Anonymisierungsbeschreibung für Paneldaten der Kostenstrukturerhebung der Unternehmen des Verarbeitenden Gewerbes sowie des Bergbaus und der Gewinnung von Steinen und Erden (1999 – 2002)

1. Einleitung

Im Jahre 1987 wurde der Wissenschaft im Bundesstatistikgesetz[1] mit dem § 16 Abs. 6 ein privilegierter Zugang zu Mikrodaten der amtlichen Statistik eingeräumt. Dieser Paragraph erlaubt die Übermittlung von Mikrodaten an die Wissenschaft, sofern diese nur mit unverhältnismäßig hohem Aufwand reidentifiziert werden können. „Unverhältnismäßig" bedeutet hier, dass die Kosten einer Reidentifikation deren Nutzen übersteigen (faktische Anonymität). Dies impliziert, dass die Enthüllung von Einzelangaben in einem faktisch anonymen Datensatz nicht mit absoluter Sicherheit ausgeschlossen werden muss.

Durch die Arbeiten des Projektes „Wirtschaftsstatistische Paneldaten und Faktische Anonymisierung" kann nun erstmals eine faktisch anonyme Datei (Scientific Use File) von Paneldaten aus dem Bereich der Wirtschaftsstatistiken für die Wissenschaft angeboten werden. Der Scientific Use File wurde aus den Daten der Kostenstrukturerhebung im Verarbeitenden Gewerbe von 1999 bis 2002 generiert.[2] Der vorliegende Scientific Use File enthält vier zusammenhängende Wellen mit 13 294 Unternehmen.

Während der Projektphase war die größte Herausforderung die Sicherstellung der faktischen Anonymität bei gleichzeitig bestmöglichem Erhalt des Potenzials für wissenschaftliche Analysen. Die Ergebnisse haben gezeigt, dass in der Regel eine Unterdrückung oder Vergröberung von Informationen bei den qualitativen Merkmalen wie z.B. die Zusammenfassung von Wirtschaftsabteilungen oder eine Vergröberung der Regionalangabe beachtlich zur Anonymisierung beiträgt. Bei Paneldaten über Unternehmen und Betriebe sind jedoch zusätzlich weitere datenverändernde Anonymisierungsmaßnahmen nötig, die zu einer Modifikation der im Datensatz vorhandenen quantitativen Merkmale führen. Bei den für das Scientific Use File auf die Originaldaten angewendeten Anonymisierungsmaßnahmen wurde daher ein großes Gewicht auf die Behandlung der qualitativen Merkmale gelegt.

2. Anonymisierungsmaßnahmen

Im Folgenden werden die einzelnen Anonymisierungsmaßnahmen aufgeführt, die im Rahmen der Projektarbeiten entwickelt wurden, einen ausreichenden Datenschutz sicherstellen und vor dem Hintergrund der datenschutzrechtlichen Erfordernisse das Analysepotential bestmöglich erhalten.

---

1 Gesetz über die Statistik für Bundeszwecke (Bundesstatistikgesetz – BStatG) vom 22. Januar 1987 (BGBl. I S. 462, 565), zuletzt geändert durch Artikel 16 des Gesetzes vom 21. August 2002 (BGBl. I S. 3322).

2 Aggregate dieser Erhebung werden in der Fachserie 4.3, „Produzierendes Gewerbe - Kostenstrukturerhebung der Unternehmen des Verarbeitenden Gewerbes sowie des Bergbaus und der Gewinnung von Steinen und Erden", des Statistischen Bundesamtes veröffentlicht.

2.1 Informationsreduzierende Anonymisierungsverfahren

Auf die Merkmale „Tätige Inhaber“ und „Arbeiter und Angestellte“ wurde verzichtet, da sich diese als besonders reidentifikationsgefährdend und für wissenschaftliche Analysen als wenig wertvoll herausgestellt haben.

