegeben, die um verschiedene Schwellen von den wahren Werten abweichen.

Das angewandte Verfahren der Mikroaggregation erhält den Datensatz in einem Maße dass die Abweichungen in der Regel die Brauchbarkeit der Merkmalsausprägung nur bedingt in Frage stellt. Trotzdem ist mit der Mikroaggregation ein zusätzlicher Schutz dahingehend verbunden, dass der Versuch einen Datensatz zu re-identifizieren weiter erschwert wird und dass darüber hinaus das zusätzliche Risiko entsteht, Werte zu enthüllen, die nicht den wahren Werten entsprechen. Der Nutzen eines Missbrauchs wird dadurch unkalkulierbar klein.

8 Als wahrer Wert wird der in der originalen KSE-Erhebung enthaltene Wert verstanden

Tabelle 1: Abweichungsanalyse

Merkmal	Anzahl abweichender Datensätze	Anzahl der Datensätze mit der Stärke der Abweichung von					
		min. 1%	min. 5%	min. 10%	min. 25%	min. 50%	min. 100%
Tätige Inhaber	2	2	2	2	1	1	0
Angestellte/Arbeiter	139	67	0	0	0	0	0
Teilzeit	47	47	30	13	5	1	1
Teilzeit in Vollzeiteinheiten	243	242	211	154	72	17	16
Tätige Personen	134	70	0	0	0	0	0
Umsatz eigene Erzeugnisse	500	128	11	4	0	0	0
Umsatz Handel	201	144	50	23	9	3	1
Gesamtumsatz	500	115	17	5	0	0	0
Bestandsveränderung fertigen/unfertigen Erzeugnisse	441	323	88	36	14	4	3
Gesamtleistung	500	120	13	5	0	0	0
Anfangsbestand an gemessen am Umsatz	431	173	23	11	8	2	1
Endbestand an gemessen am Umsatz	440	167	31	15	9	5	1
Anfangsbestand an gemessen am Umsatz (Rohstoffe etc.)	461	117	27	12	6	3	0
Endbestand an gemessen am Umsatz (Rohstoffe etc.)	467	146	22	13	9	5	1
Rohstoffverbrauch	500	166	31	11	3	1	0
Energieverbrauch	493	145	21	7	2	0	0
Anfangsbestand an gemessen am Umsatz (Handel)	144	104	35	17	5	3	0
Endbestand an gemessen am Umsatz (Handel)	147	115	27	16	4	1	0
Einsatz an Handelsware	201	160	41	22	5	2	0
Bruttogehalt	500	66	8	3	0	0	0
Ges. Sozialkosten	500	68	6	3	0	0	0
sonst. Sozialkosten	362	163	23	9	3	1	0
Kosten Leiharbeiter	174	130	43	16	6	5	1
Kosten Lohnarbeiter	251	164	39	20	9	6	1
Kosten Reparaturen	479	174	17	7	2	2	0
Mieten und Pachten	464	184	25	14	9	3	1
sonst. Kosten	500	161	19	9	3	1	0
Zinsen	450	190	39	19	8	3	1
Kosten gesamt	500	112	10	7	1	0	0
Bruttowertschöpfung	500	88	16	11	7	5	2
Nettowertschöpfung	499	90	15	11	6	3	3
Aufwendungen FuE	96	79	31	19	6	1	1
Beschäftigte FuE	17	17	17	8	2	1	0

Quelle: eigene Berechnungen auf Basis des Public Use File und der Kostenstrukturhebung

5 Analyseaspekte

Die in Kapitel 4 dargestellten Abweichungen der anonymisierten Werte gegenüber den Originalwerten bleiben nicht ohne Auswirkungen auf die Analysefähigkeit der Daten. Um diese Auswirkungen exemplarisch darzustellen, sind in Tabelle 2 die arithmetischen Mittel und die Standardabweichungen der stetigen Merkmale des Public-Use-Files den arithmetischen Mittel und den Standardabweichungen der Originaldaten⁹ gegenübergestellt. Die Unterschiede liegen in den

⁹ Die Ergebnisse der Originalwerte beziehen sich auf sämtliche rund 13.000 Unternehmen der KSE mit weniger als 250 Beschäftigten.

meisten Fällen innerhalb eines „normalen“ Stichprobenfehlers. Aufgrund des geringen Umfangs dieser Stichprobe können ökonomische Schlussfolgerungen natürlich nur stark eingeschränkte Aussagekraft besitzen.