Besonders geeignet für Reidentifikationen sind regionale Angaben. Der Erhalt solcher Merkmale in einem Scientific Use File stellt daher für die Anonymisierung ein schwieriges Unterfangen dar. Für den erstellten Scientific Use File wurde daher eine Vergröberung auf eine Ost-West-Klassifizierung vorgenommen, wobei Berlin dem Westen zugeordnet wurde. Für die Variante der Mikroaggregation werden nur die kleineren und mittleren Unternehmen mit weniger als 250 Beschäftigten ausgegeben. Bei der Variante der stochastischen Überlagerung wurde für die größten Unternehmen der Beschäftigtengrößenklasse mit 1000 und mehr die Regionalangabe (Ost, West) entfernt, da größere Unternehmen häufig nicht trennscharf nach Ost und West aufgeteilt sind und um für die größeren Unternehmen eine zusätzliche Schutzwirkung zu erreichen.

Die Daten der Kostenstrukturerhebung im Verarbeitenden Gewerbe werden nach der Klassifikation der Wirtschaftszweige, Ausgabe 1993 (WZ93), auf der Vierstellerebene (Klasse) erhoben und aufbereitet. Diese Klassifikation ist von der europäischen Klassifikation NACE Rev.1 abgeleitet, die aufgrund der NACE-Verordnung des Rates der Europäischen Gemeinschaften seit 1995 in allen Mitgliedstaaten der Europäischen Union sowohl für die Erhebung als auch für die Darstellung der statistischen Daten anzuwenden ist.[3] Das Kodierungssystem der WZ93 unterscheidet zwischen Abschnitten (Buchstaben A-Q), Unterabschnitten (Buchstabenkombinationen AA-QA), Abteilungen (Zweisteller), Gruppen (Dreisteller), Klassen (Viersteller) und Unterklassen (Fünfsteller). Der Wirtschaftsbereich „Verarbeitendes Gewerbe sowie Bergbau und Gewinnung von Steinen und Erden“ erstreckt sich über die Abschnitte C und D bzw. – in der numerischen Gliederung – über die Abteilungen 10 bis 37. Im Projekt „Wirtschaftsstatistische Paneldaten und Faktische Anonymisierung“ haben sich Datenschützer und Datennutzer darauf verständigt, bei dem hierarchischen Merkmal WZ93 die Gliederungstiefe 3 (Zweistellerebene) nicht zu unterschreiten, da hierdurch zum einen eine beachtliche Schutzwirkung und zum anderen nach Einschätzung der beteiligten Wissenschaftler für einen Scientifc-Use-File eine ausreichende Breite an Analysemöglichkeiten erhalten wird.

In den Veröffentlichungen der statistischen Ämter werden aufgrund von Geheimhaltungsaspekten die Ergebnisse einiger Wirtschaftsabteilungen nicht veröffentlicht. Es handelt sich dabei um Unternehmen der Abteilungen *10, 11, 14, 16, 23, 30, 32, 35* und *37* der WZ93. Bei den im Projekt durchgeführten Simulationen hat sich bestätigt, dass diese Abteilungen neben den Abteilungen *15, 17, 18, 19*, 22 und *34* größerer Geheimhaltung bedürfen. Um diese kritischen Abteilungen im Scientific Use File belassen und weitgehend auf datenverändernde Verfahren bei den quantitativen Merkmalen verzichten zu können, werden kritische Abteilungen nur auf der Ebene der Unterabschnitte dargestellt. Damit ergeben sich folgende Zusammenfassungen: *10* (Kohlenbergbau, Torfgewinnung), *11* (Gewinnung von Erdöl und Erdgas, Erbringung damit verbundener Dienstleistungen) und *14* (Gewinnung von

3 Für neuere Erhebungen ab dem Jahr 2003 gilt mit dem Branchenschlüssel WZ 2003 wiederum eine neue Klassifikation.

Steinen und Erden, sonstiger Bergbau) zum Abschnitt C, die Abteilungen *15* (Ernährungsgewerbe) und *16* (Tabakverarbeitung) zum Unterabschnitt DA, die Abteilungen *17* (Textilgewerbe) und *18* (Bekleidungsgewerbe) zum Unterabschnitt DB, die Abteilungen *21* (Papiergewerbe) und *22* (Verlags- und Druckgewerbe, Vervielfältigung) zum Unterabschnitt DE, die Abteilungen *30* (Herstellung von Büromaschinen, DV-Geräten und -einrichtungen) und 31 (Herstellung von Geräten der Elektrizitätserzeugung, -verteilung u.ä.) zum Unterabschnitt DL sowie die Abteilungen *34* (Herstellung von Kraftwagen und Kraftwagenteilen) und *35* (Sonstiger Fahrzeugbau) zum Unterabschnitt DM zusammengefasst. Bei den Abteilungen *19* (Ledergewerbe) und *23* (Kokerei, Mineralölverarbeitung, Herstellung von Brutstoffen) wurden zur Abteilung „Sonstige" zusammengelegt. Außerdem wurde die Abteilung *37* (Recycling) aus inhaltlichen und aus Geheimhaltungsgründen herausgenommen. Obwohl die Abteilungen 32 (Rundfunk-, Fernseh- u. Nachrichtentechnik) und 33 (Medizin-, Mess-, Steuer- u. Regelungstechnik, Optik) ebenfalls zum Unterabschnitt DL zu zählen sind, werden Sie im Datensatz separat aufgeführt, da hier die Weitergabe der Zweisteller aus Sicht des Datenschutzes unbedenklich ist. Eine zusammenfassende Aufstellung der im Datensatz vorhandenen Ausprägungen des Merkmals WZ93 enthält nachfolgende Tabelle:

Wirtschaftsgliederung	Wert	WZ93-Angabe
Bergbau und Gew. v. Steinen u. Erden	1	C (10, 11 und 14)
Ernährungsgewerbe u. Tabakverarbeitung	2	DA (15 und 16)
Textil- u. Bekleidungsgewerbe	3	DB (17 und 18)
Holzgewerbe (oh. H. v. Möbeln)	4	20
Papier-, Verlags- u. Druckgewerbe	5	DE (21 und 22)
Chemische Industrie	6	DG (bzw. 24)
H. v. Gummi- u. Kunststoffwaren	7	DH (bzw. 25)
Glasgewerbe, Keramik, Ver. V. Steinen u. Erden	8	DI (bzw. 26)
Metallerzg. u. –bearbeitung	9	27
H. v. Metallerzeugnissen	10	28
Maschinenbau	11	DK (bzw. 29)
H. v. Büromasch., Dv-Gerät. u. einr., Gerät. d. Elektriz.erzg., -verteilung u. ä.	12	DL (30 und 31)
Rundfunk-, Fernseh- u. Nachrichtentechnik	13	32
Medizin-, Mess, Steuer- u. Regelungstechnik,Optik	14	33
Fahrzeugbau	15	DM (34 und 35)
H.v.Möbeln,Schmuck,Musikinstr., Sportger. usw	16	36
Sonstige: Ledergewerbe; Kokerei, Mineralölverarbeitung, H. v. Brutstoffen	17	S (19 und 23)

2.2 Datenverändernde Anonymisierungsmaßnahmen

2.2.1 Eindimensionale Mikroaggregation

Die im Datensatz verbleibenden 30 quantitativen Merkmale wurden eindimensional für jedes Merkmal

separat mikroaggregiert.[4] Bei dieser Variante der Mikroaggregation werden zunächst die Merkmalsausprägungen je Merkmal absteigend sortiert. Dann werden tripelweise (aus den Merkmalsausprägungen dreier benachbarter Merkmalsträger) die Durchschnittswerte ermittelt, die Originalwerte durch diese Durchschnittswerte ersetzt und wieder an die ursprüngliche Position zurücksortiert. Dieses Verfahren wurde im Rahmen des Projektes in ein Verfahren der Varianzerhaltenden Mikroaggregation so erweitert, dass es den Vorteil hat, dass es den für Mikroaggregationsverfahren üblichen Varianzverlust innerhalb der Aggregationsgruppen korrigiert.