Tabelle 2: Vergleich der Mittelwerte und der Standardabweichungen zwischen den Originaldaten und den Daten des Public-Use-Files

Merkmal	Originalwerte		Werte des Public-Use-Files	
	Mittelwert	Standardabweichung	Mittelwert	Standardabweichung
Tätige Inhaber	0,415	0,8080108	0,396	0,774845
Angestellte/Arbeiter	80,02	56,1745917	83,52	58,481635
Teilzeit	6,898	12,2091212	7,548	16,5015241
Teilzeit in Vollzeiteinheiten	3,336	5,8358685	3,566	6,8453899
Tätige Personen	80,442	56,1709395	83,926	58,5182133
Umsatz eigene Erzeugnisse	21.561.626,77	35.902.920,59	24.475.565,34	40.692.205,94
Umsatz Handel	1.938.966,69	10.082.134,12	1.694.099,71	7.195.025,2
Gesamtumsatz	23.797.036,84	39.828.607,31	26.530.964,81	44.360.091,59
Bestandsveränderung fertigen/unfertigen Erzeugnisse	55.083,26	2.021.378,64	236.813,24	1.910.643,77
Gesamtleistung	23.886.697,72	39.827.910,61	26.794.750,14	44.717.252,01
Anfangsbestand gemessen am Umsatz	0,08472	0,1294568	0,0867255	0,1445042
Endbestand gemessen am Umsatz	0,09211	0,2178090	0,0987421	0,2091237
Anfangsbestand gemessen am Umsatz (Rohstoffe etc.)	0,05895	0,0751860	0,0535389	0,0595362
Endbestand gemessen am Umsatz (Rohstoffe etc.)	0,06075	0,0763173	0,0548073	0,0609251
Rohstoffverbrauch	10.525.121,31	26.624.760,67	11.338.373,67	23.566.723,46
Energieverbrauch	410.636,77	1.953.317,16	366.687,69	764.271,89
Anfangsbestand an gemessen am Umsatz (Handel)	0,0898	0,9522405	0,0741232	0,2818941
Endbestand an gemessen am Umsatz (Handel)	0,0925	0,9346530	0,0710890	0,2627486
Einsatz an Handelsware	1.532.760,18	9.205.588,00	1.333.250,27	6.068.640,36
Bruttogehalt	4.491.243,55	3.854.998,43	4.734.669,40	4.207.862,11
Ges. Sozialkosten	883.239,97	744.692,45	927.105,71	787.046,76
sonst. Sozialkosten	114.547,31	341.392,08	106.131,60	245.816,87
Kosten Leiharbeiter	113.769,31	484.720,91	119.066,29	503.326,43
Kosten Lohnarbeiter	748.479,57	3.421.776,23	1.066.910,19	4.460.339,38
Kosten Reparaturen	431.535,12	1.729.621,88	371.624,04	619.918,06
Mieten und Pachten	480.528,26	1.043.629,99	543.888,41	1.201.057,33
sonst. Kosten	2.311.836,01	5.153.023,83	2.914.644,78	7.598.538,20
Zinsen	253.734,22	485.060,65	254.195,91	415.881,97
Kosten gesamt	11.004.208,24	1.3707.613,52	13.155.027,22	21.085.817,32
Bruttowertschöpfung	7.399.139,86	7.509.900,55	7.866.877,25	8.922.980,98
Nettowertschöpfung	6.592.795,68	6.754.768,41	7.004.603,16	8.078.064,46
Aufwendungen FuE	137.106,33	626.643,86	116.534,62	437.410,30
Beschäftigte FuE	1,20544	4,3824402	1,1700000	3,6900315

Quelle: eigene Berechnungen auf Basis des Public Use File und der Kostenstrukturerhebung

6 *Fazit*

Dem Anwender liegt mit den Daten dieser CD erstmalig ein absolut anonymisierter Datensatz aus dem Bereich der amtlichen Unternehmensstatistik vor. Er kann aufgrund seiner starken Anonymisierung zwar kaum zu belastbaren wissenschaftlichen Analysen herangezogen werden, dies ist aber auch nicht der Zweck dieses Public-Use-Files. Vielmehr soll er Interessierten die Gelegenheit bieten anhand von relativ "echten" Daten empirische Forschungsmethoden einzuüben. Für diesen Zweck könnte der Datensatz in der Lehre eingesetzt und damit Studierenden die Möglichkeit geboten werden, statistische Methoden anhand real existierender Daten zu erlernen.

Darüber hinaus erlaubt es der Datensatz, sich mit der Kostenstrukturerhebung im Verarbeitenden Gewerbe vertraut zu machen. Dadurch soll aufgezeigt werden, welche Analysepotenziale in der KSE enthalten sind. Das Statistische Bundesamt strebt an, im Rahmen des erwähnten Forschungsprojektes für die Wissenschaft einen wesentlich schwächer anonymisierten Scientific-Use-File zu entwickeln.

7 *Ansprechpartner*

Für Fragen zur KSE allgemein:	Ottmar Hennchen	Tel.: 0611/75-2308
	Gerald Göbel	Tel.: 0611/75-2301
	E-Mail:	kse-industrie@destatis.de
Für Fragen zum Public-Use-File:	Roland Sturm	Tel.: 0611/75-2580
	Dr. Daniel Vorgrimler	Tel.: 0611/75-3486
	E-Mail:	roland.sturm@destatis.de