Bei der varianzerhaltenden eindimensionalen Mikroaggregation wird die Datenspalte absteigend sortiert und so in Gruppen mit variabler Gruppengröße zusammengefasst, dass die gruppeninterne Varianz minimiert wird (Verfahren nach *Hansen, S. L. und Mukherjee, S. 2003*). Die variable Gruppengröße liegt dabei zwischen der minimalen Größe m und 2m-1 (üblicherweise ist m=3). Um die Null- und Missingwerte im Datenbestand erhalten zu können, wurden diese Werte im Verfahren nicht berücksichtigt sondern behalten auch nach der Anonymisierung ihren originalen Zustand.

Üblicherweise werden die Einzelwerte in den Mikroaggregaten durch den Durchschnitt der Werte in der Gruppe ersetzt. Damit werden die Summen- und Durchschnitte des Gesamtbestandes fehlerfrei reproduziert. Dieses Vorgehen hat jedoch auch zur Folge, dass die gruppeninterne Varianz Null wird, während die Varianz zwischen den Gruppen unverändert bleibt. Im Ergebnis wird die Gesamtvarianz eines Merkmals durch Mikroaggregationsverfahren systematisch nach unten verzerrt. Bei varianzerhaltender Mikroaggregation (mit minimaler Gruppengröße m=4) erfolgt die Bestimmung der anonymisierten Werte über folgende Regel. Der Mittelwert und die Standardabweichung der Mikroaggregationsgruppe werden zuerst bestimmt über:

$$\overline{x^{o}_{l,j}} = \frac{\sum_{i=0}^{m_l-1} x^{o}_{l+i,j}}{m_l} \qquad ;l = 1, m_1+1, m_1+m_2+1, \ldots \quad ;j = 1,2,3,\ldots,k$$

$$\sigma(x^{o}_{l,j}) = \sqrt{\frac{\sum_{i=0}^{m_l-1}\left(x^{o}_{l+i,j} - \overline{x^{o}_{l,j}}\right)^2}{m_l}}$$

Mit $x^{o}_{i,j}$ werden die originalen Werte der Einheit *i* beim Merkmal *j* definiert. Analog sind im folgenden $x^{a}_{i,j}$ die anonymisierten Werte der Einheit *i* beim Merkmal *j*. Die Reihe $l=1,m_1+1,m_1+m_2+1,\ldots$ beschreibt die Reihe der jeweils ersten Elemente einer Mikroaggregationsgruppe, während m_l die Größe der jeweiligen Gruppe beschreibt (z.B. l=1,5,9,…).

Zusätzlich wird für jede Gruppe aus m_l Elementen festgelegt, wie diese in zwei möglichst gleich große Gruppen geteilt wird. Dazu wird die Gruppengröße für die größeren Werte (g_l) bestimmt (mit $g_l=int(m_l/2)$). Die Teilgruppe der größeren Werte soll nach der Anonymisierung auch mit größeren

4 Zur Methode der Mikroaggregation und anderen im Projekt untersuchten Methoden siehe Höhne, J.: „Anonymisierungsverfahren für Paneldaten“ in Wirtschaft- und Sozialstatistisches Archiv 2/2008.

aber einheitlichen Werten und die Teilgruppe der kleineren Werte mit kleineren aber einheitlichen Werten belegt werden. Diese anonymen Werte lassen sich ermitteln als

$$x^a_{l+i,j} = \overline{x^o_{l,j}} + \sqrt{\frac{m_l - g_l}{g_l}}\,\sigma(x^o_{l,j}) \quad ;i = 0,1,....,g_l - 1$$

$$x^a_{l+i,j} = \overline{x^o_{l,j}} - \sqrt{\frac{g_l}{m_l - g_l}}\,\sigma(x^o_{l,j}) \quad ;i = g_l, g_l + 1,....,m_l - 1$$

Bei symmetrischen Gruppen ($m_l = 2*g_l$) vereinfachen sich die Formeln zu:

$$x^a_{l+i,j} = \overline{x^o_{l,j}} + \sigma(x^o_{l,j}) \qquad ;i = 0,1,....,g_l - 1$$

$$x^a_{l+i,j} = \overline{x^o_{l,j}} - \sigma(x^o_{l,j}) \qquad ;i = g_l, g_l + 1,....,m_l - 1$$

Damit wird die eine Hälfte der Gruppe durch Mittelwert plus Standardabweichung und die andere durch Mittelwert minus Standardabweichung ersetzt.

Bei extrem schief verteilten Daten kann es vorkommen, dass $\overline{x^o_{l,j}} < \sigma(x^o_{l,j})$ gilt.

Dann erhalten die Elemente der unteren Teilgruppe jedoch negative anonyme Werte, was im Widerspruch zu ihrer inhaltlichen Definition stehen kann (z.B. negative Anzahl „tätiger Personen“ in einem Unternehmen). Da solche Werte in vielen Auswertungen hinderlich sind (wenn z.B. Regressionen mit logarithmierten Werten erfolgen), wird für diese Gruppe die Größe der oberen Teilgruppe mit g_l=2 festgelegt. Das bewirkt eine Aufteilung der extremen Abweichungen nach oben auf mehr Elemente in der unteren Gruppe (m_l-g_l), so dass das Auftreten von negativen Werten i.d.R. verhindert werden kann, da die kleineren anonymisierten Werte um weniger als die Standardabweichung vom Mittelwert abweichen ($g_l/(m_l-g_l)<1$).

2.2.2 Stochastische Überlagerung

Im Ergebnis der Erfahrungen aus mehreren Projekten zur faktischen Anonymisierung haben sich multiplikative Überlagerungen gegenüber additiven Überlagerungen bei wirtschaftsstatischen Daten durchgesetzt. Wegen ihrer größenabhängigen Datenveränderung erzeugen sie eine entsprechend gleichmäßige relative Schutzwirkung für alle stetigen Merkmale des Datenbestandes.

Im Gegensatz zur additiven Überlagerung hat die multiplikative Überlagerung jedoch den Nachteil, dass diese Verfahren keine Einflussmöglichkeiten auf die Kovarianzen bzw. Korrelationen zwischen den Merkmalen ermöglichen. Da aber gerade Wirtschaftsstatistiken oft eine Schiefe aufweisen, werden für wirtschaftsstatistische Analysen oft logarithmierte Werte verwendet. Eine direkte additive

Überlagerung der logarithmierten Werte hat einerseits einen positiven Einfluss auf den Erhalt der Analysefähigkeit der Daten. Andererseits ist sie zu einer multiplikativen Überlagerung der Originalwerte äquivalent. Werden die logarithmierten Werte additiv zufallsüberlagert, sind somit die Vorteile der additiven Überlagerung nutzbar. Deshalb wurde im Projekt „Faktische Anonymisierung von Paneldaten“ ein auf dieser Idee basierendes Verfahren zur „stochastischen Überlagerung der logarithmierten Werte“ entwickelt, welches hier kurz dargestellt werden soll.

Die additive Zufallsüberlagerung von logarithmierten Werten wurde von Kim und Winkler (2003) nach folgender Schrittfolge empfohlen.

1.) $X^1=\log(X^0)$

2.) $X^2=X^1+U$

3.) $X^a=\exp(X^2)$

mit:

$X^{o,1,2,a}$ – Datenmatrix der Einzelwerte mit
o – im Original
1 – nach 1. Anonymisierungsschritt
2 – nach 2. Anonymisierungsschritt
a – anonyme Datenmatrix
$E(U)=\underline{0}$ – alle Werte der Matrix U haben den Erwartungswert 0
log(X), exp(X) – die log-Funktion bzw. antilog-Funktion wird für jedes Element einzeln angewendet.

Für die Varianz-/Kovarianzmatrix Σ der Zufallsmatrix U (Σ(U)) wurde durch Kim und Winkler eine Proportionalität zur Varianz-/Kovarianzmatrix der logarithmierten Originaldaten unterstellt $\Sigma(U)=a\,\Sigma(X^1)$.

Für die anonymen Einzelwerte gilt dann:

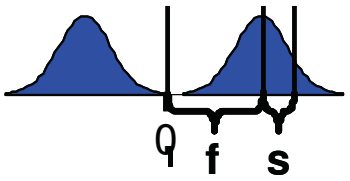
$$x_{ij}^a = e^{u_{ij}+\log(x_{ij}^o)} = x_{ij}^o e^{u_{ij}}$$

Damit führt die Veränderung der logarithmierten Werte zu erwartungstreuen anonymen logarithmierten Werten, ihre Varianz ist allerdings von der Varianz der Zufallszahlen und somit indirekt von der Varianz der Originalwerte in den Merkmalsspalten abhängig. Je größer also die Streuung der Originalwerte ist, umso größer ist die Veränderung/Schutzwirkung durch die Anonymisierung. Außerdem ergibt sich eine systematische Verzerrung in den Erwartungswerten der Originalwerte, so dass Auswertungen korrigiert werden sollten (siehe Kim und Winkler 2003).

Ziel des Verfahrens sollte jedoch eine möglichst gleichwertige Schutzwirkung für alle Variablen und Werte sein. Außerdem ist für gute ökonometrische Analysen der logarithmierten Werte (X^1) nicht der

Erhalt der proportionalen Varianz-/Kovarianzmatrix notwendig, sondern bereits der Erhalt der Korrelationsmatrix ausreichend. Deshalb wird eine Zufallsmatrix U verwendet, deren Varianz in allen Spalten gleich ist, aber eine Korrelation analog der Originaldaten besitzt $R(U)=R(X^1)$ (mit R-Korrelationsmatrix von U bzw. X^1).

Um bei der Überlagerung eine Mindestveränderung für die meisten Werte sicherzustellen, wird die Überlagerungsmatrix so gebildet, dass sie aus einer zweigipfeligen Mischungsverteilung besteht (siehe Abb.). Die Schutzwirkung der Mischungsverteilung wird dabei durch den Verschiebungsfaktor f und die Streuung in den Mischungskomponenten s bestimmt. Die zwei Zufallsmatrizen für die beiden Mischungskomponenten sind so zu erzeugen, dass für die Mischungsverteilung wieder gilt $R(U)= R(X^1)$.



Die beiden Parameter f und s bestimmen einerseits die Schutzwirkung des Verfahrens, sind andererseits wegen der nichtlinearen Transformation nicht mehr inhaltlich interpretierbar. Deshalb hat sich für die Programmsteuerung die Vorgabe des Parameters f für die Verschiebung in den Mischungskomponenten und die Angabe eines Parameters s für die Standardabweichung insgesamt in den Spalten der Matrix U bewährt.

Diese Überlagerung sichert den guten Erhalt der Mittelwerte und der Korrelation der logarithmierten Werte. Es treten kleine Abweichungen auf, die vor allem darauf basieren, dass im Datenbestand fehlende Werte enthalten sein können, deren Logarithmus sich einerseits nicht bestimmen lässt, die andererseits aber auch nicht überlagert werden sollen um logische Zusammenhänge in den Daten nicht zu gefährden. So sollen z.B. bei Paneldaten keine Angaben für Einzelwerte in den Jahren entstehen, in denen das Unternehmen nicht existierte. Fehlende Werte verzerren die Berechnung der Korrelation.

Das Verfahren wurde so konzipiert, dass es nicht nur für die Logarithmen annähernd erwartungstreu ist, sondern auch die Ausgangsvariablen möglichst erwartungstreu erhält. Dies setzt allerdings voraus, dass der stochastische Prozess teilweise aufgehoben wird.

Um die Eigenschaften bei den Ausgangsvariablen ebenfalls möglichst gut zu erhalten, ist es erforderlich, dass die multiplikativen Überlagerungen eine stärkere gegenseitige Kompensationswirkung haben als bei einem reinen stochastischen Prozess. Dieser würde sonst zu systematischen Verzerrungen führen, wie sie z.B. bei Kim und Winkler (2003) beschrieben sind. Deshalb muss die Werte verkleinernde bzw. vergrößernde Wirkung der beiden Mischungskomponenten gezielt für die Kompensationswirkung eingesetzt werden. Werden aus den beiden Mischungskomponenten jeweils Paare von Zufallsvektoren gebildet, so haben bereits diese Zufallsvektoren eine kompensierende Wirkung. Die eine Überlagerung vergrößert die Werte, während die andere die Werte verkleinert. Die Kompensation ist umso größer je ähnlicher sich die beiden zu überlagernden Datensätze des Originaldatenbestandes sind, da durch die Mischungskomponenten möglichst ähnliche aber entgegengesetzt gerichtete Überlagerungsfaktoren ($\exp(u)>1$ genau dann

wenn u>0 und exp(u)<1 genau dann wenn u<0) verwendet werden. Außerdem soll bei der Überlagerung die Fehlerwirkung im Datenbestand von bereits anonymisierten Werten berücksichtigt werden. Das funktioniert am Besten, wenn die Überlagerung von den großen Werten (größere absolute Veränderung bei gleichen relativen Fehlern) zu den kleineren Werten erfolgt (siehe Verfahren zur kontrollierten multiplikativen Überlagerung nach Höhne in Ronning et al. 2005).

Deshalb werden die Originaldaten nach ihrer Größe absteigend sortiert und Paare gebildet. Anschließend werden diese Paare der logarithmierten Originaldaten mit je einem Zufallsvektor aus den beiden Mischungskomponenten überlagert, womit ein Merkmalsträger vergrößert und der andere verkleinert wird. Die konkrete Paarung wird dabei anhand eines Abstandsmaßes entschieden, das die Gesamtqualität des Datenbestandes maximiert (z. Zt. ist der Erhalt der Mittelwerte implementiert).

Algorithmus:

Die nächsten zu bearbeitenden Datensätze werden bestimmt, indem

1) der Datensatz i mit dem größten Abstand zum Zentroiden (Vektor der Mittelwerte) des noch zu bearbeitenden Datenbestandes bestimmt wird.

$$\max_i \left(\sum_{k=1}^{m} \left(\frac{\left| x_{i,k}^o - \overline{x_k^o} \right|}{\left| \overline{x_k^o} \right|} \right)^2 \right)$$

2) Der Datensatz j mit dem geringsten Abstand zu dem in Schritt 1) ausgewählten Datensatz gesucht wird.

$$\min_j \left(\sum_{k=1}^{m} \left(\frac{\left| x_{j,k}^o - x_{i,k}^o \right|}{\left| \overline{x_k^o} \right|} \right)^2 \right)$$

Die beiden Datensätze i und j werden an das Ende des noch zu bearbeitenden Datenbestandes sortiert (Position n und n-1). Für die beiden ausgewählten Datensätze des Originaldatenbestandes x_n und x_{n-1} und gegebene Zufallsvektoren u_n und u_{n-1}, wird aus den beiden unten skizzierten Überlagerungsvarianten v1 und v2 diejenige ausgewählt, die den relativen Fehler in den Mittelwerten/Summen der Merkmale minimiert, d.h.:

$$\min_{v1,v2} \left(\sum_{k=1}^{m} \left(\frac{\sum_{p=1}^{N} x_{p,k}^a - x_{p,k}^o}{1/N \sum_{p=1}^{N} x_{p,k}^o} \right)^2 \right)$$

mit :

v1:

$$x^a_{n,k} = x^o_{n,k} e^{u_{n,k}}$$
$$x^a_{n-1,k} = x^o_{n-1,k} e^{u_{n-1,k}}$$

v2:

$$x^a_{n,k} = x^o_{n,k} e^{u_{n-1,k}}$$
$$x^a_{n-1,k} = x^o_{n-1,k} e^{u_{n,k}}$$

Der noch zu anonymisierende Datenbestand wird dann um diese beiden Einheiten reduziert (n=n-2) und analog weiter bearbeitet.

Da der Originaldatenbestand nach der Größe absteigend bearbeitet wird, werden neben der gezielten Kompensationswirkung auch immer kleinere „Restfehler“ erzeugt. Damit besteht nach vollständiger Anonymisierung der Daten auch nur noch eine minimale Abweichung in den ersten Momenten zwischen originalen und anonymen Daten.

3. Beurteilung der Schutzwirkung

Zur Messung der Schutzwirkung wurde im Statistischen Bundesamt ein Programm zur Simulation von so genannten Massenfischzügen entwickelt. Bei einem Massenfischzug versucht ein Datenangreifer, möglichst viele Einheiten seiner externen Datenbank den Zieldaten (vertrauliche, anonymisierte Daten) zuzuordnen. Die beiden nachfolgenden Tabellen zeigen die Simulationsergebnisse unter Verwendung der kommerziell erhältlichen MARKUS-Datenbank (etwa 10.000 überprüfbare Einheiten für die Kostenstrukturerhebung) als mögliche externe Datenbank eines potenziellen Datenangreifers. Hier wird der Effekt der oben angesprochenen Vergröberung der Regionalinformation deutlich. Auch eine Reidentifikation eines Merkmalsträgers kann erfolglos sein, wenn die zugeordneten Einzelwerte um mehr als 10 Prozent von dem zugehörigen Originalwert abweichen.Zur Beurteilung der faktischen Anonymität wurde in einer Sitzung der Projektgruppe „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten“ beschlossen, die Risikobetrachtung auf Merkmalskombinationen von (Ost-West/Beschäftigtengrößenklasse/Wirtschaftsabteilung) herunterzubrechen, wodurch in einzelnen Wirtschaftszweigen weit höhere Risiken als in den obigen Tabellen entstehen können. Es wurde beschlossen, als Obergrenze für das Risiko einer jeden Merkmalskombination (Ost-West/Beschäftigtengrößenklasse/Wirtschaftsabteilung) den Wert 0,5 zu setzen.

Bei der Anonymisierungsvariante der stochastischen Überlagerung (siehe oben) war eine Reidentifikation von Merkmalsträgern aufgrund der stärkeren Datenveränderung noch schwieriger als bei der Mikroaggregationsvariante. Hierzu wurden umfangreiche Datenangriffsszenarien durchgeführt. Hinzu kommt, dass durch eine durchschnittliche Datenverfremdung von 10,5 Prozent bereits a priori ein rationaler Datenangreifer abgeschreckt werden könnte, da weniger als 50 Prozent der Einzelwerte

brauchbar sind und ein Datenangreifer selbst bei den seltenen erfolgreichen Zuordnungen mit unbrauchbaren Einzelinformationen rechnen muss.

Die mit den probeweise anonymisierten Daten durchgeführten Simulationen haben also auch hier gezeigt, dass eine Enthüllung von Einzelwerten nur mit unverhältnismäßig großem Aufwand möglich ist und damit eine Weitergabe dieser Daten an die Wissenschaft unbedenklich ist.

4. Stellungnahme des Wissenschaftlichen Begleitkreises

In den Rückmeldungen des für das Projekt „Wirtschaftsstatistische Paneldaten und faktische Anonymisierung“ eingerichteten Wissenschaftlichen Begleitkreises wurden die methodischen Arbeiten zur Beurteilung des Analysepotenzials, welche sowohl deskriptive Maße als auch inferenzstatistische Auswertungen in Form linearer und nichtlinearer ökonometrischer Modellierung beinhalten, als überzeugend beurteilt. Die konkrete Anonymisierungsstrategie für die Kostenstrukturerhebung im Verarbeitenden Gewerbe (1999 – 2002) wurde als sorgfältig und umfassend eingestuft und der Empfehlung zur Erstellung eines Scientific Use Files voll zugestimmt.

Literatur:

Höhne, J. (2008) Anonymisierungsverfahren für Paneldaten. In: Springer, Wirtschafts- und Sozialstatistisches Archiv (2008) Bd. 2: S. 259-275

Kim, J.J. und Winkler, W. E. (2003) Multiplicative Noise for Masking Continuous Data. Research Report Series (Statistics #2003-01), Statistical Research Division U.S. Bureau of the Census Washington D.C. 20233

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

Hansen, S. L. und Mukherjee, S. (2003): A polynomial algorithm for optimal univariate microaggregation, in IEEE Transactions on Knowledge and Data Engineering, Vol. 15 No. 4:1043-1044, 2003